What I should have learned in ECON 30330

Bill Evans
Spring 2009

ECON 30330 is prerequisite for ECON 30331. Below is a 19 page handout that outlines most of the important concepts from ECON 30330 that will be utilized in this course, ECON 30331. If you are not familiar with these concepts, you will need to review quickly and get up to speed.

Random variables are variables whose outcome is determined in part by chance. There are two types of random variables

> *Discrete* – a countable number of possible outcomes. Examples include a coin flip, throw of a die, a grade (A,B,C…) in a class, or scores on an SAT test, etc.
>
>> Suppose for a discrete random variable X there are n potential outcomes, $(x_1, x_2….x_n)$. Each outcome has a probability associated with it, p_i with the set if probabilities being $(p_1, p_2, p_3…p_n)$ These probabilities must satisfy two properties
>>
>> $$p_i \geq 0 \ for \ all \ i$$
>>
>> $$\sum_{i=1}^{n} p_i = 1$$
>>
>> The second condition simply states that the sum of all probabilities must equal 1.
>>
>> Example: Suppose that going into an undergraduate class, your chances of getting an A, B, C, D or F are 0.20, 0.25, 0.40, 0.10, and 0.05, respectively. Notice that the sum of the probabilities are 1.

> *Continuous* – an infinite number of possible values, such as measurements like height, weight and temperature.
>
>> A variable x is said to be a continuous random variable if it can be described with a function f(x) called a *probability density function* or *pdf,* that satisfies the following conditions
>>
>> $(a) \quad f(x) \geq 0 \ for \ all \ x$
>>
>> $(b) \quad \int_{-\infty}^{\infty} f(x)dx = 1$
>>
>> $(c) \quad for \ all \ a \ and \ b \ with -\infty < a < b < \infty$
>>
>> $$\Pr ob(a \leq x \leq b) \quad = \quad \int_{a}^{b} f(x)dx$$
>>
>> For continuous distributions, the area under the curve represents a probability, as in condition (c). For ANY continuous random variable, the integral over all possible values must be 1 (condition b).
>>
>>> Example: The exponential distribution describes the probability a product/person/object will fail before a particular time, so an object can fail any time between 0 and ∞. If x is the time to failure, the *pdf* is defined as
>>>
>>> $f(x) = \lambda e^{-\lambda x}$ for x>0 and λ>0

where λ is a non-zero parameter that describes the shape of the distribution. It is easy to demonstrate that f(x) is a proper *pdf*, that is, it integrates to 1

$$\int_{O}^{\infty} f(x)dx = \int_{0}^{\infty} \lambda e^{-\lambda x}dx = -e^{-\lambda x}\mid_{0}^{\infty} = -e^{-\lambda \infty} -- e^{-0} = 0--1 = 1$$

In figure 1a below, I graph the exponential distribution for λ=0.5. Notice the PDF declines continuously as X increases.
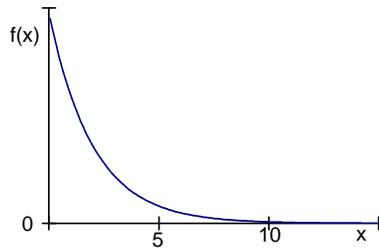

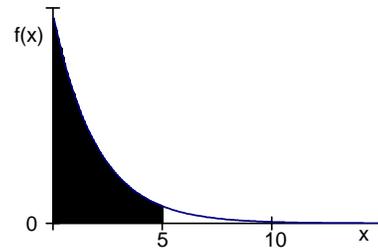Figure 1a   Figure 1b

Cumulative distribution functions (CDF): Let x be a random variable, either continuous or discrete. The CDF of the random variable is the probability the random variable is below a particular value "a"

$$\Pr ob(x \le a) = F(a)$$

For a discrete random variable, F(a) is

$$F(a) = \sum_{i=1}^{x} p_i \quad for \; x_i \le a$$

For a continuous random variable, F(a) is

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

And using the exponential distribution above, it is easy to show that the *cdf* is

$$F(a) = \int_{0}^{a} \lambda e^{-\lambda x}dx = \int_{0}^{a} -e^{z}dz = -e^{z}\mid_{0}^{a} = -e^{-\lambda a} - (-e^{0}) = 1 - e^{-\lambda a}$$

The shaded area in Figure 1b above is the CDF for the exponential distribution where λ=0.5. This particular CDF is Prob(X$\le$5)=F(5). Using the definition of the exponential CDF above F(5) when λ=0.5 is therefore $1-e^{-0.5(5)} = 1-0.082$ =0.918.

Expected value is a measure of central tendency. It represents what we expect the "average" outcome of a random process to be if we had a large number of experiments that generated the random variable. The expectation operator for a random variable X is represented by the statement E[X] and we usually represent the expected value by the Greek letter mu (u_x). The calculation of the expected value differs for discrete and continuous distributions.

Discrete distributions:

Expected values are simply weighted averages – the sum of the possibilities times the probabilities

Suppose a random variable has n possible outcomes $X = \{x_1, x_2, \ldots x_n\}$ and the probabilities associated with these are $P = \{p_1, p_2, \ldots p_n\}$.

$$E[X] = x_1 p_1 + x_2 p_2 \ldots + x_n p_n = \sum_{i=1}^{n} x_i p_i = \mu_x$$

Example1: a die has six sides, all equally likely with probabilities 1/6. The expected value of a throw of a die is

$$E[X] = (1/6)(1) + (1/6)(2) + (1/6)(3) + (1/6)(4) + (1/6)5 + (1/6)6 = 3.5$$

Example 2: Suppose that going into an undergraduate class, your chances of getting an A, B, C, D or F are 0.20, 0.25, 0.40, 0.10, and 0.05, respectively. The expected quality points from the class are therefore

$$E[X] = 0.2(4) + 0.25(3) + 0.4(2) + 0.1(1) + 0.05(0) = 2.45$$

Continuous Distribution

Given a continuous random variable x with PDF f(x), the expected value is kind of a weighted average where the weights are the values of the PDF. Since the variable is continuous, we integrate rather than sum

$$E[x] = \int_{\infty}^{\infty} xf(x)dx = u_x$$

Example 1: Suppose that the daily high temperatures T in a city are uniformly distributed over the interval [40,90]. Therefore, the PDF is defined as

f(T) = 1/50 for 40≤T≤90

It is easy to demonstrate this is a proper PDF, i.e., it integrates to 1

$$\int_{40}^{90} \frac{1}{50} dt = \frac{T}{50}\Big|_{40}^{90} = \frac{90 - 40}{50} = 1$$

And the expected value is

$$E[x] = \int_{40}^{90} \frac{1}{50} T dT = \frac{T^2}{2(50)}\Big|_{40}^{90} = \frac{90^2 - 40^2}{100} = 65$$

Example 2: The Pareto distribution is used to describe the distribution of income in the top portion of the income distribution. Therefore, if x is income, the pdf is defined for incomes in excess of T. It is usually used to describe the income in higher income groups. If x is the random variable (income), the PDF for the Pareto distribution is

3

$f(x) = kT^k x^{-(k+1)}$ for k>0 for the values of x in the range (T, ∞)

The parameter k describes the shape of the distribution and it is required to be positive. It is easy to show f(x) integrates to 1

$$\int_T^\infty kT^k x^{-(k+1)} dx = -T^k x^{-k} \Big|_T^\infty = -T^k \infty^{-k} - -T^k T^{-k} = 0 - -1 = 1$$

The expected value is therefore

$$E[x] = \int_T^\infty xf(x)dx = \int_T^\infty xkT^k x^{-(k+1)} dx = \int_T^\infty kT^k x^{-k} dx =$$
$$- kT^k x^{-k+1}/(k-1) \Big|_T^\infty = -kT^K \infty^{-k+1}/(k-1) - -kT^k T^{-k+1}/(k-1) = kT/(k-1)$$

Example: Consider a sample of people with incomes in excess of $50,000 (X>$50,000) and suppose K=5. In this case, E[X] for X>$50,000 equals kT/(k-1) = 5(50,000)/4=$62,500

Properties of expectations:

If x is a random variable and a and b are constants, then:

E[ax] = aE[x]
E[x + b] = E[x] + b
E[ax + b] = aE[x] + b

Example: Temperatures in Fahrenheit (F) can be converted to Celsius (C) using the simple linear transformation C = -17.78 + 0.556F. If the average daily high temperature in F is u_f=72 degrees, the average daily high in C is therefore u_c = -17.78 + 0.556u_f = 22.2

Estimation: The expected value is estimated by the sample average. Let x_i be a random and independent draw to a distribution with expected value u_x and suppose there are n such draws, $\{x_1, x_2, ...x_n\}$.

$$\bar{x} = (1/n) \sum_{i=1}^n x_i$$

In a random sample of size n, the people selected for inclusion in the sample are drawn randomly. Since $\bar{x}$ is a function of the n people picked at random, then $\bar{x}$ is itself a random variable.

Suppose you sample 100 adult males and record their weight. For every sample of 100, you would get a different estimate for $\bar{x}$.

It can easily be shown that the sample average is an unbiased estimate of the expected value. And unbiased estimate is one that returns, on average, the true value. So if the random variable is draw from a population with a mean of u_x then

$$E[\bar{x}] = u_x$$

4

This does not mean your estimate is the correct one – it simply means that on average, the procedure you used to produce the estimates will generate the correct estimate.

Variance: is a measure of dispersion. It is a measure of how much volatility there is in a random variable from one observation to the next. Variances are usually represented by the lower case sigma squared (σ^2). The definition of a variance for a random variable x is

$$E[(x - E(x))^2] = E[(x - u_x)^2] = \sigma^2_x$$

The variance is the expected squared deviation in values from the expected value (u_x). Given the properties of expectation defined above, it is straightforward to show that

$$\sigma^2_x = E[(x - u_x)^2] = E[x^2] - u_x^2$$

Unfortunately, the units of measure for the variance is in units squared. So if x measures dollars, the units σ^2_x are dollars squared. Therefore, we typically report the *standard deviation* of variance (σ_x), which is the square root of variance because the units of measure of σ_x are in the same units of u_x.

To calculate the variance, note that $\sigma^2_x = E[x^2] - u_x^2$ so one needs to first calculate the expected value u_x, and square it, then subtract from this value $E[X^2]$ which is simply

$$E[x^2] = \int_{-\infty}^{\infty} x^2 f(x)dx = \sigma^2_x$$

Properties of variances:

It is easy to show that if x is a random variable and a and b are constants:

$V[ax] = a^2V[x]$
$V[x + b] = V[x]$
$V[ax + b] = a^2V[x]$

Example: Temperatures in Fahrenheit (F) can be converted to Celsius (C) using the simple linear transformation C = -17.78 + 0.556F. If the standard deviation of the daily high temperature in F is σ_f=18, what is the standard deviation in Celsius? $\sigma^2_c = 0.556^2\sigma^2_f = 100.16$, therefore $\sigma_c = 10.01$.

Estimation: Given n observations for X, the sample variance is given by the equation

$$s^2 = \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

The estimated sample variance is usually represented by s^2.

The Standard Normal Distribution: The most used distribution is of course the standard normal distribution. The standard normal is a symmetric, unimodal distribution that has a 'bell curve' shape. The PDF is defined as

$$\phi(z) = \frac{1}{(2\pi)^{0.5}} e^{-0.5z^2} \ \ for \ \ -\infty < z < \infty$$

And by definition, the CDF is

$$\Phi(a) = \int_{-\infty}^{a} \phi(z)dz = \int_{-\infty}^{a} \frac{1}{(2\pi)^{0.5}} e^{-0.5z^2} dz$$

The standard normal E[z]=u=0 and Var(z) = σ^2=1. The CDF is the area under the curve so graphically, the CDF $\Phi(a)$ is represented as the shaded area in Figure 2a below.

Figure 2a Figure 2b Figure 2c
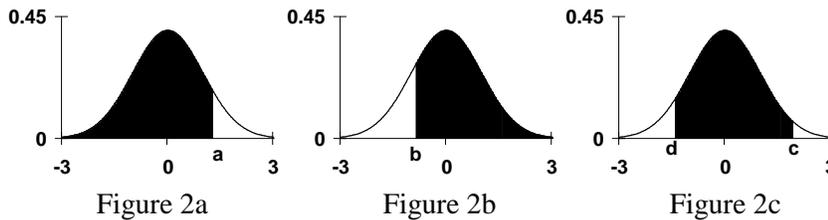
Most software packages including Excel have functions that evaluate the standard normal distribution. However, in statistics classes, the evaluations are typically done with tables. Below in Table 1 is a typical table that evaluates the standard normal CDF. This table reports the area Prob(Z≤a) for "a" reported out to two digits. To conserve space, I only report the positive values for "a" – most books also report results for negative values of "a" as well.

To use the table, suppose Z is normally distributed with a mean 0 and variance 1 and you are asked to evaluate $\Phi(1.65)$ = Prob(Z≤1.65). The left vertical axis provides the left most digits for "a" and the top horizontal axis provides the second digit after the decimal. To calculate the answer for this question, read down to 1.6, read over to 0.05 and $\Phi(1.65) = 0.9505$.

Suppose you wanted to estimate the opposite of the CDF, namely, the Prob(Z>b). This area is represented as the shaded portion of figure 2b above. Recognizing that Prob(Z≤b) + Prob(Z>b) = 1, then Prob(Z>b) is simply calculated as 1- Prob(Z≤b). Example: Prob(Z>2.27) = 1 – Prob(Z≤2.27) = 1 – 0.9884 = 0.0116.

Finally, you can calculate the probability a random variable falls between two values, such as Prob(d < Z ≤ c). This probability is represented as the shaded area in Figure 2c and using the results from the previous two examples, this area is $\Phi(c) - \Phi(d)$. Prob(0.93 < Z ≤ 2.59) = $\Phi(2.59) - \Phi(0.93)$ = 0.9952 -0.8238 = 0.1714.

The Normal Distribution: The normal distribution has a very similar shape to the standard normal but in this instance the mean (u) and the variance (σ^2) need not be zero and/or one respectively. The normal distribution is a two parameter distribution which means that the entire shape of the PDF can be described if we know two parameters: the mean (u) and the variance (σ^2). If a random variable Y is normally distributed, we usually represent this with the statement Y~ N(u, σ^2).

We can transform any normally distributed random variable into a standard normal distribution to solve problems. Suppose that Y~ N(u, σ^2) then the transformed random variable Z=[(Y-u)/ σ]~N(0,1). REMEMBER: DIVIDE BY σ, NOT σ^2.

> *Example 1*: Suppose Y~N(10,9). What is Prob(Y≤16)? Prob[Z≤(16-u)/σ] = Prob[Z≤(16-10)/3] = $\Phi(2)$=0.9772.

6

Example 2: A school district defines students as having 'special educational needs' if their IQ scores are 80 points or lower or in excess of 130 points. In this school district, the IQ $\sim N(100,15^2)$. What fraction of students will have special educational needs in this district?

$$\text{Prob(Special needs)} = \text{Prob(IQ} \leq 80) + \text{Prob(IQ} > 130)$$
$$\text{Prob(IQ} \leq 80) = \text{Prob}(Z = \leq (80-100)/15) = \text{Prob}(Z \leq -1.33) = 0.0918$$
$$\text{Prob(IQ} > 130) = 1 - \text{Prob(IQ} \leq 130) = 1 - \text{Prob}(Z \leq (130-100)/15) =$$
$$1 - \text{Prob}(Z \leq 2) = 1 - 0.9772 = 0.0228$$

$$\text{Prob(IQ} \leq 80) + \text{Prob(IQ} > 130) = 0.0918 + 0.0228 = 0.1146$$

Table 1: Cumulative area under the Standard Normal Distribution

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Conditional probabilities and independent events

Consider two random outcomes X and Y. Sometimes we are interested in the conditional probability of an event, that is, what is the probability that a value for Y is observed conditional on observing some value of X? Will it rain today given that is rained yesterday? What is the chance I get an A in physics given that I earned a D in Calculus?

Mathematically, conditional probabilities are stated as

$$Pr(Y=b \mid X=a).$$

Or what is the probability of observing a value of b for Y given we already observe a value of a for X? Example: when you apply to college, there are two potential outcomes, you can get in, Y=yes, or not, Y=no. At Notre Dame last year, there were 14,000 applicants for undergraduate admission and roughly 25% were accepted. The unconditional probability of admission is 25%, or $Pr(Y=yes) = 0.25$. However, not all people have the same probability of admission because people have different characteristics. You may want to ask what is the probability of admission if your total SAT scores T= M+V were less than 1000? $Pr(Y=yes \mid T<1000)$? In this instance, $Pr(Y=yes \mid T<1000)$ would be far less than 25%. In contrast, $Pr(Y=yes \mid T>1550)$ is much greater than 0.25. Therefore, in this instance, the realization of T conveys lots of information about Y.

Sometimes the realization of X conveys no information about the likelihood of admission. In that case

$$Pr(Y=b \mid X) = Pr(Y=b)$$

In this instance, the unconditional probability is the same as the conditional probability. When these two probabilities are equal, the variables are considered "independent" and the realization of X conveys no information about the likely value of Y.

Example 1: Among all births in the US, 51.4% are males and 48.6% are females. Consider a mom who has one girl already. What is the chance she will have a boy on the second birth?

The data demonstrates that among all second births, 51.4% are boys, or $Pr(Y=$Boy on 2nd birth$)=.514$. Not surprisingly, the conditional probability does not differ. In this case

Prob(Y=Boy on 2nd birth | girl on 1st birth) = 0.514
Prob(Y=Boy on 2nd birth | boy on 1st birth) = 0.514

Because the unconditional probability Prob(Y=Boy on 2nd birth) equals the conditional, Prob(Y=Boy on 2nd birth | girl on 1st birth), the sex of children on the 1st and 2nd births are independent – that is, the realization of sex on the 1st birth conveys no likely information about the sex on the second.

Example 2: Among traffic accidents where an unbelted passenger is killed (an accident severe enough to kill someone in the car), the unconditional probability the driver dies is roughly 24%. However, the probability varies considerably based on whether the driver was wearing a seat belt or not.

Prob(Y=driver dies in crash) = 0.24

Prob(Y=driver dies in crash | X=wearing a seat belt) = 0.17

Prob(Y=driver dies in crash | X= not wearing a seat belt) = 0.34

In this case, there is a lot of information contained in the fact that a driver is wearing a seat belt. In fact, Prob(Y=driver dies in crash | X= not wearing a seat belt)/ Prob(Y=driver dies in crash | X=wearing a seat belt) = 2 – unbelted drivers are twice as likely to die than belted drivers.

Conditional probabilities can be calculated used Baye's Law which is simply Prob(Y | X) = Prob(Y ∩ X)/Prob(X). The statement Prob(Y ∩ X) is the joint probability that events Y and X occur.

These calculations are easily illustrated by a simple 2 x 2 table. Below is some information about adults aged 50-64 in the US. There are 2 potential dichotomous outcomes: X measures whether the person is a smoker at the time of the survey and this can take two values (smoker and non smoker) while Y measures whether the person died within five yes and again, there are 2 values (yes and no).

| | | Currently smokes (X) | | Column totals |
		Yes	No	
Will die within 5 years (Y)	Yes	0.023	0.027	0.050
	No	0.227	0.723	0.950
	Row totals	0.250	0.750	1.000

The row totals in the table reveal that 25% of the population currently smokes [Pr(X=smoke) = 0.25] while the column indicate that 5% will die within the next 5 years [Pr(Y=Die in 5 years=0.050]. The column elements are joint probabilities. For example Pr(X=yes ∩ Y=yes) = 0.023, or 2.3% of the population currently smokes and will die in the next 5 years while Pr(Y=no ∩ X=no) = 0.723 or 72.3 percent of those surveyed are neither a smoker nor will they die in 5 years.

Note that by Baye's Law, Pr(Y=yes | X=yes) = Pr(Y=yes ∩ X=yes)/Pr(X=yes) = 0.023/0.250 = 0.092, or 9.2 percent of smokers will die within the next 5 years

Note also that Pr(Y=yes | X=no) = Pr(Y=yes ∩ X=no)/Pr(X=no) = 0.027/0.750 =0.036, or among non-smokers, only 3.6% will die in the next five years.

Note further that Pr(Y=yes | X=yes)/Pr(Y=yes | X=no) = 0.092/0.036 = 2.55 or among adults aged 50-64, smokers have 2.55 times the five-year mortality rates that non-smokers have.

In this case Pr(Y=yes)=0.050 is different from Pr(Y=yes | X=yes) = 0.092 so smoking reveals a lot about future mortality and they are *NOT* independent events.

Covariance and correlation:

Covariance: Measures the co-movement between two variables. When one variable is above average, what is the expected magnitude of the other?
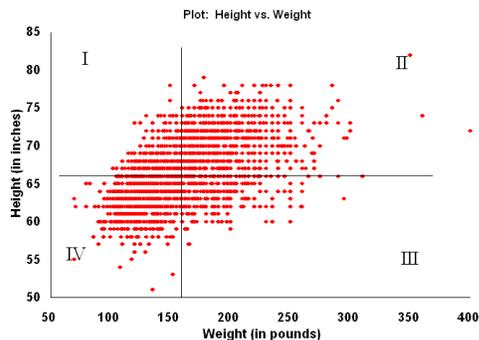
The picture above is a scatter plot of heights and weights of adult females. The horizontal and vertical lines in the middle of the table represent the average height and weight respectively, so the average adult female is 66 inches (5'6") and weights 160 pounds. Note that in quadrant II the dots represent people who have both above average height and weight while people in quadrant IV have below average height and weight. In contrast, quadrant I represents people with above average height and below average weight. Note that the bulk of the observations are in quadrants II and IV, so when height is below average, chances are, weight is below average. In this case, X and Y (height and weight) are positively correlated.

Mathematically, covariance is defined as follows.

$$\text{Cov}(X,Y) = \sigma_{xy} = E[(X - u_x)(Y - u_y)]$$

if $\sigma_{xy} > 0$ then when $X > u_x$, we also expect $Y > u_y$

if $\sigma_{xy} < 0$ then when $X > u_x$, we also expect $Y < u_y$ and vice versa

Estimation: Given n observations for both X and Y

$$\hat{\sigma}_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

The covariance is scale dependent – change the scale of X and/OR Y, you change the value of the covariance. So for example, the covariance between weight in kilograms and height in centimeters will be different than the covariance between weight in pounds and height in inches. Scale dependence is not a good thing. The extended example below illustrates this result.

For example, suppose height in weight (X and Y) are measured in inches and pounds, are transformed into centimeters and kilograms (W and Z). These transformations are linear combinations so we can represent them as:

W = a + bX
Z = c + dY

Where a, b, c, and d are constants and let the original covariance be $\text{Cov}(X,Y) = \sigma_{xy}$.

What is Cov(W,Z)?

$$\text{Cov}(W,Z) = \sigma_{wz} = E[(W - u_w)(Z - u_z)]$$

Note from above that $u_w = a + bu_x$ and $u_z = c + du_y$

$$\text{Cov}(W,Z) = \sigma_{wz} = E[(a + bX - a - bu_x)(c + dY - c - du_y)]$$

$$= E[(bX - bu_x)(dY - du_y)]$$

$$= E[b(X - u_x)d(Y - u_y)]$$

$$= E[bd(X - u_x)(Y - u_y)]$$

$$= bdE[(X - u_x)(Y - u_y)] = bd\, \sigma_{xy}$$

10

Therefore, when we changed how X and Y are measured, the value of the covariance changed as well.

Correlation coefficient: Because covariances are scale dependent, their value changes based on how X and Y are measured. A scale independent measure of co-movement between two variables is the correlation coefficient

$$\rho(x, y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

By construction, the correlation coefficient is bounded between -1 and 1

$$-1 \leq \rho(x,y) \leq 1$$

Suppose you have two variables X and Y. How does the correlation coefficient between X and Y change if one changes the scale of the variables? For example, suppose height in weight (X and Y) are measured in inches and pounds, are transformed into centimeters and kilograms (W and Z). These transformations are linear combinations so we can represent them as:

$$W = a + bX$$
$$Z = c + dY$$

What happens to the correlation coefficient? Recall from the statements above about the properties of variances and covariances when variables are changed by linear combinations.

$$Var(W) = b^2 Var(X) = b^2 \sigma_x^2$$
$$Var(Z) = d^2 Var(Y) = d^2 \sigma_y^2$$
$$Cov(W,Z) = bd Cov(X,Y) = bd\sigma_{xy}$$

and therefore:

$$\rho(w, z) = \frac{\sigma_{wz}}{\sigma_w \sigma_z} = \frac{bd\sigma_{xy}}{b\sigma_x d\sigma_y} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \rho(x, y)$$

Unlike the covariance, the correlation coefficient is therefore scale invariant. The correlation coefficient between height and weight in inches and pounds is the same as the value when height and weight are measured in centimeters and kilograms.

Estimation of correlation coefficients. With estimates of the covariance and variances, the estimate of the correlation coefficient is:

$$\hat{\rho}_{xy} = \frac{\hat{\sigma}_{xy}}{s_x s_y}$$

Linking back to independence: two variables with zero covariance will also have a zero correlation coefficient. A zero correlation coefficient means that the two variables are statistically independent, that is, the realization that x is

above average conveys no information about the likely magnitude of y (and vice versa). A test that x and y are statistically independent is therefore a test that the estimated correlation coefficient is zero.

Linear combination of random variables

Many times, we can construct new random variables by combining two or more random variables. For example, your math (M) and verbal (V) SAT scores are both random variables with particular properties. Therefore, the combined score (T=M+V) is a random variable as well. The question is, what are the properties of the new random variable T?

What are the expected value and variance of a linear combination of random variables?

Consider the general case where X and Y are random variables with means of u_x and u_y and standard deviations σ_x and σ_y. Suppose that a third random variable Z is constructed using a linear combination of X and Y

$$Z = a + bX + cY$$

We can easily show, using the properties of expectations and variances above that

$$E[Z] = a + bu_x + cu_y$$

$$V[Z] = b^2\sigma^2_x + c^2\sigma^2_y + 2\ bc\sigma_{xy}$$

Expected value of Z:

$$E[Z] = E[a + bX + cY] = a + bE[X] + cE[Y] = a + bu_x + cu_y$$

Variance of Z:

$$V[Z] = E[(Z - u_z)^2] = E[(a+bX+cY - a-cu_x-cu_y)^2] = E[(b(X-u_x)+c(Y-u_y))^2]$$

$$=E[b^2(X-u_x)^2 + c^2(Y-u_y)^2 + 2bc(X-u_x)(Y-u_y)]$$

$$= b^2E[(X-u_x)]^2 + c^2E[(Y-u_y)]^2 + 2bcE[(X-u_x)(Y-u_y)]$$

$$= b^2\sigma^2_x + c^2\sigma^2_y + 2\ bc\sigma_{xy}$$

Note that given the definition of the correlation coefficient $\rho_{xy} = \sigma_{xy}/(\sigma_x\sigma_y)$, we can also write the V(Z) as

$$V[Z] = b^2\sigma^2_x + c^2\sigma^2_y + 2\ bc\sigma_x\sigma_y\,\rho_{xy}$$

Example 1: Suppose the expected values of SAT M and V are 550 and 570 respectively with standard deviations of $\sigma_m = 100$ and $\sigma_v = 96$ and assume that $\rho_{mv}=0.45$. What is the expected value and the variance of the total SAT score T=M+V? In this case T=a+bM+cV so a=0, b=c=1, so

$$T = a+bM+cV = M+V$$

$$E[T] = u_t = a+bu_m +c\,u_v = u_m +u_v = 550 + 570 = 1020$$

$$V[T] = \sigma_t^2 = b^2\sigma_m^2 + c^2\sigma_v^2 + 2\,bc\rho_{mv}\sigma_m\sigma_v = \sigma_m^2 + \sigma_v^2 + 2\,\rho_{mv}\sigma_m\sigma_v$$

$$= 100^2 + 96^2 + 2(0.45)(100)(96) = 27{,}856$$

$$\sigma_t = 166.9$$

Example 2: Suppose that 60% of a person's portfolio is in asset 1 and 40% are in asset 2. Expected returns for these assets are $u_1 = 6.1\%$ and $u_2 = 5.2\%$, the standard deviations of returns are $\sigma_1 = 2.1\%$ and $\sigma_2 = 2.6\%$ respectively, and assume that the correlation coefficient in returns is $\rho_{12} = -0.25$. What is expected returns and the standard deviation in returns for the portfolio? In the case, a=0, b=0.6 and c=0.4 so

$$r_t = \pi_1 r_1 + \pi_2 r_2 = 0.6r_r + 0.4r_2$$

$$E[r_t] = u_t = E[\pi_1 r_1 + \pi_2 r_2] = \pi_1 u_1 + \pi_2 u_2 = 0.6(6.1) + 0.4(5.2) = 5.74$$

$$V[r_t] = \sigma_t^2 = \pi_1^2\,\sigma_1^2 + \pi_2^2\,\sigma_2^2 + 2\pi_1\,\pi_2\,\rho_{12}\,\sigma_1\,\sigma_2$$

$$= 0.6^2(2.1^2) + 0.4^2(2.6^2) + 2(0.6)(0.4)(-0.25)(2.1)(2.6) = 2.014$$

$$\sigma_t \qquad = 1.42$$

Notice in this example that the standard deviation of the portfolio is much smaller than the standard deviation of either asset. This is because the expected returns are negatively correlated – the investor has picked assets with negative correlation to reduce the risk in the portfolio.

Variance of the sample mean

The mean or the sample average is the most frequent statistic constructed in applied work. Above we showed that the sample mean x̄ is a random variable (its value is dependent on the n observations selected at random to be in the sample) but we also demonstrated that the mean is an unbiased estimate of the underlying expected value.

In this section, we show another property of the sample mean, namely, its variance.

Recall $\bar{x} = (1/n)(x_1 + x_2 + x_3 \ldots.. + x_n)$, it is a linear combination of random variables. Remember also that each observation is independent of all the others and therefore $\text{Cov}(x_i, x_j) = 0$ for $i \neq j$. Using the definition of the variance of a linear combination above, if Z=a+bX+cY, and $\rho_{xy}=0$ then $V(Z) = b^2 V(X) + c^2 V(Y)$ – the complicated covariance term drops out. With this result we can generalize and construct the variance of the mean

Write x̄ as $x_1/n + x_2/n + x_3/n + \ldots.x_n/n$ and recall that the draws are independent, so $\sigma_{12}=0$, $\sigma_{23}=0$, etc.

Therefore, $V(\bar{x}) = (1/n)^2 V(x_1) + (1/n)^2 V(x_2) + \ldots.. (1/n)^2 V(x_n)$

Each random draw to x is made from the same distribution, one with mean u_x and a variance σ_x^2. Therefore

$$V(\bar{x}) = \sigma_x^2/n^2 + \sigma_x^2/n^2 + \sigma_x^2/n^2 \ldots + \sigma_x^2/n^2$$

There are n terms of σ_x^2/n^2 and then $V(\bar{x})$ collapses to $n\sigma_x^2/n^2 = \sigma_x^2/n$.

$$V(\bar{x}) = \sigma_x^2/n$$

As the sample size grows, the variance of the sample mean falls proportionally.

Central Limit Theorem

Suppose a sample of n observations is drawn from a population with mean u_x and variance σ_x^2. The Central Limit Theorem states that as the sample size gets large, the distribution of $\bar{x}$ approaches that of a normal distribution with mean of u_x and a variance of σ_x^2/n.

As n gets large, $\bar{x} \sim N(u_x, \sigma_x^2/n)$

Testing hypotheses about a sample mean

Suppose you constructed a sample mean $\bar{x}$, and you wanted to use this information to test a hypothesis about the underlying population.

For example, suppose that a cereal packaging machine is calibrated to distribute "a" ounces of cereal in each box. You want to test the machine to make sure it is working properly. Therefore, you select a sample of n boxes at random, calculate the average weight $\bar{x}$ and test whether the machine is, on average, dispensing "a" ounces of cereal. The null and alternative hypotheses are:

Ho: $u_x = a$
H_a: $u_x \neq a$

To test the hypothesis, we need to know the underlying distribution of $\bar{x}$. Unfortunately, we cannot directly use the results of the Central Limit Theorem because we do not know σ_x^2 but instead, we must estimate it. Therefore, we must use another distribution.

One can show that if the null hypothesis is true ($u_x = a$) then the ratio

$$\hat{t} = (\bar{x} - a)/[s/n^{0.5}]$$

is distributed as a student-t distribution with n-1 degrees of freedom. The student-t is a symmetric distribution with a mean of zero. It's shape looks a lot like the normal distribution and as n approaches ∞, the student-t approaches a normal distribution. The tests of hypotheses are based on $\hat{t}$. To test hypotheses, we can use either confidence intervals or t-test.

Confidence intervals: the confidence interval (CI) calculates the most likely values of the underling expected value given the a) sample size, b) estimate of $\bar{x}$, and c) the sample variance s^2. The confidence interval usually specifies the 95% most likely values of the expected value. The level of confidence that you want is usually measured as 1-α where α is the Type I error rate, that is, the chance you make the wrong decision. So for a 95% confidence interval, α=0.05 and for a 99% confidence interval, α=0.01.

Below in Figure 3 is a graph of the student t-distribution – it approximates the normal when the degrees of freedom (DOF) are large. Notice that the distribution is symmetric and unimodal.

Because $\hat{t}$ is symmetric, if the null hypothesis is true, then (1- α) % of the distribution for $\hat{t}$ will be between two numbers -$t_{\alpha/2}$(DOF) and $t_{\alpha/2}$(DOF) where $t_{\alpha/2}$ (DOF) is the (1-α) percent critical value of the t-distribution with a certain degree of freedom. (1- α) percent of the distribution lies between -$t_{\alpha/2}$ and $t_{\alpha/2}$ while $\alpha/2$ percent lies below –t and $\alpha/2$ percent lies above t. Given this fact and the definition of $\hat{t}$ above, the (1- α)% confidence interval for a is

$\bar{x} \pm t_{\alpha/2}$ (n-1)$s/n^{0.5}$

If the value of "a" lies outside the confidence interval, you reject the null hypothesis. If "a" lies inside the confidence interval, you cannot reject the null. In most instances, we construct a 95% confidence interval so α=0.05 and

14

α/2=0.025. To find the critical value of the t-distribution, we use a table like Table 2. In this Table, the vertical axis provides the degrees of freedom and the horizontal axis provides α/2. So for a t-test with 14 degrees of freedom and a 95% confidence level, we would look down to 14, over to 0.025 (α/2) and the critical value is therefore 2.145.
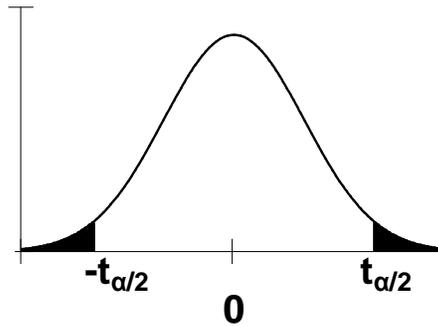

Figure 3: Critical values of a student t-distribution

Table 2

Critical values of student t-distribution

		α/2			
	0.100	**0.050**	**0.025**	**0.010**	**0.005**
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
90	1.291	1.662	1.987	2.368	2.632
120	1.289	1.658	1.980	2.358	2.617
infinity	1.282	1.645	1.960	2.326	2.576

Degrees of freedom

Example: A cell phone manufacturer advertises that the battery in their phone will allow for 90 minutes of talk time before running out. To test this claim, a famous consumer magazine ran a test on 20 fully charged cell phones and found that, on average ($\bar{x}$), the phones ran out after only 84.05 minutes of talk time. The standard deviation (s) of this estimate is 15.59 minutes. Test the hypothesis that the mean equals 90 using a 95% confidence interval.

Below are results from STATA for this experiment. In the data set, there is a variable titled "battery_life" and to test the hypothesis that mean battery life equals 90 minutes, you need to invoke the statement

```
ttest battery_life=90
```

With this data, can you reject the manufacturer's claim? The null and alternative hypotheses are:

$$\text{Ho: } u_x = 90$$
$$H_a: \ u_x \neq 90$$

The results from STATA are reported below. Looking on the table for the t-distribution, in this case, there are 20 observations and 19 degrees of freedom so the critical value of the t-value is $t_{.025}(n-1) = t_{0.025}(19) = 2.093$. Note that n=20, $\bar{x}$=84.05 and s=15.59, and plugging these values into the definition of the confidence interval, we find that

$$\bar{x} \pm t_{0.025}(n-1)s/n = 84.05 \pm 2.093(15.59/20^{.05}) = 84.05 \pm 7.30 = (76.75, 91.34)$$

Although the mean talk is below 90 minutes, 90 is a "likely" outcome and within the 95% confidence interval so one cannot reject the null that mean talk time is 90 minutes.

```
ttest battery_life=90

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean     Std. Err.    Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
batter~e |      20     84.0461     3.485876     15.58931     76.75008    91.34212
------------------------------------------------------------------------------
    mean = mean(battery_life)                                 t =  -1.7080
Ho: mean = 90                                  degrees of freedom =      19

    Ha: mean < 90                 Ha: mean != 90                 Ha: mean > 90
 Pr(T < t) = 0.0520          Pr(|T| > |t|) = 0.1039          Pr(T > t) = 0.9480
```

t-tests: As we noted above, if the null hypothesis is correct, the t-ratio is defined as

$$\hat{t} = (\bar{x} - a)/[s/n^{0.5}]$$

is distributed as a student-t with n-1 degrees of freedom. The student-t has a mean of zero and therefore, if the null is true, then $\hat{t}$ should be 'close' to zero. If $\hat{t}$ is a large positive or negative number, the null

hypothesis is most likely not true. 95% of the student-t distribution lies between $-t_{0.025}(n-1)$ and $t_{0.025}(n-1)$ these are the likely values if the null is true. Therefore, we can use the value of $\hat{t}$ in relation to these cutoffs to establish the null hypothesis.

if $|\hat{t}| < t_{0.025}(n-1)$ we cannot reject the null

if $|\hat{t}| \geq t_{0.025}(n-1)$ we can reject the null

Example: Continuing with the cell phone example above, $t_{0.025}(n-1) = 2.093$ and $\hat{t} = (84.05 - 90)/[15.59/20^{0.5}] = -5.95/3.49 = -1.71$. Therefore, we cannot reject the null that $u_x = 90$. Again, although the mean talk time is below 90 minutes, because of sampling variance, we cannot reject the null that it is in fact 90 minutes

Testing for the equality in means across samples

In many cases, we are interested in testing the null hypothesis that the expected value is the same across two samples. For example, in an experiment, we are interested in testing whether the outcome is the same in the treatment and control samples. In another example, you might want to examine whether alcohol use rates are the same at two different schools.

There are two populations, groups 1 and 2, and we want to test the null hypothesis that the expected values of a particular outcome are the same across the two groups. Therefore, if u_1 and u_2 are the expected values in the two groups, the null and alternative hypotheses are:

H_o: $u_1 = u_2$
H_a: $u_1 \neq u_2$

To perform this test we sample n_1 and n_2 observations from the two populations, construct the sample means ($\bar{x}_1$ and $\bar{x}_2$) and the sample standard deviations (s_1 and s_2). If the null hypothesis is correct than the difference in means should equal zero. Therefore, we construct

$$\hat{\Delta} = \bar{x}_1 - \bar{x}_2$$

And the null and alternative hypotheses are

H_o: $\hat{\Delta} = 0$
H_a: $\hat{\Delta} \neq 0$

Remember that $\bar{x}_1$ and $\bar{x}_2$ are random variables because they are a function of the observations selected at random to be in the samples. Therefore, their difference should be a random variable as well. If the null hypothesis is true, the expected value of Δ divided by its standard errors should be distributed as a Student t-distribution with n_1+n_2-2 degrees of freedom

$$\hat{t} = \frac{\hat{\Delta}}{s_p \left[\dfrac{1}{n_1} + \dfrac{1}{n_2} \right]^{0.5}} \sim t(n_1 + n_2 - 2)$$

The denominator in the statistic above is the square root of the variance of $\hat{\Delta}$ or the standard error of $\hat{\Delta}$

17

$$Var(\hat{\Delta}) = s_p^2 \left[\frac{1}{n_1} + \frac{1}{n_2} \right]$$

$$\text{where } s_p^2 \text{ is called the pooled var} iance \text{ and}$$

$$s_p^2 = [(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/[n_1 + n_2 - 2]$$

Where s^2_1 and s^2_2 are the estimated variances of X in the two samples, 1 and 2. As with the tests of hypotheses for single means, there are two ways to test the null hypothesis above: with confidence intervals and with t-tests.

Confidence intervals:

If the null is correct than the 95% confidence interval for the difference in means is

$$\hat{\Delta} \pm t_{0.025}(n_1 + n_2 - 2)s_p \left[\frac{1}{n_1} + \frac{1}{n_2} \right]^{0.5}$$

Example: Students in an electrical engineering class ran a series of experiments to examine whether an expensive name-brand battery (like Eveready or Duracell) had a longer life than a low-cost generic battery. The students purchased sets of AA batteries and placed them in a toy and measured the hours until the batteries wore out. The results of the experiment from STATA are reported below. The variable battery_type=1 if name brand and =0 of generic and the variable toy_life measured average hours until failure.

To invoke the test of difference in means across samples in STATA, we invoke the command

```
ttest toy_life, by(battery_type)
```

which generates the following results

```
ttest toy_life, by(battery_type)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.    Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      14    6.357092    .0771071     .2885083    6.190513    6.523672
       1 |      16    6.713234    .1008812     .4035248    6.498211    6.928257
---------+--------------------------------------------------------------------
combined |      30    6.547034    .0716923     .3926749    6.400407    6.693662
---------+--------------------------------------------------------------------
    diff |             -.3561414   .1298406                -.6221077    -.090175
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   -2.7429
Ho: diff = 0                                     degrees of freedom =        28

    Ha: diff < 0                 Ha: diff != 0                  Ha: diff > 0
 Pr(T < t) = 0.0052        Pr(|T| > |t|) = 0.0105          Pr(T > t) = 0.9948
```

Use these results to test the null hypothesis that there is no difference in battery life between the name brand generic batteries.

The null hypothesis h_0: $\hat{\Delta} = \bar{x}_g - \bar{x}_n = 0$. The value of $\hat{\Delta} = 6.357 - 6.713 = -0.356$

The pooled variance is calculated as

$$s_p^2 = \frac{(n_n - 1)s_n^2 + (n_g - 1)s_g^2}{n_n + n_g - 2} = \frac{(15)0.4035^2 + (13)0.2885^2}{28} = 0.12586$$

and therefore

$$s_p = 0.3548$$

Confidence intervals: The critical value of the t-statistic with 95% confidence interval for this test is as follows

$$\hat{\Delta} \pm t_{0.025}(n_n + n_g - 2)s_p[\frac{1}{n_n} + \frac{1}{n_g}]^{0.5} = -0.356 \pm 2.048(0.3548)(0.3660) = (-0.62, -0.09)$$

In this instance, because 0 is not within the confidence interval, we can conclude that the expected life of a name brand and generic battery differ.

t-tests: recall that when the null is true, the estimated difference in means divided the standard error is distributed as a t-distribution with n1+n2-2 degrees of freedom

$$\hat{t} = \frac{\hat{\Delta}}{s_p\left[\frac{1}{n_1} + \frac{1}{n_2}\right]^{0.5}} \sim t(n_1 + n_2 - 2)$$

If the null is correct, the estimated value of $\hat{t}$ should be close to zero – but how close? As was the case for confidence intervals, the 95% most likely values for $\hat{t}$ are between $-t_{\alpha/2}$ and $t_{\alpha/2}$ and therefore, if $|\hat{t}| > t_{\alpha/2}$ then we can reject the null hypothesis. From the printout, $\Delta = -0.3561414$, and we have calculated that $s_p=0.3458$ and $[(1/n_n) + (1/n_g)]^{0.5} = 0.3660$, so $\hat{t} = -0.35613414/[0.3458*0.3660] = -2.74$. The critical value of the t-distribution $t_{0.025}(28)=2.048$ and again, we can reject the null and conclude that the expected life of name brand and generic batteries differ.