

Quantitative Methods Lecture Notes
Levin et al. Chapters 1, 3, 6, 7, 8, 9, and 11

Chapter 1 – Introduction

I. This course deals with statistics. Two types:

Descriptive Statistics

Inferential Statistics – makes possible the estimation of a characteristic of a population based only on sample results.

The first part of this course will deal with descriptive statistics but we will quickly move to inferential stats – more interesting and useful for what we are after.

The second class (HCAI 5221) deals with the application of statistics to decision making in the business setting.

II. Types of data

Categorical random variables – yield categorical responses – yes/no, agree/disagree, etc.

Numerical RV can be discrete or continuous

III. Data Collection

Typically data are collected through sampling. The most common sample is a Simple Random Sample – every individual in the population has the same probability of being chosen. This can be done with replacement or without replacement.

Stratified Sample – The N individuals are first subdivided into separate subpopulations, or strata, according to some common characteristic. – this allows for oversampling of some groups to ensure representation.

Cluster Sample – The N individuals are divided into several clusters so that each cluster is representative of the entire population. Then a random sample of clusters are taken and all individuals in the selected clusters are studied. Counties, election districts, families, etc.

IV. Problems with Sampling

1. Selection Bias – occurs when certain groups in the population were not properly included. – black/white wage differential, medical trials (pneumonia study).
2. Nonresponse Bias – occurs when certain groups in the population do not respond to the survey. Mail or Telephone surveys.
3. Sampling Error – reflects the heterogeneity or chance differences from sample to sample, ± 4 percentage points, margin of error

4. Measurement Error – reflects inaccuracies in the recorded responses. Due to weakness in wording of survey, interviewer’s effect on respondent, effort made by respondent

We’ll get back to many of these issues later in the semester.

Chapter 3 –Measures of Central Tendency and Dispersion

This chapter deals with some basic descriptive statistics. Much of this you may already know, but we need to be careful to make sure we’re on the same page.

Consider the following: Data are Total Patient Care Revenues for a sample of hospitals in Duval County (Jacksonville, FL)

<u>Hospital</u>	<u>Revenue (in millions)</u>
1	414.6
2	358.6
3	439.8
4	64.8
5	159.2
6	130.5
7	395.3

Note: Hospitals all have different level of revenues
 The spread ranges from 64.8 million to 414.6 million
 Hospital 4 appears to have unusually low revenues – outlier?

We want to better understand how to describe such data.

I. Measures of Central Tendency, Variation, and Shape

Three measures of central tendency –arithmetic mean, the median, the mode

The Arithmetic Mean

$$\bar{X} = (\sum X_i)/n$$

Where \bar{X} is the sample mean
 n is the number of observations in the sample
 X_i is the ith observation of the variable X
 $\sum X_i$ is the summation of all X_i values in the sample

given our data we get:

$$(414.6 + 358.6 + 439.8 + 64.8 + 159.2 + 130.5 + 395.3) / 7 = 280.4$$

The average revenue in 1996 was 280.4 million

The Median

The median is the middle value in an ordered array of data. If there are no ties, half the observations will be smaller than the median and half will be larger. The median is unaffected by any extreme observations in a set of data.

64.8 130.5 159.2 358.6 395.3 414.6 439.8

So 358.6 is the median

If there are an odd number of observations then the median is represented by the numerical value corresponding to the middle value.

If there are an even number of observations then the middle is between two observations. The median in this case is the average of the two values.

The Mode

The mode is the value in a set of data that appears most frequently. There is no mode in the above data.

Suppose we have the following

County	# Hospitals
A	5
B	1
C	2
D	1
E	6
F	1
G	5
H	2
I	1
J	4

Now we can see that the mode is 1 -- it is most common for there to be 1 hospital in these counties. Note that the mean is 2.8 and the median is 2

Quartiles

Quartiles split the observations into four equal parts

The first quartile $Q_1 = (n+1)/4$ ordered observations

The Third quartile $Q_2 = 3(n+1)/4$ ordered observations

64.8 130.5 159.2 358.6 395.3 414.6 439.8

$$\text{So } Q_1 = 8/4 = 2$$

$$Q_3 = 24/4 = 6$$

So the second observation is Q_1 Q_3 is obs 6

Measures of Variation

Variation is the amount of dispersion or spread in the data. This gives us an idea of how spread out the data are.

Range

The range is the difference between the largest and smallest observations in the data. In our hospital revenue data the range is $414.8 - 64.8 = 350$.

This measures the total spread in the set of data. Note that while this is a simple measure to calculate it really is not all that helpful. Since one outlier can really affect the range.

The Interquartile Range

The interquartile range is obtained by subtracting the first quartile from the third quartile.

Variance and Standard Deviation

These are measures of how the values fluctuate about the mean. The sample variance is roughly the average of the squared differences between each of the observations in a set of data and the mean:

$$S^2 = (\sum_{i=1}^n (X_i - \bar{X})^2) / (n-1)$$

$i=1$ to n ,

So for our data we would get: 24,396.3 million

If the denominator was n instead of $n-1$ then the var would be the average of the squared differences around the mean. However, $n-1$ is used here because of certain desirable mathematical properties that using $n-1$ gives.

The Sample Standard Deviation is just the square root of the variance

$$S = \sqrt{S^2}$$

So $S = 156.2$

This implies that hospitals are clustering within 156.2 million of the mean.

Note that in the squaring process, observations that are further from the mean get more weight than are observations that are closer to the mean. Also if we did not square and just took the deviation from the mean that it would always equal zero. An alternative that one sometimes sees is the mean absolute deviation from the mean

Draw some “polygons” with same mean but with different variances to make the point.

Coefficient of Variation

The coefficient of variation is a *relative measure* of variation. It is always expressed as a percentage rather than in terms of the units of the data. The CV measures the scatter in the data relative to the mean

$$CV = (S/\bar{X})100\%$$

From our example $S = 156.2$, $\bar{X} = 280.4$, so $CV = 156.2/280.4 * 100 = 55.7$

This statistic is most useful when making comparisons across different types of data that might use different scales or different units of measurement. It makes it easier to compare apples to oranges.

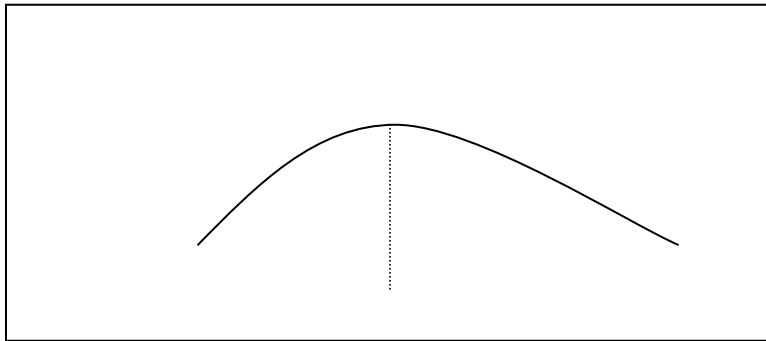
Eg, wait times for three different types of lab tests.

Comparing three different lab tests			
	test a	test b	test c
	12	54	105
	15	31	95
	20	54	110
	10	60	135
	7	60	187
	9	51	115
mean	12.17	51.67	124.50
st dev	4.71	10.75	33.37
cv	0.39	0.21	0.27

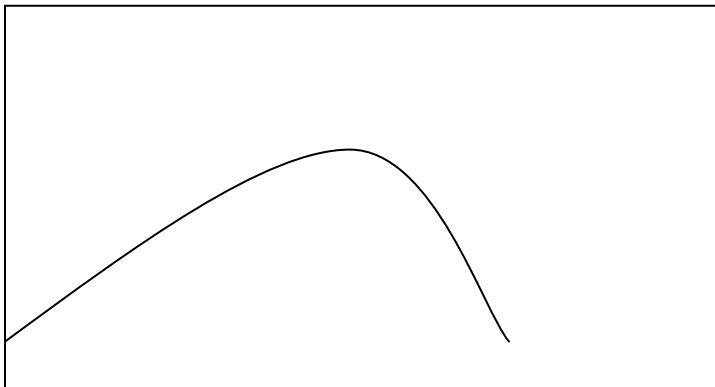
Note that if you are interested in comparing the variation across the three types of tests, the standard deviation is somewhat misleading since the scale of test a is much smaller than b or c. The coefficient of variation, however, shows that test a actually has the highest degree of variability, test b the lowest.

Shape

Shape deals with whether the data are symmetrical or not.



Positive or right skewed Mean > Median



If Mean = Median – then it is symmetrical.

Positive skewness occurs when the mean is increased by some unusually high values,
Negative occurs when extreme low values occur.

In our example Mean = 280.4 Median = 358.6 So we have negative skewness, The 65 is pulling the mean down relative to the median.

II. Obtaining Descriptive Summary Measures from a Population

What we did in the above section dealt with descriptive statistics from a sample. Suppose, however, that we have information about the entire population.

Year	Hospital Patient Discharges (1,000)
------	-------------------------------------

94	30,843
95	30,722
96	30,545
97	30,914
98	31,827
99	32,132

The Population Mean

$$\mu = \Sigma X/N$$

The only difference here is that we are dividing by the population sample size.
=31,164

The Population Variance and Standard Deviation

The population variance is represented by the symbol σ^2 (sigma) and the standard deviation is σ

$$\sigma^2 = (\Sigma(X_i - \mu)^2)/N$$

$$\text{and } \sigma = \sqrt{\sigma^2}$$

Note here that we are dividing by N and not N-1

WARNING: EXCELL ASSUMES SAMPLE So if you try to do =VAR(xxx) or STDEV it will divide by n-1 and not N. So make sure that is what you want. Need VARP or STDEVP if want population parameters.

$$\sigma^2 = 353,443$$

$$\sigma = 594.5$$

Note that there is not much variation here Pretty small CV 1.6 The observations are tightly clustered around the mean.

III. The Correlation Coefficient

This builds on the scatter diagram. It gives a numerical measure to the relationship between two variables.

<u>Hospital</u>	<u>Revenue (in millions)</u>	<u>Technology Index</u>
1	414.6	7.45
2	358.6	8.19

3	439.8	7.37
4	64.8	1.68
5	159.2	5.06
6	130.5	4.20
7	395.3	5.97

It looks like there is some correlation between the revenue of a hospital and the rarity of their technology

The strength of the relationship is measured by the coefficient of correlation, ρ , whose values range from -1 for a perfect negative correlation to 1 for a perfect positive correlation. Perfect implies that the scatter diagram would be a straight line.

Note that this does not imply anything about causation, only tendencies. That is just because there is a correlation there is not necessarily a causation – me and basketball!

The Sample Correlation coefficient, r , can be calculated as

$$r = \frac{\sum(X - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2} \sqrt{\sum(Y_i - \bar{Y})^2}}$$

When I do this for the above data I get .89

This suggests that there is a positive correlation between the technology a hospital uses and the size of total patient revenues, note that it is not a pretty strong relationship. Also, note we cannot imply causation. All we know is that they go together.

We will see later on that there is more to it than this, ρ (the population statistic) could well be closer to 0 and we just happened to get an r of .89, etc. We need to figure out how to account for sampling error. That is where we are headed.

Chapter 6 –Probability Distributions and the Normal Distribution

A **Random Variable** is a function or rule that assigns a number to each outcome of an experiment. This could be the number of heads observed in an experiment that flips a coin 10 times, or the annual revenue of a randomly selected hospital. Random variables can be discrete or continuous.

A **Probability Distribution** is a listing of all possible numerical outcomes from a random variable. There are discrete and continuous probability distributions. The binomial distribution and Poisson distribution are examples of discrete probability distributions. Chapter 5 in the text deals with these. We are skipping for sake of time.

Continuous Distributions

When we have a continuous RV we have another set of distributions to draw from. Since there are an infinite number of possible outcomes, one characteristic of a continuous distribution is that the probability of one exact value is zero, (probability of being exactly 6 feet tall is zero) thus we tend to measure things in intervals (between 5'9" and 6'0").

The Normal Distribution

One particularly useful distribution is the normal distribution. We will to a lot with this distribution throughout the remainder of the semester, thus the book devotes an entire chapter to it.

Properties of the normal distribution:

1. It is bell shaped (and thus symmetrical) in appearance
2. its measures of central tendency are all identical
3. its “middle spread” is equal to 1.33 standard deviations. This means that the interquartile range is contained within an interval of two-thirds of a standard deviations below and above the mean.
4. Its associated random variable has an infinite range.

Draw some normal distributions

We use the expression $f(X)$ to denote the probability density function. For the normal distribution the density function is:

$$f(X) = \frac{1}{\sqrt{2\pi}\sigma} * e^{-(1/2)[(X-\mu)/\sigma]^2}$$

So if we want to know the probability that a normally distributed random variable is between $-\infty$ and X we would just plug X into the above equation and it would tell us.

Note that since μ and σ will change for each possible distribution there are an infinite number of these.

$=\text{normdist}(x,\mu,\sigma,\text{cumulative})$ Is the Excel function to do this

So we would have to make this calculation every time we want to know something. One way around this is to transform the variable X to a specific normal distribution.

Suppose we Let:

$$Z = (X - \mu) / \sigma$$

Now this new variable Z will be normally distributed with a mean of zero and a standard deviation of 1.

Note that what this is doing is putting the variable in standard deviation units.

So if we have a normal distribution with a mean of 5 and a SD of 2 and we are concerned with the outcome 3, then

$Z = (3 - 5) / 2 = -1$. In other words 3 is one standard deviation below the mean.

Or if we are interested in the outcome 9:

$Z = (9 - 5) / 2 = 2$ or 9 is two standard deviations above the mean.

Now the density function for the standard normal distribution simplifies to:

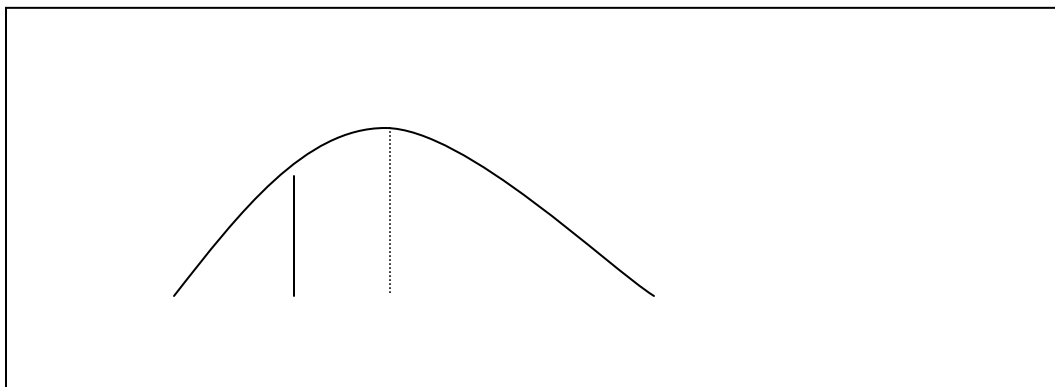
$$f(Z) = (1/\sqrt{2\pi})e^{-(1/2)Z^2}$$

This is easier to deal with since everything but Z is a constant.

EX:

Given a normal distribution with a mean of 50 and a sd of 4, what is the probability that:

1. $X < 43$



First convert to Z:

$$Z = 43 - 50/4 = -7/4 = -1.75$$

We can either look in a standard normal table or use NORMSDIST(Z) in Excel

We get .0401, or there is about a 4 percent chance that x is less than 43.

2. what is the probability that $x > 43$?

$$1 - .0401 = .9599$$

3. $42 < X < 48$

First draw the picture:

Calculate Zs

$$Z_{42} = 42 - 50/4 = -2$$

$$Z_{48} = 48 - 50/4 = -.5$$

$$P(X < 42) = .0228$$

$$P(X < 48) = .3085$$

$$\text{So } P(42 < X < 48) = .3085 - .0228 = .2858$$

4. $X > 57.5$?

$$Z = 57.5 - 50/4 = 1.875$$

$$P(X < 57.5) = .9696$$

$$\text{So } P(X > 57.5) = 1 - .9696 = .0304$$

Sampling Distributions

One of the main goals of data analysis is to make statistical inferences – that is to use information about a sample to learn something about the population.

Census vs CPS

To make life easier we often take a sample and use the sample statistics to say something about the population statistics, but in order to do this we need to say something about the sampling distribution. That is, there are lots of potential samples we could take of a population and each would give us slightly different information. Dealing with this sampling distribution helps us to account for this.

The Sampling Distribution of the Mean

Now we want to talk about making inferences about the mean of a population

The sample mean is an **unbiased** estimator of the population mean – this means that if we took all possible sample means and averaged them together, the average would equal the population mean.

Population		
Student	GPA	
A	3.82	
B	3.55	
C	2.89	
Mu	3.42	
Sigma	0.3906405	
	Sample	Xbar
AA		3.82
AB		3.685
AC		3.355
BA		3.685
BB		3.55
BC		3.22
CA		3.355
CB		3.22
CC		2.89
	Mux-	3.42
	Sigma(xbar)	0.276225
	Sigma/Sqrt(n)	0.276225

So $(6.5-6.25)/(3.62/5) = .345$, converting this to a probability ($\text{normsdist}(.345)=.635$, $1-.635=.365$). This tells us that there is a 36.5% chance of getting a sample mean of 6.5 hours or more.

Note that since the square root of n is in the denominator of the standard error, as the sample size increases the standard error will decrease, making any given sample mean less likely.

If $n=100$

$Z = (6.5-6.25)/3.65/10 = .691$, or only a 24.5% chance.

Sampling from Non-Normally Distributed Populations

What we've done above assumed that we were sampling from a normally distributed population. But often the population distribution is not normal or is unknown. The good news is the central limit theorem:

Central Limit Theorem -- as the sample size gets large, the sampling distribution of the mean can be approximated by the normal distribution.

As a general rule of thumb sample sizes of at least 30 will tend to give a sampling distribution that is approximately normal.

The sampling distribution of the proportion

Sometimes the random variable we are interested in is based on a categorical response – yes/no, male/female, etc. What is done here is to assign one group 1 and the other 0 then the sample mean would be the sum of the ones, which gives the proportion of the sample in that category

The sample proportion $P_s = X/n$
 X is the number of successes, n is the sample size

Again the sample proportion is an unbiased estimator of the sample proportion and the standard error of the proportion is given by:

$$\sigma_{ps} = \sqrt{(p(1-p)/n)}$$

The sampling distribution of the proportion follows the binomial distribution, but the good thing here is that we can use the CLT and use the normal distribution

So that
$$Z = \frac{P_s - P}{\sqrt{(p(1-p)/n)}}$$

Historically 10% of a large shipment of machine parts are defective. If random samples of 400 are selected, what proportion of the samples will have: less than 8% defective?

$$Z = \frac{.08 - .10}{\sqrt{(.1(.9)/400)}}$$

$= -.02 / .015 = -1.33 \sim$ that about 9.1 % of samples will have less than 8% defective

$$24 - (1.96)(30/\sqrt{100}) < \mu < 24 + (1.96)(30/\sqrt{100})$$

$$1.96(30/10) = 5.88$$

$$\text{so } 18.12 < \mu < 29.88$$

So we are 95% confident that the population mean wage is between 18.12 and 29.88.

What if we want a 90% confidence interval?

Draw this

Need to find z value corresponding to .05 = use = Normsinv(.05) = 1.645

$$24 - (1.645)(30/\sqrt{100}) < \mu < 24 + (1.645)(30/\sqrt{100})$$

$$19.065 < \mu < 28.935$$

A tighter interval. Note though that the cost of the tighter interval is a lower level of confidence.

Confidence Interval Estimation of the Mean when σ is unknown.

The above examples assumed, maybe unrealistically, that the population standard deviation was known. In most cases if we do not know the population mean we will not know the standard deviation either. How do we handle this.

If a random variable X is normally distributed, then the following statistic has a t distribution with $n-1$ **degrees of freedom**:

$$t = (X - \mu) / (S / \sqrt{n})$$

The t -distribution looks much like the normal distribution except that it has more areas in the tails and less in the center. Draw this.

Because σ is unknown and we are using S to estimate it, then we are more cautious in our inferences, the t -distribution takes this into account.

As the sample size increases, so do the degrees of freedom, and so the t -distribution becomes closer and closer to the normal distribution

Degrees of Freedom

Recall that in calculating the sample variance we did:

$$\sum (X_i - \bar{X})^2$$

note that in order to estimate S^2 we first had to estimate \bar{X} , so only $n-1$ of the sample values are free to vary.

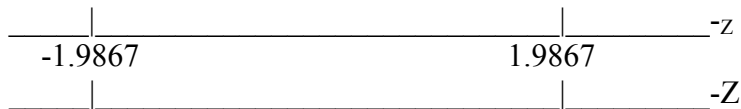
Suppose a sample of five values has a mean of 20. we know $n=5$ and $\bar{X}=20$ so we know that $\sum X_i=100$

Thus once we know four of the five values, the fifth one will not be free to vary because the sum must add to 100. if four of the values are 18, 24, 19, and 16 the fifth value has to be 23.

Estimating confidence intervals with the t-distribution

In a study of costs, it was found that the hospital costs of managing 90 cases of heart disease averaged 1120 with a standard deviation of 75. Construct a 95% confidence interval for the true average cost of managing this condition.

First draw the picture



Excel Commands:

`tdist(x,df,tails)`
`tin(prob,df)`

$$\text{so } 1120 - (-1.9867) * (75/\sqrt{90}) < \mu < 1120 + (-1.9867) * (75/\sqrt{90})$$

$$\text{or } 1,104.5 < \mu < 1,135.5$$

Suppose we know the average length of stay experience by 100 patients was 3.5 days and the standard deviation was 2.5 days. Construct a 98% CI for the true mean stay.

$$Z = 2.3642 \text{ so}$$

$$3.5 - (2.3642) * (2.5/\sqrt{100}) < \mu < 3.5 + (2.3642) * (2.5/\sqrt{100})$$

$$\text{or } 2.91 < \mu < 4.09$$

Confidence interval estimation for the proportion

Here we can do exactly the same thing

$$P_s - Z\sqrt{(P_s(1-P_s)/n)} < P < P_s + Z\sqrt{(P_s(1-P_s)/n)}$$

Suppose we are interested in the proportion of the RN population who are members of a union. A sample of 200 RNs is taken and 5% are found to be unionized. Construct the 95% CI for the true proportion of unionized RNs.

$$.05 - 1.96\sqrt{(.05)(.95)/200}$$

$$.05 + 1.96\sqrt{(.05)(.95)/200}$$

or we are 95% confident that the population proportion of RNs who are unionized is between .020 and .080

Determining the Sample Size

Up to now the sample size has been predetermined, but in reality the researcher can often set the desired sample size. Note that the larger the sample size the lower your sampling error, but also the higher the cost of obtaining the information.

Recall: $Z = (X - \mu) / (\sigma / \sqrt{n})$

Or $X - \mu = Z(\sigma / \sqrt{n})$

Or this second term is known as the error – how far off the sample mean is from the pop mean.

$$e = Z(\sigma / \sqrt{n})$$

We can solve this for n to get:

$$n = Z^2 \sigma^2 / e^2$$

so you can pick your sample size to get the desired level of error. It is a function of:

1. the desired level of confidence, the Z critical value
2. the acceptable sampling error, e
3. the standard deviation

Suppose we want to estimate the average salary paid to RNs and we want our estimate to be off by less than \$100 with a probability of .95. Suppose we think the population standard deviation $\sigma = \$1,000$. How big of a sample do we need?

$$n = 1.96^2 (1,000)^2 / 100^2 = 384.2. \text{ So a sample of 384 will work.}$$

Introduction to Hypothesis Testing

This deals with an issue highly similar to what we did in the previous chapter. In that chapter we used sample information to make inferences about the range of possibilities for the population statistic. That is, the confidence interval says that “we are pretty sure” that the population parameter is within this range somewhere. In this chapter we use sample information to either refute or fail to refute a given conjecture or hypothesis.

I. Basics of Hypothesis Testing

Typically we begin with some theory or claim about a particular parameter. For example, suppose a hospital administrator claims it requires 20 minutes to perform a given procedure. This claim by the administrator is referred to as the **null hypothesis**. The null hypothesis is typically given the symbol: H_0

$$H_0: \mu=20$$

That is, we assert that the population parameter is equal to the null. If the null turns out to be false, then some alternative must be specified. The **alternative hypothesis** is specified as:

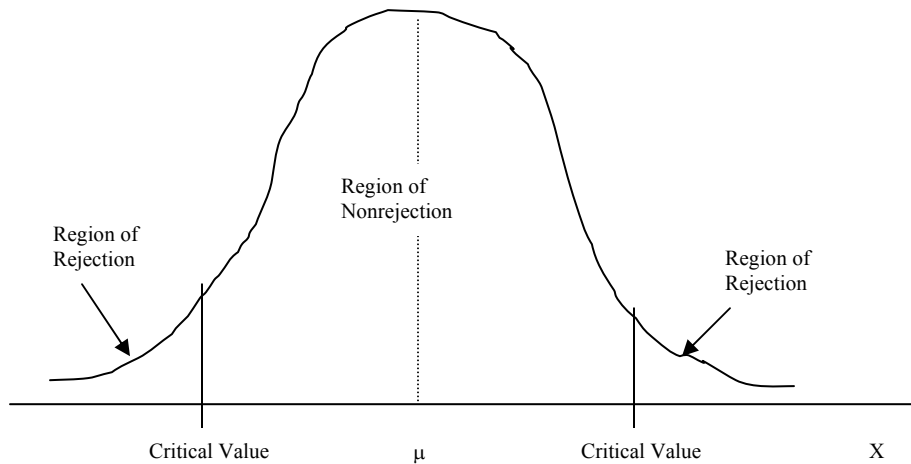
$$H_1: \mu \neq 20$$

This represents the conclusion we would reach if we reject the null. At this point it is possible to reject the null as being too low or too high. This is what is known as a **two-tailed test**, later we will see **one-tailed tests**, where the alternative is in one direction or the other.

Generally the null always refers to some specific value (it has an equal sign in it), whereas the alternative is not specific (it is the opposite of the null)

The way the null hypothesis is tested is to take a sample from the population and compare the sample results to the null. If the sample statistic is “far away” from the null then we conclude that the null must not be true, whereas if the sample statistic is “close” to the null then we conclude that there is no reason to reject the null.

NOTE: we never accept a null hypothesis, we only fail to reject. This is a big deal! If we say we accept the null, this implies that we have proven the null is true. We cannot do this, all we can do is refute the null or fail to refute. Failing to refute is not the same as accepting. Be careful with your wording.



The idea is that the sampling distribution of the test statistic is normally distributed (either because the population is normally distributed or because of the Central Limit Theorem), which means that it looks something like the above. The idea is if we have a population with mean μ then most of our sample means ought to be in the region of nonrejection. If we take a sample and get a mean outside this region then either it is a really odd sample or (more likely) the mean really is not μ .

So to make this decision we first need to determine the **critical value** of the test statistic. This divides the regions.

Type I and Type II errors

There are some risks involved in doing this. That is, there are two kinds of mistakes that we can make when conducting hypothesis testing.

Type I Error – this occurs if the null hypothesis is rejected when in fact it is true. The probability of a type I error is α

α is referred to as the level of significance. Typically this is what is chosen by the researcher. This is generally selected to be .01, .05, or .1. A critical value of .05 is equivalent to a 95% confidence interval.

Type II Error – this occurs if the null hypothesis is not rejected when in fact it is false and should be rejected. The probability of a Type II error is denoted as β .

The probability of a Type II error is dependent on the difference between the hypothesized and actual values of the population parameter.

The **Power of a Test** is denoted as $1-\beta$. This is the complement of a Type II error – it is the probability of rejecting the null hypothesis when, in fact, it is false. Generally statisticians attempt to maximize the power of a test, but as the power of the test increases so also does α . More on this below.

Decision	H ₀ True	H ₀ False
Do Not Reject H ₀	Correct Decision Confidence = $1-\alpha$	Type II Error Prob = β
Reject H ₀	Type I Error Prob = α	Correct Decision Power = $1-\beta$

II. Z tests of hypothesis for the mean -- σ known.

Going back to our example suppose we take a sample of 36 occasions on which the procedure was performed with a mean of 15.2 minutes. If the population standard deviation is 12 minutes, can the administrator's claim be substantiated?

So:

$$H_0: \mu=20$$

$$H_1: \mu \neq 20$$

1. Using the Critical Value

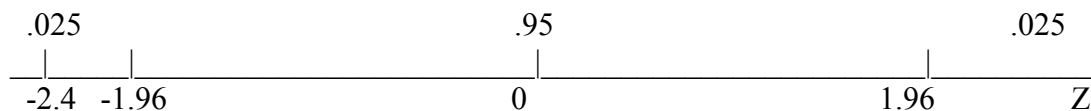
Given that the population standard deviation is known, the Z test statistic is:

$$Z = (X-\mu)/(\sigma/\sqrt{n})$$

For our example: $z = (15.2-20)/(12/6) = -2.4$

Or 15.2 is 2.4 standard errors away from the mean. Is this too unlikely?

It depends on our choice of critical value. If we choose a level of significance of .05 then the size of the rejection region is .05. Since this is a two-tailed test we divide this into parts, .025 each.



using **=normsinv(.025)** we get the Z value of -1.96 . That is there is a .025 probability of getting a value that is 1.96 standard deviations below the mean. Likewise there is a 2.5 percent chance of getting a z value 1.96 standard deviations above the mean. Thus our critical value for Z is 1.96. That is if our Z-statistic is more than 1.96 standard deviations away from the mean (in absolute value) then we say that is too unlikely and we reject our null hypothesis.

In our example, our Z-statistic is -2.4 which is more than 1.96 so we reject our null hypothesis. There is evidence from our sample that the population mean is NOT 20.

Suppose, however, our sample of 36 procedures revealed a mean of 18, how would this change the answer?

Now $Z = 18 - 20 / 2 = -1$. So now our z-statistic falls within the region of nonrejection. That is, the sample mean is only 1 standard deviation away from the hypothesized population mean. This is not too unlikely. **So we do not reject our null hypothesis. There is not sufficient evidence in our sample to conclude that the average procedure time is not 20 minutes.**

Again, note that we do not conclude that the mean time IS 20 minutes. We only fail to find evidence to disprove it.

2. Using the p-value

An alternative to choosing some cut off and comparing the z-statistic to this cut off, is to simply calculate the probability that one could obtain the given sample mean assuming the null hypothesis is true. For example.

In the above example with a mean of 15.2 we got a z statistic of -2.4 . If we use `=normsdist(2.4)`, we get .992, or there is a .008 probability that we could obtain a sample mean of 15.2 or lower if indeed the population mean is 20. This probability (.0082) is known as the p-value.

The p-value is the probability of obtaining a test statistic equal to or more extreme than the result obtained from the sample data, given that the null hypothesis is true.

The advantage of this approach is that it lets the reader decide what is acceptable. In this case we would likely conclude that this is such a low probability of occurrence that the null hypothesis is probably not true. Note that it is possible, but highly unlikely.

If instead our sample mean was 18, then the z-value of -1 translates into a p-value of .1587, or there is a 15.9 percent chance of getting a sample mean of 18 or lower if the population mean really was 20. This is not too unlikely and so we would fail to reject our null.

Note that both approaches are doing the same thing; we have to decide what is “too unlikely.” Often you will see the p-value reported, these are easy to interpret and do not need much other information to make inferences.

Note that there is a connection between hypothesis testing and confidence intervals. If we construct a 95% confidence interval for the average procedure time:

$$15.2 \pm 1.96(12/\sqrt{36})$$

$$15.2 \pm 3.924$$

or $11.27 < \mu < 19.124$ with 95 percent confidence

Note that since this interval does not contain 20, we conclude with 95% confidence (or with a 5% chance of a Type I error) that the true mean is not 20.

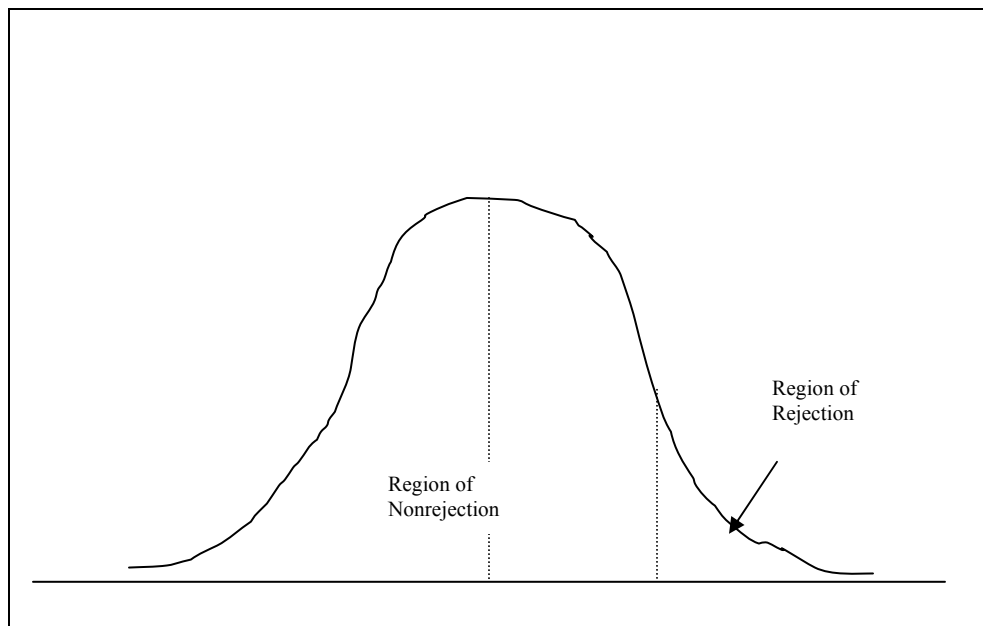
III. One-Tail Tests

Sometimes, the alternative hypothesis focuses in some specific direction. For example, suppose that a third party payer claims that a patient suffering from heart disease should experience a length of stay of no more than 6.5 days. Suppose we review a random sample of 40 cases of heart disease and get a mean of 8 days. If the population standard deviation is 5.2 days what can we conclude about these data at the 5% level of significance?

So here our null and alternative hypotheses are.

$$H_0: \mu \leq 6.5$$

$$H_1: \mu > 6.5$$



Now there is a single region of rejection. We are only concerned with “high” values of the variable. **This is known as a one-tailed test.**

Our z-statistic is the same:

$$Z = 8 - 6.5 / (5.2 / \sqrt{40}) = 1.5 / .822 = 1.82$$

But for our critical value we look up .05 since we are not dividing α into two regions.

So the critical value is 1.645. Thus we would reject our null hypothesis and conclude from this sample that there is evidence that the length of stay is **more than 6.5**.

The p-value associated with this Z is .0344 – there is about a 3.4% chance of getting a sample mean of 8 if the population mean is below 6.5.

IV. t-tests of Hypothesis for the Mean -- σ Unknown.

Generally, we do not know the population standard deviation. Just like we did in Chapter 6, when this happens we can use the sample standard deviation as an estimate for the population standard deviation. We then need to use the t distribution to account for this.

The same example above:

For example, suppose that a third party payer claims that a patient suffering from heart disease should experience a length of stay of no more than 6.5 days. Suppose we review a random sample of 40 cases of heart disease and get a mean of 8 days with a standard deviation of 5.2. What can we conclude about these data at the 5% level of significance?

So here our null and alternative hypotheses are.

$$H_0: \mu \leq 6.5$$

$$H_1: \mu > 6.5$$

$$t = 8 - 6.5 / (5.2 / \sqrt{40}) = 1.5 / .822 = 1.82$$

Now we look under the t-distribution under .05 and 39 d.f. Our critical value is 1.6849. Thus we (barely) reject our null hypothesis and conclude there is evidence that the true mean is greater than 6.5. Notice that the critical value here is slightly larger than the critical value for when σ was known. The logic is that since we are using an estimate for σ , we are a little more careful and need a bigger deviation from the null hypothesis to reach the same conclusion.

We can also calculate the p-value associated with this statistic. Note that the construction of the t-table at the end of the book does not work well in calculating the p-value. But we can do this easily in Excel with the command:

TDIST(x,degrees_freedom,tails)

Where X is your t-statistic, degrees_freedom is the degrees of freedom, and tails 2 for a two tailed test and 1 for a one tailed test. So if I do:
 $=\text{Tdist}(1.82, 39, 1)$, excel returns .03822. This is the p-value. There is a 3.8 percent chance of getting the sample mean of 8 or more if the population mean is 6.5. Thus if we use the 5% level of significance we would reject the null and conclude there is evidence the mean is greater than 6.5.

V. Z Test of Hypothesis for the Proportion

This is not much different from tests for the mean. Our Z statistic is:

$$Z = \frac{P_s - P}{\sqrt{\frac{P(1-P)}{n}}}$$

Where $P_s = X/n$ or the proportion of successes in the sample, and P is the hypothesized proportion of successes in the population.

Ex: In 1990 it was determined that 20 percent of high school seniors smoked cigarettes. In a recent survey taken in 2001 of 60 high school seniors it was found that 23 percent of the sample smoked cigarettes. Is there sufficient evidence at the 5% level of significance to conclude that the proportion of high school seniors who smoke has changed?

$H_0: P = .20$

$H_1: P \neq .20$

$$Z = \frac{.23 - .20}{\sqrt{\frac{.2(1-.2)}{60}}} = .03/.052 = .5809$$

This will be within the nonrejection region, so we do not reject H_0 and conclude that there is not evidence that the proportion has changed.

Or the p-value associated with this z-statistic is .28 – that is there is a 28% chance we could get a sample proportion of 23 or more if the population proportion is 20.

VI. Hypothesis test for the correlation coefficient.

Recall that the sample correlation coefficient is:

$$r = \frac{\sum(X - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2} \sqrt{\sum(Y_i - \bar{Y})^2}}$$

This indicates how the variables X and Y move together. Note that r is an estimator for the population correlation coefficient ρ , which is unknown. So if we calculate a correlation coefficient from a sample we would like to be able to make some inferences about the population correlation coefficient.

Suppose we calculate a correlation coefficient between two variables (X and Y) from a sample of 14 observations and obtain $r=.737$. Can we conclude that the population correlation coefficient is different from zero?

Our null and alternative hypothesis are:

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

The test statistic here is

$$t = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

This follows a t-distribution with n-2 degrees of freedom

So in our example:

$$t = \frac{.737 - 0}{\sqrt{\frac{1 - .737^2}{14 - 2}}} = 3.777$$

Using the .05 level of significance the critical value (under 12 degrees of freedom) is 2.1788. So we would reject our null hypothesis and conclude that there is evidence of an association between the variables X and Y.

Note that this could be done as a one tailed test in the same way as before. It depends on how the question is framed.

VII. More on the Power of a Test

Recall, the **Power of a Test**, denoted as $1-\beta$. is the complement of a Type II error – this is the probability of rejecting the null hypothesis when in fact it is false.

Typically the calculation of power is used to plan a study before any data have been obtained.

Assume the standard deviation is known.

Example:

Suppose we want to test the hypothesis that mothers with low socioeconomic status (SES) deliver babies whose birth weights are lower than normal. To test this, a list is obtained of birth weights from 100 consecutive, full-term, live born deliveries from the maternity ward of a hospital in a low SES area. The mean birth weight is found to be 115

with a sample standard deviation of 24 oz. Suppose we know from nationwide surveys based on millions of deliveries that the mean birth weight in the US is 120oz. Assume the pop standard deviation is 24. What is the power of the test assuming the alternative is 115.

To determine this we need to do a power calculation. The power of the study is the probability that we will be able to declare a significant difference with a sample size of 100 if the true mean is 115. Usually we want a power of at least 80% to perform a study.

The Null and Alternative Hypothesis:

$$H_0: \mu = \mu_0$$

$$H_1: \mu = \mu_1 < \mu_0$$

We know from the Z-statistic, that H_0 is rejected if $Z < Z_\alpha$ and H_0 is not rejected if $Z > Z_\alpha$ (recall the Z's will be negative here). The probability of a Type I error is α (rejecting the null when it is true).

The power of the test $(1-\beta)$ can be written as:

$$\text{Power} = P(\text{Reject } H_0 | H_0 \text{ false}) = P(Z < Z_\alpha | \mu = \mu_1)$$

$$= P \left[\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} < Z_\alpha \mid \mu = \mu_1 \right]$$

$$= P \left[\bar{X} < \mu_0 + Z_\alpha \frac{\sigma}{\sqrt{n}} \mid \mu = \mu_1 \right]$$

If H_1 is true then we know $\bar{X} \sim N(\mu_1, \frac{\sigma^2}{n})$

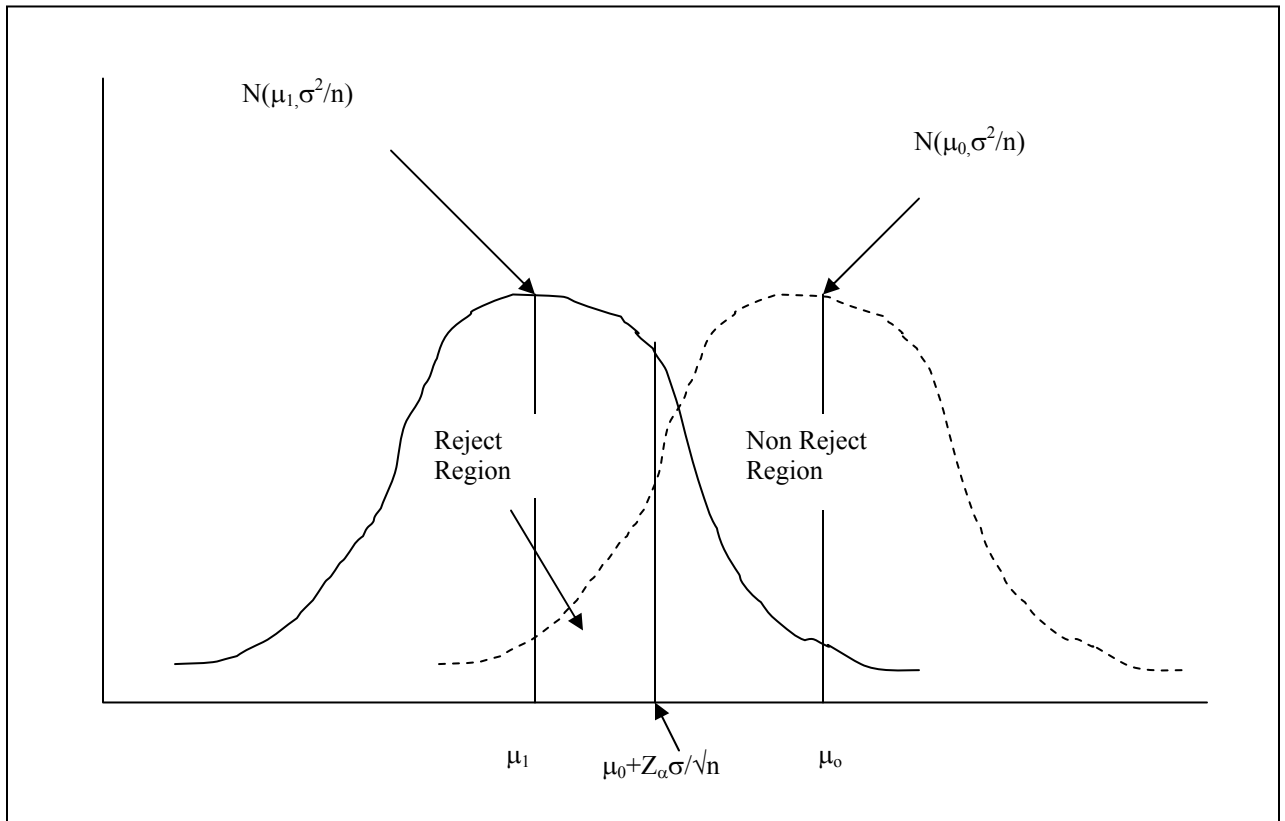
So if we standardize $\mu_0 + Z_\alpha \sigma / \sqrt{n}$ - subtract μ_1 and divide by the standard error. We get:

$$\text{Power} = \Phi \left[(\mu_0 + Z_\alpha \frac{\sigma}{\sqrt{n}} - \mu_1) / \frac{\sigma}{\sqrt{n}} \right]$$

The symbol Φ represents the normal distribution or the normal density function.

Which rearranged becomes:

$$= \Phi \left[Z_\alpha + \frac{(\mu_0 - \mu_1) \sqrt{n}}{\sigma} \right]$$



So if you can figure this out, you are doing great!

The dashed distribution is the distribution assuming the mean is μ_0 the solid distribution assumes μ_1 is the true mean. Under the null (μ_0 is true) we would pick our critical value based on α this is the term $\mu_0 + Z_\alpha \sigma / \sqrt{n}$. The area to the left of this but still under μ_0 distribution is α , the probability of a type I error. But suppose that μ_1 is true. Note now that the area to the left of $\mu_0 + Z_\alpha \sigma / \sqrt{n}$ under the μ_1 distribution is the probability rejecting H_0 when in fact H_0 is false. This is the power of the test. The area to the right of $\mu_0 + Z_\alpha \sigma / \sqrt{n}$ is β (prob of Type II). So to figure out the power of the test we calculate

$$P(X < \mu_0 + Z_\alpha \sigma / \sqrt{n}) = \Phi \left[Z_\alpha + \frac{(\mu_0 - \mu_1) \sqrt{n}}{\sigma} \right]$$

So back to our example,

$$\mu_0 = 120, \mu_1 = 115, \alpha = .05, \sigma = 24, n = 100$$

$$\text{So Power} = \Phi \left[-1.645 + \frac{(120 - 115) \sqrt{100}}{24} \right] = \Phi(-1.645 + (5/24)10) = \Phi(.438)$$

Using NORMSDIST in excel I get prob= .669

Or there is about a 67% chance of detecting a significant difference using a 5% significance level with this sample size.

A standard rule of thumb is to have the power of the test be at least .8, but note that it is difficult to do since we really do not know μ_1 . If the null is not true, what specifically is the alternative? We generally do not know this.

Factors affecting the Power:

1. If the significance level is made smaller (α decreases), Z_α decreases and hence the power decreases – and therefore β increases!
2. If the alternative mean is shifted further away from the null mean, then the power increases.
3. if the standard deviation of the distribution of individual observations increases, the power decreases
4. If the sample size increases then the power increases.

Generally what one does is selects α and then makes the sample size big enough to get the power to some acceptable level. Of course this is all assuming the alternative is known.

VIII. Some Issues in Hypothesis Testing

1. Random Samples – it must be the case that the data being used is obtained from a random sample. If the sample is not random then the results will be suspect. Respondents should not be permitted to self-select (selection bias) for a study, nor should they be purposely selected.
2. One Tail or Two? – if prior information is available that leads you to test the null against an alternative in one direction, then a one tailed test will be more powerful. Previous research or logic often will establish the difference in a particular direction, allowing you to conduct a one-tailed test. If you are only interested in differences from the null, but not the direction of the difference, then the two-tailed test is the appropriate procedure.
3. Choice of Significance α -- typically it is standard to use $\alpha=.05$, but this is somewhat arbitrary. The important thing is to establish α before you conduct your test. Suppose you conduct your test and get a p-value of .075. If you use $\alpha=.05$ you will not reject the null, whereas if you use $\alpha=.1$ then you will reject the null. The danger is to select the level of significance that allows you to make the conclusion you would like. This is where the p-value is useful in that you can simply report the p-value. Alternatively you will often see results presented where they specify some

results are significant at the .05 level while others are only significant at the .10 level, etc.

4. Cleansing Data – Here the danger is throwing out observations you do not like under the grounds that they are “outliers.” In order to throw out observations you must verify that they are not valid, mistakes, incomplete, etc. But just because an observation is “extreme” doesn’t mean you can delete it.

Chapter 9: Two-Sample Tests

There are many problems in which we have to decide whether an observed difference between two means is attributable to chance or whether they represent different populations.

Suppose we are interested in cardiovascular risk factors in children and want to know if the average heart rate among newborns is different between whites and blacks.

When comparing two means, our objective is to decide whether the observed difference is statistically significant or whether the difference is attributable to chance fluctuation.

We let $\bar{X}_1 - \bar{X}_2$ represent the difference between the means of sample 1 and sample 2.

I. Test between Means: Large Samples

Suppose we selected two large independent samples of size n_1 and n_2 having the means \bar{X}_1 and \bar{X}_2 .

We **assert** (rather than prove) that the difference $\bar{X}_1 - \bar{X}_2$ can be approximated by a standard normal curve whose mean and standard deviation are given by:

$$\mu = \mu_1 - \mu_2$$

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Where μ_1 and μ_2 are the population means from the respective populations and σ^2 represents the variance. We refer to $\sigma_{\bar{X}_1 - \bar{X}_2}$ as the *standard error of the difference between two means*.

If the population standard deviations are not known, we substitute S_1 for σ_1 and the same for S_2 and estimate the standard error of the difference between two means by:

$$\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

When comparing two means, the null hypothesis usually assumes the form:

$$H_0: \mu_1 = \mu_2 \quad \text{or} \quad \mu_1 - \mu_2 = 0$$

Which implies that there is no difference between the means

The alternative for a two tailed test:

$$H_1: \mu_1 \neq \mu_2 \quad \text{or} \quad \mu_1 - \mu_2 \neq 0$$

If it is a one-tailed test:

$$H_1: \mu_1 > \mu_2 \quad \text{or} \quad \mu_1 - \mu_2 > 0$$

The Z value in this case (recall we can use the standard normal here since either we know the population standard deviation, or the sample size is large so that the normal distribution can be used as an approximate.

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_1^2 / n_1 + S_2^2 / n_2}}$$

Which is a random variable having the normal distribution.

Back to our example: Suppose we collect the following data on heart rate among newborns:

Race	Mean (beats per minute)	sd	n
White	129	11	218
Black	133	12	156

And we want to test the hypothesis that the average heart rate for white newborns is different than the average heart rate for black newborns.

$$H_0: \mu_w = \mu_b \quad \text{or} \quad \mu_w - \mu_b = 0$$

$$H_1: \mu_w \neq \mu_b \quad \text{or} \quad \mu_w - \mu_b \neq 0$$

$$Z = \frac{129 - 133}{\sqrt{11^2 / 218 + 12^2 / 156}} = \frac{-4}{\sqrt{.555 + .923}} = \frac{-4}{1.216} = -3.289$$

Using Excel: =normsdist(-3.289) gives the p-value associated with this: .00050.

There is only a .05% chance that we could get this large of a difference between the sample means if, indeed, the population means were equal. Thus we would reject the null hypothesis and conclude that there is evidence from our sample that the average heart rates are different between white and black newborns.

II. Test between Means: Small Samples

If our samples come from populations that can be approximated by the standard normal curve and if $\sigma_1 = \sigma_2 = \sigma$ (i.e., the standard deviations of the two populations are equal), a

small sample test of the differences between two means may be based on the t distribution. Then our test statistic becomes:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\Sigma(x_1 - \bar{X}_1)^2 + \Sigma(x_2 - \bar{X}_2)^2}{n_1 + n_2 - 2} * \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Note that this “kind of” looks like $(X-\mu)/\sigma/\sqrt{n}$. In the denominator note that there is something like the variance of each variable in there, and we are dividing by n.

By definition $\Sigma(x_1 - \bar{X}_1)^2 = (n_1 - 1)S_1^2$ and $\Sigma(x_2 - \bar{X}_2)^2 = (n_2 - 1)S_2^2$
This equation can be “simplified” to be:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} * \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

This is the equation in your book (page 375) This is known as the “pooled variance t-test” for the difference between two means. Note that we are calculating something that looks like the two variances combined in the equation.

Since $n_1 - 1$ of the deviations from the mean are independent in S_1^2 and $n_2 - 1$ of the deviations from the mean are independent in S_2^2 , we have $(n_1 - 1) + (n_2 - 1)$ or $(n_1 + n_2 - 2)$ degrees of freedom. Thus the above equation represent the t distribution with $(n_1 + n_2 - 2)$ degrees of freedom.

Suppose we want to know if there is a difference in turnover between RNs who work in the ICU and those who work in the med/surg unit. We take a sample of both groups and calculate the average number of years spent on the job.

Type	Average Term Years	Sd	N
ICU	20	5	12
Med/Surg	22	4	15

Is there evidence that Med/Surg RNs stay on the job longer than ICU RNs?

$$H_0: \mu_{icu} = \mu_{med} \quad \text{OR} \quad \mu_{icu} - \mu_{med} = 0$$

$$H_1: \mu_{icu} < \mu_{med} \quad \text{OR} \quad \mu_{icu} - \mu_{med} < 0$$

$$t = \frac{20 - 22}{\sqrt{\frac{(12 - 1)5^2 + (15 - 1)4^2}{12 + 15 - 2} * \left(\frac{1}{12} + \frac{1}{15}\right)}} = \frac{-2}{\sqrt{\frac{499}{25} * \sqrt{.15}}} = \frac{-2}{1.73} = -1.156$$

The critical value for $\alpha=.05$ and 25 degrees of freedom is -1.7081 so we would fail to reject the null and conclude that there is no evidence from this sample that Med/Surg RNs stay on the job longer than ICU RNs.

Or using Excel: =tdist(-1.56, 25, 1) gives a pvalue of .129, so there is about a 13 percent chance of getting two sample means this far apart if the population means were really equal, so if we use $\alpha=.05$ we would fail to reject the null and conclude there is no evidence that med/surg RNs stay on the job longer than ICU RNs.

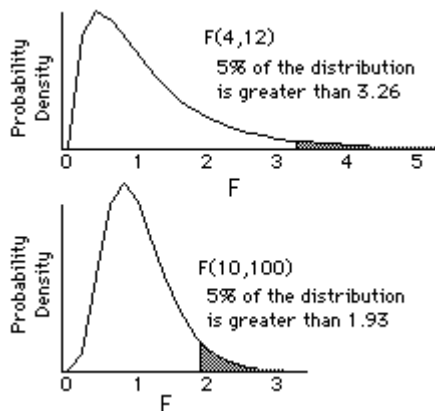
Note that in Excel under the Tools, data analysis, tabs there are canned formulas for the z-test for different sample means, t-tests under equal variances and t-tests under unequal variances. These are OK to use as a check on your work, but for this section of the homework, I want you to calculate the z/t values on your own. Afterwards you are free to use them.

III. F-Test for the difference between two variances

The above t-test made the assumption that the two samples have a common variance. In order to be complete this assumption must be tested. Assuming that independent random samples are selected from two populations, test for equality of two population variances are usually based on the ratio S_1^2/S_2^2 or S_2^2/S_1^2 . If the two populations from which the samples were selected are normally distributed, then the sampling distribution of the variance ratio is a continuous distribution called the F distribution.

The F distribution depends on the degrees of freedom given by n_1-1 and n_2-1 where n_1 and n_2 are the sample sizes that S_1^2 and S_2^2 are based.

The F- distribution looks as follows:



It is skewed to the right, but as the degrees of freedom increases it becomes more symmetrical and approaches the normal distribution.

If you really have lots of time to waste and are curious you can go to the following web site and play with df and how the F changes shape.

<http://www.econtools.com/jevons/java/Graphics2D/FDist.html>

The easiest way to conduct the test is to construct the ratio such that the large of the two variances are on the top. That way you are always looking at the upper tail and do not need to worry about the lower tail.

Basically what is going on is you are comparing the two variances and asking “is one bigger than the other?” If the variances are equal then this will equal 1, if the variances are “different” then this ratio will be something greater than 1. So the question is how far away is too far away? That is what the F distribution tells us. Under the Null the variances are equal and so the ratio should equal one, but there is some probability that the ratio will be larger than 1, thus if it is too unlikely for the variances to be equal we reject the null and conclude the variances are not equal:

Suppose we are interested in examining the mean stays experienced by two populations from which we have drawn a random sample of size $n_1=6$ and $n_2=8$. Based on this random sample, suppose we found that $S_1^2=12$ days and $S_2^2=10$ days. Before testing the significance of the observed difference between the two mean stays, we should examine the assumption that $\sigma_1^2 = \sigma_2^2$ by:

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

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

$$F = 12/10 = 1.2$$

Using $\alpha=.05$. Here there are 5 degrees of freedom in the numerator and 7 degrees of freedom in the denominator. Using the Excel function **FDIST(F,df1,df2)** or =FDIST(1.2,5,7)=.301. Thus if the null is true that the variances are equal, there is a 30 percent chance we could get two sample variances that were this far apart from each other. Thus, we would not reject our null hypothesis and conclude that there is no evidence that the two variances are different.

IV. Tests between two samples: Categorical Data

This section deals with how we can test for the difference between two proportions from different samples. We might be interested in the difference in the proportion of smokers and nonsmokers who suffer from heart disease or lung cancer, etc.

Our Null Hypothesis would be: $H_0: P_1=P_2$

Where P_1 and P_2 are the population proportions. Letting p_1 and p_2 represent the sample proportions, we can use $p_1 - p_2$ to evaluate the difference between the proportions.

1. Z test for the difference between two proportions

One way to do this test is using the standard normal distribution. Here our Z statistic is:

$$Z = \frac{\frac{x_1}{n_1} - \frac{x_2}{n_2}}{\sqrt{P^*(1-P^*)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Where: x_1/n_1 the proportion of successes from sample 1 and x_2/n_2 is the sample of successes from sample 2

and

$$P^* = \frac{x_1 + x_2}{n_1 + n_2}$$

Note that this is similar to the formula we used to test hypotheses on a single sample proportion, the denominator is using the combined proportion to estimate the standard error of the difference in the sample proportions.

EX: in 1997 a sample of 200 low income families (less than 200 percent of the poverty line) was taken and it was found that 43 of them have children who have no health insurance. In 2003 a similar survey of 250 families was taken and 48 were found not to have insurance. Is there evidence from these samples that the proportion of low income children without insurance has changed?

$$H_0: P_{97} = P_{03}$$

$$H_1: P_{97} \neq P_{03}$$

$$P^* = \frac{43 + 48}{200 + 250} = .2022$$

So:

$$Z = \frac{\frac{43}{200} - \frac{48}{250}}{\sqrt{.2022(1-.2022)\left(\frac{1}{200} + \frac{1}{250}\right)}} = \frac{.215 - .192}{\sqrt{.1613 * .009}} = \frac{.023}{.038} = .6052$$

Thus we would reject the null hypothesis and conclude that there is no evidence that the proportion of uninsured children from low income families has changed.

2. χ^2 test for the difference between two proportions

An alternative to the Z test, is the χ^2 (chi-square) test. This starts by laying out a contingency table of the outcomes:

	1997	2003	Total
Successes	43	48	91
Failures	157	202	359
Totals	200	250	450

We want to test the Null: $H_0: P_{1997} = P_{2003}$
 Against $H_1: P_{1997} \neq P_{2003}$

The χ^2 test statistic takes the following form:

$$\chi^2 = \sum_{AllCells} \frac{(f_0 - f_e)^2}{f_e}$$

where f_0 is the observed frequency, in a particular cell of the 2x2 table
 f_e is the theoretical, or expected frequency in a particular cell if the null hypothesis is true. This test approximately follows a chi-square distribution with 1 degree of freedom.

I will discuss the chi-square distribution in a minute.

To understand what f_e is, assume the null hypothesis is true, then the sample proportions computed from each of the two groups would differ from each other only by chance and would each provide an estimate of the common population parameter, p . In such a situation, a statistic that pools or combines these two separate estimates together into one overall estimate of the population parameter provides more information than either of the two separate estimates could provide by itself.

This statistic, P^* , is:

$$P^* = \frac{x_1 + x_2}{n_1 + n_2}$$

To obtain the expected frequency for each cell pertaining to successes (the cells in the first row of the table), the sample size for a group is multiplied by P^* to obtain the expected frequency for each cell. For each cell in the second row the sample size is multiplied by $(1-P^*)$.

In our example:

$$P^* = \frac{43 + 48}{200 + 250} = .2022$$

	1997 actual	1997 expected	2001 actual	2001 expected	Total
Successes	43	40.44	48	50.55	91
Failures	157	159.56	202	199.45	359

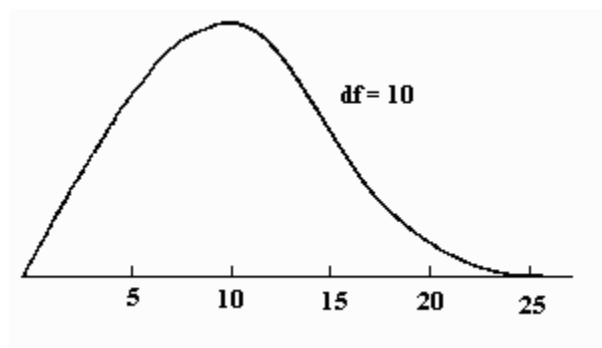
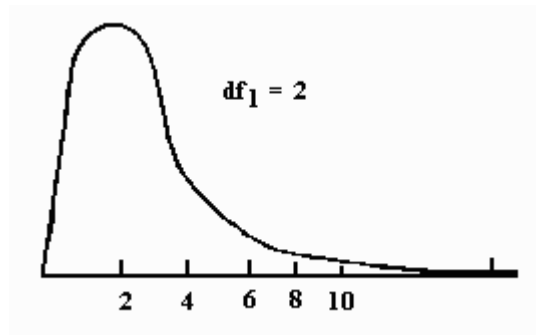
Totals	200	200	250	250	450
--------	-----	-----	-----	-----	-----

f_0	f_e	$(f_0 - f_e)$	$(f_0 - f_e)^2$	$(f_0 - f_e)^2 / f_e$
43	40.44	2.56	6.55	.1621
48	50.55	-2.55	6.50	.1286
157	159.56	-2.56	6.55	.0411
202	199.45	2.55	6.50	.0326
			Sum=	.3644

This is our test statistic, .3644, note that what this is doing is taking the sum of the squared difference between the actual proportion in each cell from the predicted proportion *assuming the null is true*. So if any of these are “way off” then the test statistic will be “large” while if they are not off the test statistic will be “small”.

The question now is what is large and what is small. This is where the χ^2 distribution comes in. It turns out that if we have the sum of n statistically independent standard normal variables, then this sum follows the χ^2 distribution.

The probability density function for the chi-square distribution looks as follows:



The degrees of freedom are calculated as follows:

$DF = (r-1)(c-1)$, where r is the number of rows in the table, and c is then number of columns.

So in our 2x2 table $df = 1$.

The rejection region is the right hand tail of the distribution, that is we reject H_0 if $\chi^2 > \chi^2_{1-\alpha}$

In this case the critical value of the we use =CHIDIST(x,df) or =CHIDIST(.3644,1)=.546 to get the pvalue. So if the null is true and the two proportions are equal, then there is a 54 percent chance of getting sample proportions this far from each other. Thus we fail to reject the null hypothesis and conclude that there is no evidence that the two populations have a different rate of uninsured.

The F-test and χ^2 tests will usually give the same answers; they are just two different ways of doing the same thing.

The nice thing about the χ^2 test is it is generalizable to more than two proportions. The logic is exactly the same. I won't go through this; you can do it on your own if you get bored.

V. χ^2 test for independence

What we did above was show the χ^2 test for the difference between two (or more) proportions. This can be generalized as a test of independence in the joint responses to two categorical variables. This is something similar to the correlation coefficient (but applied to categorical data) in that it tells you if there is a relationship between two variables.

The null and alternative hypotheses are as follows:

H_0 : the two categorical variables are independent (there is no relationship between them)

H_1 : The variables are dependent (there is a relationship between them)

We use the same equation as before:

$$\chi^2 = \sum_{AllCells} \frac{(f_o - f_e)^2}{f_e}$$

Suppose we are interested in the relationship between education and earnings for registered nurses. We collect data for 550 randomly chosen RNs and get the following table:

Earnings	Level of Education			Total
	High School	Some College	College grad+	
< 30k	80	65	44	189
30k-40k	72	34	53	159
40k+	48	31	123	202
Total	200	130	220	550

So our null would be that there is no relationship between education and earnings, and the alternative is that there is a relationship between the two.

To obtain f_e in this case we go back to the probability theory. If the null hypothesis is true, the multiplication rule for independent events can be used to determine the joint probability or proportion of responses expected for any cell combination.

That is, $P(A \text{ and } B) = P(A|B)*P(B)$ – the joint probability of events A and B is equal to the probability of A, given B, times the probability of B.

So, suppose we want to know the probability of drawing a Heart and a face card from a normal deck of cards, let Heart = A, Face =B, then

$P(H \text{ and } F) = P(H|F)*P(F)$, there are 4*4 or 16 face cards in a deck, 4 of them hearts so $P(H|F) = 4/16$ or $1/4$. Then the probability of Face is $16/52=4/13$.

So $P(H \text{ and } F) = 1/4*4/13 = 4/52$ or $1/13$.

But note that the events Heart and Face are independent of each other since $P(H) = 13/52=1/4$ and $P(H|F)=1/4$. That is, knowing the card is a face card does not change the probability of getting a heart. Then the joint probability formula simplifies to:

$P(A \text{ and } B) = P(A)*P(B)$ when A and B are independent.

But suppose we had a deck of cards that had an extra king of hearts and the king of spades was missing. Then $P(H) \neq P(H|F)$. Knowing that the card is a face card would change the probability of drawing a heart.

So back to the example above: assuming independence, if we want to know the probability of being in the top left cell – representing having a high school degree and earning less than 30k – this will be the product of the two separate probabilities:

$$P(\text{high school and } < 30k) = P(\text{high school}) * P(<30k)$$

$$P(\text{high school}) = 200/550 = .3636$$

$$P(<30k) = 189/550 = .3436$$

So if these two guys are independent then $P(\text{high school and } < 30k) = .3636 * .3436 = .1249$. This is the predicted proportion.

The observed proportion is $80/550 = .1455$. Basically we want to compare the actual vs. predicted proportions for each cell.

If the predicted proportion was .1249, we would expect $550 * .1249 = 68.7$ RNs to be in this cell

To find the expected number in each cell:

$$f_e = (\text{row total} \times \text{column total}) / n$$

$$\text{For this example: } 200 * 189 / 550 = 68.7$$

So to calculate the χ^2 statistic:

	f_0	f_e	(f_0-f_e)	$(f_0-f_e)^2$	$(f_0-f_e)^2/f_e$
HS/<30k	80	68.7	11.3	127.7	1.86
HS/30-40	72	57.8	14.2	201.6	3.49
HS/40k+	48	73.5	-25.5	650.3	8.85
Some/<30k	65	44.7	20.3	412.1	9.22
Some/30-40	34	37.6	-3.6	13.0	0.34
Some/40k+	31	47.7	-16.7	278.9	5.85
Coll/<30k	44	75.6	-31.6	998.6	12.21
Coll/30-40	53	63.6	-10.6	112.4	1.77
Coll/40k+	123	80.8	42.2	1780.8	22.04
				Sum=	65.63

In this case we have $c=3$ and $r=3$, so there are $2*2=4$ degrees of freedom.

The pvalue is very close to zero. Thus we would reject the null hypothesis and conclude that there is evidence of a relationship between earnings and education among RNs.

Note that just like the correlation coefficient this test does not tell you anything about the relationship between the two variables, just that one appears to exist.

Note that the Independent Project asks you to do two of these: one for Y and X_1 , and one for Y and X_2 and to divide the variables into three categories. Depending on what you are doing it may not be possible to divide the data into three categories. For example if you are looking at the effect of gender on earnings then it will be difficult to break gender into more than two categories. That is fine, just do two!