Primer on Statistical Analysis

(Level 1)

Table of Contents

1. Introduction

Operations analysts are usually interested in determining technical information about operational issues. The type of information generally falls under one of the following categories:

1. Does a system meet technical specifications?
2. Is a new system or modification significantly better than a previous one?
3. Which of several competitive systems is significantly better than a previous one?
4. Was a test "good", i.e.,
    a. was there bias in the data?
    b. was there much scatter in the data?
    c. was there enough data?
    d. how confident is the analyst in the conclusions about the data?

Since budget, schedule, and other constraints usually exist, the point is to design an efficient test that gets maximum results with a minimum number of trials. This series of statistical primers (3 levels) will present the basic concepts and definitions required to perform such tests. It is laid out in such a way to be a handy reference. The reader should not attempt to memorize the information, just understand the concepts. No prior knowledge of statistics is required. Level 1 covers basic probability and statistics as well as some simple tools. Level 2 moves on to more advanced statistical techniques and Level 3 introduces design of experiments (DOE).

2. Learning the Language

Statistics, like any other specialized field, uses it's own language to communicate unique concepts. The following list defines the most common terms used in statistics (in alphabetical order for easy reference):

Continuous - random variable that can have any real value within some interval

Cumulative Distribution Function (cdf) - relates any outcome of interest to probability that any outcome will be less than or equal to that value; cdf is usually written as a function of a random variable using an upper case letter (e.g., $F(x)$)

Dependent - knowing the outcome of one random variable changes the probabilities of possible outcomes of another random variable; variables that are functions of independent variables (e.g., the height of ballistic projectile is a function of independent variable time and deterministic parameters (initial velocity, starting angle, & acceleration due to gravity))

Deterministic - can be predicted with certainty (e.g., current resulting from V volts and R ohms (V/R); force F applied to projectile of mass M results in acceleration (in vacuum) of F/M)

Discrete - random variable that can only take on distinct values

Exhaustive - there are no other possible outcomes

Experiment - observation of a process and noting the outcomes

Event - occurrence of one or more of the possible outcomes; any subset of a sample space

Independent - knowing outcome of one random variable has no effect on probabilities of possible outcomes of another random variable; variables that are not determined by other variables (e.g., time & x,y,z coordinates for encounter between aircraft & air defense gun)

Mutually Exclusive - only one of the alternative outcomes may occur at any given time

Outlier - data point in sample that is extreme enough to have a large influence on a statistic computed from the sample; normally identified by $|z\text{-score}| > 3$ or using a box plot

Parameter - numerical, descriptive measure of a population; fixed (non-stochastic) constant whose value we usually don't know (e.g., Mean & Variance)

Population - data set that is target of interest

Census - complete enumeration of a population (Descriptive Statistics)

Sample - subset of data selected from a population (Inferential Statistics)

Probability - given the population parameters, predicting the likelihood (frequency of occurrence) that certain outcomes of interest would occur in a sample; "what might be"

Probability Density Function (pdf) - relates any outcome of interest to probability that it occurs in a continuous distribution; the probability that an outcome is located within an interval is equal to the area under the pdf; pdf is usually written as a function of a random variable using a lower case letter (e.g., $f(x)$); continuous equivalent of a pmf;

Probability Distribution - table, graph, or formula that gives the probability $P(x)$ associated with each possible value of x

Probability Mass Function (pmf) - relates each discrete outcome to a probability; discrete equivalent of a pdf

Process - timewise sequence or combination of interrelated actions involving one or more parameters & variables associated with a system

Random - cannot be predicted with certainty (e.g., roll of a die; projectile may hit aircraft or it may not); most things are technically random, but for ease we treat them as deterministic or neglect them

Random Sample - n elements selected from a population such that every set of n elements has an equal probability of being selected

Random Variable - real valued function whose domain is a sample space; in simple terms, it's an unknown quantity; RV's are usually assumed to be *iid* (independent & identically distributed) and are written as capital letters

Sample Space - set of all possible basic outcomes of an experiment listed in a mutually exclusive and exhaustive manner

Statistic - numerical descriptive measure computed from sample data, usually to estimate a parameter

Statistics - science of collecting, analyzing, & interpreting numerical data relating to an aggregate of individuals; "what is estimated to be"

Descriptive - organization, summarization, & description

Inferential - use sample to make inference on population

Stochastic - contains one or more random variables

Trial - a single repetition of an experiment

3. Set Theory

Before diving into the realm of statistics, most textbooks traditionally focus on probability. It is always easier to take snap shots of the whole picture (i.e., probabilities) than it is to build a puzzle of that picture from several pieces or samples (i.e., statistics). The best way to learn the difficult concept of probability is to use pictures. Venn Diagrams and set theory are usually incorporated for this purpose. The following will give a brief review of set theory.

3.1 Definitions

Element - member of a set
Null Set - set that contains no elements; subset of all sets; denoted by ∅
Set - collection of elements or members that have some property in common (e.g., event that aircraft survived, or was killed, or was hit and survived)
Universal Set - set of all elements in an experiment; sample space; denoted by *U*
Venn Diagram - represents collection of sets; rectangular boundary represents universal set (*U*); ovals are drawn & labeled to represent other sets (see Figure 1)

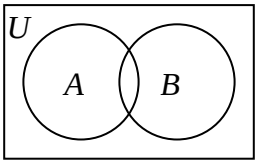


Figure 1. Venn Diagram of employees at a company who speak additional languages. Set *A* contains the people who speak Spanish. Set *B*, those who speak French.

3.2 Set Operations

Exclusive Union - *A* XOR *B*; set of elements that belong to *A* only or to *B* only, but not to both *A* & *B*; from Figure 1, *A* XOR *B* would be those employees who speak only Spanish or only French, but not both (see Figure 2)

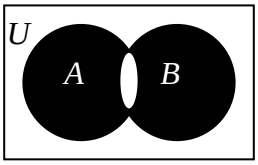


Figure 2. *A* XOR *B*

Intersection - *A*∩*B*; *A* AND *B*; *AB*; set of those elements in *A* & *B* that are common to both *A* & *B*; *A*∩*B* would be those employees who speak both Spanish and French (see Figure 3)

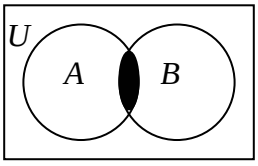


Figure 3. *A*∩*B*

Union - $A \cup B$; *A* OR *B*; set of those elements in *A* & *B* that are in *A* only, *B* only, or both *A* & *B*; from Figure 1, $A \cup B$ would be those employees who speak Spanish, French, or both (see Figure 4)

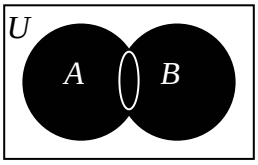


Figure 4. $A \cup B$

Complement - A^c; *A*', NOT *A*; set of elements in universal set that do not belong to *A*; from Figure 1, A^c would be those employees who do not speak Spanish (see Figure 5)

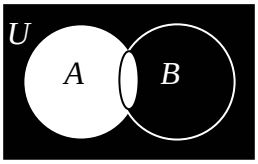


Figure 5. A^c

Difference - $A \backslash B$; *A* AND (NOT *B*); set of elements that belong to *A* and not to *B*; from Figure 1, $A \backslash B$ contains those employees whose only additional language is Spanish (see Figure 6)

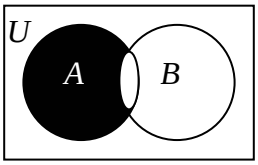


Figure 6. $A \backslash B$

Product - $A \times B = \{(a,b)$ where $a \in A$ and $b \in B\}$; set whose elements are all possible ordered pairs of elements in *A* & *B* (NOTE "$a \in A$" means *a* is an element [member] of the set *A*)

Laws of Algebra of Sets:

Idempotent Laws -	$A \cup A = A$	$A \cap A = A$
Associative Laws -	$(A \cup B) \cup C = A \cup (B \cup C)$	$(A \cap B) \cap C = A \cap (B \cap C)$
Commutative Laws -	$A \cup B = B \cup A$	$A \cap B = B \cap A$
Distributive Laws -	$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
Identity Laws -	$A \cup \varnothing = A$; $A \cup U = U$	$A \cap U = A$; $A \cap \varnothing = \varnothing$
Complement Laws -	$A \cup A^c = U$; $(A^c)^c = A$	$A \cap A^c = \varnothing$; $U^c = \varnothing$; $\varnothing^c = U$
DeMorgan's Laws -	$(A \cup B)^c = A^c \cap B^c$	$(A \cap B)^c = A^c \cup B^c$

The majority of the operations and laws in this section are rarely seen in practice and are reserved for theoretical mathematicians. The union, intersection, and compliment are the most common (and practical) operations. They carry over directly to probability which can be seen in the next section.

4. Probability

Probability is actually a simple concept, but it has been obscured over the years by Mathenese. This section will attempt to explain probability in simple terms much to the dismay of mathematicians everywhere. All of probability is based on three simple axioms and one basic equation.

4.1 Basics

Axioms:

1. $P(A) \geq 0$ (Translation: The probability of an event A is nonnegative)
2. $P(S) = 1$ (Translation: The probability of the sample space S is equal to 1)
3. If $A_1, A_2, A_3, \ldots$ are mutually exclusive events

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

(Translation: The chance of any of several mutually exclusive events occurring is equal to the sum of the probabilities of each event occurring on its own)

Basic Equation:

$P(A) = m/n$ (Translation: If event A occurred m times in n trials, the probability that event A will occur in a future experiment is estimated by the ratio m/n)

4.2 Probability with Pictures

Venn Diagrams can easily be applied to probabilities by following a few simple steps. First, the universal set is the sample space. A traditional example involves the use of a deck of cards. Thus, the sample space or universal set includes all 52 cards and the probability of drawing one of those 52 cards is equal to 1. As a reminder, a deck of cards has four suites (spades, hearts, clubs, & diamonds) each with 13 cards (2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen, King, Ace).

The second step involves viewing each event as a set in the Venn Diagram. For example, redefine the sets in Figure 1 as follows: Event A is represented by the set of Hearts and Event B by the set of Queens. Can you identify the different areas of the Venn Diagram based on the set operations discussed in the previous section ? Here is what it should look like:

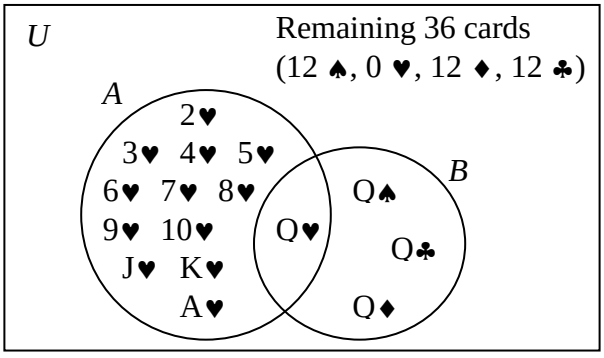


Figure 7. Venn Diagram of Cards

Before moving on to the actual probabilities associated with Figure 7, there are a few more definitions and equations that are required (statisticians don't make this stuff easy do they?):

Event Probabilities - associated with each set is a probability; probability of occurrence of any event *A*, denoted by P_A or $P(A)$ is (# of elements in *A*)/(# of elements in *U*)

Joint Probabilities - associated with event formed by intersection of any two or more events;

$P(A \cap B) = P(AB) = P_{AB}$ = (# of elements in $A \cap B$)/(# of elements in *U*)

$P(A \cup B) = P(A \text{ OR } B) = P_{A \cup B} = P_A + P_B - P_{AB}$

$P(A \text{ XOR } B) = P_{A \text{XOR} B} = P_A + P_B - 2P_{AB}$

$P(A^c) = P(A') = P(\text{NOT } A) = 1 - P_A$

$P(A \backslash B) = P(A \text{ AND (NOT } B)) = P_A - P_{AB}$

Conditional Probabilities - occurrence is dependent or conditional upon the occurrence of another event; $P_{A|B}$ = probability that event *A* occurs, given event *B* has occurred;

$P(A|B) = P_{A|B} = P(AB)/P(B) = P_{AB}/P_B$ (the "|" is read "given")

Special Cases:

$P(AB) = P_{AB} = P_B P_{A|B} = P_A P_{B|A}$ (from definition of conditional probability)

$P(A|B) = P_{A|B} = P_A$ (if *A* & *B* are independent)

$P(A \cap B) = P_{AB} = P_A P_B$ (if *A* & *B* are independent)

$P(A \cap B) = P_{AB} = 0$ (if *A* & *B* are mutually exclusive)

From the material so far it is easy to determine the probability of drawing a Heart and the probability of drawing a Queen ($P(\heartsuit) = P_A = 13/52 = 1/4$; $P(\text{Q}) = P_B = 4/52 = 1/13$). There are other questions that you may want to answer, however, such as the probability of drawing the Queen of Hearts or the probability of drawing a Queen given that you drew a Heart. These can be found by applying the equations given above ($P(\text{Q}\heartsuit) = 1/52$; $P(\text{Q}|\heartsuit) = (1/52)/(1/4) = 1/13$; $P(\heartsuit|\text{Q}) = (1/52)/(1/13) = 1/4$).

4.3 Example (Part I)

OK, so you understand the concept, but you aren't a card dealer or a gambler and you want to know what probability is good for in the real world. The following scenario will be used throughout all levels of the primer to show how the tools presented can be used in operational test and evaluation (OT&E).

> **Scenario:** You are on the OT&E team for the new F-31 Viper, a replacement for the Air Force's F-16 Fighting Falcon and the Navy and Marine Corps' F-18 Hornet and AV-8B Harrier. This joint strike fighter incorporates stealth characteristics with vertical takeoff and landing (VTOL) and many other high technology developments. It's primary mission will be air interdiction (AI) and close air support (CAS), but it will also be capable of fulfilling the offensive and defensive counter air missions (OCA & DCA).

A major area where probability is used in OT&E is survivability analysis. Survivability is generally viewed as two separate characteristics, susceptibility and vulnerability. The first is the

inability of an aircraft to avoid being hit and the latter is the inability of the aircraft to sustain damage when it is hit. Bells should be screaming "conditional probability" because that is exactly how survivability is viewed. Susceptibility is expressed by the probability that the aircraft will be hit by a certain weapon (P_H) and vulnerability is expressed by the probability that the aircraft will be killed given that it was hit ($P_{K|H}$). With these numbers it is easy to compute the survivability of the aircraft, $P_S = 1 - P_K = 1 - P_H \cdot P_{K|H}$.

4.4 Probability Trees

Another way to view probability with pictures is to use a probability tree. Basically a tree is made up of nodes which represent events and branches which show the possible outcomes of each event. The tree can be drawn from left to right or top to bottom with the probability of each outcome written next to the corresponding branch. After the first branch, however, the probabilities are conditional on the previous branches. Each event in any level of branches must be mutually exclusive. The tips of all the final branches are unique events, the probability of each being the product of all the conditional probabilities on the path.

A good example for using a probability tree is a bomb run on a target. Assume the F-31 discussed in the example above is going to attack a surface to air missile (SAM) site. The SAM is it's own defense and no other enemy defenses pose a threat to the F-31. The plane has a three in four chance making it to the target undamaged. There is a 15 percent chance of minor damage and a 10 percent chance the plane will be destroyed. A fully functional F-31 can destroy the SAM site with probability 0.7, but only 0.4 if the F-31 is damaged. If the SAM site is not destroyed, the chances of an undamaged F-31 surviving the egress is the same as the ingress. There is an additional 10 percent chance of a damaged F-31 not making it home regardless of the condition of the SAM site. If the target is destroyed, the F-31 will always make it home unless it was damaged or destroyed (see Figure 8).

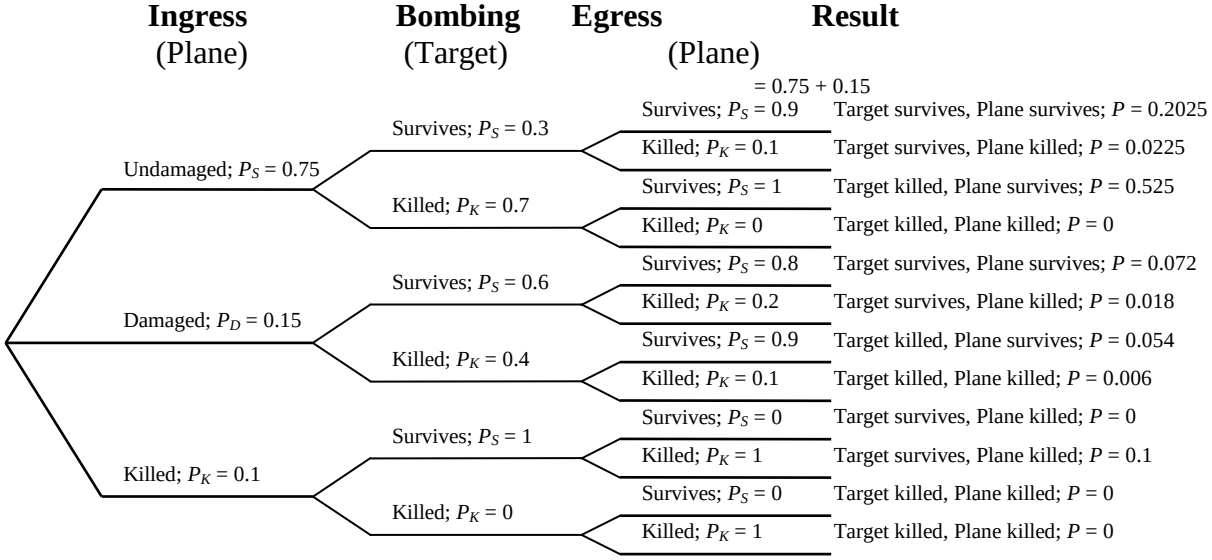


Figure 8. Probability Tree of Bomb Run

Using the information in Figure 8, several questions can be answered. For example, the probability that the F-31 successfully accomplishes its mission undamaged is 0.525. The probability that the SAM site is destroyed regardless of the outcome to the F-31 is 0.525 + 0.054 + 0.006 = 0.585. The probability that the target survives is 0.2025 + 0.0225 + 0.072 + 0.018 + 0.1 = 0.415 (note that this is also 1 - 0.585).

Probability can get much more complicated than what is presented in this section. With this basic understanding, however, you should be fine going through the rest of the primer. If you need more involved probability theory, consult a textbook (the Mendenhall book is recommended).

5. Distributions

Diagrams and probability trees work well for small examples, but eventually they get cumbersome. For this reason statisticians developed mathematical probability distributions to approximate real world probabilities. Unfortunately, to understand what they did, some math is required.

5.1 Discrete Distributions

It is always easiest to start with discrete distributions, the simplest of which is the discrete uniform. As defined is Section 2, a discrete probability distribution is a table, chart, or formula that gives the probability associated with each value of a random variable. It is also called a probability mass function (pmf). Figure 9 shows the possible outcomes of the toss of a six-sided die, i.e., the pmf of a discrete uniform with values in the set {1,2,…,6}.

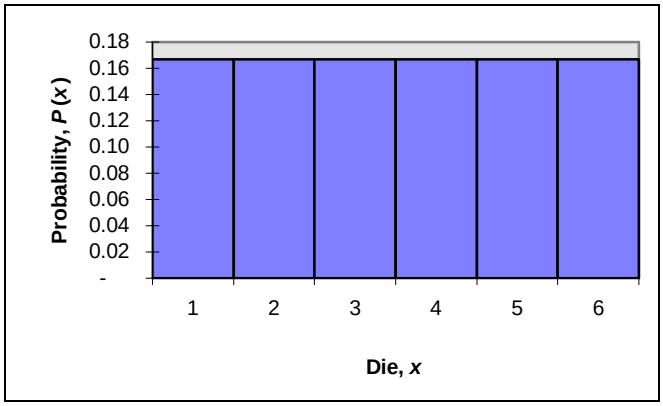


Figure 9. PMF of Uniform {1,2,…,6}

A more mathy way to write the pmf for the Uniform {1,2,…,6} shown in Figure 9 is:

$$P(X = x) = 1/6, \quad x = 1,2,...,6$$

The formula says that the probability of some random variable X being equal to each integer from 1 to 6 is 1/6. Although this was an easy example, all discrete distributions use a formula similar to this.

Before moving on, it is important to understand the cumulative distribution function (cdf) which applies to both discrete and continuous distributions. It is a formula or graph that represents the probability that any outcome of the distribution will be less than or equal to some specified outcome, that is $F(x) = P(X \leq x)$. Figure 10 shows the cdf for the pmf in Figure 9. Section 6 discusses other ways to describe probability distributions, but either the pmf (pdf for continuous) or cdf is required to calculate those quantities.

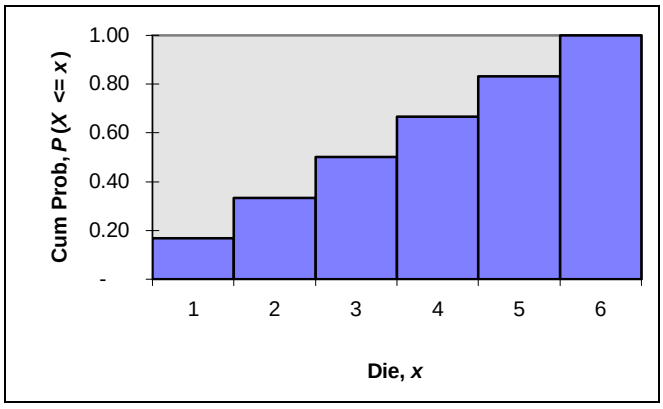


Figure 10. CDF of Uniform {1,2,…,6}

Appendix A contains pmf's for several common discrete distributions along with descriptions of the distributions.

5.2 Central Limit Theorem

An interesting phenomenon occurs when random variables are added. First, look at Figure 11 which shows the possible outcomes of tossing two dice. Note that the pmf is highest in the middle and it appears symmetric (the same on both sides).

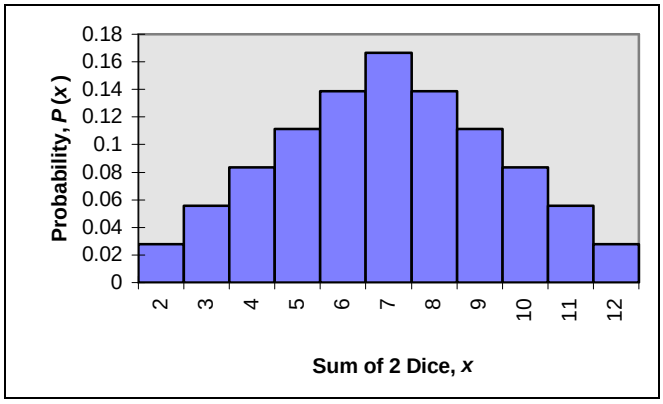


Figure 11. PMF of Sum of Two Uniform {1,2,…,6}

Figure 12 shows the cdf of the same distribution. Note how it is beginning show an "S" shape that accumulates probability slowly. Then around the middle value (7) the probability is building up at the fastest rate. Beyond that value, the amount of probability added by each possible outcome begins to diminish. Note also, that the cdf also reveals the symmetric nature of the distribution.

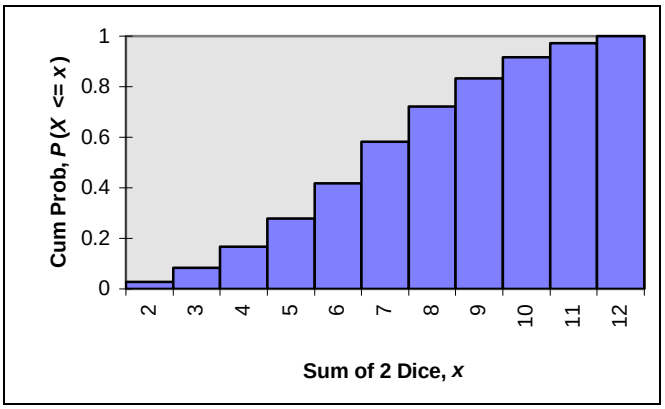


Figure 12. CDF of Sum of Two Uniform {1,2,…,6}

The distribution of the sum looks suspicious, and indeed, adding the values of 12 dice rolls reveals something interesting. The pmf looks like the famous bell curve which is generated by the Normal distribution (see Figure 13). The marvel is a mathy sounding thing called the Central Limit Theorem (CLT). In English, the CLT says that if you take *n* observations from a population with a finite mean μ and variance σ^2, and if *n* is large enough, the sampling distribution of the sample mean is approximately normal with mean μ and variance σ^2/n. In other words, $\bar{x} \sim N(\mu, \sigma^2/n)$ (~ is read "is distributed"). The CLT forms the basis of most of the inferential statistics that will be discussed in this primer. There is no set definition of how big *n* should be for the CLT to apply. Modern statistics says anything greater than or equal to 30, but more traditional statisticians prefer 60.

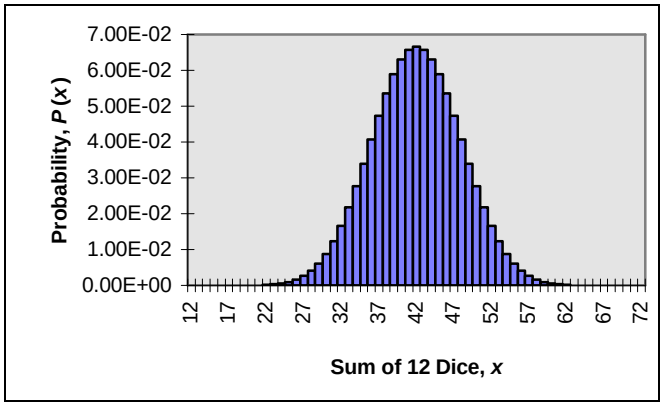


Figure 13. PMF of Sum of 12 Uniform {1,2,…,6}

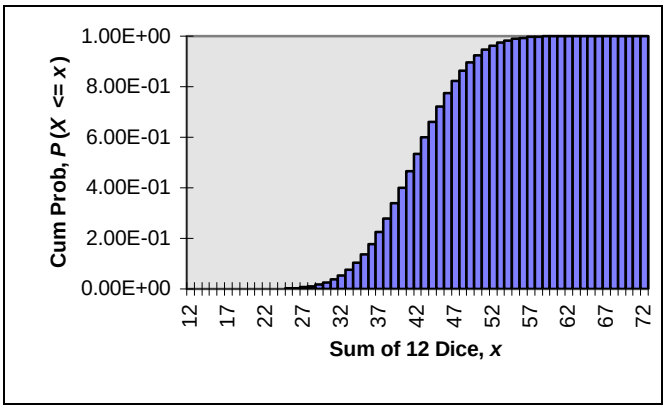


Figure 14. CDF of Sum of 12 Uniform {1,2,…6}

5.3 Continuous Distributions

The normal distribution is very important because it occurs frequently in many natural processes. In order to work with the normal distribution, however, the basics of continuous distributions must first be understood. The most important thing to understand about continuous distributions is that the probability of a continuous random variable taking on a specific value is zero. This is not too unimaginable because there are always more decimal points than we can measure. For example, consider the thickness of the paint on the F-31 from Section 4. Someone may measure it and record 4 mm. Later another person comes with a more precise measuring device and records 3.98 mm. Still another person measures 3.9814. When you consider all the other possible values 3.9814 can take on beyond the fourth decimal point, it is easy to believe the probability of any one specific number occurring is very close to zero. Incidentally, if you can predict the number out to 10 or 12 decimal points, you probably will not be allowed to play the state lottery.

The name continuous probability distribution may sound frightening, but it is very much like a discrete distribution… plus a little calculus (sorry). Just like a discrete distribution, the total area under a continuous distribution function is equal to one. Rather than adding up all the probabilities of each event like we could with discrete distributions, however, calculating the area under a probability density function (pdf) is done with an integral:

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

where $f(x)$ is the pdf of a random variable X. Similarly, finding the probability that the random variable takes on any of several values is done with an integral rather than a summation as it is in the discrete case:

$$P(a \le X \le b) = \int_{a}^{b} f(x)dx$$

A special kind of normal distribution is the one with mean zero and variance one (N(0,1)). This is called the standard normal. Any normal distribution can be converted into a standard normal

by using the following formula: (the conversion is used to save paper because text books will only need to include one normal table in the appendix)

$$z = \frac{x - \mu}{\sigma}$$

Traditionally the pdf for a standard normal is written $\phi(x)$ and the cdf as $\Phi(x)$ rather than the generic $f(x)$ and $F(x)$, respectively.

One way to test the normality of the data is to use a Q-Q Plot or Normal Probability Plot. Basically you plot $x_{(i)}$ versus m_i and look for a straight line with y-intercept μ and slope σ. (NOTE: x_n is the n^{th} observation; $x_{(n)}$ is an **order statistic**, the n^{th} observation when sorted.) For the plot, m_i is determined by

$$m_i = \Phi^{-1}\left(\frac{i - 3/8}{n + 1/4}\right)$$

Usually, the analyst will standardize the data to look for a line through the origin with slope one. Most statistical packages have a feature to generate Q-Q Plots so you don't really have to worry about making one (hard), just know how to read it (easy).

Another interesting feature of the normal distribution is how most of the data lies near the mean. Using the standard deviation, σ, 68% of the data lies within $\pm\sigma$ of the mean; 95% within $\pm 2\sigma$; 99.6% within $\pm 3\sigma$. Therefore, any observations that lie more than $\pm 3\sigma$ from the mean are very rare. This is the basis of confidence intervals and hypothesis testing. Appendix A contains pdf's for several other common continuous distributions.

5.4 Example (Part II)

Review the description of the F-31 given in Section 4.3. On a trip out to the manufacturing facility, the plant manager comes to you and suggests a way of improving the painting process for the F-31. She claims the average number of defects will decrease to 1 per 20 ft^2. You recognize the random variable of the number of defects to be approximated by a Poisson distribution. Assuming the F-31 has roughly 490 ft^2 of surface area to cover, you tell the manager that the expected number of defects per plane will be

$$\lambda = \frac{1}{20 \text{ ft}^2} \cdot \frac{480 \text{ ft}^2}{480 \text{ ft}^2} = \frac{24}{480 \text{ ft}^2}$$

or 24 defects per aircraft. (Note that λ must be in the same size units you want to work with in order to use the Poisson distribution.) Furthermore, you calculate the probability of a plane having no defects to be (using the formula for the Poisson distribution in Appendix A)

$$P(X = 0) = \frac{\lambda^x}{x!} e^{-\lambda} = \frac{24^0}{0!} e^{-24} = 3.775 \text{ x } 10^{-11}$$

With such a poor job of controlling the paint application process you inquire about other methods of improving defects. The plant manger says she thinks there may be another way, but it will

delay the production process. She predicts the average time to get an F-31 through the production line will be 10 months. Being ever so analytical, you assume the production time is distributed normally with a standard deviation of one month and determine the probability of an F-31 being built in under a year to be

$$P(X \le 12) = P\left(\frac{X - 10}{1} \le \frac{12 - 10}{1}\right) = P(z \le 2) = 0.9772$$

Note that $P(z \le 2)$ can be computed several ways. The best way is to let a computer do it for you, but if you insist on making life difficult, you can use a standard normal table in any statistics textbook. Be careful how the table is presented. Some tables will give values for $P(z \ge x)$ and others will give it for $P(0 \le z \le x)$. In the first case $P(z \le 2)$ can be calculated using $1 - P(z \ge 2)$. The second case is a little trickier, $P(z \le 2) = P(-\infty < z < 0) + P(0 \le z \le 2) = 0.5 + P(0 \le z \le 2)$.

6. Descriptive Statistics

It has been said, "There are three kinds of lies: lies, damned lies and statistics" (Disraeli). Unfortunately, that statement has been proven true too many times by people trying to support their claims through the improper use of statistics (mainly with polling data). Descriptive statistics attempt to describe some attribute of a population, however, depending on how the sample is drawn, the statistic may not be a good estimator of that attribute. An exaggerated example would be to do a poll on alcoholism by asking your questions at a bar. It is imperative that all samples be completely random from the population of interest (exceptions will be discussed in Level 3). Anytime a statistic is reported, the analyst should include information on the sample used. The following list runs though some of the more common descriptive statistics:

6.1 Measures of Location

Mean - average; μ for population; $\bar{x}$ for sample; also called expected value, $E(X)$

$$\mu = \begin{cases} \sum_i x_i P(x_i) \text{ , discrete} \\ \int x f(x) \mathrm{d}x \text{ , continuous} \end{cases} \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Some rules for $E(X)$: (c is a constant, X & Y are random variables)

$E(c) = c$
$E(cX) = cE(X)$
$E(X + c) = E(X) + E(c) = E(X) + c$
$E(X + Y) = E(X) + E(Y)$

Median - middle observation in ascending order; m for population; $\tilde{x}$ for sample

$$m: P(X \le m) = P(X \ge m) = 0.5 \qquad \tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & , n \text{ odd} \\ \frac{1}{2}\left(x_{\left(\frac{n+2}{2}\right)} + x_{\left(\frac{n}{2}\right)} \right) & , n \text{ even} \end{cases}$$

Mode - value that occurs with greatest frequency

100α% Trimmed Mean - average adjusted to be robust to outliers

$$\bar{x}_\alpha = \frac{x_{([n\alpha]+1)} + \ \ldots \ + x_{(n-[n\alpha])}}{n - 2[n\alpha]}$$

6.2 Measures of Variation

Range - largest measurement minus smallest measurement; very sensitive to outliers

Variance - defined as the average squared deviation from the mean; σ^2 for population; s^2 for sample

$$\sigma^2 = E(X-\mu)^2 = E(X^2) - \mu^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$$

(This method isn't good for computers because of floating point errors. See Appendix B.)

Some rules for $V(X)$: (c is a constant, X and Y are random variables)

$V(c) = 0$
$V(cX) = c^2V(X)$
$V(X + c) = V(X)$
$V(X + Y) = V(X) + V(Y) + 2\text{Cov}(X,Y)$

Standard Deviation - the square root of the variance; note that the variance is a squared deviation so it is not usually in the same units as the mean (what are lb^2?); the standard deviation takes care of that problem

Covariance - a measure of the strength of a linear relationship between two random variables; if Y tends to increase as X increases, Cov(X,Y) will be positive; if Y tends to decrease as X increases, Cov(X,Y) will be negative; this significant relationship is used in simulation to decrease overall variance and reduce the number of runs (e.g., antithetic random numbers)

$\text{Cov}(X,Y) = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - \mu_X\mu_Y$
$\text{Cov}(X,Y) = \text{Cov}(Y,X)$

Cov(X,Y) = 0 if X & Y are independent

Correlation - standardized version of the covariance; values of -1 and 1 imply perfect straight-line relationships between X and Y; a value of zero indicates no linear relationship between X and Y

$$\rho = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

6.3 Measures of Relative Standing

$100p^{th}$ Percentile - $100p$% of area under pdf or pmf lies to left (i.e., ≤) of $100p^{th}$ percentile and $100(1-p)$% lies to the right (≥); three special cases:

Midquartile - median

Lower Quartile - $Q_L = x_{\left(\frac{1}{4}(n+1)\right)}$ (round up)

Upper Quartile - $Q_U = x_{\left(\frac{3}{4}(n+1)\right)}$ (round down)

Interquartile Range - IQR = $Q_U - Q_L$

***z*-Score** - method of standardizing the data; $z = (x - \bar{x}) / s$

6.4 Histogram

One of the most common techniques for preliminary analysis of data is the histogram. It is a graphical way to represent the data which, if there is enough data, will represent the underlying distribution. Basically, a histogram contains a number of bins which contain the number of data points that lie in each bin. A relative frequency histogram contains the fraction of data points that lie in each bin which better approximates a distribution.

Most computer packages can generate histograms for you, but here are the steps just in case you like doing things the old fashioned way:

1. Calculate the **range** of the data
2. Divide the range into bins of equal width. Mendenhall presents a Rule of Thumb for determining the number of bins:

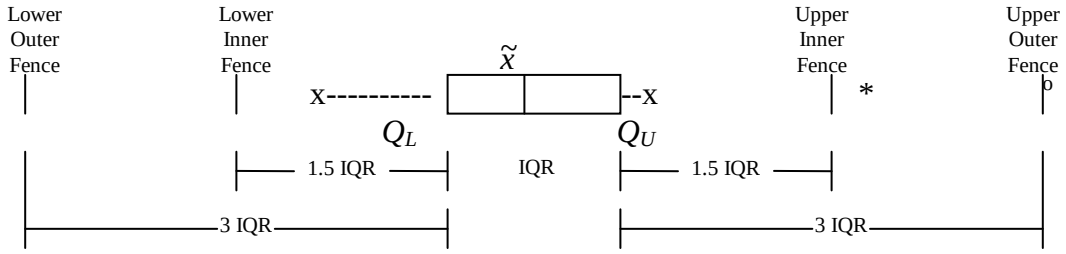
# Observations	# Bins
< 25	5 or 6
25-50	7-14
> 50	15-20

   The lower boundary of the first bin should be less than the smallest data point. Similarly, the upper boundary of the last bin should be greater than the largest data point. Finally, no data point should fall on a class boundary.
3. For each bin, count the number of observations that fall into it

4. If using relative frequencies, calculate the relative frequency of each bin as the number of data points in the bin divided by the total number of observations
5. The histogram is essentially a bar graph where the categories are the bins. For a frequency histogram, the bar heights are determined by the bin counts. In a relative frequency histogram, the heights are given by the relative frequencies of the bins

6.5 Box Plot

Whenever data is collected, the first thing an analyst should do is look for outliers. These may have an exaggerated influence on any statistics the analyst computes. An easy check is looking for |*z*-scores| > 3. Most computer packages, however, will also form box plots. Figure 15 shows an example of a box plot.



x - whiskers; most extreme points within 1.5 IQR
* - suspect outlier
o - highly suspect outlier

Figure 15. Box Plot

6.6 Example (Part III)

Review the description of the F-31 given in Section 4.3. You are diligently working on the design of your next flight test in order to make the most efficient use of your time when you are interrupted by a call from the F-31 program manager (PM). He is preparing to attend a hearing on the F-31 and wants to know the latest information on the plane's top speed. Luckily, you have some data from your previous flight test. The top speeds posted on each flight were (knots):

900, 1047, 938, 896, 1102, 973, 1015, 924, 991, 1072

All of the speeds were the maximum sustained speeds while using the afterburner and the differences are attributed to random influences of temperature, pressure, wind, weight, etc. Using this data you tell the PM that the top speed is estimated to be:

$$\frac{1}{10}(900+1047+938+896+1102+973+1015+924+991+1072) = 985.8 \text{ knots}$$

with a standard deviation of:

$$\sqrt{\frac{1}{9}\left[(900-985.8)^2+\ldots+(1072-985.8)^2\right]} = 72.5 \text{ knots}$$

Maximum speed in the air is not the only important consideration for fighter aircraft. There is also the amount of time it takes to get the plane in the air. Therefore, you were wise to collect data for the time (in minutes) it took the test pilots to run the preflight checklist:

1.17	1.61	1.16	1.38	3.53
1.23	3.76	1.94	.96	4.75
1.15	2.41	.71	2.02	1.59
1.19	.82	2.47	2.16	2.01
.92	.75	2.59	3.07	1.40

The first step in building a histogram is to calculate the range: 4.75 - 0.71 = 4.04
Now the approximate bin width is $4.04/7 = 0.577 \approx 0.6$
In order to build the histogram, you need to count up how many observations fall into each bin (the Analysis ToolPak in Excel can do this for you):

Bin	Interval	Count	Rel. Freq.
1	0.6-1.2	9	0.36
2	1.2-1.8	5	0.20
3	1.8-2.4	4	0.16
4	2.4-3.0	3	0.12
5	3.0-3.6	2	0.08
6	3.6-4.2	1	0.04
7	4.2-4.8	1	0.04

Given the table above, it's simple to finish the histograms with any graphing tool. The graphs in Figure 16 come from Microsoft Excel. Glancing at the histograms, you guess the data may be distributed exponentially although you would need more data to confirm that. Knowing the distribution of the preflight time could be useful when developing a simulation because adding the uncertainty will yield much better results than using a deterministic preflight time.

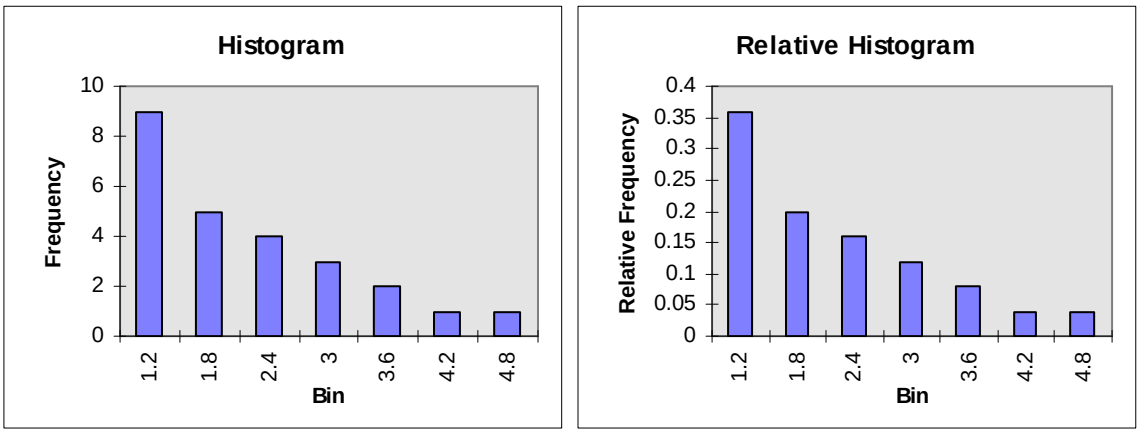


Figure 16. Histograms

7. Inferential Statistics

Inferential statistics takes descriptive statistics to the next level. In the former, the analyst merely describes the sample that was observed. Using inferential statistics, however, the analyst uses the sample to draw conclusions about the population. This is the area where it is imperative to report the process used as well as the error in the conclusion. There are two main types of inferential statistics: confidence intervals and hypothesis tests.

7.1 Hypothesis Tests

Just as important as the result of an inferential statistic are the assumptions that are made and the errors related to the statistic. A typical assumption is that the data is normally distributed. Errors are easier to view in the context of hypothesis tests, but they also apply to confidence intervals. The following lists the six elements of a hypothesis test:

Null Hypothesis (H_0) - assumption made about a population parameter (e.g., $\mu = 0$)

Alternate Hypothesis (H_a) - opposite of H_0; accepted as true if H_0 is rejected; in practice, the analyst should make H_a what is desired to be proven because Type I error is easier to quantify. The following table lists the three cases of H_0 and H_a that are normally used (the = sign can go on H_0 or H_a, but not both):

H_0	H_a
$=$	$\neq$
$\leq$	$>$
$\geq$	$<$

Test Statistic - computed from data assuming H_0 is true; the statistic will be compared to some theoretical distribution and a decision will be made as to the validity of H_0

Rejection Region (RR) - values of the test statistic that imply rejection of H_0; URL is the upper rejection limit; LRL is the lower rejection limit

Errors - Type I = $\alpha = P$(reject $H_0|H_0$ true); Type II = $\beta = P$(fail to reject $H_0|H_0$ false); closely related to Type II error is the **power** = 1 - $\beta = P$(reject $H_0|H_0$ false). Something like Figure 17 is typically used to explain statistical errors. The column headings are the actual cases while the row headings are the test results. A classical example is a medical test for a certain disease. The test can indicate that you have the disease, but there is some chance that you really do not (Type II error). Similarly, the test can say you do not have the disease when if fact you do (Type I error).

Test / Actual	TRUE	FALSE
TRUE	Good Test	Type II Error
FALSE	Type I Error	Good Test

Figure 17. Type I & II Errors

***p*-value** - observed significance level; probability, assuming H_0 is true, of observing a value of the test statistic at least as contradictory to H_0 (and supportive of H_a) as the one computed from the sample data; if $\alpha > p$-value, the analyst should reject H_0

In English, Type I error (α) is the probability of rejecting a conjecture about the data when it is actually true. Type II error (β) is the probability of failing to reject a conjecture when it is actually false. Figure 18 graphically depicts a hypothesis test with two tails. The theorized population parameter is represented by $\hat{x}$ and the rejection region is anything less than LRL or greater than URL. The actual population parameter is μ.

In the figure, the Type I error is the area in the rejection region that is underneath the curve centered on $\hat{x}$. The *p*-value is the area of all observations further than the test statistic from $\hat{x}$ under the curve centered on $\hat{x}$. The Type II error is the area outside the rejection region (i.e., between LRL and URL) that lies under the curve centered on μ.

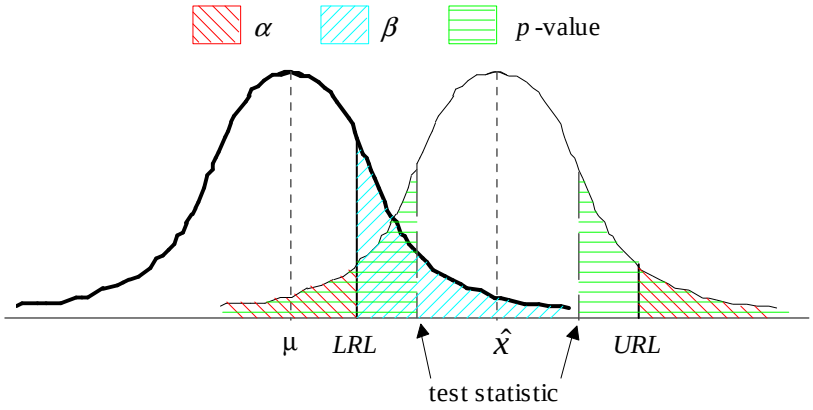


Figure 18. Errors & *p*-value

From Figure 18, it appears the null hypothesis will not be rejected because the *p*-value is greater than α. Note, however, that this conclusion will actually be incorrect because the true parameter is μ. The probability of making such a false assertion is given by β. Unfortunately, β is difficult to compute for real world problems (if you knew what the actual parameter was, you wouldn't be doing the test!). Figure 18 is a pretty extreme case and is purely hypothetical to help you picture the errors and the *p*-value.

7.2 Confidence Intervals

A confidence interval accomplishes the same as a hypothesis test, but with less formality. A CI basically computes two points (confidence limits) between which the true population parameter may be expected to lie $(1-\alpha)100\%$ of the time. A popular way to present CIs is to report the point estimate and the half-width as well as the Type I error. A simple way to perform tests is to form the CI and check if the value desired for the parameter (e.g., zero) lies in the interval. If the value does not lie in the interval, the conclusion (with $(1-\alpha)100\%$ risk) is that the desired value is not the actual value. To be statistically correct, if the value does lie in the interval, the conclusion is that the data is insufficient to say the desired value is not the actual value (since β is not known).

The rejection region for a hypothesis test is determined in much the same way as the half-width of a confidence interval. The basic starting point is the generic equation

$$P(\hat{\theta}_L \le \theta \le \hat{\theta}_U) = 1 - \alpha$$

where θ is the actual parameter and the hatted (^) θ s are the confidence limits. The trick from here is using the correct algebra to get to the desired formula. It is not the most exciting thing to do in statistics and as far as practicality goes, it's near the bottom since the generous statisticians of the past have already computed formulas for many situations. Appendix C introduces some of the definitions of theoretical distributions which are handy for developing half-widths. The appendix also covers the most common types of CIs and test statistics.

Most statistical programs are capable of generating confidence intervals and hypothesis tests. The reader should be focused on understanding the concept and the computer output than actually being able to crunch out some of the tedious calculations. As always, it is important to report the assumptions and the error (at least Type I or *p*-value) when presenting the results of CIs or hypothesis tests.

7.3 Example (Part IV)

Review the description of the F-31 given in Section 4.3. After computing the mean and standard deviation for the top speed of the F-31, you begin to suspect giving the PM such information could be hazardous. Instead you decide to give him a confidence interval in which the true top speed would be expected to lie 95 percent of the time (i.e., $\alpha = 0.05$). In other words you calculate:

$$\bar{x} \pm t_{\alpha/2,(n-1)}\left(\frac{s}{\sqrt{n}}\right) = 985.8 \pm 2.262\left(\frac{72.5}{\sqrt{10}}\right) = [933.9,\ 1037.7]$$

After the hearing, the PM calls you up and says there has been a proposal to use a new engine. Experts say the new engine will outperform the old one. It has an estimated top speed of 1034 knots with a standard deviation of 136 knots (based on 10 flights). You set up the following test using the most conservative comparison of means there is ($\sigma_1^2 \neq \sigma_2^2$ but are unknown and $n <$ 30):

H_0: $\mu_1 \ge \mu_2$
H_a: $\mu_1 < \mu_2$ (Note that you want to show the second engine is better)
For a 95% hypothesis test the rejection region comes from a *t*-statistic with

$$v = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} = \frac{\left(72.5^2/10 + 136^2/10\right)^2}{\frac{\left(72.5^2/10\right)^2}{9} + \frac{\left(136^2/10\right)^2}{9}} = 13.7 \approx 13\,df$$

Therefore the rejection region is $|t| > t_{0.025,13df} = 2.160$.
The test statistic *t* is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = \frac{985.8 - 1034}{\sqrt{72.5^2/10 + 136^2/10}} = -0.989$$

Thus, you fail to reject H_0 and say the data is insufficient to support the claim that the new engine has better performance.

This concludes the beginning of your wonderful journey into the useful, yet dangerous, world of statistics. This level attempted to present some of the basic concepts that form the foundations of statistical analysis. From here on out, statistics become much more complicated with more mathy sounding terms the further you go. Never fear! There are still two more levels of this primer to guide you through the perilous waters. Both will use this level as a springboard so make sure you understand the concepts (especially the examples). There is also a self-assessment in Appendix D with solutions in Appendix E.

Appendix A. Common Distributions

Most of the following results can be computed using simple calculus and the definitions of expected value and variance given in the main text. Some additional formulas that come in handy for the derivations are: (WARNING: Play with these at your own risk!)

$\binom{n}{r} = \frac{n!}{r!(n-r)!}$ (Combination - number of ways to group r of n objects, order not important)

$\sum_{n=0}^{\infty} ar^n = \frac{a}{1-r}$ (a is constant, -1 < r < 1)

$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$

$E(X^2) = E(X(X\text{-}1)) + E(X)$

A.1 Discrete Distributions

Bernoulli - simple yes/no or success/failure; P(success) = p; $x \in \{0, 1\}$

pdf: $p^x(1-p)^{1-x}$ $E(X) = p$ $V(X) = p(1-p)$

Binomial - x successes in n trials; P(success) = p; sum of n *iid* Bernoulli's; $x \in \{0, 1, 2, \ldots n\}$

pdf: $\binom{n}{x} p^x(1-p)^{n-x}$ $E(X) = np$ $V(X) = np(1-p)$

Geometric - x trials until first success; P(success) = p; $x \in \{1, 2, 3, \ldots\}$

pdf: $p(1-p)^{x-1}$ $E(X) = \frac{1}{p}$ $V(X) = \frac{1-p}{p^2}$

Negative Binomial - x trials until r^{th} success; P(success) = p; $x \in \{r, r+1, \ldots\}$

pdf: $\binom{x-1}{r-1} p^r(1-p)^{x-r}$ $E(X) = \frac{r}{p}$ $V(X) = \frac{r(1-p)}{p^2}$

Poisson - x successes in λ (continuous area, volume, time, etc.); $x \in \{0, 1, 2, \ldots\}$

pdf: $\frac{\lambda^x}{x!} e^{-\lambda}$ $E(X) = \lambda$ $V(X) = \lambda$

Hypergeometric - binomial with dependent trials; m = # of possible successes; x = # of desired successes; N = total # of trials possible; n = # of trials used

pdf: $\dfrac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}$ $\quad E(X) = n\left(\dfrac{m}{N}\right)$ $\quad V(X) = n\left(\dfrac{m}{N}\right)\left(1-\dfrac{m}{N}\right)\left(\dfrac{N-n}{N-1}\right)$

A.2 Continuous Distributions

Uniform - equal probability of occurrence anywhere in range [a,b]

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \le x \le b \\ 0, & \text{else} \end{cases} \qquad E(X) = \frac{a+b}{2} \qquad V(X) = \frac{(b-a)^2}{12}$$

Exponential - time until first occurrence; closely related to Poisson

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \ge 0 \\ 0, & \text{else} \end{cases} \qquad E(X) = \frac{1}{\lambda} \qquad V(X) = \frac{1}{\lambda^2}$$

Gamma (Erlang) - time until k^{th} occurrence; sum of k *iid* Exponential(λ)

$$f(x) = \frac{\lambda^k}{\Gamma(k)} x^{k-1} e^{-\lambda x} = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} \qquad E(X) = \frac{k}{\lambda} = \alpha\beta \qquad V(X) = \frac{k}{\lambda^2} = \alpha\beta^2$$

NOTE: $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} \mathrm{d}x = (\alpha-1)\Gamma(\alpha-1) = (\alpha-1)!$ (if α is integer)

Normal - symmetric about mean

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad E(X) = \mu \qquad V(X) = \sigma^2$$

Appendix B. Golub's One-Pass Method for Variance
(From OR 253 (Simulation), Stanford University, Spring 1996; Prof. Don Iglehart)

a_n = Sample Mean based on n observations
V_n = Sample Variance based on n observations

Standard Method:

$$V_n = \frac{1}{n\text{-}1}\sum_{i=1}^{n}(x_i - a_n)^2$$

$$= \frac{1}{n\text{-}1}\left(\sum_{i=1}^{n} x_i^2 - n a_n^2\right) \text{ , this method is not good computationally on a computer}$$

Golub's One-Pass Method:

1. $\tilde{V}_1 = 0$

2. $\tilde{V}_n = \tilde{V}_{n\text{-}1} + \frac{1}{n(n+1)}\left[\left(\sum_{i=1}^{n\text{-}1} x_i\right) - (n\text{-}1)x_n\right]^2$, n = 1, 2, … # of observations

   Note: $\sum_{i=1}^{n\text{-}1} x_i$ can be updated within a loop & doesn't need to be recomputed

3. $V_n = \frac{\tilde{V}_n}{n\text{-}1}$

Appendix C. Confidence Intervals and Test Statistics

As mentioned in the text, confidence intervals and hypothesis testing are closely related. Both are derived from the basic equation

$$P(\hat{\theta}_L \le \theta \le \hat{\theta}_U) = 1-\alpha$$

where θ is the actual parameter and the hatted (^) θs are the confidence limits. Here is an example of how the equation is used:

C.1 Deriving CI of mean

Parameter = μ; Estimator = $\bar{x}$; Assume the data is normally distributed (either because we know it is or because of the Central Limit Theorem).

$$\begin{aligned} P(-z_{\alpha/2} \le z \le z_{\alpha/2}) &= 1-\alpha \\ &= P\left(-z_{\alpha/2} \le \frac{\bar{x}-\mu}{\sigma_{\bar{x}}} \le z_{\alpha/2}\right) \\ &= P\left(-z_{\alpha/2}\sigma_{\bar{x}} \le \bar{x}-\mu \le z_{\alpha/2}\sigma_{\bar{x}}\right) \\ &= P\left(-\bar{x}-z_{\alpha/2}\sigma_{\bar{x}} \le -\mu \le -\bar{x}+z_{\alpha/2}\sigma_{\bar{x}}\right) \\ &= P\left(\bar{x}-z_{\alpha/2}\sigma_{\bar{x}} \le \mu \le \bar{x}+z_{\alpha/2}\sigma_{\bar{x}}\right) \end{aligned}$$

Therefore, the (1-α)100% CI can be written $\bar{x} \pm z_{\alpha/2}\,\sigma/\sqrt{n}$ (note that by CLT $\sigma_{\bar{x}} = \sigma/\sqrt{n}$)

C.2 Tricks for Half-Width

Not all derivations are as straight forward as the CI for the mean. Other parameters of interest tend to not be distributed normally. Also, the normality assumption for the mean itself may not be valid, especially if there are not many data points. For this reason, statisticians over the years have developed other theoretical distributions. The following four equations are used to derive the rest of the CIs and test statistics in this appendix. Recall that σ^2 is the population variation, s^2 is the sample variance, and *n* is the number of trials.

$\chi^2_{(n-1)} = \dfrac{(n-1)s^2}{\sigma^2}$, where $\chi^2_{(n-1)}$ is a chi-square random variable with (*n*-1) degrees of freedom (*df*)

$t_v = \dfrac{z}{\sqrt{\chi^2_v / v}}$, where t_v is a student-*t* random variable with *v df*

$F_{v_1,v_2} = \dfrac{\chi^2_{v_1}/v_1}{\chi^2_{v_2}/v_2}$, where $F_{v1,v2}$ is a random variable from an *F* distribution with (v_1,v_2) *df*

$\chi^2_{v_1} + \chi^2_{v_2} = \chi^2_{(v_1+v_2)}$

C.3 (1-α)100% Confidence Intervals

(Normal Population)

Mean:

$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$ assumes σ^2 is known

$\bar{x} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$ assumes σ^2 unknown and $n \geq 30$ (or 60)

$\bar{x} \pm t_{\alpha/2,(n-1)}\left(\frac{s}{\sqrt{n}}\right)$ assumes σ^2 unknown, $n < 30$ (or 60), and normal population

NOTE: To get a desired half-width of H, the sample size should be $n = \left(\frac{z_{\alpha/2}\sigma}{H}\right)^2$

Difference of Means:

$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ assumes σ_1^2 & σ_2^2 are known

$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ assumes σ_1^2 & σ_2^2 are unknown and $n \geq 30$ (or 60)

$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,v} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ assumes $\sigma_1^2 = \sigma_2^2$ but are unknown and $n < 30$ (or 60)

where $v = n_1 + n_2 - 2$ and $s_p = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$

$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,v}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ assumes $\sigma_1^2 \neq \sigma_2^2$ but are unknown and $n < 30$ (or 60)

where $v = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$ (round down)

$\bar{D} \pm t_{\alpha/2,(n-1)}\left(\frac{s_D}{\sqrt{n}}\right)$ matched pairs with $\bar{D}$ equal to mean of differences

NOTE: To get a desired half-width of H, make sample size $n_1 = n_2 = \left(\frac{z_{\alpha/2}}{H}\right)^2 (\sigma_1^2 + \sigma_2^2)$

Variance:

$\left[\frac{(n-1)s^2}{\chi^2_{\alpha/2,(n-1)}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,(n-1)}}\right]$ assumes normal population

Ratio of Variances:

$$\left[\frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{\alpha/2,(\nu_1,\nu_2)}},\frac{s_1^2}{s_2^2}\cdot F_{\alpha/2,(\nu_2,\nu_1)}\right]$$ assumes normal population

where $\nu_1 = n_1\text{-}1$ and $\nu_2 = n_2 - 1$

NOTE: To check for equivalent variances, test statistic should be approximately 1.

(Binomial Population)
Proportion:

$$\hat{p}\pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\quad \hat{p}=\frac{x}{n}=\frac{\text{successes}}{\text{trials}} \qquad \hat{p}\pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$ doesn't contain 0 or 1

NOTE: To get a desired half-width of H, make sample size $n=\left(\frac{z_{\alpha/2}}{H}\right)^2 p(1-p)$

($p = 0.5$ is conservative guess if p is unknown)

Difference of Proportions:

$$(\hat{p}_1-\hat{p}_2)\pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{\hat{p}_2\hat{q}_2}{n_2}},\ \hat{q}_i=1-\hat{p}_i \qquad \hat{p}_1\pm 2\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}}\ \&\ \hat{p}_2\pm 2\sqrt{\frac{\hat{p}_2\hat{q}_2}{n_2}}$$ don't contain 0 or 1

NOTE: To get a desired half-width of H, make sample sizes $n=\left(\frac{z_{\alpha/2}}{H}\right)^2 (p_1q_1 + p_2q_2)$

C.4 Hypothesis Tests

Section 7.3 gives an example of a hypothesis test. Although there are many choices for the null and alternate hypotheses, this section will only present a generic H_0 with an equality (=) to introduce the notation. The equations immediately following the test type (e.g., Mean) will be the test statistics for different circumstances. Test statistics will be assigned the symbol of a distribution (e.g., z, t, F). Actual values from the distributions will be distinguished by subscripts (e.g., $z_{\alpha/2}$, $t_{\alpha/2}$, F_α). The rejection regions will be listed for three cases ($^1 =$, $^2 \le$, & $^3 \ge$).

(Normal Population)
Mean: H_0: $\mu = \mu_0$

$$z=\frac{\bar{x}-\mu_o}{\sigma/\sqrt{n}}$$ assumes σ^2 is known

$$z=\frac{\bar{x}-\mu_o}{s/\sqrt{n}}$$ assumes σ^2 is unknown & $n > 30$ (60)

$$t=\frac{\bar{x}-\mu_o}{s/\sqrt{n}}$$ assumes σ^2 is known

RR: $^1 |z| > z_{\alpha/2}$, $^2 z > z_\alpha$, $^3 z < z_\alpha$ (or use $t_{\alpha/2,(n-1)df}$ instead of $z_{\alpha/2}$)

Difference of Means: H_0: $\mu_1 - \mu_2 = D_0$

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_o}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$ assumes σ_1^2 & σ_2^2 are known

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_o}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$ assumes σ_1^2 & σ_2^2 are unknown and $n \geq 30$ (or 60)

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_o}{s_p\sqrt{1/n_1 + 1/n_2}}$$ assumes $\sigma_1^2 = \sigma_2^2$ but are unknown and $n < 30$ (or 60)

where $s_p = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$; use $v = (n_1 + n_2 - 2)$ *df* for *t*-statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_o}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$ assumes $\sigma_1^2 \neq \sigma_2^2$ but are unknown and $n < 30$ (or 60)

use $v = \left(\frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} \right)$ *df* for *t*-statistic (round down)

$$t = \frac{\overline{D} - D_o}{s_D/\sqrt{n}}$$ matched pairs with $\overline{D}$ equal to mean of differences

use $v = (n - 1)$ *df* for *t*-statistic

RR: [1] $|z| > z_{\alpha/2}$, [2] $z > z_\alpha$, [3] $z < z_\alpha$ (or use $t_{\alpha/2,vdf}$ instead of $z_{\alpha/2}$)

Variance: H_0: $\sigma^2 = \sigma_0^2$

$$\chi^2 = \frac{(n-1)s^2}{\sigma_o^2}$$ assumes normal population

use $v = (n - 1)$ *df* for χ^2 statistic

RR: [1] $\chi^2 < \chi^2_{1-\alpha/2,v}$ or $\chi^2 > \chi^2_{\alpha/2,v}$, [2] $\chi^2 > \chi^2_{\alpha,v}$, [3] $\chi^2 < \chi^2_{1-\alpha,v}$

Ratio of Variances: H_0: $\sigma_1^2/\sigma_2^2 = 1$

$$F = \frac{s_1^2}{s_2^2}$$ assumes independent, normal populations; $s_1^2 > s_2^2$ (flip if needed)

use $v_1 = (n_1 - 1)$ *df* & $v_2 = (n_2 - 1)$ *df* for *F*-statistic

RR: [1] $F > F_{\alpha/2}$, [2] $F > F_\alpha$, [3] $F > F_\alpha$

(Binomial Population)

Proportion: H_0: $p = p_0$

$$z = \frac{\hat{p} - p_o}{\sqrt{p_o q_o / n}},\ q_0 = 1 - p_0,\ \hat{p} = \frac{x}{n}$$

$$\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$ doesn't contain 0 or 1

RR: [1] $|z| > z_{\alpha/2}$, [2] $z > z_{\alpha}$, [3] $z < z_{\alpha}$

Difference of Proportions: H_0: $p_1 - p_2 = D_0$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(1/n_1 + 1/n_2\right)}},\ \hat{p} = \frac{x_1 + x_2}{n_1 + n_2} \qquad D_0 = 0$$

$$z = \frac{\left(\hat{p}_1 - \hat{p}_2\right) - D_o}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}} \qquad D_0 \neq 0$$

for both cases: $\hat{p}_1 \pm 2\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}}$ & $\hat{p}_2 \pm 2\sqrt{\frac{\hat{p}_2\hat{q}_2}{n_2}}$ don't contain 0 or 1

RR: [1] $|z| > z_{\alpha/2}$, [2] $z > z_{\alpha}$, [3] $z < z_{\alpha}$

Appendix D. Self Assessment (Level 1)

1. What is the basic equation of probability?

2. Using the basic equation in 1 and knowing that the Dolphins are 1-6 for the games I have watched and 5-0 for games I haven't watched, what do you estimate their record to be in a season where I watch half of the 16 games?

3. Given a well shuffled deck of cards, what is the probability of drawing a queen of hearts given you just drew the king of hearts?

4. You are one of three officers competing for a great assignment, A, B, & C (you're A of course). You know exactly one of you will get the job and the other two will be stuck in their same old boring jobs. The people at MPC refuse to announce who will get the job. You're naturally curious and ask the personnel officer which if the other two will not get the assignment (you already know at least one of them will not get the job). You're assuming each of you has an equal chance of getting the job. Also, if neither B or C gets the job, the personnel officer is equally likely to name either one in response to your questions. You are told that B will not get the job. Assuming the personnel officer doesn't lie, what probably is there for you to get the job? For officer C?

5. You're running DIME to set the scope for flight test. You figure it crashes every tenth time you try to run a model (i.e., $P(\text{success}) = p = 0.1$). The computer support personnel are all away from the office. How many models can you expect to run before the system crashes?

6. Refer to question five. Instead of keeping track of the number of models you run, you figure the computer crashes every 20 minutes (following a Poisson distribution). You think you can be done in an hour. What is the probability that you will finish without the system crashing?

7. Solve question six using an Exponential distribution.

8. Calculate the mean, median, and mode for the following data:

125	135	324	675	122	246
310	194	276	132	463	123
157	198	243	157	284	300
105	95	189	357	156	164
131	227	309	415	502	306
204	264	342	175	160	215

9. Are there any outliers in the data for question 8?

10. Construct a histogram for the data in question 8.

11. When interpreting computer output for a hypothesis test, what is the general rule for rejecting the null hypothesis? (Hint: Involves *p*-value)

12. The flight surgeon comes to you concerned about the cleaning solvents being used on the F-31 (see Section 4.3). He has collected 20 air samples over the period of a month and examined them for the harmful chemical. His results show:

$$\bar{x} = 2.1 \text{ ppm (parts per million)}; \quad s = 1.7 \text{ ppm}$$

The doctor claims the workers are in danger of contracted a harmful disease if the chemical level is greater than 1 ppm. Using $\alpha = 0.05$, are the workers safe?

13. In preparation for the Super Bowl, you decide to conduct an experiment on the price fluctuations of your favorite premium beer. You use the base Class Six and Smith's to collect you data. Here is the information you collected over the last couple months:

	Class Six	Smith's
Sample Size	18	13
Mean	5.90	5.60
Standard Deviation	1.93	3.10

You want to know if there is a difference in the price variability (which will influence when you should buy your beer).

14. You read in the base paper that many of the unit commanders want to raise the speed limit to 35 MPH. The SPs, however, are fighting to keep the limit at 30 for "safety reasons." The base commander decides to put an end to the discussions by bringing you in to conduct a test. You observe 100 randomly selected vehicles going down a 30 MPH street and notice 49 violations of the speed limit. After the limit was raised to 35 MPH, you again observe 100 randomly selected vehicles. This time only 19 violated the speed limit. Construct a 99% confidence interval for (p_1 - p_2), where p_1 is the true proportion of vehicles that exceed the lower speed limit and p_2 is the true proportion of vehicles that exceed the higher speed limit.

15. Why are statistics difficult?
    a. Job Security
    b. Mathenese
    c. They aren't, it's just supposed to look hard
    d. All of the Above

Appendix E. Self Assessment Solutions (Level 1)

1. ***P(A) = m/n*** (If even *A* occurred *m* times in *n* trials, the probability that event *A* will occur in a future experiment is estimated by the ratio *m*/*n*). [7]

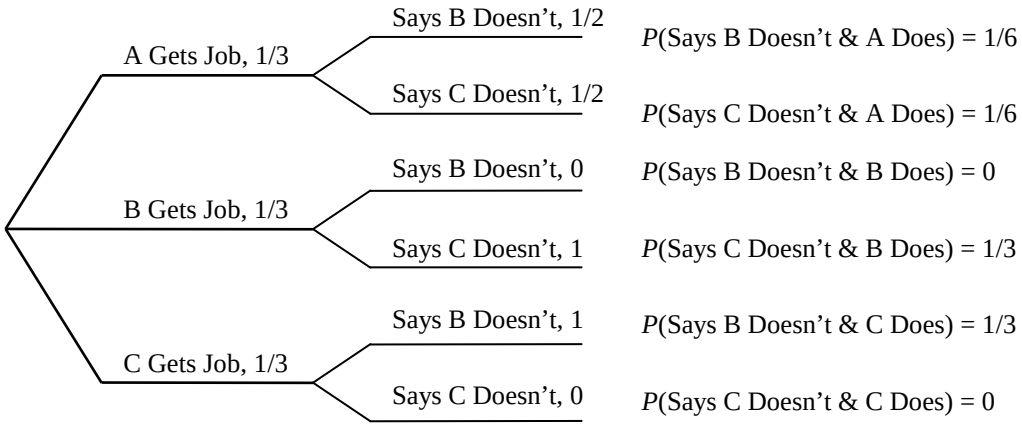
2. 8 games are played with 100% victories (8-0). The remaining games are played with a 1/6 or 16.67% chance of victory. That translates to .1667*8 = 1.33. The estimate would be about **9-7**... maybe a slim chance at a Wild Card slot in the playoffs. [7]

3. This is a trick question. There are 51 cards in the deck after drawing the king of hearts. That means the probability of drawing any one card is **1/51** regardless of what the card is (except the king of hearts which has already been drawn). [8]

4. This is a classical problem of conditional probability (originally titled "Curious Prisoners"… fitting?). Basically, from the information you can determine the following probabilities:

P(A Gets Job) = *P*(B) = *P*(C) = 1/3 (each of you has an equal chance of getting the job)
P(Says B Doesn't|A) = (if neither B or C gets the job, the personnel officer
P(Says C Doesn't|A) = 1/2 is equally likely to name either one)
P(Says B Doesn't|B) = 0, *P*(Says C Doesn't|B) = 1 (personnel officer doesn't lie)
P(Says C Doesn't|B) = 1, *P*(Says C Doesn't|C) = 0

Now build a probability tree:



Now *P*(A Gets Job|Says B) = *P*(Says B Doesn't & A Does)/*P*(Says B Doesn't) = (1/6)/(1/3+1/6) = **1/3**.
P(C Gets Job|Says B) = *P*(Says B Doesn't & C Does)/P(Says B Doesn't) = (1/3)/(1/3+1/6) = **2/3** [9]

5. The number of trials until the first success is a Geometric distributions (See Appendix A). Therefore, the expected number of runs is $E(X) = 1/p$ = **10**. [24]

6. Remember that a Poisson has to have λ in the same units you want to work with. You have $\lambda_{20} = 1$ (that is, the average number of crashes in 20 minutes is 1) and you need to convert that to λ_{60}.

$$\frac{1 \text{ crash}}{20 \text{ minutes}} \cdot \frac{3 \cdot 20 \text{ minutes}}{60 \text{ minutes}} = \frac{3 \text{ crashes}}{60 \text{ minutes}}, \text{ therefore, } \lambda_{60} = 3.$$

Now define a random variable X = number of system crashes in 60 minutes. You want to know $P(X = 0) = (\lambda^0/0!)e^{-\lambda} = e^{-3} =$ **0.0498** or roughly only 5% of the time will you finish without the system crashing. [14, 24]

7. You know an Exponential is the time until the first occurrence. Therefore, the probability that you finish is equal to the probability that the first system crash occurs after 60 minutes. You use $\lambda = 1/20$ to make the standard unit a minute.

$$P(X > 60) = \int_{60}^{\infty} \lambda e^{-\lambda x} dx = \int_{60}^{\infty} \frac{1}{20} e^{-\frac{1}{20}x} dx = \left[-e^{-\frac{1}{20}x} \right]_{60}^{\infty} = 0 - \left(-e^{-\frac{60}{20}} \right) = e^{-3} = \mathbf{0.0498} \text{ [14,25]}$$

8. Here's the sorted data:

95	105	122	123	125	131
132	135	156	**157**	**157**	160
164	175	189	194	198	**204**
215	227	243	246	264	276
284	300	306	309	310	324
342	357	415	463	502	675

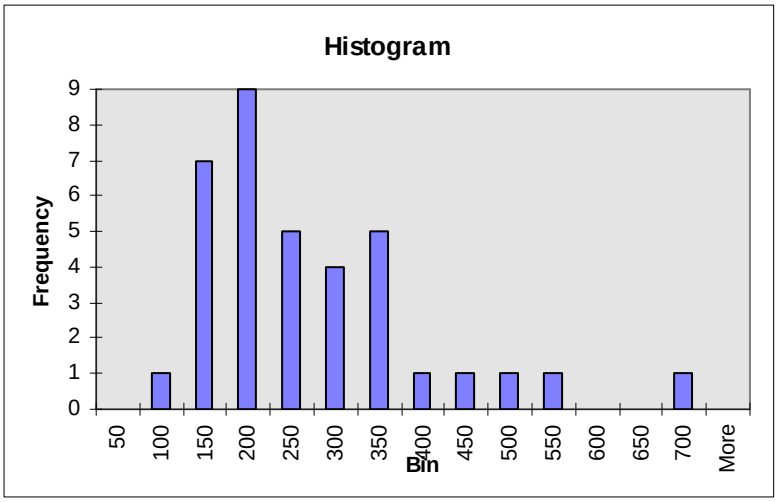
Mean = **243.8889**
Median = Average of 18th & 19th observations = .5(204 + 215) = **209.5**
Mode = **157** (only # that appears twice) [15]

9. Using s = 125.4434, you can compute the z-scores ($z = (x - \bar{x}) / s$)

-1.18690	-1.10718	-.97166	-.96369	-.94774	-.89991
-.89194	-.86803	-.70062	-.69265	-.69265	-.66873
-.63685	-.54916	-.43755	-.39770	-.36581	-.31798
-.23029	-.13463	-.00708	.01682	.16032	.25598
.31975	.44730	.495132	.51904	.52701	.63862
.78211	.90169	1.36405	1.74669	2.05759	**3.43669**

From here the last observation (675) appears to be an outlier. [17]

10. Using Microsoft Excel with bins of size 50: [17]



11. α > ***p*-value** [20]

12. This is a simple hypothesis test with a small sample ($n < 30$) and unknown σ

H_o: $\mu \leq 1$

H_a: $\mu > 1$

$$t = \frac{\bar{x} - \mu_o}{s\sqrt{n}} = \frac{2.1 - 1}{1.7\sqrt{20}} = 2.89$$

From Table 6, Appendix A (Level 2), or any stat program (including Excel) $t_{.05,19} = 1.729$
Since 2.89 > 1.729, you **reject H_o** and immediately ask the flight surgeon for suggestions on dealing with the problem. [20, 29]

13. This is a hypothesis test of a ratio of variances:

H_o: $\sigma_1^2/\sigma_2^2 = 1$

H_a: $\sigma_1^2/\sigma_2^2 \neq 1$

$$F = \frac{s_2^2}{s_1^2} = \frac{(3.10)^2}{(1.93)^2} = 2.58$$ (Note: The larger sample variance goes on top)

$F_{.05/2,\,(12,17)df} = 2.38$
Since 2.58 > 2.38, you **reject H_o** and conclude that the price variability between Smith's and the Class Six differs. [20, 30]

14. This is a simple confidence interval so all you have to do is follow the equation in Section C.3

$$\hat{p}_1 = \frac{49}{100} = 0.49; \quad \hat{p}_2 = \frac{19}{100} = 0.19$$

To make sure the technique in Section C.3 is valid, you first check the assumption:

$$\hat{p}_1 \pm 2\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}} = 0.49 \pm 2\sqrt{\frac{(0.49)(0.51)}{100}} = 0.49 \pm 10$$

$$\hat{p}_2 \pm 2\sqrt{\frac{\hat{p}_2\hat{q}_2}{n_2}} = 0.19 \pm 2\sqrt{\frac{(0.19)(0.81)}{100}} = 0.19 \pm 0.08$$

Neither interval contains 0 or 1 so you're good to go:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} = (0.49 - 0.19) \pm 2.58\sqrt{\frac{(0.49)(0.51)}{100} + \frac{(0.19)(0.81)}{100}} = 0.30 \pm 0.164$$

Since (p_1 - p_2) falls between **0.136 and 0.464**, you are fairly confident that the proportion of all vehicles in violation of the lower speed limit exceeds the corresponding proportion in violation of the higher speed limit by at least 0.136. Then you ask yourself, does the base make money off speeding tickets? [21, 29]

15. d [7]

Primer on Statistical Analysis

(Level 2)

Table of Contents

1. Introduction

Level 1 of this statistics primer reviewed many of the basic concepts covered in any college level statistics course. Hopefully, if this primer accomplished its purpose, the topics were easier to understand than having to read them in a book. Having a firm grasp of the basic terms and procedures in Level 1 is essential for any type of analysis. Granted, the material itself is a little dry and theoretical, but those concepts form the building blocks of the more practical statistics that will be covered in the remainder of the primer.

Using confidence intervals is fine for answering simple questions without dedicating a lot of resources. In the real world, however, few things are ever simple. The hypothesis tests discussed in Level 1 are also simple techniques, but they are more general and will arise over and over again in advanced statistical procedures. Some of those procedures discussed in Level 2 include nonparametric statistics, analysis of variance, categorical data, and regression. Most software packages (including spreadsheets) can perform many of these techniques. As with Level 1, this primer is intended for your reference. Don't try to memorize this stuff… people will think you're weird if you do.

2. Nonparametric Statistics

An unfortunate drawback of many of the procedures from Level 1 is the assumption that the random variable of interest comes from a normal population. In many cases, the normality assumption is either invalid (not enough data points) or incorrect (data comes from a different distribution). In other situations, the data may not be measurable, quantitative results. In such instances, as long as the data can be ranked in some order, the analyst can use statistical tests that don't rely on any assumptions about the underlying distribution. Such tests are logically called distribution-free tests. They make fewer assumptions than the tests discussed in Level 1 so they are not as powerful, but they are more applicable. Nonparametric statistics is a branch of inferential statistics devoted to distribution free tests. This section will cover a few common nonparametric techniques. If these techniques are not suitable, consult a statistics book for other tests.

2.1 Sign Test for Median

The simplest of nonparametric tests, the sign test, is specifically designed for testing hypotheses about the median of any continuous population. The test is based on taking the difference between each observation and the desired median, M_o. Two special values are defined to account for the differences: S_+ is equal to the number of positive differences (i.e., observations greater than M_o) and S_- is equal to the number of negative differences (i.e., observations less than M_o). Note that nothing is done about observations that are equal to M_o. Under the null hypothesis, S_+ and S_- come from a binomial distribution with parameters $p = 0.5$ and $n = S_+ + S_-$. The following summarizes the test:

H_o	H_a	Test Statistic	Rejection Region
$M \geq M_o$	$M < M_o$	$S = S_-$	p-value = $P[\text{Bin}(n,0.5) \geq S] < \alpha$
$M \leq M_o$	$M > M_o$	$S = S_+$	p-value = $P[\text{Bin}(n,0.5) \geq S] < \alpha$
$M = M_o$	$M \neq M_o$	$S = \max(S_+,S_-)$	p-value = $P[\text{Bin}(n,0.5) \geq S] < \alpha/2$

For large samples ($n \geq 25$), the sign test is simplified by taking advantage of the fact that the binomial distribution can be approximated by a normal. For such cases, the test statistic is

$$z = \frac{S - E(S)}{\sqrt{V(S)}} = \frac{S - 0.5n}{\sqrt{(0.5)(0.5)n}} = \frac{S - 0.5n}{0.5\sqrt{n}}$$

The rejection region is $z > z_\alpha$ for one-tailed tests and $z > z_{\alpha/2}$ for two-tailed tests. This approximation is especially handy when binomial tables (like Table 2 of Appendix A) are not available.

The sign test can also be used to test two <u>paired</u> populations by using the median of the difference between the observations. In other words, the null hypothesis will be something like H_o: $M(X - Y) = M_o$ (where $M(X - Y)$ is the median of the difference of two random variables). Note if $M_o = 0$, the test is equivalent to H_o: $X \sim Y$ (i.e., X and Y come from the same distribution).

Example. The F-31 (Section 4.3 in Level 1) is undergoing bomb range tests. Because of the tight budget, only 10 flights were performed. For each mission, the number of bombs required to penetrate a hardened bunker were: 3.5, 2.5, 3.75, 6.0, 2.5, 4.0, 5.0, 3.25, 7.0, 5.0 (units are 2,000 pound bomb equivalents). As the lead analyst on the range, you are asked to determine if 4 bombs will be sufficient for real bombing missions. Since you do not want to assume normality and 10 data points are insufficient to invoke the Central Limit Theorem, you perform the sign test as follows:

H_o: $M \leq 4$
H_a: $M > 4$ (Remember, you want to put the result you want to prove as H_a.)
The differences are: -0.5, -1.5, -0.25, 2.0, -1.5, 0.0, 1.0, -0.75, 3.0, 1.0
Test Statistic: $S_+ = 4$ and $S_- = 5$ ($n = 4 + 5 = 9$)
$P[\text{Bin}(0.5, 9) \geq S_+ = 4] = 1 - P[\text{Bin}(0.5,9) < 4] = 1 - 0.2539 = 0.7461$
Since 0.7461 is greater than any respectable α, you fail to reject the null hypothesis and conclude an F-31 carrying four bombs will take out a hardened bunker at least 50 percent of the time. You normally wouldn't make a claim that strong by failing to reject the null hypothesis, but with a p-value of 0.7461 it's a pretty safe bet.

2.2 Wilcoxon Rank Sum Test

When <u>independent</u> samples are taken to compare two populations which cannot be assumed to be normal, the Wilcoxon rank sum test is usually used. The test is also called the Mann-Whitney rank sum test by some computer programs. The first step in conducting the rank sum test is to rank the data from 1 to n with ties getting the average rank (e.g., if 4 observations tie for the 7th

smallest observation, each would be given the rank of (7+8+9+10)/4 = 8.5 and the next observation is given a rank of 11). The rankings include observations from both samples. For convenience, define population 1 to be the one with fewer observations (i.e., $n_1 \le n_2$). Also, let D_1 and D_2 represent the relative frequency distributions for populations 1 and 2, respectively. The rank sum statistics are T_1 and T_2 which, as the name implies, are merely the sums of the ranks for the observations in samples 1 and 2. The test can be summarized as follows with <, =, and > referring to the shapes of D_1 and D_2 (e.g., "$D_1 > D_2$" is read "population 1 is shifted to the right of population 2"):

H_o	H_a	Test Statistic	Rejection Region
$D_1 \ge D_2$	$D_1 < D_2$	$T = T_1$	$T \le T_L$
$D_1 \le D_2$	$D_1 > D_2$	$T = T_1$	$T \ge T_U$
$D_1 = D_2$	$D_1 \ne D_2$	$T = T_1$	$T \le T_L$ or $T \ge T_U$

where T_L and T_U come from Table 14 of Appendix A. Some statistics books will go one step further and introduce a *U*-statistic, but it's basically the same test (no sense complicating it further).

For large sample sizes (n_1 and $n_2 \ge 10$), the rank sum test can take advantage of a normal approximation similar to the one discussed in the previous section. This is useful when the T_L and T_U you need are not on the table or you don't have the table at all. The new test statistic is:

$$z = \frac{T_1 - E(T_1)}{\sqrt{V(T_1)}} = \frac{T_1 - \left[\dfrac{n_1 n_2 + n_1(n_1+1)}{2}\right]}{\sqrt{\dfrac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

The rejection regions for the three cases given in the table above are $z < -z_\alpha$, $z > z_\alpha$, and $|z| > z_{\alpha/2}$.

Example. The PM for the F-31 comes to you wondering which of two techniques is best for patching defects in the paint. According to range tests, both techniques are equally effective, so you are looking for the quicker one. The times (in minutes) for method one are 35, 50, 25, 55, 10, 30, 20. For method two the times are 45, 50, 40, 35, 46, 45, 32. At first glance you guess the first technique is going to be faster so you set up your test to prove that.

H_o: $D_1 \ge D_2$
H_a: $D_1 < D_2$
The ranks for the observations are:

Technique 1		Technique 2	
Time (min)	Rank	Time (min)	Rank
35	7.5	45	10
50	12.5	50	12.5
25	3	40	9
55	14	35	7.5
10	1	46	11
30	5	28	4
20	2	32	6
$n_1 = 7$	$T_1 = 45$	$n_2 = 7$	$T_2 = 60$

Test Statistic: $T = T_1 = 45$
Rejection Region: Using an $\alpha = 0.05$, the $T_L = 39$ and $T_U = 66$
Since 45 > 39 you cannot reject the null hypothesis and conclude that there is insufficient evidence to suggest the first technique is faster than the second. The decision must be postponed until more data is collected or must be based on other factors (e.g., cost, ease of setup, hazardous materials, etc.).

2.3 Wilcoxon Signed Rank Test (Matched Pairs)

The Wilcoxon signed rank test is similar to the rank test discussed above except that it's used for matched pairs instead of independent samples. In technical terms, a matched pairs design is a randomized block design with $k = 2$ treatments (see Section 3). In English, that means the data for the two populations considered are collected in pairs. For example, a taste test looking for differences between Coke and Pepsi will have a judge submit a score for each drink (i.e., pairs of data). The first step in the signed rank test is to get the differences between the matched pairs. Differences equal to zero are eliminated and the number of pairs n is reduced accordingly. The differences are then ranked by absolute value with ties getting the average rank. As in the sign test, special values are defined for the ranks: T_- is the sum of the ranks for the negative differences and T_+ is the sum of the ranks for the positive differences. Using the same notation as Section 2.2 ("$D_1 > D_2$" is read "population 1 is shifted to the right of population 2"), the test can be summarized as follows:

H_o	H_a	Test Statistic	Rejection Region
$D_1 \geq D_2$	$D_1 < D_2$	$T = T_+$	$T \leq T_o$
$D_1 \leq D_2$	$D_1 > D_2$	$T = T_-$	$T \leq T_o$
$D_1 = D_2$	$D_1 \neq D_2$	$T = \min(T_-, T_+)$	$T \leq T_o$

where T_o comes from Table 15 in Appendix A.

Just like the sign test in Section 2.1 was adapted for pairs, the signed rank test can be adapted for a single population median. The only adjustment is to make the differences discussed above be the differences between the observations and the theorized median M_o. Note that the signed rank test does not make the assumption of a continuous distribution as the sign test did.

Just like the other nonparametric tests, the signed rank test can also take advantage of the normal approximation for large sample sizes ($n \geq 25$). In such cases, the test statistic is:

$$z = \frac{T - E(T)}{\sqrt{V(T)}} = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

The rejection region is $z < -z_{\alpha}$ for one-tailed tests and $z < -z_{\alpha/2}$ for two-tailed tests.

Example. Review the example in Section 2.2. Assume that the data was collected in a controlled environment so each patch was done on the same type and size of defect (i.e., collected in pairs). Using the Signed Rank Test results in:

H_o: $D_1 \geq D_2$
H_a: $D_1 < D_2$
The ranks for the observations are:

	Technique			
Patch	1	2	Difference	Rank
1	35	45	-10	2
2	50	50	0	N/A
3	25	40	-15	4
4	55	35	**20**	**5**
5	10	46	-36	6
6	30	28	**2**	**1**
7	20	32	-12	3
$n = 6$				$T_+ = 6$

Test Statistic: $T = T_+ = 6$
Rejection Region: Using an $\alpha = 0.05$, the value for T_o is 2.
Since 6 > 2 you cannot reject the null hypothesis. As in Section 2.2, you conclude that there is insufficient evidence to suggest the first technique is faster than the second.

2.4 Kruskal-Wallis *H*-Test

The previous tests are only good for comparing two populations at a time. The Kruskal-Wallis *H*-test, however, is designed to compare the means of *k* populations. The test is the nonparametric equivalent of the analysis of variance (ANOVA) *F*-test (see Section 4). It is important to look at the assumptions of the Kruskal-Wallis *H*-test before trying to use it. The first assumption is a completely randomized design. This simply means that the data used comes from independent random samples of $n_1, n_2, ..., n_k$ observations from the *k* populations. Other assumptions are that each sample has at least five measurements and the observations can be ranked. Just like the Wilcoxon rank sum test, the $n = n_1 + n_2 + ... + n_k$ observations must be sorted according to rank (with ties getting the average rank). Also like the rank sum test, the ranks for each sample *i* are added to form the rank sums T_i. If the assumptions stated above hold,

the test statistic H will be approximated by a chi-square distribution with (k - 1) degrees of freedom. Here are the specifics of the test:

H_o: The k population probability distributions are identical
H_a: At least two of the k population probability distributions differ in location

Test statistic: $H = \dfrac{12}{n(n+1)} \sum_{i=1}^{k} \dfrac{T_i^2}{n_i} - 3(n+1)$

Rejection Region: $H > \chi^2_{\alpha,(k-1)}$

Example. Going back to the example in section 2.2 again, assume there are actually four techniques and you want to know if any of them is significantly better (i.e., faster) than the others. The data for the first two are the same. For method three you have 9 observations: 25, 35, 30, 45, 20, 40, 23, 34, 28. There are 8 observations for technique four: 55, 40, 46, 35, 54, 50, 32, 42. You conduct an H-test by creating the following table:

Tech. 1	Rank	Tech. 2	Rank	Tech. 3	Rank	Tech. 4	Rank
35	16.5	45	23	22	5	57	31
50	26.5	50	26.5	35	16.5	40	19.5
25	7	40	19.5	30	10.5	46	24.5
55	30	35	16.5	44	22	35	16.5
10	1	46	24.5	20	3.5	54	29
30	10.5	28	8.5	13	2	51	28
20	3.5	32	12	23	6	33	13
				34	14	42	21
				28	8.5		
T_1 =	95	T_2 =	130.5	T_3 =	88	T_4 =	182.5

From the table you can calculate H:

$$H = \frac{12}{31(32)}\left(\frac{95^2}{7} + \frac{130.5^2}{7} + \frac{88^2}{9} + \frac{182.5^2}{8}\right) - 3(32) = 9.797$$

Also, from Table 7 in Appendix A, you know that a χ^2 with $\alpha = 0.05$ and 3 *df* is 7.8147. Since 9.797 > 7.8147, you reject H_o and conclude that at least one of the four techniques is different than the others. From here you can go back and perform some of the two population tests to get more information.

3. Basics of Design of Experiments

Up to this point in the primer there has been mention of completely randomized designs and other data collection techniques. The focus has been more on what to do with the data, rather than how to collect it. That's where design of experiments (DOE) comes in. Basically DOE is a procedure for selecting sample data. If done correctly, DOE can save time and resources by

obtaining more information from smaller samples. Here are a few definitions that should be enough to get through this level of the primer (Level 3 will go more in depth on DOE):

Block - relatively homogeneous (similar) group of experimental units; observing treatments within blocks is a method of eliminating known sources of data variation (see Section 4.3)

Experimental Design - method used to assign treatments to experiment units; 4 steps:

[1] Select Factors

[2] Decide How Much Information You Want

[3] Choose Treatments & Number of Observations

[4] Choose Experimental Design

Experimental Unit - object upon which measurements are made

Factors - independent variables related to the response variable(s); factors are correlated with the response(s), hence their importance, but they do not necessarily have direct influence on the response(s) (see Section 6.1)

Level - different levels or settings of a factor; also called the factor's intensity

Replication - number of observations per treatment

Treatment - particular combination of levels for the factors involved in an experiment

Setting up an experimental design requires four steps as mentioned above. The first step involves selecting the factors. This means identifying the parameters that are the object of the study and investigating what factors have an influence on them. Usually, the target parameters are the population means associated with the factor-level combinations. Once you know what you are looking for, the next step is to decide how much you want to know about it. That is, decide on the magnitude of the standard error(s) that you desire. (The standard error of a statistic is the standard deviation of its probability distribution; e.g., for the sample mean $\bar{x}$, the standard error is $s/\sqrt{n}$).

The third step in an experimental design is to choose the factor-level combinations (i.e., treatments). Usually, each factor is only tested at two levels if its effect on the response(s) can be assumed to be linear. If the assumption cannot be made, the factor is set at three levels. Occasionally, factors may be assigned more than three levels, but it may complicate the design. Once all the treatments are decided for each factor, they are put into a design which will accomplish the desired objectives. Level 3 goes more into detail about specific designs.

4. Analysis of Variance

Once data for a designed experiment has been collected, it must be analyzed. The usual technique is some form of analysis of variance (ANOVA). The basic idea behind an ANOVA is to see whether two (or more) treatment means differ based on the means of the independent random samples. Figure 1 shows the plots for two cases with five measurements for each sample. The open circles on the left side are from the first sample and the solid circles on the right are from the second sample. Horizontal lines pass through the means for the two samples,

$\bar{y}_1$ and $\bar{y}_2$. For Case A, it seems a fair statement to say that the sample means differ. It seems right because the distance (variation) between the sample means is greater than the variation within the y values for each of the two samples. The opposite is true in Case B which suggests the sample means do not differ.

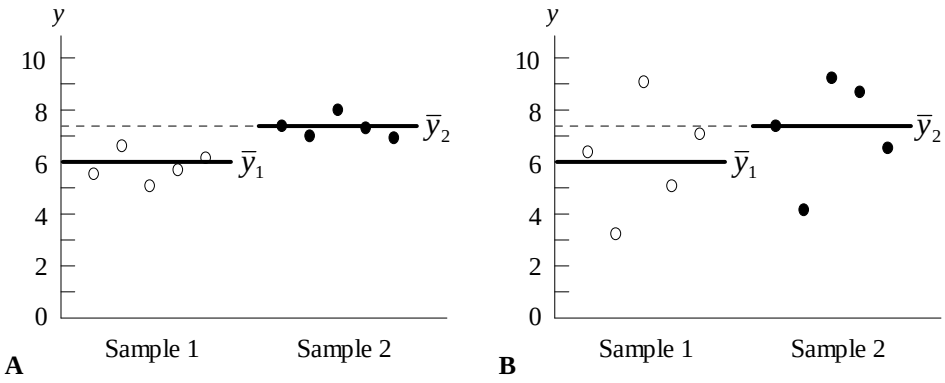


Figure 1. Plots of Data for Two Cases

4.1 One-Way Classification

As explained above, the basic idea behind an ANOVA is pretty simple. Unfortunately, when it comes time to actually do one, some math is required. Luckily, most software packages do all the calculations for you so you only need to worry about understanding the concepts in the remainder of this section. The simplest type of ANOVA is the one-way classification of a completely randomized design. Basically, that means there are *a* possible treatments to which experimental units are assigned randomly (with the same probability as the other treatments). Each treatment *i* has n_i observations $x_{i1}, x_{i2}, ..., x_{ini}$. The population mean for treatment *i* is represented by $\mu_{i\cdot}$ and the population variance by σ^2 (note there is no subscript because the variance is assumed to be constant for each treatment). Therefore, the overall population mean is given by:

$$\mu_{\cdot\cdot} = \frac{\sum_{i=1}^{a} n_i \mu_{i\cdot}}{\sum_{i=1}^{a} n_i}$$

In order to confuse you with the typical Mathenese you will find in a text book, here is what the sample mean and variance look like for treatment *i*:

$$\bar{x}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

$$s_{i\cdot}^2 = \frac{\sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_{i\cdot} \right)^2}{n_i - 1}$$

Those equations look pretty complicated, especially with the little · everywhere (it's used for two-way classifications). They are basically the same equations for sample mean and variance given in Level 1, except you only use the responses that pertain to treatment *i*. If you understand that, ANOVA will be no problem for you. Luckily, if you don't understand it, you can let a computer do all the number crunching so it doesn't matter.

All the fancy equations are nice, but what do you do with them? You use them in other fancy equations! Before listing those, however, it is probably best to take a step back and look at where they fit into the ANOVA. A one-way classification has two basic assumptions:

[1] All $x_{ij} \sim N(\mu_{i\cdot}, \sigma^2)$, $i = 1, 2, ..., a$; $j = 1, 2, ..., n_i$

[2] Model: $x_{ij} = \mu_{\cdot\cdot} + \tau_i + \varepsilon_{ij}$

where $x_{ij} = \mu_{i\cdot} + \varepsilon_{ij}$, $\varepsilon_{ij} \sim N(0, \sigma^2)$, $\tau_i = \mu_{i\cdot} - \mu_{\cdot\cdot}$, $\sum_{i=1}^{a} n_i \tau_i = 0$

Talk about some fancy equations to impress your friends! Basically, the first assumptions says that each observation comes from a normal distribution with a specific mean for the respective treatment ($\mu_{i\cdot}$) and a constant variance (σ^2). The second assumption states that the basic model for each observation is the overall mean ($\mu_{\cdot\cdot}$) plus the deviation from the mean for the respective treatment (τ_i) plus some random error term (ε_{ij}). All ANOVAs use some form of these two assumptions (unfortunately, they only get more complicated). The ANOVA for a one-way classification looks something like the following:

Source	*SS*	*df*	*MS*	*F*
Treatment	SS_T	$a - 1$	$SS_T/(a - 1)$	MS_T/MS_E
Error	SS_E	$n - a$	$SS_E/(n - a)$	
Total	SS	$n - 1$		

Figure 2. ANOVA for One-Way Classification

The SS column is for the sums of squares which represents the variability in the data caused by the source (treatment or error). If the treatment error (SS_T) is very small relative to the random error (SS_E), you would conclude that the treatment is not significant to the response variable. To put that quantitatively, there is the *F* statistic computed in the last column. It is the ratio of the mean square error for treatment (MS_T) to the mean square error (MS_E). Before moving on, it is important to note that the MS_E is an estimate of the population variance σ^2. Also, you should be warned that statisticians are notorious for developing their own special notation, especially when it comes to regression and ANOVA. Some classical statisticians will even use something called the correction for the mean which changes SS to $SS_{total(corrected)}$ and adds SS_{mean} and SS_{total}. The terms and symbols used in this primer may not be exactly what you see in a text book or computer output, but the basic concepts are the same.

Here is the formal hypothesis test for the *F* statistic:

$\boldsymbol{H_0}$**:** $\mu_{1\cdot} = \mu_{2\cdot} = ... = \mu_{a\cdot}$ (i.e., the treatments have no affect on the response)

***H*a:** $\mu_{i\cdot} \neq \mu_{j\cdot}$ for some $i \neq j$
(This is equivalent to: H_o: $\tau_1 = \tau_2 = ... = \tau_a = 0$ and H_a: some $\tau_i \neq 0$)
Test Statistic: $F = MS_T/MS_E$
Rejection Region: $F > F_{\alpha, (a-1, n-a)}$

If for some unfortunate reason, you do not have access to a computer and you have to compute the ANOVA by hand, here are the equations you will need:

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{n_i}\left(x_{ij} + \bar{x}_{i\cdot}\right)^2 = \sum_{i=1}^{a}\left(n_i - 1\right)s_{i\cdot}^2$$

$$SS_T = \sum_{i=1}^{a} n_i\left(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot}\right)^2$$

$$SS = SS_T + SS_E = \sum_{i=1}^{a}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{x}_{\cdot\cdot}\right)^2 = \sum_{i=1}^{a}\sum_{j=1}^{n_i} x_{ij}^2 - n\bar{x}_{\cdot\cdot}^2$$

Example. Refer to the F-31 description in Section 4.3 of Level 1. After new range testing, the PM brings you data which he wants analyzed. The contractor has been experimenting with different painting techniques to get a better sortie rate. The three techniques take an equal amount of time to paint the F-31, but they seem to differ in how long they last before defects form during flight. The following table lists the hours of flight time before the paint needs to be fixed or reapplied:

	Painting Technique		
	1	2	3
	148	513	335
	76	264	643
	393	433	216
	520	94	536
	236	535	128
	134	327	723
	55	214	258
	166	135	380
	415	280	594
	153	304	465
Totals	2,296	3,099	4,278

The PM wants to know if there is any significant difference between the techniques. Being an analyst on a big program like the F-31, you're very happy that you have access to computers so you don't have to do the tedious calculations by hand. First you set up the formal hypothesis test:

H_o: $\mu_{1\cdot} = \mu_{2\cdot} = \mu_{3\cdot}$
H_a: At least two of the three means differ

Then you let the computer crunch some numbers and get:

Source	*SS*	*df*	*MS*	*F*	$F_{.05,(2,27)}$	*p-value*
Treatment	198772	2	99386.2	3.48	3.35	0.045
Error	770671	27	28543.4			
Total	969443	29				

From here you conclude that you must reject H_o because $F > F_{\alpha, (a-1,n-a)}$ (the same conclusion is drawn by noting that *p*-value = 0.045 > α = 0.05). Therefore, with a 5 percent (α) chance of being wrong, you tell the PM that at least two of the three techniques differ in duration. The next section will expand on what can be done from here.

4.2 Pairwise Comparison of Means (with Extrapolation Example)

If the null hypothesis of a one-way classification is rejected, there are several other tests that can be done to gain additional information about the treatments. The most common of these is the pairwise comparison of means. Basically, the comparison checks all or some of the possible $\binom{a}{2}$ pairs of treatment means to see which ones are not equal. There are three methods normally used:

Least Significant Difference (LSD):

H_o: $\mu_{i\cdot} = \mu_{j\cdot}$ (repeat as desired for all i = 1, 2, ..., a & j = 1, 2, ..., a with $i \neq j$)

H_a: $\mu_{i\cdot} \neq \mu_{j\cdot}$

Test Statistic: $$t = \frac{\bar{x}_{i\cdot} - \bar{x}_{j\cdot}}{\sqrt{\mathrm{MS_E}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

Rejection Region: $|t| > t_{\alpha/2, (n-a)}$

Recall that a hypothesis test can also be performed as a confidence interval:

$$\left(\bar{x}_{i\cdot} - \bar{x}_{j\cdot}\right) \pm t_{\alpha/2,(n-a)} \sqrt{\mathrm{MS_E}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

Simultaneous Bonferroni CIs:

H_o: $\mu_{i\cdot} = \mu_{j\cdot}$ (test any m linear combinations up to $\binom{a}{2}$)

H_a: $\mu_{i\cdot} \neq \mu_{j\cdot}$

Test Statistic: $$t = \frac{\bar{x}_{i\cdot} - \bar{x}_{j\cdot}}{\sqrt{\mathrm{MS_E}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$ (same as LSD)

Rejection Region: $|t| > t_{(\alpha/2)/m, (n-a)}$ (note change in level of confidence)

Studentized Range (Tukey):

H_o: $\mu_{i\cdot} = \mu_{j\cdot}$ (tests all $\binom{a}{2}$ pairwise combinations)

$\boldsymbol{H_a}$: $\mu_{i\cdot} \neq \mu_{j\cdot}$

Test Statistic: $t = \dfrac{\bar{x}_{i\cdot} - \bar{x}_{j\cdot}}{\sqrt{\frac{1}{2}\mathrm{MS_E}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$

Rejection Region: $|t| > q_{\alpha,\,(a,\,n-a)}$

Requires $n_1 = n_2 = \ldots = n_a$ to be an exact test

The percentage points of the studentized range, $q(p,v)$, can be found in Tables 12 and 13 of Appendix A for $\alpha = 0.05$ and 0.01, respectively

There are other tests, such as the contrast of means test, but they are rarely used in practice. Some computer packages and text books may cover them, but it's no big loss if they don't.

Example. After performing the ANOVA in the previous section, you wisely stop yourself before going to the PM. You know at least two of the means differ, but you figure you should know which ones before presenting your findings. Since there are 10 observations for each treatment, the Tukey method for a pairwise comparison of means will be exact. You also realize that the computer software you originally used to do the ANOVA can also do the Tukey comparisons (if you tell it to). You run back to the computer and get the following results:

Tukey Comparisons

Comparison	Estimate ($\bar{x}_{i\cdot} - \bar{x}_{j\cdot}$)	95% CI	Significant
Techs 1 & 2	-80.3000	[-267.705, 107.105]	No
Techs 1 & 3	-198.200	[-385.605, -10.7949]	Yes
Techs 2 & 3	-117.900	[-305.305, 69.5051]	No

The results from the software show that only one confidence interval does not contain zero. Therefore, techniques 1 and 3 differ significantly (technique 3 lasts longer than 1 since all points in the CI are negative). Notice, however, that there is no significant difference between 1 and 2 or between 2 and 3. For the benefit of those underprivileged officers who may not have access to high tech computers, you decide to repeat the comparison of techniques 1 and 3 by hand:

H_o: $\mu_{1\cdot} = \mu_{3\cdot}$

H_a: $\mu_{1\cdot} \neq \mu_{3\cdot}$

Test Statistic: $t = \dfrac{229.6 - 427.8}{\sqrt{\frac{1}{2}28543.4\left(\frac{1}{10} + \frac{1}{10}\right)}} = -3.710$

Rejection Region: From Table 12 of Appendix A, you **extrapolate** $q_{.05,\,(3,\,27)}$ by solving:

$$\frac{x - 3.53}{3.49 - 3.53} = \frac{27 - 24}{30 - 24} \quad \Rightarrow \quad x = 3.51$$

Since 3.71 > 3.51, you reject H_o and conclude there is a significant difference between the mean duration of the paint applied by techniques 1 and 3. Note that you really did not

need to extrapolate since 3.71 is also larger than the more conservative 3.53. The extrapolation was done as a demonstration.

4.3 Two-Way Classification

A natural extension of the one-way classification is to add a second factor. Statisticians have creatively called this a two-way classification. An important application of the second factor is to account for subject variability, which will be driven home with an example. Now, just to confuse the readers, most statistics books change notation from one-way to two-way classifications. In order to avoid upsetting the statistics world too much, this section will use the most common notation. For a two-way classification there are *r* treatments and *c* blocks with one observation per block per treatment (replications will be considered later). Similar to the one-way classification, each treatment *i* has a population mean represented by $\mu_{i\cdot}$ and the population variance by σ^2 (note there is no subscript because the variance is assumed to be constant for each treatment). Also, each block *j* has a population mean $\mu_{\cdot j}$ and variance σ^2. Block and treatment sample means and variances are computed as they are for the one-way classification. The population mean for each treatment-block pair is μ_{ij} which really can't be estimated since there are no replications. A two-way classification has two basic assumptions:

[1] All $x_{ij} \sim N(\mu_{ij}, \sigma^2)$, $i = 1, 2, ..., r$; $j = 1, 2, ..., c$

[2] Model: $x_{ij} = \mu_{\cdot\cdot} + \tau_i + \beta_j + \varepsilon_{ij}$

where $x_{ij} = \mu_{ij} + \varepsilon_{ij}$, $\varepsilon_{ij} \sim N(0, \sigma^2)$, $\tau_i = \mu_{i\cdot} - \mu_{\cdot\cdot}$, $\sum_{i=1}^{r} \tau_i = 0$, $\beta_j = \mu_{\cdot j} - \mu_{\cdot\cdot}$, $\sum_{j=1}^{c} \beta_j = 0$

The interpretation of the assumptions is similar to that of a one-way classification, but a little more complicated. The two-way ANOVA looks something like the following:

Source	*SS*	*df*	*MS*	*F*
Treatment	SS_T	$r - 1$	$SS_T/(r - 1)$	MS_T/MS_E
Block	SS_B	$c - 1$	$SS_B/(c - 1)$	MS_B/MS_E
Error	SS_E	$(r-1)(c-1)$	$SS_E/(r-1)(c-1)$	
Total	SS	$rc - 1$		

Figure 3. ANOVA for Two-Way Classification

The equations given in Section 4.1 for SS_T and SS_E are the same (after adjusting the ranges of the summations). The formula for SS only requires one modification: $SS = SS_T + SS_B + SS_E$. The equation for the new term is:

$$SS_B = r\sum_{j=1}^{c}\left(\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot}\right)^2$$

Here are the formal hypothesis tests for the two *F* statistics:

$\boldsymbol{H_o}$**:** $\tau_1 = \tau_2 = ... = \tau_r = 0$ (i.e., the treatments have no affect on the response)
$\boldsymbol{H_a}$**:** some $\tau_i \neq 0$

Test Statistic: $F = \text{MS}_\text{T}/\text{MS}_\text{E}$
Rejection Region: $F > F_{\alpha,\,(r\text{-}1,(r\text{-}1)(c\text{-}1))}$

$\boldsymbol{H_o}$**:** $\beta_1 = \beta_2 = \ldots = \beta_c = 0$ (i.e., the blocks have no affect on the response)
$\boldsymbol{H_a}$**:** some $\beta_j \neq 0$
Test Statistic: $F = \text{MS}_\text{B}/\text{MS}_\text{E}$
Rejection Region: $F > F_{\alpha,\,(c\text{-}1,(r\text{-}1)(c\text{-}1))}$

You may be wondering why someone would go through the extra trouble of doing a two-way classification. There are more equations, but by blocking the data, you can remove known (or suspected) sources of variation. That means you can get the same sensitivity (MS_E) as a one-way classification using less data. If collecting data consumes a lot of resources, this is a good thing. The relative efficiency *R* of a two-way classification tells how many times as many observations you would need to obtain the same sensitivity with a one-way versus a two-way classification. The value can be found by:

$$R = \frac{\hat{\sigma}^2_{\text{one-way}}}{\hat{\sigma}^2_{\text{two-way}}} = \frac{(c-1)\text{MS}_\text{B} + c(r-1)\text{MS}_\text{E}}{(rc-1)\text{MS}_\text{E}}$$

Example. Supposed you are busy reviewing bomb run data from the F-31. You notice that there are three difference methods used to deploy the bomb in question. While organizing the data, you also notice that there are four different pilots during the test flights. The data for the bomb miss distances in meters is listed here:

		Pilot				Totals	Means
		1	2	3	4		
Method	1	4.6	6.2	5.0	6.6	22.4	5.60
	2	4.9	6.3	5.4	6.8	23.4	5.85
	3	4.4	5.9	5.4	6.3	22.0	5.50
Totals		13.9	18.4	15.8	19.7	67.8	
Means		4.63	6.13	5.27	6.57		

You decide to perform a two-way classification ANOVA to determine if either the bombing method or the pilots have a significant impact on the bomb miss distances. As shown in the table above, the bombing method is the treatment and the pilots are the blocks.

Source	*SS*	*df*	*MS*	*F*	$F_{.05,(-,-)}$	*p-value*
Method	0.26000	2	0.130000	4.18	5.14	0.073
Pilot	6.76333	3	2.25444	72.46	4.76	0.000
Error	0.186667	6	0.0311112			
Total	7.21000	11				

According to the computer output, the bombing method itself does not appear to have a significant impact on the bomb miss distance (*p*-value = 0.073 > α = 0.05). On the other hand,

the pilots have enough variation between them that it does matter which pilot is flying the mission as to what the bomb miss distance is. The critical *F* statistics in the table are computed with (2,6) *df* and (3,6) *df* for the bombing methods and pilots, respectively (in case you really feel the urge to verify the table by hand; Table 9 of Appendix A). Without a two-way classification, the impact of the pilots would have been mixed in with the bombing methods. In other words, it may have appeared that the methods themselves were significantly different, when in fact, they aren't.

4.4 Two-Way with Replications

An easy way to impress your friends and complicate the notation is to add replications to a two-way classification. Having replications is actually a good thing because you gain more information about the population. A two-way classification with $m \geq 2$ observations per cell can have the same information described above in addition to a term for interactions (a cell is a treatment-block pair). In the bomb miss distance example just discussed, the interaction term can tell whether there is a significant relationship between the pilots and the bombing methods. For example, pilots 1 and 3 might be best with method 1, but pilot 2 is best with method 3. In such a situation, there would be some interaction between the pilots and the bombing methods. If all pilots had similar standings among the methods, the interaction would not be significant.

The notation is complicated by adding a third subscript to denote the replication. Therefore, x_{ijk} is the k^{th} observation of treatment i and block j. The equations given thus far can be modified to use the new subscript by just adding over all k. Also, the equation for SS now includes the SS_I (Sum of Squares for Interaction) term as part of the sum. The changes begin with the two basic assumptions:

[1] All $x_{ijk} \sim N(\mu_{ij}, \sigma^2)$, $i = 1, 2, ..., r$; $j = 1, 2, ..., c$, $k = 1, 2, ..., m$

(Note it is μ_{ij} because the observation k does not affect the mean.)

[2] Model: $x_{ijk} = \mu_{..} + \tau_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$

where $x_{ijk} = \mu_{ij} + \varepsilon_{ijk}$, $\varepsilon_{ijk} \sim N(0, \sigma^2)$, $\tau_i = \mu_{i.} - \mu_{..}$, $\sum_{i=1}^{r} \tau_i = 0$, $\beta_j = \mu_j - \mu_{..}$,

$$\sum_{j=1}^{c} \beta_j = 0, \ \sum_{i=1}^{r} \gamma_{ij} = 0 \ \forall j = 1, 2, ..., c, \ \sum_{j=1}^{c} \gamma_{ij} = 0 \ \forall i = 1, 2, ..., r$$

Again, the interpretations of the assumptions are similar to before (but much more complicated). The important parts of the assumptions will be discussed in Section 4.6 so you don't have to worry if you can't recite these in your sleep. The revised ANOVA looks something like the following:

Source	*SS*	*df*	*MS*	*F*
Treatment	SS_T	r - 1	$SS_T/(r - 1)$	MS_T/MS_E
Block	SS_B	c - 1	$SS_B/(c - 1)$	MS_B/MS_E
Interaction	SS_I	$(r-1)(c-1)$	$SS_I/(r-1)(c-1)$	MS_I/MS_E
Error	SS_E	$rc(m-1)$	$SS_E/rc(m-1)$	
Total	SS	$rcm - 1$		

Figure 4. ANOVA for Two-Way with Replications

The equation for the new term is:

$$SS_I = m\sum_{i=1}^{r}\sum_{j=1}^{c}\left(\bar{x}_{ij\cdot} - \bar{x}_{i\cdot\cdot} - \bar{x}_{\cdot j\cdot} + \bar{x}_{\cdot\cdot\cdot}\right)^2$$

Here are the formal hypothesis tests for the three *F* statistics:

$\boldsymbol{H_o}$**:** $\tau_1 = \tau_2 = ... = \tau_r = 0$ (i.e., the treatments have no affect on the response)
$\boldsymbol{H_a}$**:** some $\tau_i \neq 0$
Test Statistic: $F = MS_T/MS_E$
Rejection Region: $F > F_{\alpha, (r-1, rc(m-1))}$

$\boldsymbol{H_o}$**:** $\beta_1 = \beta_2 = ... = \beta_c = 0$ (i.e., the blocks have no affect on the response)
$\boldsymbol{H_a}$**:** some $\beta_j \neq 0$
Test Statistic: $F = MS_B/MS_E$
Rejection Region: $F > F_{\alpha, (c-1,\ rc(m-1))}$

$\boldsymbol{H_o}$**:** $\gamma_{ij} = 0 \ \forall\ i = 1, 2, ..., r, j = 1, 2, ..., c$ (i.e., there is no interaction)
$\boldsymbol{H_a}$**:** some $\gamma_{ij} \neq 0$
Test Statistic: $F = MS_I/MS_E$
Rejection Region: $F > F_{\alpha, ((r-1)(c-1),\ rc(m-1))}$

If you are attempting to do a two-way classification with replications, you had better have a computer or you'll spend all your time performing calculations and you'll never get to the analysis part. Hopefully by now you understand how to interpret the output from an ANOVA so there's really no need for another example on two-way classifications (it also saves time, paper, and ink).

4.5 Latin Square

If you understand one and two-way classifications, it's time to step into something a little further out there (there being defined as anywhere you don't want to be). A latin square is similar to a one-way classification in that there is one factor with *a* possible treatments. In addition to that, there are two other factors with *a* treatments each (that's a total of three factors for the mathematically challenged). The extra factors set up the data in such a way that it is possible to reduce the MS_E without increasing the number of observations. It gets even better… these

factors don't necessarily have to be under your control as will be demonstrated shortly. Figure 5 shows a typical latin square design with $a = 4$.

		Factor *B*			
		1	2	3	4
Factor *A*	1	1	2	3	4
	2	4	1	2	3
	3	3	4	1	2
	4	2	3	4	1

Figure 5. Latin Square with $a = 4$

To read the table, select a cell. The number in the cell indicates what level to set the main factor. The row and column indicate the levels for the additional two factors. The basic things to remember in setting up a latin square is that there will be a^2 observations and each treatment occurs only once in each row and once in each column. As you may suspect, collecting so little data (a^2 observations) when there are so many possible combinations of factors (a^3) means that gaining additional information like interactions is not possible. On the other hand, a latin square by definition will give you an orthogonal design which is highly desirable and will be discussed in more detail in Level 3. Some text books may show designs for several values of *a*, but there is no unique way to set up a latin square.

A classic example for latin squares is a field on a farm. In order to test several different techniques, the farmer must spread them evenly over the field in order to reduce the variation caused by the conditions in the field. (Parts of the field may receive more or less water, or more or less sunlight, or have different nutrients in the soil, etc.) In this example, the main factor could be the type or amount of fertilizer or the type of seeds used. The secondary factors would be the grid coordinates of the physical location in the field. Notice that the farmer does not directly control the secondary factors, but uses them to his advantage anyway.

Hopefully you understand the basic set up and purpose for a latin square because it's time to hit the math again. Just like the previous types of ANOVA discussed, latin squares have certain assumptions:

[1] All $x_{ijk} \sim N(\mu_{ijk}, \sigma^2)$, $i = 1, 2, ..., a$; $j = 1, 2, ..., a$, $k = 1, 2, ..., a$ (not all are observed)

[2] Model: $x_{ijk} = \mu_{...} + \tau_i + \beta_j + \delta_k + \varepsilon_{ijk}$

where $x_{ijk} = \mu_{ijk} + \varepsilon_{ijk}$, $\varepsilon_{ijk} \sim N(0, \sigma^2)$, $\tau_i = \mu_{i..} - \mu_{...}$, $\sum_{i=1}^{a} \tau_i = 0$, $\beta_j = \mu_{j.} - \mu_{...}$,

$\sum_{j=1}^{a} \beta_j = 0$, $\delta_k = \mu_{..k} - \mu_{...}$, $\sum_{k=1}^{a} \delta_k = 0$

The revised ANOVA looks something like the following:

Source	*SS*	*df*	*MS*	*F*
Rows	SS_R	a - 1	$SS_R/(a - 1)$	MS_R/MS_E
Columns	SS_C	a - 1	$SS_C/(a - 1)$	MS_C/MS_E
Treatment	SS_T	a - 1	$SS_T/(a - 1)$	MS_T/MS_E
Error	SS_E	$(a-2)(a-1)$	$SS_E/(a-2)(a-1)$	
Total	SS	$a^2 - 1$		

Figure 6. ANOVA for Latin Square

Here are the formal hypothesis tests for which the three F statistics:

$\boldsymbol{H_0}$**:** $\tau_1 = \tau_2 = ... = \tau_a = 0$ (i.e., the row treatments have no affect on the response)
$\boldsymbol{H_a}$**:** some $\tau_i \neq 0$
Test Statistic: $F = MS_R/MS_E$
Rejection Region: $F > F_{\alpha, (a-1,(a-2)(a-1))}$

$\boldsymbol{H_0}$**:** $\beta_1 = \beta_2 = ... = \beta_a = 0$ (i.e., the column treatments have no affect on the response)
$\boldsymbol{H_a}$**:** some $\beta_j \neq 0$
Test Statistic: $F = MS_C/MS_E$
Rejection Region: $F > F_{\alpha, (a-1, (a-2)(a-1))}$

$\boldsymbol{H_0}$**:** $\delta_1 = \delta 2 = \ldots = \delta_a = 0$ (i.e., the main treatment has no affect on the response)
$\boldsymbol{H_a}$**:** some $\delta_k \neq 0$
Test Statistic: $F = MS_T/MS_E$
Rejection Region: $F > F_{\alpha, ((a-1), (a-2)(a-1))}$

Even though no one in their right mind would try to do stuff like this by hand, there is the occasional masochist so here are the necessary equations:

$$SS_R = a\sum_{i=1}^{a}\left(\bar{x}_{i\cdot\cdot} - \bar{x}_{\cdots}\right)^2$$

$$SS_C = a\sum_{j=1}^{a}\left(\bar{x}_{\cdot j\cdot} - \bar{x}_{\cdots}\right)^2$$

$$SS_T = a\sum_{k=1}^{a}\left(\bar{x}_{\cdot\cdot k} - \bar{x}_{\cdots}\right)^2$$

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{a}\left(x_{ijk} - \bar{x}_{i\cdot\cdot} - \bar{x}_{\cdot j\cdot} - \bar{x}_{\cdot\cdot k} + 2\bar{x}_{\cdots}\right)^2$$

$$SS = SS_R + SS_C + SS_T + SS_E = \sum_{i=1}^{a}\sum_{j=1}^{a}\left(x_{ijk} - \bar{x}_{\cdots}\right)^2$$

Don't those look like fun? Unfortunately, many software programs still do not incorporate latin squares so you may have to find out just how much fun it really is. If you have access to software that can perform two-way ANOVAs though, you might be able to persuade it do

perform a latin square… if you ask it nicely. The way to do that is to duplicate and rearrange the data so the rows correspond to the main treatment. In other words, reorganize the data so all the values in the first row correspond to the first treatment, all the data points in the second row correspond to the second, etc. Now run a two-way using the original data (i.e., rows and columns are the treatment and block for the two-way). From here you can extract the values for SS, SS_R, and SS_C. Next you have to run another two-way on the rearranged data. The new term will be SS_T (if you're quick, you'll notice the other numbers are the same ones you got in the previous ANOVA). The final step is to compute $SS_E = SS - SS_R - SS_C - SS_T$. With all these values it's pretty easy to fill in the rest of the table. The example below shows how to do this trick in Microsoft Excel.

Oh, things can get more complicated if you wish. Just like the one-way ANOVA discussed in Section 4.1, once you determine a treatment is significant, you probably want more information. Back then there was something called pairwise comparisons. That is exactly what you will do with the data from the latin square… see some things in the statistics world are simple. The way statisticians hold their job security is by not telling you what the subtle differences are. For example, you use MS_E as the estimated variance for all the comparisons which means you now use [a,(a-2)(a-1)] degrees of freedom instead of what is written is Section 4.2.

Example. You are probably thinking that all this stuff doesn't sound easy, but when you have the computing power, a latin square is actually your friend. It was already shown that a latin square can be used for physical areas (farm example), but it can also account for time. Continuing the F-31 example, you are approached by the program director who is concerned about the safety of the workers in the paint shop. They wear protective gear, but it is only designed to protect up to certain tolerances and regardless of the suits, it is always best to keep levels of dangerous chemicals as low as possible (as to not offend the environmentalists). There are five distinct processes that involve the use of a certain hazardous chemical. You need to do some preliminary analysis before tackling such a large problem so you look over historical data of recorded amounts of the chemical in the air (in parts per million, ppm) and come up with a latin square design with $a = 5$:

	Day					
Week	M	T	W	Th	F	Mean
1	18 (D)	17 (C)	14 (A)	21 (B)	17 (E)	$\bar{x}_{1\cdot\cdot} = 17.4$
2	13 (C)	34 (B)	21 (E)	16 (A)	15 (D)	$\bar{x}_{2\cdot\cdot} = 19.8$
3	7 (A)	29 (D)	32 (B)	27 (E)	13 (C)	$\bar{x}_{3\cdot\cdot} = 21.6$
4	17 (E)	13 (A)	24 (C)	31 (D)	25 (B)	$\bar{x}_{4\cdot\cdot} = 22.0$
5	21 (B)	26 (E)	26 (D)	31 (C)	7 (A)	$\bar{x}_{5\cdot\cdot} = 22.2$
Mean	$\bar{x}_{\cdot1\cdot} = 15.2$	$\bar{x}_{\cdot2\cdot} = 23.8$	$\bar{x}_{\cdot3\cdot} = 23.4$	$\bar{x}_{\cdot4\cdot} = 25.2$	$\bar{x}_{\cdot5\cdot} = 15.4$	$\bar{x}_{\cdots} = 20.6$

The designations A through E label the different processes. The various sample means are included for those sick people who like to try things by hand. In addition to those, you will need to go through the table and get sample means for the process treatments. Those are:

$\bar{x}_{\cdot\cdot 1} = 11.4$ $\bar{x}_{\cdot\cdot 4} = 23.8$
$\bar{x}_{\cdot\cdot 2} = 26.6$ $\bar{x}_{\cdot\cdot 5} = 21.6$
$\bar{x}_{\cdot\cdot 3} = 19.6$

Now you can go through all those equations or you can skillfully demonstrate your prowess on the computer and come up with the following results (see Appendix B for the Excel calculations):

Source	*SS*	*df*	*MS*	*F*	$F_{.05,(4,12)}$	*P-value*
Rows	82	4	20.5	1.30573	3.25916	0.32257
Columns	477.2	4	119.3	7.59873	3.25916	0.00273
Treatment	664.4	4	166.1	10.5796	3.25916	0.00066
Error	188.4	12	15.7			
Total	1412.0	24				

The information above indicates that both the processes (treatments) and the days of the week (columns) cause significant variation in the data. If a simple one-way classification was done instead, the fact that days of the week are significant would not have been noticed. There are several explanations why the days may be important. Two simple ones would be that the workers aren't as productive on Mondays and Fridays. Another one is that the chemical may build up during the week. These situations would have to be investigated further. Knowing that there is also variation caused by the processes, you can perform more analyses to determine which processes are releasing the largest amounts of the hazardous chemical. The Tukey comparisons discussed in Section 4.2 can provide that information. Remember that the degrees of freedom will be different (5 & 12 in this case).

4.6 Failure of Assumptions

After mention of it in Section 2, the many assumptions made for each ANOVA should have been dreadfully obvious. Luckily, those assumptions are valid in most cases. If they do not hold, however, the tests may not be accurate. This section will review the assumptions of normality and constant variance for the error terms (ε). Of course, working with the actual error terms is impossible because you need to know the actual population parameters to compute ε. As you may already suspect, we have to estimate the error terms. Those estimates are called **residuals** (e) and are defined to be the difference between the observed values and the fitted values. An **observed value** is the data that is collected while **fitted values** are those computed by the model developed from the data.

The normality assumption is used to derive the *F*-tests for an ANOVA. The easiest way to verify the assumption is to compute the residuals and then calculate standardized residuals (e/MS_E) which should be distributed as a standard normal. From here, there are three techniques. The first is to check for standardized residuals greater than 3 in absolute value (recall that 99 percent of them should fall within the interval from -3 to 3). The second test involves forming a histogram and examining the shape. The width of each bin is extremely important because it has a serious impact on the shape of the histogram (some software programs can determine widths

for you). The most difficult and most accurate test is to calculate a Q-Q plot as described in Section 5.3 of Level 1. Luckily, this option requires just as many keystrokes on a computer as the other ones.

If the checks for normality indicate that the residuals are not normal there are two options. The first is to ignore the problem. Before you get too excited, this option is only available if it's only a moderate departure from normality. The reason for ignoring the problem is that the test statistics will only differ slightly from what they would be if the assumption was valid. If there is a serious departure from normality, the second option requires the nonparametric *F*-test. The first step in performing this test is to rank all the observations in increasing order with ties getting the average rank (see Section 2.2). Then repeat the ANOVA using the ranked data.

If the normality assumption turns out to be valid based on the checks discussed above, you're not out of hot water yet because you still have to check for constant variance. That assumption is used to prove that the MS_E is an unbiased estimate of the population variance for the error terms. One way to test the assumption is to perform Bartlett's Test which extends a simple comparison of two variances (see Section C.4 in Level 1) to *a* variances:

H_0: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_a^2$

H_a: some $\sigma_i^2 \neq \sigma_j^2$

Test Statistic: $X^2 = M/C$

$$M = \sum_{i=1}^{a}\left[\left(n_i - 1\right)\ln\left(\mathrm{MS_E}\right)\right] - \sum_{i=1}^{a}\left[\left(n_i - 1\right)\ln\left(s_i^2\right)\right]$$

$$C = 1 + \frac{1}{3(a-1)}\left(\sum_{i=1}^{a}\frac{1}{n_i - 1} - \frac{1}{\sum_{i=1}^{a}(n_i - 1)}\right)$$

Rejection Region: $X^2 > \chi^2_{\alpha,\,(a\text{-}1)}$

Unfortunately, Bartlett's Test only tells whether the constant variance assumption is valid or not. It does not suggest any ways to remedy the situation. The Bartlett Test is also unthinkable without a computer so a graphical method is sometimes employed to verify the assumption. The graphical technique basically tries to find a pattern in the plot of the residuals versus the fitted values. Figure 7 summarizes the most common departures and corrections for the constant variance assumption. There are many other departures as well as many other tests, including some that get nasty enough to use derivatives. The material in this section should be enough for most cases. If you're unfortunate enough to encounter a case where this is not enough, consult a statistics text book and request divine intervention.

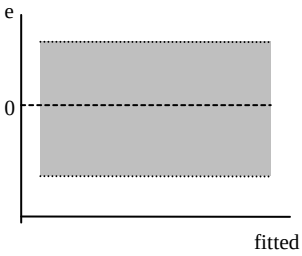
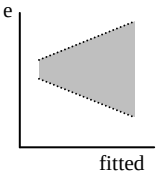
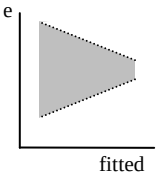
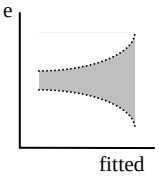
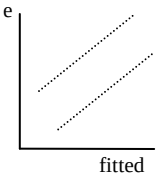
Type of Residuals	Plot	Correction
$e \sim N(0,\sigma^2)$		None Needed (it's the assumption)
σ^2 increases with fitted values		ln transform data; redo ANOVA
σ^2 decreases with fitted values		ln transform data; redo ANOVA
Poisson		$\sqrt{\ }$ transform data; redo ANOVA
Binomial	There is no set plot for binomial residuals. 	$\sin^{-1}\sqrt{\ }$ transform; redo ANOVA

Figure 7. Graphical Method for Non-Constant Variance

5. Categorical Data

The ANOVA just discussed is a very powerful technique, but there is a large class of data for which it is invalid. Because of the normality assumption, ANOVA technically cannot be performed on discrete data, like counts for surveys. The most common type of categorical data comes from a multinomial distribution. That's a fancy name for a generic finite discrete probability distribution with k possible outcomes (e.g., binomial has $k = 2$). The population parameters of interest are $p_1, p_2, ..., p_k$, where p_i is the probability of the i^{th} outcome. As you

would expect, $p_1 + p_2 + ... + p_k = 1$ (see Level 1, Section 4.1). A multinomial variable may also be called a qualitative variable because the only information it provides is which of the k bins it belongs to. The bin itself can provide further information (e.g., Pepsi drinker, Coke drinker, etc.).

5.1 One-Way Tables

If there is only one qualitative variable for an experiment, the data is arranged in a one-way table as shown in Figure 8.

Category	1	2	...	k	Total
Count	n_1	n_2	...	n_k	n
Proportion	$\hat{p}_1$	$\hat{p}_2$	...	$\hat{p}_k$	1

Figure 8. One-Way Table of Category Counts

The values $n_1, n_2, ..., n_k$ represent the category counts and $n = n_1 + n_2 + ... + n_k$ is the total number of observations. It is simple to estimate the population probabilities discussed above because a multinomial experiment can always be reduced to a binomial experiment by isolating one category. For example, the estimate for the i^{th} category is

$$\hat{p}_i = \frac{n_i}{n}$$

Similar to a binomial distribution, when n is large, $\hat{p}_i$ will be approximately normally distributed with

$$E(\hat{p}_i) = p_i$$

and

$$V(\hat{p}_i) = \frac{p_i(1-p_i)}{n}$$

You may recall from Level 1 that you can do simple confidence intervals (or hypothesis tests) for individual population proportions as well as for differences between any pair of proportions. As a reminder, if n is large, the $(1 - \alpha)100\%$ confidence interval for p_i is

$$\hat{p}_i \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{n}}$$

which is exactly the same as the CI for a binomial proportion given in Section C.3 of Level 1. A difference of proportions is a little more difficult because they are no longer independent. It can be shown that $\text{Cov}(n_i,n_j) = -np_ip_j$ and $\text{Cov}(\hat{p}_i, \hat{p}_j) = -p_ip_j/n$. These come in handy in calculating the variance of the difference between $\hat{p}_i$ and $\hat{p}_j$

$$\text{V}(\hat{p}_i - \hat{p}_j) = \text{V}(\hat{p}_i) + \text{V}(\hat{p}_j) - 2\text{Cov}(\hat{p}_i, \hat{p}_j)$$

Putting in the values and applying your knowledge form Level 1 (don't panic), you can derive the CI for $(p_i - p_j)$

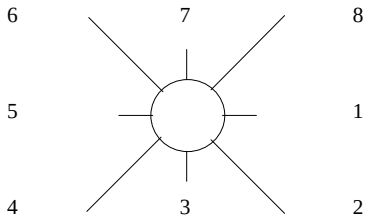
$$\left(\hat{p}_i - \hat{p}_j\right) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_i(1-\hat{p}_i) + \hat{p}_j(1-\hat{p}_j) + 2\hat{p}_i\hat{p}_j}{n}}$$

The previous two confidence intervals are useful, but it can be pretty tedious to calculate one for all k population proportions and all $\binom{k}{2}$ pairs of proportions. That's where some nasty math comes into play with a weighted sum of squared deviations between observed and expected cell counts… a good topic for conversation at parties. It sounds impressive, but it's really not that difficult (especially if a computer is doing all the number crunching). Here is a summary of the hypothesis test for all population proportions:

H_o: $p_1 = p_{1,o}, p_2 = p_{2,o}, ..., p_k = p_{k,o}$ ($p_{i,o}$ is hypothesized value for category i)
H_a: some $p_i \neq p_{i,o}$
Test Statistic: $X^2 = \sum_{i=1}^{k} \frac{\left[n_i - E(n_i)\right]^2}{E(n_i)}$, where $E(n_i) = np_{i,o}$
Rejection Region: $X^2 > \chi^2_{\alpha,\,(k-1)}$

The only assumption this test makes is that $E(n_i) \geq 5$ for all i. That's not asking too much is it? Again, it may look complicated (all things with cool formulas do), but it's pretty simple. The following example will prove it.

Example. Continuing with the F-31 example, you are looking at the Viper's susceptibility to detection. The plane was designed to have a typical four spike signature as shown here:



That is, the strongest reflected signal is at 45 degrees from the F-31's heading (marker 1). For simplicity you only consider eight bearings for the tracking radar sites: 1 is the F-31's heading and each bearing is 45 degrees from the next. The contractor claimed that 90 percent of all detections would come from the even bearings (45 degrees off). You want to verify that claim because it will make mission planning easier (you'll know how to fly the missions to avoid detections). Assuming a uniform spread of detections between the four large spikes (and four smaller spikes), the expected proportions of detections are:

	Bearing							
	1	2	3	4	5	6	7	8
P(Detect)	0.025	0.225	0.025	0.225	0.025	0.225	0.025	0.225

You have data from the flight range that tells you how many detections there were from each bearing:

	Bearing								
	1	2	3	4	5	6	7	8	Total
# Detects	6	17	5	19	6	22	5	18	98

You set up the hypothesis test as discussed in this section:

H_o: $p_1 = 0.025$, $p_2 = 0.225$, ..., $p_8 = 0.225$
H_a: some $p_i \neq p_{i,o}$
Rejection Region: $X^2 > \chi^2_{\alpha, 7df} = 14.0671$ (Table 7, Appendix A)

$$X^2 = \sum_{i=1}^{k} \frac{\left[n_i - E(n_i)\right]^2}{E(n_i)} = \frac{(6 - 98 \cdot 0.025)^2}{98 \cdot 0.025} + \frac{(17 - 98 \cdot 0.225)^2}{98 \cdot 0.225} + \ldots + \frac{(18 - 98 \cdot 0.225)^2}{98 \cdot 0.225} = 17.92$$

Since 17.92 > 14.0671, you reject the null hypothesis. Having access to those high tech computers, you'd probably just look at the *p*-value (i.e. $P(\chi^2_{.05,7}) \geq 17.92$) which in this case is 0.01234. Granted this is only a test based on 98 detections at the range under "operational" conditions so it does not necessarily prove the contractor failed to meet the specifications (if it didn't meet them you wouldn't have the plane in OT&E).

5.2 Two-Way Tables

Occasionally, you may have more than one type of category. A classic example is an election poll where the data is collected based on political party of the individuals and the candidate they plan to vote for. Another involves breaking out survey results based on demographic data. The generic layout out of such a two-way or contingency table is shown in Figure 9.

		Column				Row
		1	2	…	c	Totals
Row	1	n_{11}	n_{12}	…	n_{1c}	R_1
	2	n_{21}	n_{22}	…	n_{2c}	R_2
	⋮	⋮	⋮		⋮	⋮
	r	n_{r1}	n_{r2}	…	n_{rc}	R_r
Column Totals		C_1	C_2	…	C_c	n

Figure 9. Two-Way Table of Category Counts

where n_{ij} is the number of observed counts for row i and column j
$C_j = n_{1j} + n_{2j} + \ldots + n_{rj}$
$R_i = n_{i1} + n_{i2} + \ldots + n_{ic}$

$$n = C_1 + C_2 + \ldots + C_c = R_1 + R_2 + \ldots + R_r = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}$$

The table in Figure 9 can also be done as a contingency table of probabilities as shown in Figure 10. Of course, you will never know the actual probabilities with certainty, so they are usually estimated by the respective counts divided by the total number of observations *n* in which case all the *p*'s should be hatted (^).

		Column				Row
		1	2	…	*c*	Totals
	1	p_{11}	p_{12}	…	p_{1c}	p_{R1}
	2	p_{21}	p_{22}	…	p_{2c}	p_{R2}
Row	⋮	⋮	⋮		⋮	⋮
	r	p_{r1}	p_{r2}	…	p_{rc}	p_{Rr}
Column Totals		p_{C1}	p_{C2}	…	p_{Cc}	1

Figure 10. Two-Way Table of Category Probabilities

where p_{ij} is the expected probability of the event in cell (i,j) occurring
$p_{Ri} = p_{i1} + p_{i2} + \ldots + p_{ic}$ = probability of an event occurring in row *I* (marginal prob.)
$p_{Cj} = p_{1j} + p_{2j} + \ldots + p_{rj}$ = probability of an event occurring in column *j* (marginal prob.)

The objective of a contingency table is to determine whether the two classifications (rows & columns) are dependent. This tool is not as powerful as the one-way table which can test for specific probabilities, but it still has its uses. Going back to the definition of statistical independence discussed in Level 1 (Sections 2 and 4.2), recall that $P(A \cap B) = P(A)P(B)$ if *A* and *B* are independent. Applying that knowledge to Figure 10, if the rows and columns are independent you would expect $p_{ij} = p_{Ri}p_{Cj}$. This is the assumption of the null hypothesis for the formal test:

H_0: The two classifications (rows & columns) are independent
H_a: The two classifications are dependent

Test Statistic: $X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left[n_{ij} - E(n_{ij})\right]^2}{E(n_{ij})}$, where $E(n_{ij}) = n\hat{p}_{Ri}\hat{p}_{Cj} = n\frac{R_i}{n}\frac{C_j}{n} = \frac{R_iC_j}{n}$

Rejection Region: $X^2 > \chi^2_{\alpha,\,(r-1)(c-1)}$

The degrees of freedom in the rejection region may seem odd, but it actually comes from some tricky algebra added to the knowledge of what is known and what is estimated. Basically you lose 1 *df* because all cell counts must equal *n*. You also lose (*r*-1) and (*c*-1) *df* because all but one of the marginal probabilities for the rows and columns must be estimated. The last one is automatically determined because the sum must equal 1. So, starting with *rc* possible degrees of freedom, you end up with *rc* - 1 - (*r*-1) - (*c*-1). Now test your algebra and see if you can come up with (*r*-1)(*c*-1). As in the one-way table, there is still the assumption that $E(n_{ij}) \geq 5$ which can lead to problems as discussed in the example.

A slight change must be made to the hypothesis test above if there are fixed marginal totals (i.e., a set number of observations for either the rows or the columns). That is, $R_1 = R_2 = \ldots = R_r$ or $C_1 = C_2 = \ldots = C_c$. In such a situation, the null hypothesis changes to say that the proportions in each cell do not depend on the row or column (depending on which is fixed).

Example. Expanding on the example in Section 5.1, you realize that mission planning can also be made easier if you can distinguish which types of radar to avoid. Again, you sort through the flight test data and class it according to mission profile (low, medium, high) and which type of radar made the detection (1, 2, 3). The data is shown here

		Profile			Row
		Low	Medium	High	Totals
	1	6	17	6	29
Radar	2	5	5	13	23
	3	11	12	23	46
Col Totals		22	34	42	98

From here it's a pretty straight forward number crunching problem. If you don't have specific software that automatically computes the values for you, it's not too hard to get a spreadsheet like Excel to do it. There is no need to rewrite the null and alternate hypotheses as they are the same every time for this type of test. The rejection region, however, is test specific and in this case it's $X^2 > \chi^2_{.05,\,(2)(2)} = 9.4877$. The test statistic can be tedious to calculate by hand (and it takes up lots of space) so here is the abridged version

$$X^2 = \frac{\left[6-(29)(22)/98\right]^2}{(29)(22)/98} + \frac{\left[17-(29)(34)/98\right]^2}{(29)(34)/98} + \ldots + \frac{\left[23-(46)(42)/98\right]^2}{(46)(42)/98} = 11.844$$

Since 11.844 > 9.4877, you reject the null hypothesis and conclude that there is some interaction between the types of radar and the mission profiles with respect to the number of detections. Looking at the data, it looks like you want to avoid radar 1 at medium altitudes and radar 3 just about everywhere. There are other tests that you can conduct to be more specific (don't worry, all the tools are in the primer somewhere… you just have to find them). Incidentally, the *p*-value for this test is $P(\chi^2_{.05,\,(2)(2)} > 11.844) = 0.01855$.

6. Linear Regression

One of the most frequently used (and misused) statistical tools is linear regression. The usual nomenclature (ooo, big word) states the definition as a technique that develops a functional relationship between a response (dependent) variable and one or more explanatory (independent) variables. See Appendix C for the hard core math stuff.

6.1 The Basics

A critical subtlety about the definition is that "functional relationship" DOES NOT imply a causal relationship. That's why it is preferred to use the terms response/explanatory variables versus dependent/independent variables. In fact, in most cases the variables can be swapped however you please because regression is just accounting for the correlation between the variables. One way to look at this with a real world example is regressing the number of violent crimes with the number of police officers in a city. Of course, you expect the two to increase with one another (just watch CNN for proof). You can argue that as crimes increase the city is forced to hire more officers. On the other hand, if there are more officers, there will be more arrests which make the crime statistics go up. So which causes which? It doesn't matter… at least not as far as the regression is concerned. Assuming the response is caused by the explanatory variable(s) is probably the biggest error most people make in regression. Basically, regression can tell you what is expected to happen to a certain variable given the observed values of other variables. That's it.

With that pet peeve out of the way, a basic linear regression model looks something like the following:

$$y_i = \beta_o + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i \quad (i = 1,2,\ldots,n)$$

This shows a response variable y which is related to the k explanatory variables $x_1, x_2, \ldots, x_k$ as determined by the $k' = k+1$ parameters $\beta_o, \beta_1, \ldots \beta_k$. The term β_o is usually called the intercept and the other parameters are called partial regression coefficients or just coefficients for short. The intercept is the theoretical value that y will take on when all the explanatory variables are set to zero. In practice this definition is not valid if the data does not span the origin (see Section 6.4). The partial regression coefficient β_j shows the change in the value of y when x_j is increased by one unit and all other explanatory variables are held constant. The final term, ε, is the error term which will be discussed in the next section. The subscript i is just an index which denotes the number of the observation (out of a total of n).

One important thing to notice about the model is that it's linear in the parameters. That means there are no terms were some $\beta_j x_j$ is raised to the β_m power. Ready to be confused?… It would be fine if any $\beta_j x_j$ term were raised to some constant because then you can just introduce $\beta_l = \beta_j^c$ and $x_l = x_j^c$ which gives $\beta_l x_l$ ("linear!"). Yes, believe it or not, y can change in a nonlinear way with respect to any x_j (e.g., $y = x^2$). That's because the "linear" in linear regression means that it's linear in parameters… oh, those tricky statisticians. See Section 6.7 for more discussion on nonlinear "linear" regression.

To drive the point home, from now until Section 6.8, most concepts will be explained with the simplest linear regression model… only one explanatory variable ($y_i = \beta_o + \beta_1 x_i + \varepsilon_i$). If you run such a regression, the result can be that the response variable, y, increases linearly with x. That is, for each unit increase in x, y will increase by some constant (β_1). Similarly, y can decrease linearly with x. And as was just discussed, y can also change in a nonlinear way with respect to x. Figure 11 shows some common cases that are possible in linear regression (see Section 6.7 for the nonlinear cases).

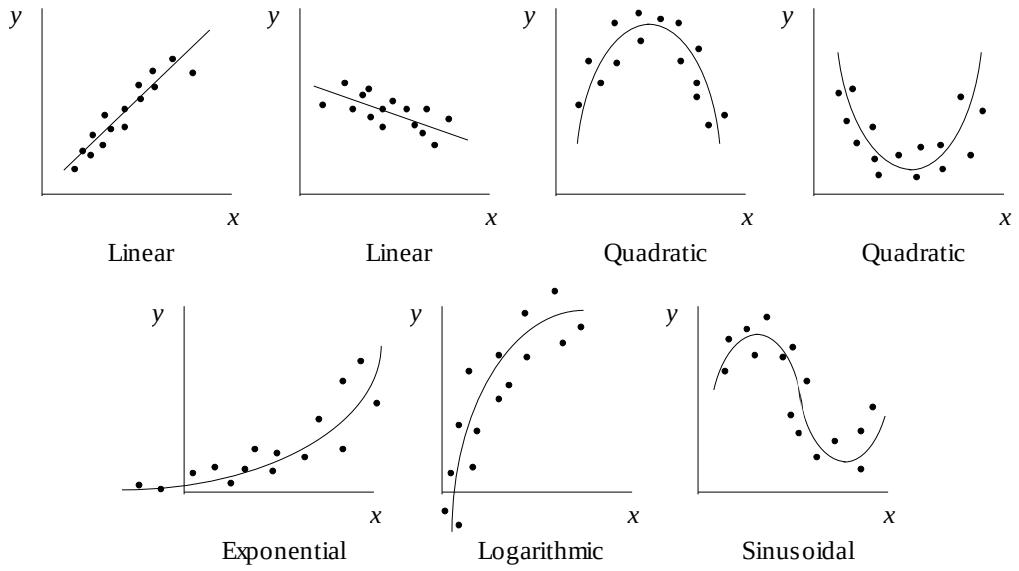


Figure 11. Examples of "Linear" Regression

6.2 Least Squares

If you understand everything so far, you're ready to move on to the next hardest concept in linear regression… how to do it. The whole point of regression is to figure out what all the parameters (the β's) are so you can figure out the relationship between the variables. Obviously you don't know the actual relationship (or you wouldn't need the regression), so you have to estimate the parameters. Once they're estimated of course, they get their party hats (^). That means the estimated regression equation will look like this

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_k x_{ki} + e_i \ \ (i = 1,2,\ldots,n)$$

Or for the one variable case:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \ \ (i = 1,2,\ldots,n)$$

Notice that the error term is now an e instead of an ε because it is also estimated. The e is called a residual and is the same as the one discussed in the ANOVA section. As a refresher, a **residual** is the difference between the **actual value** (the data point) and the **fitted value** (the one on the line). Speaking of fitted values, those get little hats (^) too because they are the result of an equation based on estimates. The easy way to picture the fitted value is to think of it being on the regression line, but it is actually the expected value of the response given the values of the explanatory variable(s). In mathy terms, that's $\hat{y}_i = E(y|x_1,x_2,\ldots,x_k)$. It's basically the same equation as before but you drop the error term:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \ \ (i = 1,2,\ldots,n)$$

There are actually many ways to perform regression, but by far the most common is least squares. Basically, least squares means you minimize the sum of the squared error terms (see Figure 12).

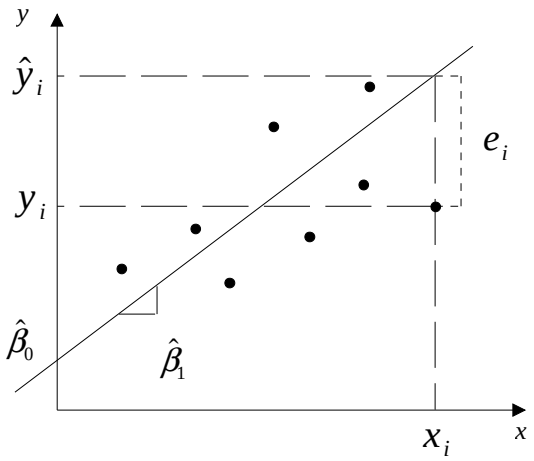


Figure 12. Least Squares Example

In other (mathy) words, least squares solves the following problem:

$$\min \sum_{i=1}^{n} e_i^2$$

$$\text{s.t.} \quad y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \ (i = 1,2,\ldots,n)$$

Operations Research types would recognize this as a form of nonlinear programming, but that's not important for solving the problem. Thanks to the miracle of computers, you can have the silicon to try a bunch of different regression lines. Then it can calculate the sum of squared errors for each and pick the one with the smallest sum. But that can take a long time if there are a lot of data points. Of course, some old guys way back before good computer games (or computers at all) had some free time so they figured out how to do this without a computer (the computer still makes it a lot easier though).

It can be shown (see Appendix C) that the least squares estimators for the one variable case are:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{yx}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{S_{xy}}{S_{xx}}$$

The main reason least squares is used for regression is that the least squares estimators are the Best Linear Unbiased Estimators (BLUE). The linear part has already been discussed. The unbiased part means the least squares estimates have an expected value equal to the true parameters they are trying to estimate. The best part means the estimators have the minimum variance of all other unbiased estimators. You can see Appendix C if you really want to learn more about the theoretical mathy stuff.

Two last points before looking at an example. First, it is important to note that least squares is only one way to perform linear regression. One drawback to it is that there is a heavy penalty for

outliers (see Level 1 Section 2) because of squaring the residuals. Another option is to minimize the sum of the absolute deviations:

$$\min \sum_{i=1}^{n} |e_i|$$

There aren't as many software packages that can do this, however, so it may be easier to just look for outliers before running a regression.

The final point for least squares (promise) is about the intercept term. For certain situations, you would expect the intercept term to be zero. For example, if you are plotting number of sorties versus the budget, you would expect that there will be no sorties flown if there is no budget. That being the case, you would force the regression line to go through the origin. Most software packages allow for this in the regression options so it will not be covered any further.

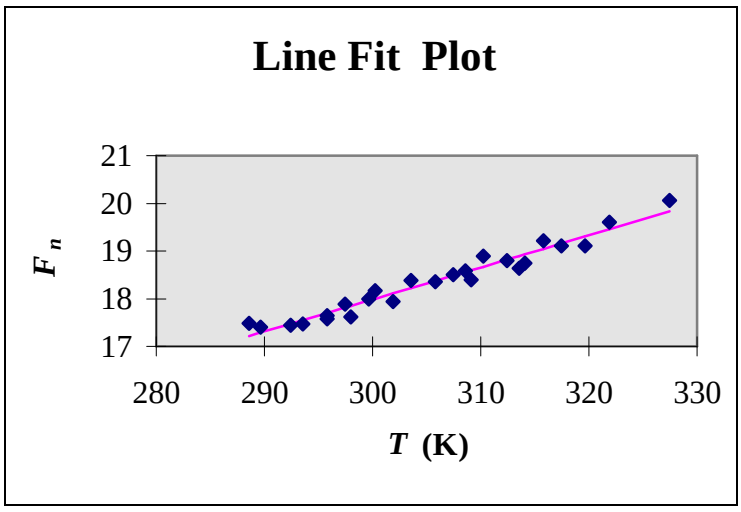
Example. The F-31 is undergoing some range testing that is returning disturbing data. Apparently, on some test flights in the desert, the radar system seems to have too much internal noise to correctly identify targets at distances it usually has no problem with. You feel there may be a text book case here to highlight why operational test is required in addition to developmental test. The radar system passed its DT testing with flying colors, but that doesn't mean it's usable in a strike fighter like the F-31 which will encounter harsh environments. Your electrical engineers tell you internal noise can be measured, but it can also be calculated using the following formula

$$N = kTBF_n$$

where N = Internal Noise of the Receiver (Watts)
k = Boltzman's Constant (1.38×10^{-23} W·s/K)
T = Temperature (K)
B = Bandwidth (Hz)
F_n = Noise Factor of Receiver (no units)

They suspect something is causing the noise factor (F_n) to increase which is creating more internal noise than originally designed. A noise factor above 20 is too extreme to make a useful system on a fighter aircraft. You look over some flight test data to uncover the culprit. You find the bandwidth is relatively constant at 1 MHz (1×10^6 Hz), but temperature may be the problem.

N	T (^{o}F)	T (K)	F_n
6.9603E-14	60	288.56	17.479
6.9575E-14	62	289.67	17.405
7.0391E-14	67	292.44	17.442
7.0748E-14	69	293.56	17.464
7.1741E-14	73	295.78	17.576
7.2018E-14	73	295.78	17.644
7.3413E-14	76	297.44	17.885
7.244E-14	77	298.00	17.615
7.4404E-14	80	299.67	17.992
7.5251E-14	81	300.22	18.163
7.4756E-14	84	301.89	17.944
7.6991E-14	87	303.56	18.379
7.7445E-14	91	305.78	18.353
7.8525E-14	94	307.44	18.508
7.9123E-14	96	308.56	18.582
7.8464E-14	97	309.11	18.394
8.0878E-14	99	310.22	18.892
8.1082E-14	103	312.44	18.805
8.0639E-14	105	313.56	18.636
8.1272E-14	106	314.11	18.749
8.3756E-14	109	315.78	19.220
8.3711E-14	112	317.44	19.109
8.4289E-14	116	319.67	19.107
8.7087E-14	120	321.89	19.605
9.0632E-14	130	327.44	20.057



You look at a plot of the noise factor versus temperature and for about a tenth of a second you think about calculating a regression by hand. Then you come to your senses and have the computer do it for you. Using the Analysis ToolPak in Excel, you get the following relationship between temperature and the noise factor (it's the line that appears in the graph above):

$$F_n = -2.1682 + 0.06717 \cdot T$$

You don't have to do this type of work in Excel. In fact, any specialized statistics package would probably be better from a statistical perspective, but Excel gives you the benefit of being able to play with your output to get the right format. Excel is also easier to learn and you can take the output and put it directly into Word for your report (like it is in this section).

6.3 Statistical Significance

Once you know the relationship between the explanatory and response variables, there are usually two things that you may be interested in. The first is making inferences about the parameters (the β's). For example, does y increase or decrease with x, and if so, by how much. The other concern is making predictions about the response variable. Some people call this forecasting.

Before hitting those practical issues though, you have to check if the regression itself is significant… sorry to spoil your fun. If you have computer software performing the regression this step boils down to looking at a couple of tables, the first of which is just an ANOVA like Figure 13. If you are unfortunate enough to have to fill in the blanks yourself see Appendix C.

Source	*SS*	*df*	*MS*	*F*
Regression	SS_R	1	$SS_R/1$	MS_R/MS_E
Residuals	SS_E	n - 2	$SS_E/(n - 2)$	
Total	SS	n - 1		

Figure 13. ANOVA for Regression (One Variable)

The formal hypothesis test checks to see if all the partial regression coefficients (in this case only β_1) are equal to zero:

H_o: $\beta_1 = 0$
H_a: $\beta_1 \neq 0$
Test Statistic: $F = MS_R/MS_E$
Rejection Region: $F > F_{\alpha,(1,n-2)}$

Once you know that the regression is significant based on the ANOVA (i.e. at least one coefficient is not equal to zero), you should look at the second table the software prints out. What you want to see are the results of the individual *t*-tests to see which ones are significant (see Figure 14). This can get to be a great deal of work if there are a lot of parameters and you have to crunch the numbers yourself. If you're still living in the dark ages, refer to Appendix C for the grueling details.

Variable	*Coefficient*	*Std Err*	*t Stat*	*P-value*
Intercept (β_0)	$\hat{\beta}_0$	$\sqrt{V(\hat{\beta}_0)}$	see below	
x (β_1)	$\hat{\beta}_1$	$\sqrt{V(\hat{\beta}_1)}$		

Figure 14. Table of *t*-Tests for Individual Parameters

The equations for the variances of the parameters can be found in Appendix C (they're too scary for the regular text!). The general hypothesis test is the same as the one discussed in Level 1 (Section C.4) and is performed for each parameter in the model:

$\boldsymbol{H_0}$**:** $\beta_i = 0$
$\boldsymbol{H_a}$**:** $\beta_i \neq 0$

Test Statistic: $t = \dfrac{\hat{\beta}_i}{\sqrt{V(\hat{\beta}_i)}}$

Rejection Region: $t > t_{\alpha/2,(n-2)}$

***P*-value:** $P\left[t_{\alpha/2,(n-2)} > \dfrac{\hat{\beta}_i}{\sqrt{V(\hat{\beta}_i)}} \right]$

If a single parameter is insignificant, it's a judgment call on whether to drop it or not. In other words, it depends. It depends on how insignificant the parameter is. It depends on what you plan to use the model for. It depends on the day of the week. Etc, etc. Actually, if you plan to use the model solely for prediction of the response, keeping the insignificant parameters won't hurt. If you are building to model to make inference on the parameters themselves (e.g., "studies show that an increase in *x* leads to a four fold increase in *y*") then you cannot include these parameters and be statistically correct (SC).

There is one other thing to look at before moving on. Just because a regression is significant, does that mean it's good? Well, again, the answer is, "It depends." (You can call this fuzzy stat.) There is an additional measure that most software packages report when performing a regression. It's called the coefficient of determination, or simply R^2.

$$R^2 = \frac{SS_R}{SS} = 1 - \frac{SS_E}{SS}$$

Put simply, R^2 measures the proportion of variation in the data that is explained by the regression model. The value can range anywhere from zero to one (or 100 percent). If you're brave enough to look at Appendix C (and you're a math-geek), you'll realize there is a problem with R^2… it is a nondecreasing function of *k* (that is, it will always stay the same or increase as you add parameters). Don't worry if you didn't catch that mathy jargon because it really only applies to multiple regression (Section 6.8). To account for that drawback there is the adjusted R^2:

$$\overline{R}^2 = 1-\left(1-R^2\right)\frac{n-1}{n-(k+1)}$$

A low adjusted R^2 can indicate specification bias in the model. That is, you may have the wrong relationship between the response and explanatory variables. You should look at a plot of *y* versus *x* to see if there is a nonlinear relationship (see Section 6.7). Then again, it could just be that *x* and *y* are not sufficiently correlated (i.e., the regression is no good).

Example. Before accepting your results in the previous section as gospel, you decide to actually read the rest of the output Excel gave you… it can be a lot of stuff if you clicked all the option boxes. After a little editing on the format, you get the following results:

Source	*SS*	*df*	*MS*	*F*	*P-value*
Regression	12.0684	1	12.0684	470.232	$8.23\text{x}10^{-17}$
Residuals	0.59029	23	0.02566		
Total	12.6587	24			

Variable	*Coefficient*	*Std Err*	*t Stat*	*P-value*
Constant	-2.1682	0.94721	-2.2891	0.03159
T (K)	0.06717	0.00310	21.6848	$8.23\text{x}10^{-17}$

From the ANOVA, you can tell the regression was significant by the *p*-value for the *F*-test. Actually, it's VERY significant (i.e., 99.999999... percent confidence!). Of course, this is just an example with clean, sterile data. You probably won't see anything this good in real life, but the idea is still the same. If you're quick, you'll notice the *t*-test for the slope parameter has the same *p*-value as the *F*-test. That's because in one variable regression (with a constant term) the two tests are the same. If you don't understand why, you probably need to re-read this section. According to the *p*-value, the constant term is also significant ($\alpha = 0.05$ is the default in the world of statistics). You wouldn't expect that since just about everything would be zero at 0 K. As mentioned in Section 6.1 (and 6.4), the value is worthless for interpreting because there is no data in that region (who'd be around to measure it?). Since it's significant though, it's best to keep it in because it will help with predictions.

Before getting ahead of yourself, though, you still have to look at the R^2. In this case it's 0.95337. That means over 95 percent of the variability in the noise factor can be explained by the temperature. That sounds pretty impressive and it is, but remember that this is just an example. In the real world, any R^2 above 0.6 is usually pretty good.

6.4 Prediction

Once you've settled on a regression model that is significant and meets the assumptions (see Section 6.5) you can finally get to the practical stuff. The first thing you may want to do is make inferences about the parameters with hypothesis tests or confidence intervals. You already have the tools required, but just in case you need your hand held here's how you do it:

H_o: $\beta_i = \beta_i^*$ (where β_i^* is the hypothesized value)
H_a: $\beta_i \neq \beta_i^*$

Test Statistic: $t = \dfrac{\hat{\beta}_i - \beta_i^*}{\sqrt{V\left(\hat{\beta}_i\right)}}$

Rejection Region: $t > t_{\alpha/2,(n-2)}$

Or you can do confidence intervals:

$$\hat{\beta}_i \pm t_{\alpha/2,(n-2)}\sqrt{V\left(\hat{\beta}_i\right)}$$

You can also make inferences on the variance of the parameters, but that's overkill for this section (see Appendix C here and in Level 1).

Making inferences about the parameters is stuff you could probably have figured out for yourself. If you can guess how to make predictions though, you probably don't need to be reading this primer. For those normal people out there, here's how you do it for a one variable regression using confidence intervals (from there you can derive the hypothesis tests):

Predict Mean Response, $\hat{y}^* = E\left(y|x^*\right)$ (x^* is the value at which you're making the prediction)

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x^*\right) \pm t_{\alpha/2,(n-2)}\sqrt{\mathrm{MS_E}\left[\frac{1}{n} + \frac{\left(x^* - \bar{x}\right)^2}{S_{xx}}\right]}$$

$$\text{where } S_{xx} = (n-1)V(x) = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

Predict Single Response, y^*

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x^*\right) \pm t_{\alpha/2,(n-2)}\sqrt{\mathrm{MS_E}\left[1 + \frac{1}{n} + \frac{\left(x^* - \bar{x}\right)^2}{S_{xx}}\right]}$$

Note from the equations how you lose confidence as the response variable's value moves further away from the mean. Figure 15 illustrates the point with a little exaggeration (see the example for a similar graph with actual data).

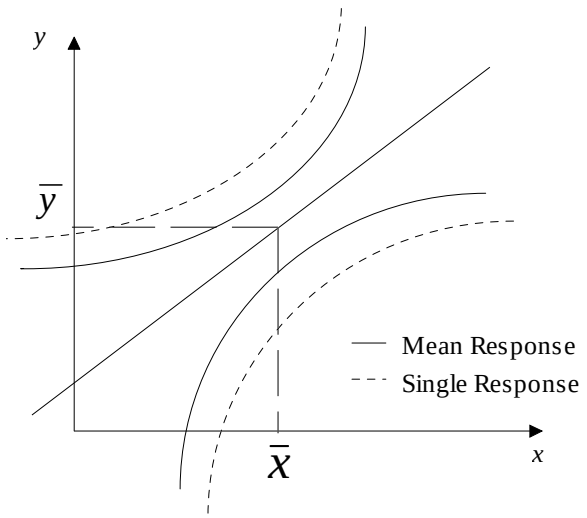


Figure 15. Prediction Intervals

A very important final point on prediction is that it's only valid within the scope of the data. That means if you want to estimate a response, each explanatory variable has to be set between its minimum and maximum in the original data. In reality predictions are more restricted than this, but this is a general rule that's easy to follow. To learn how to do predictions with multiple variables, see Section 6.8.

Example. Now you come to the practical part of the regression you ran in the previous sections. You know the noise factor of the receiver is increasing with temperature, but by how much? Well, you're sure your boss will ask that so you construct a confidence interval on the slope parameter:

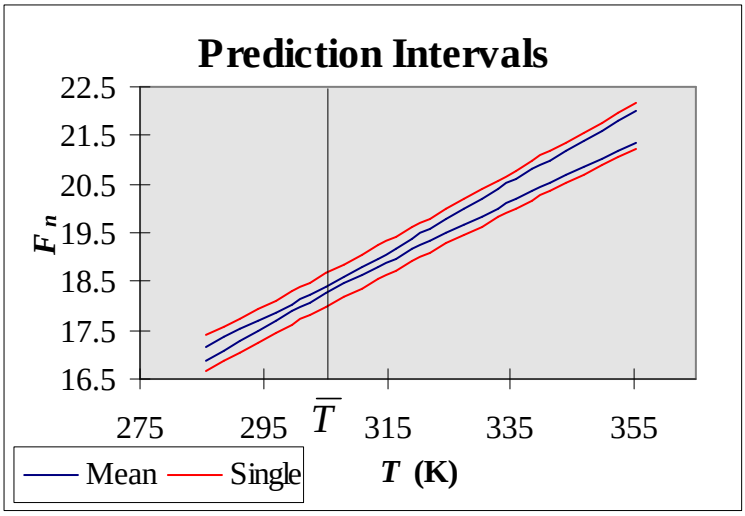
$$\hat{\beta}_1 \pm t_{\alpha/2,(25-2)}\sqrt{V\left(\hat{\beta}_1\right)} = 0.06717 \pm 2.069 \cdot 0.00310 = \left[0.06076, 0.07358\right]$$

Therefore, you know with 95 percent confidence that an increase in 1 Kelvin results in an increase of 0.060 to 0.073 in the noise factor. A more important application of the regression results would be to find out at what temperature the mean noise factor increases above 20. To do this, you can set the equation for predicted mean response equal to 20 and try to solve it for x^*, or you can demonstrate your knowledge of practical mathematics (i.e., using a computer) and create a data table in Excel. Here is a piece of the output for both mean and single response prediction intervals:

T (^{o}F)	T (K)	$\overline{F}_n$ (low)	$\overline{F}_n$ (high)	F_n (low)	F_n (high)
125	324.67	19.502	19.78	19.282	20.000
130	327.44	19.673	19.983	19.462	20.193
135	330.22	19.843	20.185	19.641	20.387
140	333.00	20.013	20.389	19.820	20.582
145	335.78	20.183	20.592	19.998	20.777

From the table, it appears the radar is not operationally effective above 135°F because the mean noise factor is predicted to be above 20. Granted the atmospheric temperature rarely gets this high, but inside the F-31's nose it can easily get above that temperature. The next step from here would be to check on the radar's cooling system.

To emphasize the point of losing confidence with predictions as you move away from the mean, you can construct a graph similar to Figure 15 using the data form this example. This can be an exercise in using Excel, but it should be easy to do if you can reproduce the table above.



6.5 Assumptions

After reading the other sections in this level, you're probably already asking yourself what all the assumptions and limitations of regression are (at least you should be). There are lots of them, but luckily most are usually valid and there a some easy fixes for the times they aren't (see Section 6.6). This section will cover the nine most discussed assumptions of least squares. There is a tenth assumption related to causal relationships, but it's only important to certain fields like econometrics.

1. **Linear in Parameters** - This has already been discussed in Section 6.1.

2. **Relationship is Inexact** - In other words, ε must exist for the regression to be valid. This is an implicit assumption of least squares because it leads to mathematical errors if there is no error term (e.g., division by zero or a singular [non-invertable] matrix).

3. **Explanatory Variables not Perfectly Correlated** - This is another implicit assumption of least squares because you will have a singular matrix in multiple regression (see Section 6.8).

4. **Explanatory Variables are Nonstochastic** - This is the first assumption that will not prevent you from calculating estimates if it does not hold. The assumption can be modified to say if the explanatory variables are stochastic, then they are independent of the error terms. In mathy terms that means $\text{Cov}(\varepsilon,x_k) = E(\varepsilon \cdot x_k) = 0$. The main reason this is used is to cancel out a lot of stuff to

make the equations nice and easy to use. The assumption is also kind of essential to have assumption nine hold.

5. **ε is a random variable with $E(\varepsilon) = 0$** - This assumption is required to prove that the least squares estimates are unbiased. Fortunately, it is only required for the intercept term which is usually invalid anyway (see Section 6.1).

6. **$\varepsilon \sim N(0,\sigma^2)$** - From here to the end are The Big Four… the most important, most frequently violated assumptions in linear regression. The normality assumption actually doesn't have too large an impact because the least squares estimates are still BLUE (see Section 6.2) regardless of the distribution. The problem comes in the *t* and *F* tests discussed in Sections 6.3 and 6.4. If *n* is large and ε does not deviate from normality too much you're OK. If not, you can try the methods discussed in Section 4.6 or resort to Bootstrapping (nonparametric regression which is way beyond the scope of this primer).

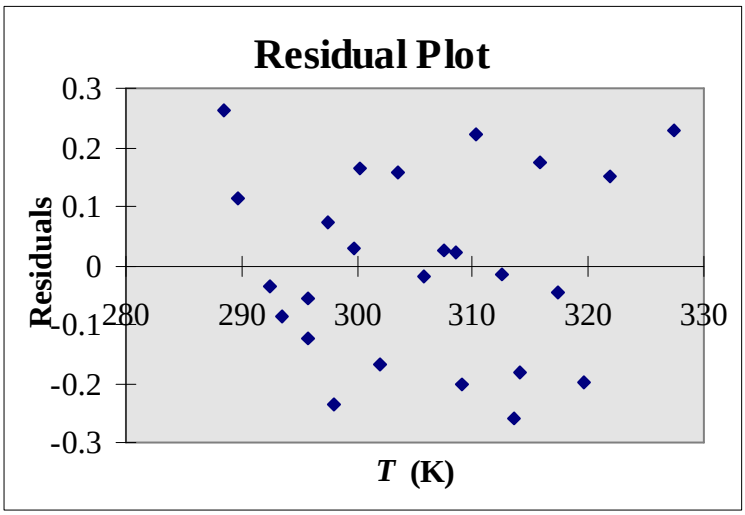
7. **Homoscedasticity** - Cool word isn't it? Actually, the next three assumptions are just things for statisticians to talk about around other people so the stat-types will sound really smart using big words. Homoscedasticity basically means constant variance (σ^2). This assumption was already discussed in Section 4.3. Additional methods for detecting heteroscedasticity (another big word!) are the Goldfeld-Quant Test, Sprearman's Rank Correlation Test, and the Park Test which can be found in any good, thick statistics book.

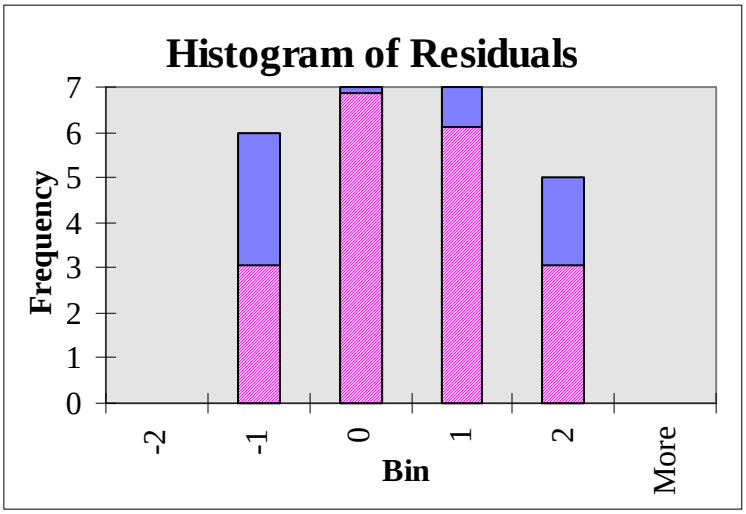
8. **No Autocorrelation** - Autocorrelation means that the error terms are correlated with each other. Another way to say no autocorrelation is to say the error terms are independent… that way some normal people may be able to understand you. This is probably the worst assumption to violate in linear regression because you can no longer guarantee that the least squares estimates are BLUE (see Section 6.2). The estimates are still linear and unbiased, but the variance of the estimate will be biased. That means the *t* and *F* tests will give inaccurate results and you will overestimate R^2. Verifying and dealing with failures of this assumption are discussed in the next section.

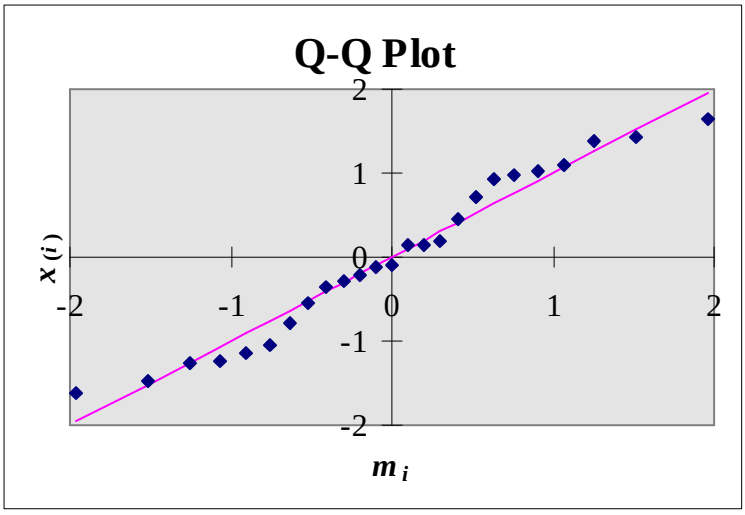
9. **No Multicollinearity** - This assumption basically means that you can't have an exact linear relationship among the explanatory variables. There are two situations: perfect (extreme) and severe. It's pretty easy to detect perfect multicollinearity because you won't be able to compute the least squares estimates (same as assumption 3). In the severe case, you can get the estimates but the variance-covariance matrix (see Section 6.8) of the estimates approaches infinity so the *t*-tests will be insignificant. The *F*-test and R^2 are unaffected.

Example. You got some good results from the regression on temperature and noise factor, but was it really a good regression (i.e., were the assumptions valid)? There's really no need to review the first three assumptions as they are inherent to this model (only one variable using least squares). The fourth is pretty easy to verify by computing the correlation or covariance between the explanatory variable (temperature) and the residuals. In this example, both values are on the order of 10^{-13} which is close enough to zero.

Verifying the regression assumptions doesn't have to be a time consuming process. Many of the remaining assumptions can be considered together by looking at a couple of plots.







All three plots suggest that assumptions five and six are valid. The first shows that the residuals are random. The second is a histogram of the standardized residuals (solid bars) overlaid by the expected histogram of 25 samples from a standard normal distribution (i.e., the same values used to generate the Q-Q plot). Although the histogram and Q-Q plot show a slight deviation from normality, that's expected in a sample of only 25 observations. The deviation is not severe so it's OK to accept the normality assumption. For added assurance, you compute the mean of the residuals to be -3.41×10^{-15} with a standard deviation of 0.1568. You don't even need to construct a confidence interval to know that zero will be contained in it.

The spread in the residual plots seems fairly constant all the way across so the homoscedasticity assumption is validated. Yes, it's that easy sometimes. If there were a slight pattern of increasing or decreasing residuals, you would perform some quantitative tests, but in this case it's fine. The final assumption to verify is no autocorrelation because multicollinearity doesn't apply to the one variable case. Autocorrelation is discussed in more detail in the next section.

As an aside to Excel users, the Q-Q plot above had to be generated the old fashioned way because Excel doesn't do it for you (see Level 1 Section 5.3).

6.6 Failure of Assumptions

The previous section gave a long list of assumptions that must be met in order to have a statistically correct regression model. Most of the time, people don't care about this stuff and they post their results just because they like them (or they're lazy). It's not quite as bad as people who automatically assume that the regression relationship is causal, but it can be depending on what the results are used for. Most of the assumptions have already been discussed in enough detail in the primer so this section will only consider the last two assumptions (no autocorrelation and no multicollinearity).

Autocorrelation. As mentioned earlier, autocorrelation is probably the worst assumption to violate so it should be one of the first ones you verify. There are several methods for checking the assumption and most software packages can do some of them for you. The simplest technique is the graphical method similar to the one for constant variance discussed in Section 4.6. Basically, you can plot the residuals versus time (e_i vs. i) or versus lagged residuals (e_i vs. e_{i-t}). A lagged residual is just one that has already been observed before the current one. Once you have a plot, all you have to do is look for patterns. It's pretty subjective, but if you don't see any patterns most likely there is no autocorrelation. As a guide, Figure 16 shows some basic results from the graphical method.

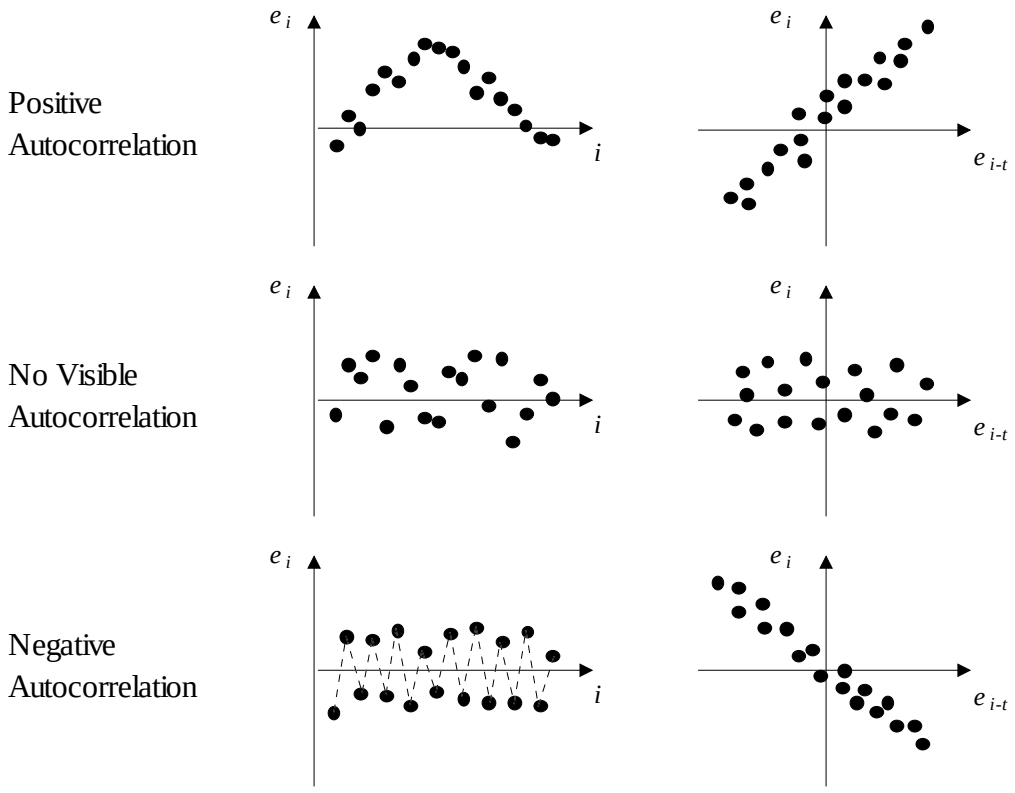


Figure 16. Detecting Autocorrelation

A more quantitative technique for detecting autocorrelation is the Durbin-Watson Test. The test is limited in that it assumes an intercept term exists and cannot have missing values. It also only tests for a lag of one time period. Still, most software packages return the Durbin-Watson statistic *d* automatically so it's not too difficult to apply the test (recall that $\rho_{(x,y)}$ is the correlation between *x* and *y*; Level 1 Section 6.2):

$\boldsymbol{H_o}$**:** $\rho_{(e_i, e_{i-1})} = 0$

$\boldsymbol{H_a}$**:** $\rho_{(e_i, e_{i-1})} \neq 0$

Test Statistic: $d = \dfrac{\sum_{i=2}^{n} \left(e_i - e_{i-1}\right)^2}{\sum_{i=1}^{n} e_i^2} \approx 2\left[1 - \rho_{(e_i, e_{i-1})}\right]$

Rejection Region: $d < d_L$ or $d > 4 - d_L$ (where d_L comes from Table 23 of Appendix A)

The test can be modified to specifically test for positive or negative autocorrelation, but you can figure out how with the information in the primer. Just to help you out though, Figure 17 gives a general look at where the significance points (d_L and d_U) fit in:

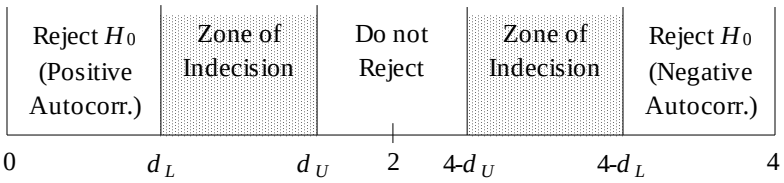


Figure 17. Regions of Durbin-Watson Statistic

There are many other tests designed to detect autocorrelation… statisticians really like getting their names in the books. Two of the more common ones are the Geary Test (also called Runs Test) and the Chi-Squared Test of Independence (see Section 5.2). You can read about these and more in the thicker statistics and econometrics books.

If you suspect there is autocorrelation in the model, there are a couple quick fixes you can try. The easiest of these is to respecify the model. Perhaps the relationship isn't linear. There could be a quadratic or logarithmic term. A second easy fix is to add a parameter for time, but that gets you into the slightly scary world of time-series forecasting. A method that generally works once the model is correctly specified is the Cochrane-Orcutt procedure. It is an iterative procedure meant to eliminate the correlation in the residuals. Basically, it starts off assuming the residuals have a lag of one (i.e., $\rho(ei,ei\text{-}1) \neq 0$) and tries to estimate the correlation in order to reestimate the parameters. Here is an outline of the procedure… you really should have a computer to do it:

Cochrane-Orcutt Iterative Procedure

1. Run least squares to get the residuals and parameter estimates (e_i, $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$)
2. Run least squares on the model $e_i = \hat{\rho} e_{i-1}$; or you can use the estimate $\hat{\rho} \approx 1 - \frac{d}{2}$
   (where d is the Durbin-Watson statistic)
3. Run least squares using n - 1 revised data points: $y_i^* = \hat{\beta}_0^* + \hat{\beta}_1^* x_{1i}^* + ... + \hat{\beta}_k^* x_{ki}^* + e_i^*$
   (where $y_i^* = y_i - \hat{\rho} y_{i-1}$ and $x_{ji}^* = x_{ji} - \hat{\rho} x_{j(i-1)}$)
4. Calculate e_i^* from step 3 and go to step 2 until the change in $\hat{\rho}$ is about 0.005

In addition to this fix, there is a Cochrane-Orcutt Two-Step procedure as well as a Durbin Two-Step procedure.

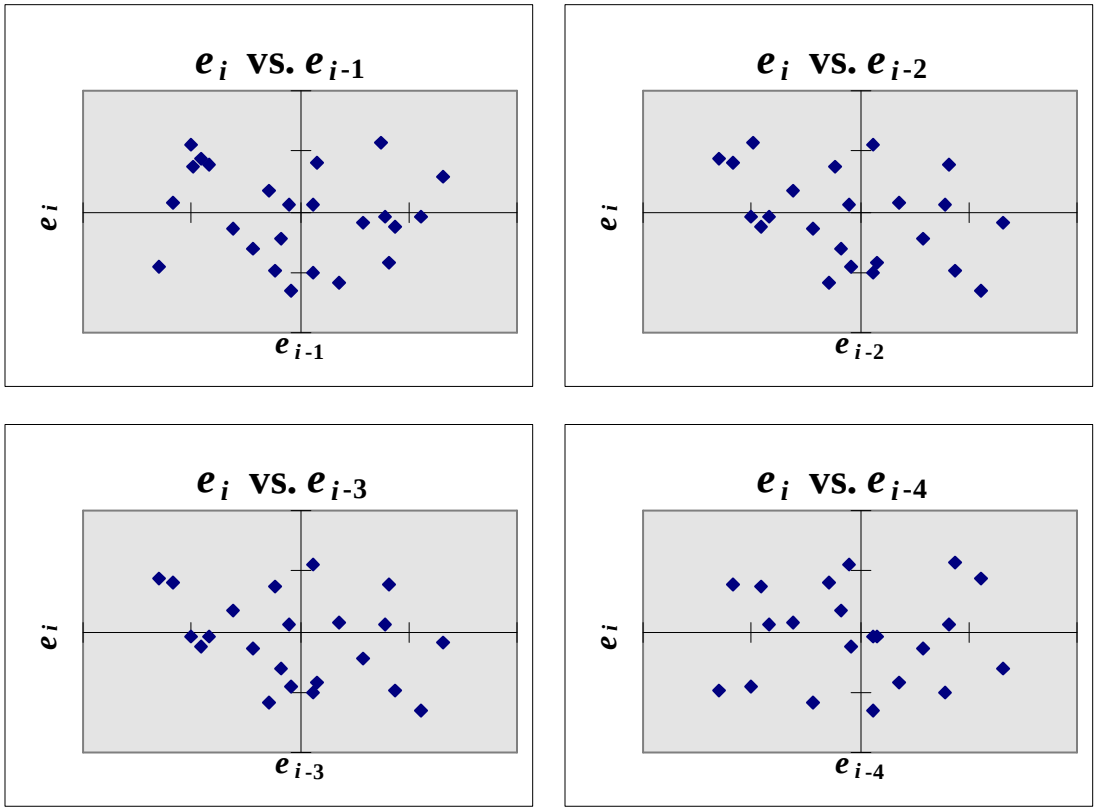
Multicollinearity. Of The Big Four assumptions, multicollinearity is one of the easiest to detect. As mentioned in the previous section, for perfect multicollinearity you cannot compute the least squares estimates because of a singular matrix. In such a situation the only fix is to remove the variable that's causing the problems. Most likely, there's just a variable that is duplicated or one variable that is a linear function of the other explanatory variables.

In severe multicollinearity, there are three things to look for to verify the assumption. The first is a low *t*-stat OR (not and) unexpected signs on all or some of the parameters. The second thing to look for is a high R^2 and significant *F*-tests. If things look bad to this point, you can compute all the pairwise correlations of the explanatory variables. Now apply some fuzzy stat. If the *t*-stat, R^2, and *F*-stat show multicollinearity may be a problem and there are high pairwise correlations

between some of the variables, there's probably some multicollinearity. How high is high? Whatever you want. Normally anything over 0.6 or 0.7 in absolute value. If that's not quantitative enough for you, there's always the condition index which has a cool formula with eigenvalues and everything. But you'll have to go to a book to get it.

Now comes the beauty of multicollinearity… the ease of fixing it. If your model does have severe multicollinearity the remedy depends on what you want the model for. Hopefully, you ran the regression to do prediction of the response variable because multicollinearity doesn't affect the predictions. There is a problem with interpreting the results of the parameters as discussed in the previous section. Unfortunately, there's really no solution to this problem except removing some explanatory variables (that's usually not desired). A simple option (on paper) is to collect more data to increase the range and variability of the explanatory variables.

Example. From the ongoing temperature and noise factor regression, the final assumption to validate is no autocorrelation. You can do that by producing a bunch of graphs like the following which show the relationship between the residuals and the ones which precede them.



The graphs look pretty good up to four times periods back. Depending on the importance of your regression and how much time you have, you can look at greater lags, but it's highly unlikely that there will be anything significant past four time periods. If you insist on numbers, you can compute the Durbin-Watson statistic using the equation above (Excel doesn't automatically do it). Doing so gives you $d = 2.22901$. Referring to Appendix A, you can figure out the rejection

region is $d < 1.206$ or $d > 2.794$. That means, according to the test discussed in this section, there is no autocorrelation (with a lag of one).

6.7 Nonlinear "Linear" Regression

As discussed in Section 6.1, there are many applications of linear regression to relationships that are obviously not linear (see Figure 11). Most of these cases really don't take much work to get into the right format. Basically, what you want to do is use some dummy variables in place of the nonlinear ones to get things to look linear. For example, any time an explanatory variable must be raised to a power, say x_j^c, just introduce $x_{k+1} = x_j^c$. The plot of $y = x_j^c$ vs. x_j is obviously not linear, but the plot y vs. $x_{k+1} = x_j^c$ will be linear (see Figure 18). The greatest thing about this fix is that there's really no real work involved. All you do is use the values for the explanatory variables that are already raised to their respective powers… all this talk of dummy variables just makes it look like you're really doing something difficult. By the way, this sly technique also works when you take the exponents, logs, reciprocals, and roots of x_j (i.e., $y = e^{x_j}$, $y = \ln(x_j)$, $y = 1/x_j$, and $y = \sqrt{x_j}$).

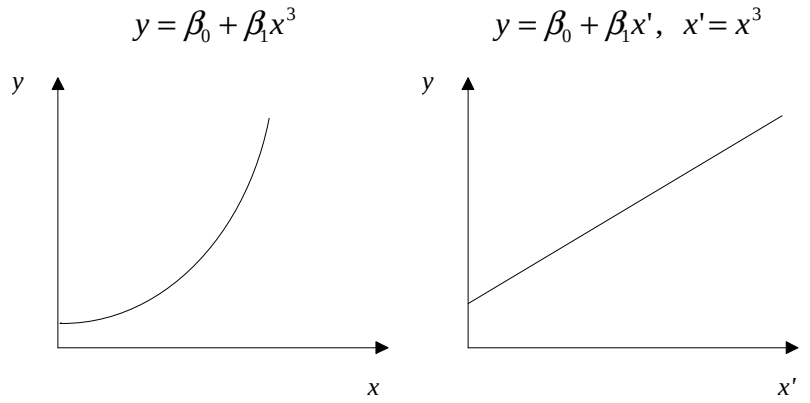


Figure 18. Making Nonlinear "Linear" Regression Linear

Another type of nonlinear regression involves qualitative variables. A common example is using gender as an explanatory variable. That is, some variable x_g which is set to 0 for females and 1 for males (or the other way around if you prefer to be statistically and politically correct). It's cleverly called a binary variable. You can also have qualitative variables with more than two values. In such a situation you have to break it up and introduce several binary variables to account for the various values ("several" being equal to one less than the number of different values). For example, if you include regions in your regression and the possibilities are northeast, southeast, southwest, and northwest, you have to use three binary variables. One way to do it is

$$x_{SW} = \begin{cases} 1 \text{ . if southwest} \\ 0 \text{ . else} \end{cases} \qquad x_{NW} = \begin{cases} 1 \text{ . if northwest} \\ 0 \text{ . else} \end{cases} \qquad x_{NE} = \begin{cases} 1 \text{ . if northeast} \\ 0 \text{ . else} \end{cases}$$

In this case, the default regression (i.e., $x_{SW} = x_{NW} = x_{NE} = 0$) is the southeast region. Of course, you need to have a restriction that at most one of the three variables can be equal to one. For more information on qualitative regression consult a textbook.

Example. A nonlinear relationship common to OT&E for all weapon systems is the survivability analysis. The general relationship between number of sorties (n), probability of survival (P_s), and attrition (A) is

$$A = P_s^n$$

That doesn't look like something you can use regression with, but you can if you know the tricks. For example, if you take the natural logarithm of both sides you get

$$\ln(A) = \ln\left(P_s^n\right) = n\ln(P_s)$$

Now you have a relationship you can work with. It's a one variable linear regression model with no intercept term. Once you get a relationship, you can use the parameter estimate to make inferences about the probability of survival (P_s). This is just one way of estimating P_s, and not necessarily the simplest way, but it emphasizes the point of doing nonlinear "linear" regression. The following table shows some output from a simulation of the F-31 in several dangerous, but typical missions at the start of a major conflict.

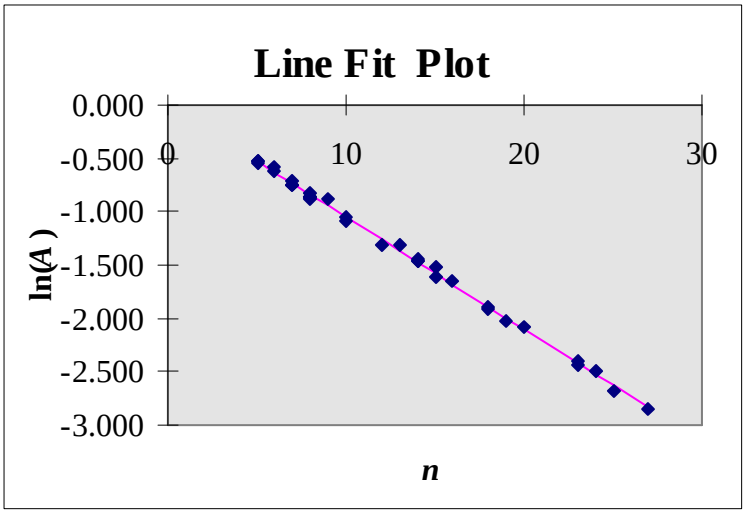
Sorties (n)	Attrition (A)	ln(A)
5	0.5904, 0.5758, 0.5898	-0.527, -0.552, -0.528
6	0.5337, 0.5549	-0.628, -0.589
7	0.4747, 0.4752, 0.4897	-0.745, -0.744, -0.714
8	0.4223, 0.4374, 0.4119, 0.4253	-0.862, -0.827, -0.887, -0.855
9	0.4148	-0.880
10	0.3386, 0.3492	-1.083, -1.052
12	0.2671	-1.320
13	0.2679	-1.317
14	0.2362, 0.2332, 0.2318	-1.443, -1.456, -1.462
15	0.2187, 0.1999	-1.520, -1.610
16	0.1919	-1.651
18	0.1509, 0.1468	-1.891, -1.919
19	0.1315	-2.029
20	0.1238	-2.089
23	0.0902, 0.0872	-2.406, -2.440
24	0.0831	-2.488
25	0.0679	-2.689
27	0.0583	-2.842

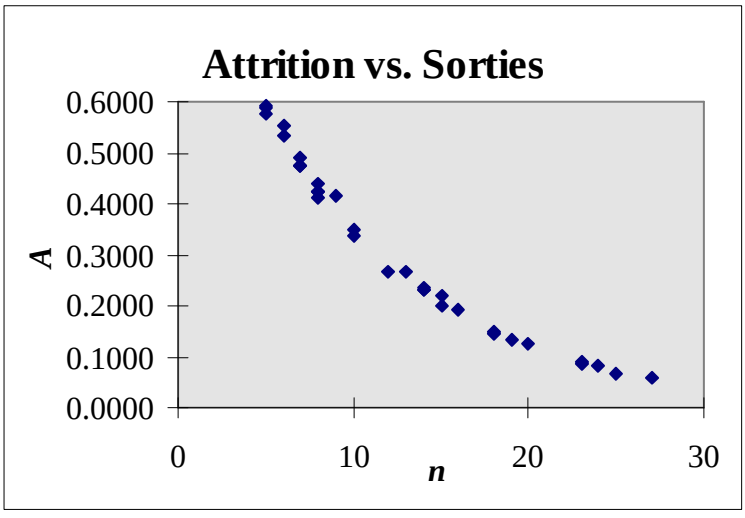
Running the regression with no intercept term you get the following results:

Source	*SS*	*df*	*MS*	*F*	*P-value*
Regression	14.9889	1	14.9889	14886.36	5.15×10^{-42}
Residuals	0.0312	31	0.001007		
Total	15.0201	32			

Variable	*Coefficient*	*Std Err*	*t Stat*	*P-value*
Constant	N/A	N/A	N/A	N/A
n	-0.1052	3.84×10^{-4}	-274.036	4.94×10^{-54}

That means you can approximate P_s by $e^{-0.1052} = 0.90014$. It's obvious from the tables above that the regression was significant. Notice this time, however, that the *F*-test and *t*-test are not the same. That's because the regression only has one parameter so the degrees of freedom for the *F*-test are different than usual. The following graphs show the modified (linear) data and the original data as a final emphasis that you can do linear regression on nonlinear data.





6.8 Multiple Linear Regression

Most of the examples and explanations so far have been based on only one explanatory variable. You will find, however, that this is rarely the case in practice. Knowing all that you do so far for regression won't help you in the real world because there is usually more than just one explanatory variable. On the plus side, since you'll probably do this type of work on a computer, the only difference is the amount of data you enter and the amount of output you get. Yeah, a couple of other things change, like some degrees of freedom on the tests discussed previously, but for the most part it's the same. As a reminder here's the general regression equation:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_k x_{ki} + e_i \ \ (i = 1,2,\ldots,n)$$

Now for ease (or just to confuse people), most people switch to matrix notation when working with multiple regression. By the way, the "multiple" just means that you have more than one explanatory variable, not that you're going to run more than one regression. Once again, those tricky statisticians are trying to confuse everyone… it's job security. Here's what the general regression equation looks like in matrix notation:

$$Y = X\hat{\beta} + E$$

Looks simple, right? Well, that's only if you like doing all that matrix stuff. This might change your mind:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \ X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}, \ \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}, \ E = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

(n x 1) (n x (k+1)) ((k+1) x 1) (n x 1)

Of course, if n and k are large, doing this kind of thing by hand would not be very fun. You can consult Appendix C for the details, but here are the basic changes to adapt the single regression stuff (this assumes you have a constant term, β_o):

Lease Squares Estimates. (Section 6.2)

$$\hat{\beta} = (X'X)^{-1} X'Y$$

X' is the transpose of the matrix X.

***F*-Test.** (Section 6.3)

Figure 19 shows the updated ANOVA for multiple regression where k is the number of explanatory variables, k+1 is the number of parameters to estimate, and n is the number of observations.

Source	*SS*	*df*	*MS*	*F*
Regression	SS_R	k	SS_R/k	MS_R/MS_E
Residuals	SS_E	n - (k+1)	$SS_E/(n$ - (k+1)	

Total	SS	n - 1	

Figure 19. ANOVA for Regression (Multiple Variables)

The rejection region for the F-test is modified to be $F > F_{\alpha,[k,n-(k+1)]}$. Now it's not as scary to look at the variance of the estimates (unless you have to calculate it by hand):

$$V\text{-Cov}\left(\hat{\beta}\right) = \text{MS}_\text{E}\left(X'X\right)^{-1} = \begin{bmatrix} V\left(\hat{\beta}_0\right) & V\text{-Cov}\left(\hat{\beta}_0,\hat{\beta}_1\right) & \cdots & V\text{-Cov}\left(\hat{\beta}_0,\hat{\beta}_k\right) \\ V\text{-Cov}\left(\hat{\beta}_1,\hat{\beta}_0\right) & V\left(\hat{\beta}_1\right) & \cdots & V\text{-Cov}\left(\hat{\beta}_1,\hat{\beta}_k\right) \\ \vdots & \vdots & \ddots & \vdots \\ V\text{-Cov}\left(\hat{\beta}_k,\hat{\beta}_0\right) & V\text{-Cov}\left(\hat{\beta}_k,\hat{\beta}_1\right) & \cdots & V\left(\hat{\beta}_k\right) \end{bmatrix}$$

$$\text{MS}_\text{E} = \hat{\sigma}_\varepsilon^2 = \frac{E'E}{n-(k+1)}$$

***t*-Tests.** (Section 6.3)

The individual t-tests for the significance of each parameter is exactly the same as the one presented in Section 6.3. The only difference is the degrees of freedom in the rejection region which is now $t > t_{\alpha/2,[n-(k+1)]}$.

Coefficient of Determination. (Section 6.3)

No change:

$$\text{R}^2 = \frac{\text{SS}_\text{R}}{\text{SS}} = 1 - \frac{\text{SS}_\text{E}}{\text{SS}} = \frac{\hat{\beta}'X'Y - n\bar{y}^2}{Y'Y - n\bar{y}^2}$$

$$\overline{\text{R}}^2 = 1 - \left(1 - \text{R}^2\right)\frac{n-1}{n-(k+1)}$$

Prediction. (Section 6.4)

Making inferences about the individual parameters is the same as discussed in Section 6.4 with the degrees of freedom modified as they are for the t-tests. That is, use $t_{\alpha/2,[n-(k+1)]}$ *df*.

Predicting a mean response is done as follows:

$$X^*\hat{\beta} \pm t_{\alpha/2,[n-(k+1)]}\sqrt{\text{MS}_\text{E}X^*\left(X'X\right)^{-1}X^*}$$

where X^* is a 1 x (k+1) row vector of explanatory variables at the levels you want to make the prediction (the value for the constant term, i.e., the first column, is always 1). Similarly, predicting an individual y value comes from:

$$X^*\hat{\beta} \pm t_{\alpha/2,[n-(k+1)]}\sqrt{\text{MS}_\text{E}\left[1 + X^*\left(X'X\right)^{-1}X^*\right]}$$

Assumptions. (Section 6.5)

Even though there are more explanatory variables than before, the linear regression assumptions are exactly the same. Actually, some of the assumptions make more sense when there are multiple explanatory variables (like no multicollinearity which is pretty obvious if you only have one variable). The first two assumptions from Section 6.5 are exactly the same. The third (explanatory variables not perfectly correlated) makes a lot of sense since you can see most equations require the inverse of the X matrix. If the columns of this matrix are linearly dependent, you will not be able to get the inverse because it will not exist (this goes into some matrix algebra stuff which isn't the purpose of this primer; consult any statistics or mathematics book).

The remaining assumptions are interpreted the exact same way as in Section 6.5... you can even use the same equations, but change the zeros to vectors of zeros. You can even get fancy with the assumptions of no autocorrelation and homoscedasticity which can be combined as $V\text{-}\mathrm{Cov}(EE') = \mathrm{MS}_{\mathrm{E}} I_n$ (where I_n is an n x n identity matrix; all zeros with ones on the main diagonal). For perfect multicollinearity, the problem is the same as assumption three: $\mathrm{rank}(X) < (k + 1) < n$ (i.e., linearly dependent columns which means the inverse does not exist). For severe multicollinearity, you have $\mathrm{rank}(X) = (k + 1) < n$, which technically means you can get the inverse… barely. The barely part is usually explained by saying you have $\lambda_1 x_{1i} + \lambda_2 x_{2i} + \ldots + \lambda_k x_{ki} + \nu_i = 0$ with not all λ's equal to zero and ν as an error term with an expected value of zero. For the sharp math-geeks out there, you'll recognize that the weird thing with the λ's is the definition of linear dependence. All the variables are linear independent (i.e., you can get an inverse to the matrix) if the only way to get the equation to hold is to set all the λ's to zero.

For more details on multiple linear regression, refer to Section C.2.

Example. Returning to the noise factor problem, you start looking for other factors that may have an impact while your engineers consider the problem of the cooling system. Before leaving, the electrical engineer suggested looking at humidity. Luckily, you have that data available.

T (K)	Humidity	F_n	T (K)	Humidity	F_n
288.56	0.5791	17.479	307.44	0.4782	18.508
289.67	0.5993	17.405	308.56	0.5311	18.582
292.44	0.5503	17.442	309.11	0.4529	18.394
293.56	0.5691	17.464	310.22	0.5215	18.892
295.78	0.3787	17.576	312.44	0.4335	18.805
295.78	0.3746	17.644	313.56	0.2051	18.636
297.44	0.5462	17.885	314.11	0.2298	18.749
298.00	0.394	17.615	315.78	0.2763	19.220
299.67	0.4607	17.992	317.44	0.0211	19.109
300.22	0.632	18.163	319.67	0.1911	19.107

301.89	0.4404	17.944	321.89	0.2957	19.605
303.56	0.4032	18.379	327.44	0.1123	20.057
305.78	0.6064	18.353			

You run back to the computers and run a multiple regression of noise factor versus both temperature and humidity. Here are the results:

Source	*SS*	*df*	*MS*	*F*	*P-value*
Regression	12.14895	2	6.074475	262.154	4.51×10^{-16}
Residuals	0.50977	22	0.023171		
Total	12.65872	24			

Variable	*Coefficient*	*Std Err*	*t Stat*	*P-value*
Constant	-4.22618	1.42435	-2.96710	0.00712
T (K)	0.07320	0.00437	16.73855	5.3×10^{-14}
Humidity	0.52330	0.28072	1.86415	0.0757

It's obvious from the ANOVA that the regression is still significant. That means that when all explanatory variables are considered together, at least one of them is significantly different than zero. You expected that though because there was a strong relationship between the noise factor and temperature anyway. The individual *t*-tests, however, show that humidity isn't a significant explanatory variable. It's close at 0.0757 so if you wanted to use the model for prediction it's possible to leave it in. Just about any regression will work better with more explanatory variables (as long as you still meet all the assumptions). In this case the R^2 jumped to 0.95973 with an adjusted R^2 of 0.95606. Since that's higher than the R^2 for the one variable regression, you'd expect the two variable model to give better predictions. That's echoed by the fact that MS_E is slighter smaller as well (0.023171 versus 0.02566).

If you look at temperature and humidity though, there's a definite correlation between them (-0.71, i.e., as temperature increases, humidity drops). Therefore, if you're going to use the model for anything other than predictions, you'll have to worry about the multicollinearity problem.

This concludes the second level of the primer. You now have a vast arsenal to accompany you in the realm of analytical analysis. There are lots of books out there that claim to help you, but hopefully, this primer was easier to understand (plus it's lighter and it's free!). It's important to understand the concepts and know what's available, but you shouldn't waste gray matter on memorizing any of this stuff. That's why you should keep the primer handy whenever you're working. To further ingrain your array of tools, there's a self-assessment in Appendix D with solutions in Appendix E.

OR 310 Probability Handout

This course enables you to create models that assist decision-making. Models are used in virtually every decision made within an organization! The most effective models are valid and reliable. Valid models effectively represent the scenario, and reliable models are replicable.

To illustrate the difference between valid and reliable, consider a "seat-of-the-pants" intuitive process of decision-making as a model. This type of decision model could be valid (in that a decision-maker may be very successful). However, unless the individual is able to transfer this success to another, the decision-making process is not reliable. So, a valid, reliable model as one that will produce successful results in your absence.

One type of model is a predictive model; these involve forecasts of future events or outcomes. Financial firms continuously strive to predict the economy. Models of events that have not yet occurred often require knowledge of probability. This handout provides enough discussion of probability to make you dangerous! So, throw caution to the wind!

Basic Probability

Consider the Weather: Tomorrow may be Snowy, Rainy, Cloudy, Sunny, or "none of the above". These events are called States of Nature. They are events beyond your control. You cannot dictate which particular state of nature that will occur, however, you can guess. When guessing, you utilize concepts of probability.

Probability is the chance that a particular state of nature will occur. The sum of probabilities for all states of nature must sum to one. (Recall that a valid state of nature is "none of the above.")

Joint Probability is the chance that multiple states of nature will occur at the same time. For OR 310, the multiple states will be mutually exclusive. The states above (Rain, Snow, etc.) can be considered the event "Precipitation". Consider another event called "Wind" with the states Calm, Breezy (5-15 mph), and Windy (>15 mph). You can use simple probabilities to guess the likelihood of Rain or Wind. Joint probabilities tell you how likely it is for both to occur.

Conditional Probabilities allow you to change your guess of the likelihood that a state of nature will occur based on additional information. Suppose you know today's weather. Does that impact Tomorrow's guess? For example, is the probability of snow tomorrow given snow today greater than the probability of snow tomorrow given sunny today? If so, then this information is helpful and you should include it within your guess of tomorrow's weather.

Math (sorry, you need some... and it builds character!)

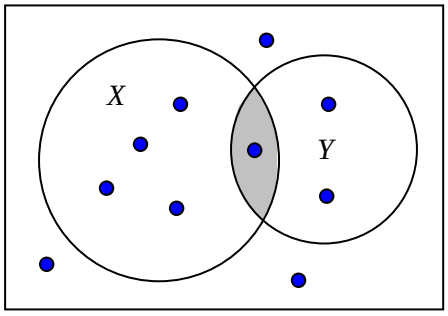
The probability described above is sometimes referred to as a marginal probability. It is represented by the $P(X)$ which is read, "the probably of X." Mathematically, the probability of a particular event X is equal to the total number of possible occurrences of X divided by the total number of all events. That is,

$$P(X) = \frac{\#X}{Total\#} \qquad \text{Eqn 1}$$

Since the sum of probabilities for all states of nature for a given event (i.e. "Precipitation") must sum to one, the probability of X and "not X" (denoted X') must be equal to 1. That is,

$$P(X)+P(X')=1$$

Before getting into joint probability, let's look at some numbers. Consider the Venn diagram below which graphically shows the probabilities of and relationship between events X and Y. The dots on the diagram represent actual occurrences.



You should be able to determine the probability of an occurrence within event X and the same for event Y. For $P(X)$, count up the total number of dots in circle X (5) and divide by the total number of dots (10).

$$P(X)=\frac{5}{10}=0.5 \quad \text{and} \quad P(Y)=\frac{3}{10}=0.3$$

The shaded region where X and Y overlap represents the joint probability of X and Y and is called the "intersection of X and Y." It can be denoted several ways:

$$P(X,Y)=P(XY)=P(X \text{ and } Y)=P(X \cap Y)$$

Given the diagram, you should be able to figure out the value of the intersection by counting dots (1/10 = 0.1). (NOTE: This example does not use mutually exclusive sets as mentioned in the previous section to keep the math simple. The next section on Bayes' Theorem will go back to mutually exclusive sets for joint probabilities.)

The total area covered by both X and Y is called the "union of X and Y," which is also denoted several ways:

$$P(X \text{ or } Y)=P(X \cup Y)$$

In order to calculate the union, you need to know the intersection. It should be intuitive: you add the area of circle X and circle Y, and then you subtract the shaded region because you counted it twice. In mathanese,

$$\boxed{P(X \cup Y)=P(X)+P(Y)-P(X \cap Y)} \qquad \text{Eqn 2}$$

From the diagram, you could easily count up the dots to get the union: 7/10 = 0.7. Here's the result using the equation:

$$P(X \cup Y)=P(X)+P(Y)-P(X \cap Y)=0.5+0.3-0.1=0.7$$

Note that the order in which you write an intersection or union does not matter:

$$P(X \cap Y)=P(Y \cap X) \quad \text{and} \quad P(X \cup Y)=P(Y \cup X)$$

Conditional probability uses the "|" symbol to denote "given". Therefore, to denote the probability that a dot is in X given it is in Y, you would write $\boldsymbol{P(X \mid Y)}$. The computation is just an application of the basic probability formula stated earlier (Eqn 1). By saying "given Y", you are restricted to only looking at circle Y. That means the total number of possible outcomes is contained in circle Y (this will be the denominator in Eqn 1). The area of interest is where dots are contained in circle X and also in circle Y. That is the shaded region in the Venn diagram, which is the intersection of X and Y. If you follow that, you should get:

$$P(X \mid Y) = \frac{1}{3} = 0.\overline{33}$$

Here's the basic equation:

$$\boxed{P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)}} \qquad \text{Eqn 3}$$

An easy way to remember this equation is to realize that whatever is to the right of "|" goes in the denominator. Just for kicks, try to figure out the probability of Y given X using the equation and then check it by looking at the diagram. Here's the solution (don't look until you try it!):

$$P(Y \mid X) = \frac{P(X \cap Y)}{P(X)} = \frac{0.1}{0.5} = 0.2$$

Note, with a little basic algebra, you can calculate a joint probability if you're given the conditional probability:

$$\boxed{P(X \cap Y) = P(X \mid Y)P(Y) = P(Y \mid X)P(X)} \qquad \text{Eqn 4}$$

Prove it to yourself:

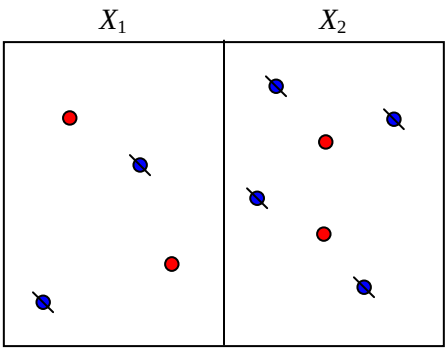
$$P(X \cap Y) = P(X \mid Y)P(Y) = 0.3\overline{3}(0.3) = 0.1$$

$$P(X \cap Y) = P(Y \mid X)P(X) = 0.2(0.5) = 0.1$$

Knowing how to modify Eqn 3 to get Eqn 4 and working with both, you should be able to answer all the probability questions in OR 310. The next section will go into other variations of Eqn 3 and build on that... hang on, this is really going somewhere.

Bayes' Theorem – we will cover Bayes' Theorem in class on Lesson 7.

Before getting into Bayes' Theorem, let's complicate the example we've been working with. Let's say the rectangle is a set X, with k subsets that are mutually exclusive and exhaustive. (Those are fancy math words. Mutually exclusive means the intersection of any two subsets is zero. Exhaustive means summing the probability of all the sets equals 1.) Each subset of X will be called X_i (for $i = 1, 2, ..., k$). To give you an idea of all that jabbering, the figure below shows an example with $k = 2$. You'll note, there are now red and blue dots in the diagram. (If you have a black and white printer, the blue dots have a line through them.) If you count up the dots and use Eqn 1, you'll find $P(X_1) = 0.4$ and $P(X_2) = 0.6$.



Using this specific example, we can use various combinations of Eqn 3 and Eqn 4 to compute a conditional probability without needing to know a joint probability. Let's try the probability that a given a red dot, it came from set X_1. From the diagram, a good guess would be 0.50 because half the red dots are in X_1 and half are in X_2. Let's see the math behind that seemingly simply observation. (This is not for the faint of heart.)

What you're trying to get (using Eqn 3) is:

$$P(X_1 \mid r) = \frac{P(X_1 \cap r)}{P(r)}$$

We're trying to build up to a case where we can compute this without having to compute the intersection. (In real world problems, the intersection is usually what we're looking for so we're trying to avoid having to compute it in advance.) You can use Eqn 4 to substitute for the intersection:

$$P(X_1 \mid r) = \frac{P(r \mid X_1)P(X_1)}{P(r)} \qquad \text{Eqn 5}$$

That wasn't so bad, was it? Let's assume we don't really know $P(r)$ either. We can split it by looking at both subsets of X as follows:

$$P(r) = P(X_1 \cap r) + P(X_2 \cap r)$$

That looks more complicated than what we need, and it will get more so because we don't want equations that use intersections. We can rewrite $P(r)$ without the intersections by applying the same trick we used before (use Eqn 4 twice):

$$P(r) = P(r \mid X_1)P(X_1) + P(r \mid X_2)P(X_2)$$

If we put that back into Eqn 5, we get:

$$\boxed{P(X_1 \mid r) = \frac{P(r \mid X_1)P(X_1)}{P(r \mid X_1)P(X_1) + P(r \mid X_2)P(X_2)}} \qquad \text{Eqn 6}$$

The trick to remembering this equation is to realize that you completely enumerate the event for the state of nature you want the probability of (i.e., **since you want to know the probability of *r*, you need to use all subsets containing *r*; so the denominator uses $P(X_1)$, $P(X_2)$, etc**.). This enumeration is being done to get the probability of the given state (r), which means you're adding the intersections so you multiply each $P(X_i)$ by the corresponding conditional probability ($P(r \mid X_i)$).

The numerator is simply the one term from the denominator that matches probability item on the left of the "|". (If you had been looking for $P(X_1)$ in the denominator, you would do it by using conditional probabilities for each type of dot, so the denominator would have terms with $P(r)$ and $P(b)$.) The general form of this equation is Bayes' Theorem:

$$P(X_1 \mid r) = \frac{P(r \mid X_1)P(X_1)}{\sum_{i=1}^{k} P(r \mid X_i)P(X_i)} \qquad \text{Eqn 7}$$

Applying it to the example we were looking at, first realize:

$$P(r \mid X_1) = \frac{2}{4} = 0.5 \quad \text{and} \quad P(r \mid X_2) = \frac{2}{6} = 0.3\overline{3}$$

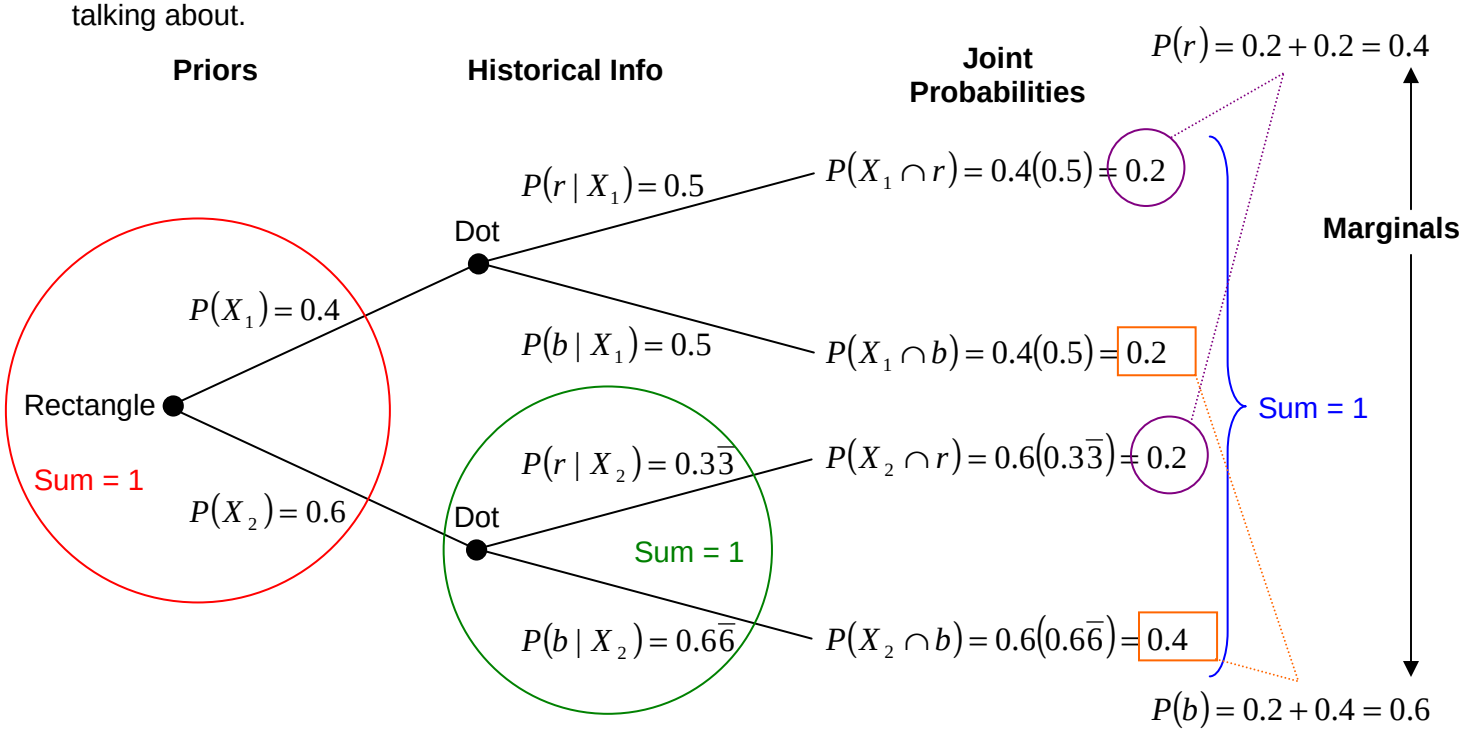
Therefore:

$$P(X_1 \mid r) = \frac{P(r \mid X_1)P(X_1)}{P(r \mid X_1)P(X_1) + P(r \mid X_2)P(X_2)} = \frac{0.5(0.4)}{0.5(0.4) + 0.3\overline{3}(0.6)} = 0.5$$

Ready for the good news? There are two other ways to figure this out without using the complicated Eqn 7. The first is the tree method.

Tree Method

The tree method starts you with the first event, which is broken down into all its states of nature (the "Priors"), each with a given probability. The second event is then broken down into all its states of nature ("Additional/Historical Info"), but this time you use probabilities conditioned on states of the preceding branch(s). When you reach the end of the tree, you simply multiply the probabilities from the beginning to the end of that branch to end up with a joint probability for all the states of nature ("Joint Probabilities"). The tree below gets this started for the example we've been talking about.



As an error check, you'll note that the sum of all the probabilities at each node is equal to 1 (as shown by the large circles). All the Joint Probabilities add up to 1 as well.

The point of using the tree (aside from Decision Trees) is to get the "Posterior Probabilities", which you may have guessed are the reverse conditional probabilities from the "Historical Info." These are calculated using Eqn 3 (i.e., divide the joint probabilities by the marginal probabilities (or just "Marginals") of each state of nature at the intermediate stem; $P(r)$ and $P(b)$). From the tree shown above, you can see how the "Marginals" are computed from the "Joint Probabilities". This is basically the denominator of Bayes' Theorem. To finish up the example:

$$P(X_1 \mid r) = \frac{0.2}{0.4} = 0.5$$

$$P(X_1 \mid b) = \frac{0.2}{0.4} = 0.5$$

$$P(X_2 \mid r) = \frac{0.2}{0.6} = \frac{1}{3} = 0.3\overline{3}$$

$$P(X_2 \mid b) = \frac{0.4}{0.6} = \frac{2}{3} = 0.6\overline{6}$$

If you think about it, each of these "Posterior Probabilities" is applying the simple version of Bayes' Theorem where $k = 2$ (Eqn 6). It works with the general case too if you have more branches.

Table Method

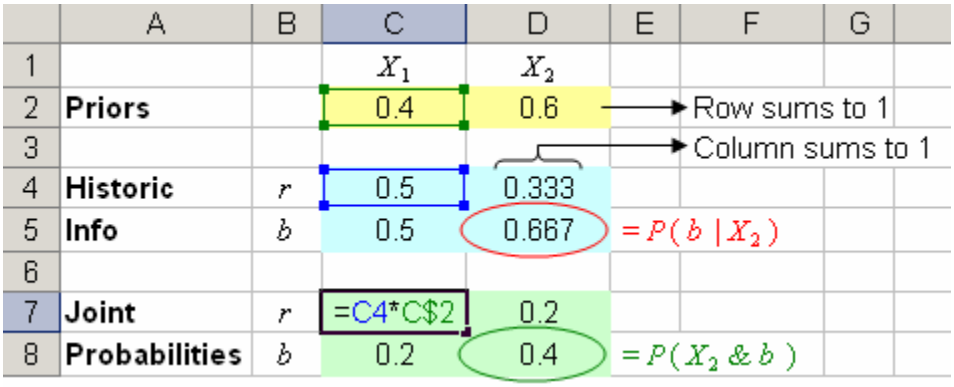
Since most of the work we'll be doing in OR 310 centers around Excel, there's one more method to learn for applying Bayes' Theorem. This is not quite as easy to follow as the tree method, but it is much easier to incorporate into Excel. First you start with your "Priors" in to adjacent columns and on the same row. If you don't want to see the derivation, skip to the last table below (just before the Example section).

	A	B	C	D	E	F	G
1			X_1	X_2			
2	**Priors**		0.4	0.6	→	Row sums to 1	

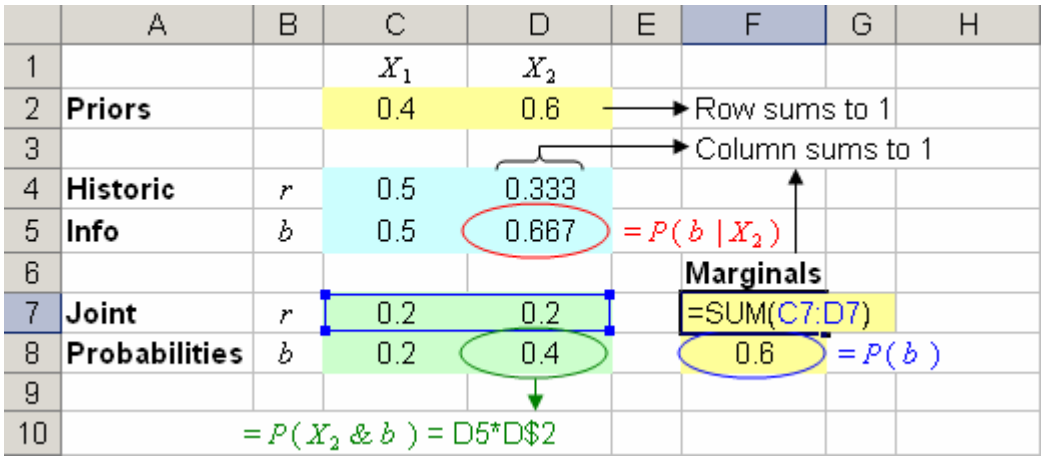
Next, you make a table of the "Historical Info" underneath these columns. (Remember that these are conditional probabilities of your second event based on the first.) Note how you can enter actual fractions into Excel. This is better than approximating by entering 0.667 even though that is what is displayed after you enter the formula.

	A	B	C	D	E	F	G
1			X_1	X_2			
2	**Priors**		0.4	0.6	→	Row sums to 1	
3					→	Column sums to 1	
4	**Historic**	r	0.5	=2/6			
5	**Info**	b	0.5	0.667	$= P(b \mid X_2)$		

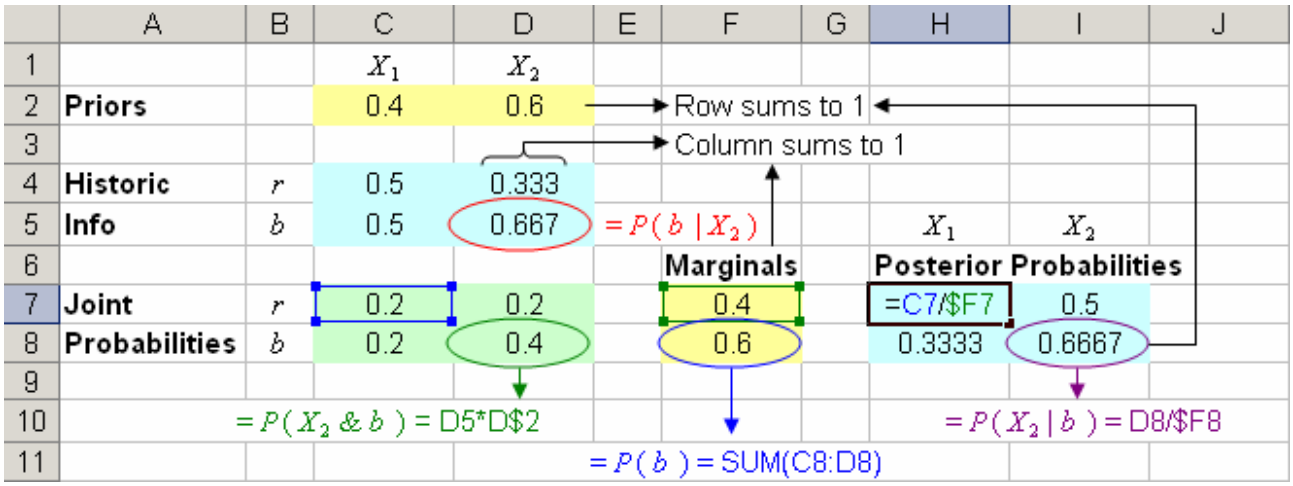
Now you have to enter equations to get the "Joint Probabilities." You're basically **multiplying** each cell in the "Historical Info" table with the respective "Prior" cell (in the same column). If you know what you're doing in Excel, you can use a combination of relative and absolute cell references in your equation to only enter it once and then copy that formula to the other three cells.



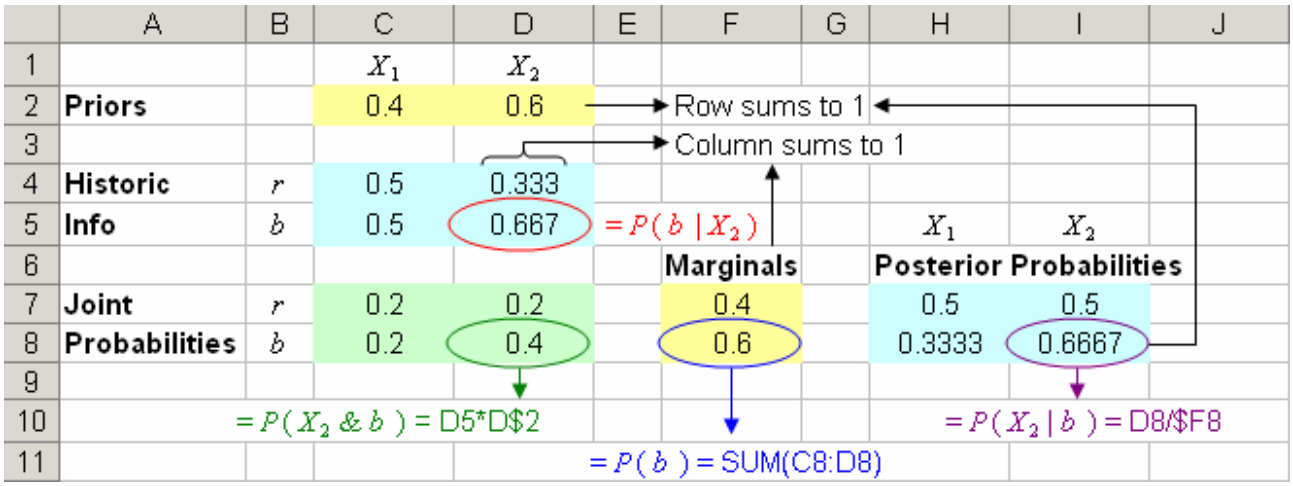
To get the denominator for Bayes' Theorem, you **add across** each row in the new "Joint Probabilities" table.



Finally, you get your "Posterior Probabilities" by **dividing** each cell from the "Joint Probabilities" table by the respective denominators you just calculated. Again, you can do this with a single formula with relative and absolute references as shown below.



This is the final product:



	A	B	C	D	E	F	G	H	I	J
1			X_1	X_2						
2	**Priors**		0.4	0.6						
3										
4	**Historic**	r	0.5	0.333						
5	**Info**	b	0.5	0.667				X_1	X_2	
6						**Marginals**		**Posterior Probabilities**		
7	**Joint**	r	0.2	0.2		0.4		0.5	0.5	
8	**Probabilities**	b	0.2	0.4		0.6		0.3333	0.6667	
9										
10										
11										

The great thing about the table method is that you only have to do it once. All you have to do now is change the information for the "Priors" and "Historical Info" and all the other numbers are updated automatically.

Example

Suppose I have red and green marbles. I put 7 red marbles and 3 green marbles in a paper bag ($bag1$). Then I put 4 red marbles and 6 green marbles in another bag ($bag2$).

1. Without looking, what is that probability that you can correctly select bag1? (This is the same as the probability of selecting bag2. Also, note if you select a marble and don't know the color, it is equally likely to come from either bag since each has a total of 10 marbles.)

$$P(bag1) = \_\_\_\_\_$$

2. Compute the joint probabilities ("Historical Info") for selecting a specific color marble given a specific bag.

$$P(r \mid bag1) = \_\_\_\_\_$$

$$P(g \mid bag1) = \_\_\_\_\_$$

$$P(r \mid bag2) = \_\_\_\_\_$$

$$P(g \mid bag2) = \_\_\_\_\_$$

3. Let's say you're in need of a nice green marble to complete your collection. Only the green marbles in $bag2$ are nice. If I hide the bags behind my back and ask you to pick left or right hand, what is the probability of randomly selecting a good green marble?

$$P(g \cap bag2) = \_\_\_\_\_$$

4. Suppose I hide the bags and select a marble without you knowing which bag it came from. If the marble is green (and you can't tell if it's nice or not), what is the probability that it came from *bag*2?

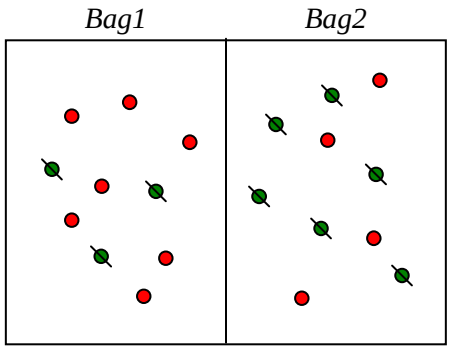
$$P(bag2 \mid g) = \_\_\_\_\_$$

5. Repeat 3 & 4 using a red marble.

$$P(r \cap bag2) = \_\_\_\_\_$$

$$P(bag2 \mid r) = \_\_\_\_\_$$

This figure may help, but try not to use it unless you really need to. (The green dots have lines through them.)



Answers

Hard way... using the math:

1. $P(bag1) = \frac{1}{2} = 0.5$

2. $P(r \mid bag1) = \frac{7}{10} = 0.7$, $P(g \mid bag1) = \frac{3}{10} = 0.3$, $P(r \mid bag2) = \frac{4}{10} = 0.4$, $P(g \mid bag2) = \frac{6}{10} = 0.6$

3. $P(g \cap bag2) = P(g \mid bag2)P(bag2) = 0.6(0.5) = 0.3$

4. $P(bag2 \mid g) = \frac{P(g \cap bag2)}{P(g)} = \frac{0.3}{9/20} = \frac{2}{3} = 0.6\overline{6}$

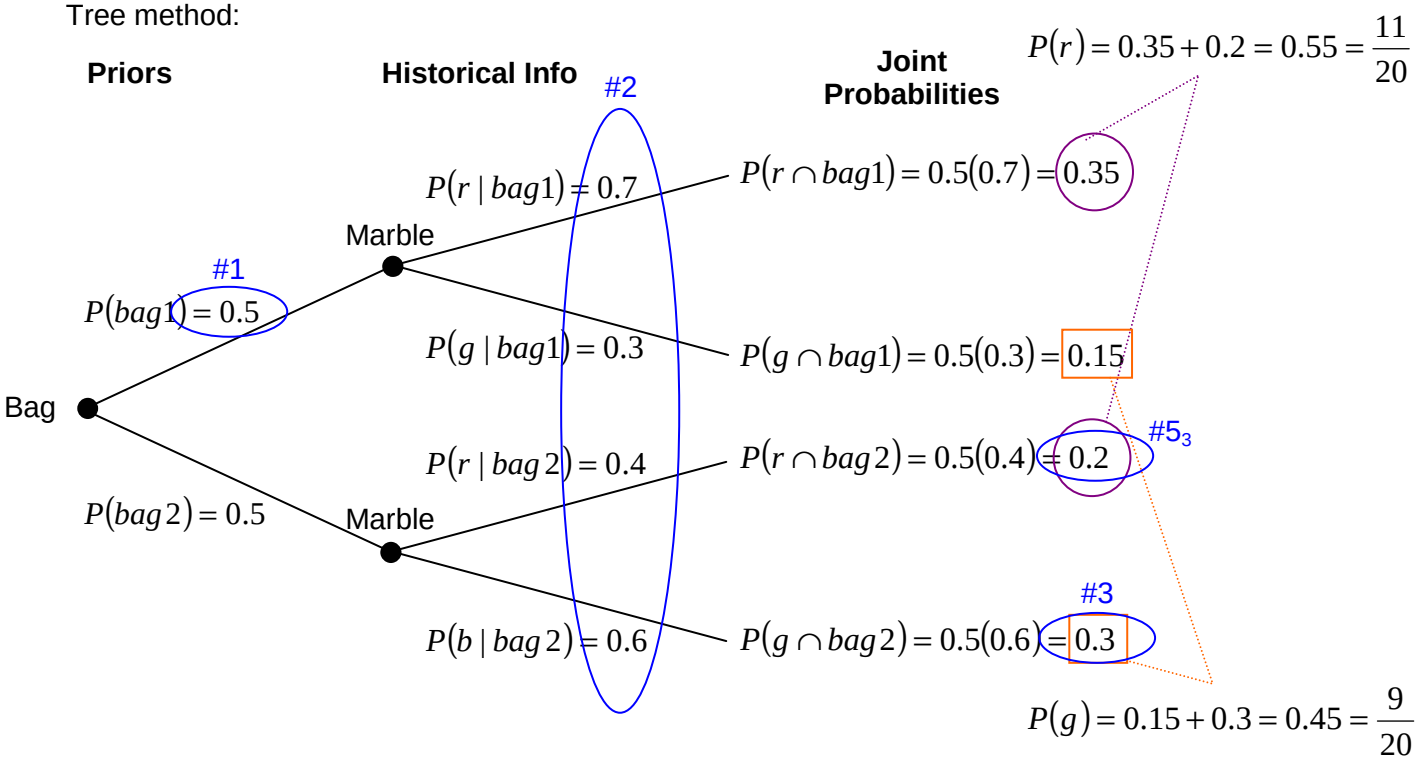
5. $P(r \cap bag2) = P(r \mid bag2)P(bag2) = 0.4(0.5) = 0.2$

   $P(bag2 \mid r) = \frac{P(r \cap bag2)}{P(r)} = \frac{0.2}{11/20} = \frac{4}{11} = 0.\overline{36}$

Easy way... table method:

	A	B	C	D	E	F	G	H	I	J
1			*bag1*	*bag2*						
2	**Priors**	#1	0.5	0.5	→	Row sums to 1	←			
3					→	Column sums to 1				
4	**Historic**	*r*	0.7	0.4	#2					
5	**Info**	*g*	0.3	0.6	= *P*(*g* | *bag2*)			*bag1*	*bag2*	
6						**Marginals**		**Posterior Probabilities**		
7	**Joint**	*r*	0.35	0.2 #5₃		0.55		0.6364	0.3636 #5₄	
8	**Probabilities**	*g*	0.15	0.3		0.45		0.3333	0.6667	
9				#3					#4	
10			= *P*(*bag2* & *g*) = D5*D$2					= *P*(*bag2* | *g*) = D8/$F8		
11				= *P*(*g*) = SUM(C8:D8)						

Tree method:



$$P(bag1 \mid r) = \frac{0.35}{0.55} = 0.\overline{63}$$

$$P(bag1 \mid g) = \frac{0.15}{0.45} = 0.3\overline{3}$$

$$P(bag2 \mid r) = \frac{0.2}{0.55} = 0.\overline{36} \quad \#5_4$$

$$P(bag2 \mid g) = \frac{0.3}{0.45} = 0.6\overline{6} \quad \#4$$