10. Hypothesis Testing

(Human) Error (in Judgement)

We usually consider that we are "either right or wrong" when we make a decision based on limited data. **For example, "Will it rain today?"** However, a careful consideration will show that there are, in fact, two different ways we are wrong and two different ways we are right for a total of 4 outcomes:

Let's consider a typical "judgement" situation. There is a knock on your door and the police arrest you for a vehicular "hit-and-run" where another car was damaged. The person driving the other car got part of the license tag number of the car that hit his and the police found your car has some damage to the left fender. You know your car was in a previous accident two weeks ago that produced that damage but it wasn't reported to the police or your insurance agent. At the time the accident occurred, you were sleeping (although no one can provide you with an alibi). You know you are innocent and if you are tried and found guilty the court and jury will have made a mistake and if they find you innocent (they BETTER) they will not have made a mistake.

BUT, the "truth" of the matter cannot be established by "taking your word for it", instead, evidence and testimony will be presented and ultimately a jury will render a verdict.

Truth➔	You are innocent	You are guilty
Jury finds you innocent	No Error	Error! This is bad for society… guilty people are let go without punishment (to commit more crimes). Other criminals see they can "get away with things" by hiring "trickster lawyers".
Jury finds you guilty	Error! This is bad for you. You will be put in prison and your personal freedom taken away. You will have "a record"	No Error

In our society, we are very aware of these two types of error. We try to make one of the types of error happen very infrequently, but we realize in doing so we make the other kind of error very frequently. Many “guilty people” are found to be “not guilty” because of the makeup of our legal system (evidence thrown out on technicalities, etc) but we rarely put innocent people in prison. Our legal system is based on “innocent until proven guilty” rather than “guilty until proven innocent”.

Robots: A Demonstration About Making Errors

Consider that you have two barrels each containing 500 metal spheres that are indistinguishable (same appearance and size) from one another except that the density of the metal in the “A” barrel is somewhat lighter than that in the “B” barrel. The weights are normally distributed with

meanA = 10 meanB = 11

stdevA = 1 stdevB = 1

There is one other important difference: The items in the A barrel are worth $20 each and the ones in the B barrel are only worth $1.

Suppose you have been assigned to move the contents of the two barrels to a new location some distance away (up on the third floor). We could load a few spheres at a time (5 = 50lbs) to a bucket and walk to the new location but another employee who has been watching from a distance comes over and tells you that there is a “plant robot” who can do jobs like this. He (the robot) works for free so all you need to do is program him and sit under a tree until he’s done.

On the positive side, he is equipped with a very sensitive “balance” that can quickly and accurately weigh what he is carrying. Also, he is very fast (better to stay out of his way!).

On the negative side, he has a faulty memory unit and isn’t able to remember which barrel he gets something from.

Since you have had statistics you know something about things with normal distributions so you devise a plan to allow him to move the spheres and place them in the destination barrels on a weight basis.

You decide to program him in the following fashion: You know that if the A barrel contains items costing $20, you don’t want to put them in the wrong destination barrel too often (where they will be mistaken for the $1 spheres). With an average weight of 10.0 pounds you know half the spheres in the barrel weigh more than 10.0 so you decide if the sphere being carried weighs 11 lbs or less the robot is to put it in a barrel marked AA and if it weighs 11 lbs or more the robot is to put it in a barrel marked BB.

(see simulation!)

Type I and Type II Error (in Hypothesis Testing)

There are two kinds of errors that can be made in significance testing:

(1) a true null hypothesis can be incorrectly rejected and

(2) a false null hypothesis can fail to be rejected.

Statistical decision	True state of null hypothesis	
	Ho True	Ho False
Reject Ho	Type I error	Correct
Do not reject H0	Correct	Type II error

The former error is called a Type I error and the latter error is called a Type II error. These two types of errors are defined in the table.

The probability of a Type I error is designated by the Greek letter alpha (α) and is called the Type I error rate; the probability of a Type II error (the Type II error rate) is designated by the Greek letter beta (β) .

A Type II error is only an error in the sense that an opportunity to reject the null hypothesis correctly was lost. It is not an error in the sense that an incorrect conclusion was drawn since no conclusion is drawn when the null hypothesis is not rejected.

A Type I error, on the other hand, is an error in every sense of the word. A conclusion is drawn that the null hypothesis is false when, in fact, it is true. Therefore, Type I errors are generally considered more serious than Type II errors.

The probability of a Type I error (a) is called the significance level and is set by the experimenter.

There is a tradeoff between Type I and Type II errors. The more an experimenter protects him or herself against Type I errors by choosing a low level, the greater the chance of a Type II error. Requiring very strong evidence to reject the null hypothesis makes it very unlikely that a true null hypothesis will be rejected. However, it increases the chance that a false null hypothesis will not be rejected, thus lowering power.

The Type I error rate is almost always set at 0.05 or at 0.01, the latter being more conservative since it requires stronger evidence to reject the null hypothesis at the 0.01 level then at the 0.05 level.

What Is The Null Hypothesis?

The null hypothesis is an hypothesis about a population parameter. The purpose of hypothesis testing is to test the viability of the null hypothesis in the light of experimental data. Depending on the data, the null hypothesis either will or will not be rejected as a viable possibility.

Consider a researcher interested in whether the time to respond to a tone is affected by the consumption of alcohol. The null hypothesis is that $\mu_1 - \mu_2 = 0$ where μ_1 is the mean time to respond after consuming alcohol and μ_2 is the mean time to respond otherwise. Thus, the null hypothesis concerns the parameter $\mu_1 - \mu_2$ and the null hypothesis is that the parameter equals zero.

The null hypothesis is often the reverse of what the experimenter actually believes; it is put forward to allow the data to contradict it. In the experiment on the effect of alcohol, the experimenter probably expects alcohol to have a harmful effect. If the experimental data show a sufficiently large effect of alcohol, then the null hypothesis that alcohol has no effect can be rejected.

It should be stressed that researchers very frequently put forward a null hypothesis in the hope that they can discredit it. For a second example, consider an educational researcher who designed a new way to teach a particular concept in science, and wanted to test experimentally whether this new method worked better than the existing method. The researcher would design an experiment comparing the two methods. Since the null hypothesis would be that there is no difference between the two methods, the researcher would be hoping to reject the null hypothesis and conclude that the method he or she developed is the better of the two.

The symbol H_0 is used to indicate the null hypothesis. For the example just given, the null hypothesis would be designated by the following symbols:

H_0: $\mu_1 - \mu_2 = 0$
or by
H_0: $\mu_1 = \mu_2$.

The null hypothesis is typically a hypothesis of no difference as in this example where it is the hypothesis of no difference between population means. That is why the word "null" in "null hypothesis" is used -- it is the hypothesis of no difference.

Despite the "null" in "null hypothesis," there are many times when the parameter is not hypothesized to be 0. For instance, it is possible for the null hypothesis to be that the difference between population means is a particular value. Or, the null hypothesis could be that the mean SAT score in some population is 600. The null hypothesis would then be stated as: H_0: $\mu = 600$.

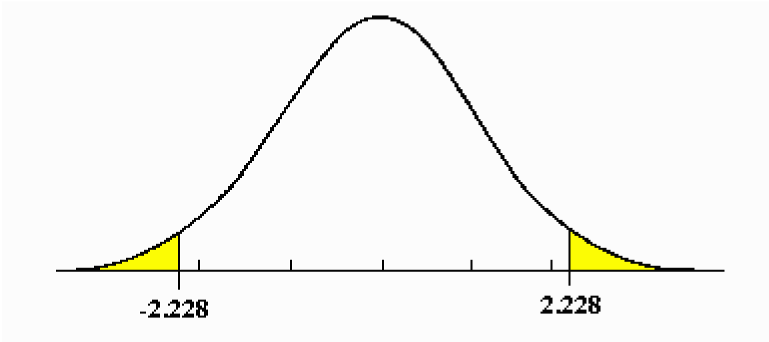
Although the null hypotheses discussed so far have all involved the testing of hypotheses about one or more population means, null hypotheses can involve any parameter. An experiment investigating the variations in data collected from two different populations could test the null hypothesis that the population standard deviations were the same or differed by a particular value.

One and Two Tailed Tests

A one- or two-tailed t-test is determined by whether the total area of α is placed in one tail or divided equally between the two tails. The one-tailed t-test is performed if the results are interesting only if they turn out in a particular direction. The two-tailed t-test is performed if the results would be interesting in either direction. The choice of a one- or two-tailed t-test effects the hypothesis testing procedure in a number of different ways.

Two-Tailed Tests (z- and t-)

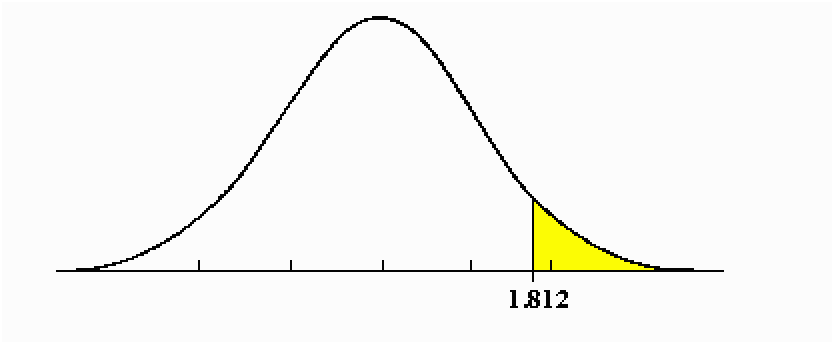
A two-tailed t-test divides α in half, placing half in the each tail. The null hypothesis in this case is a particular value, and there are two alternative hypotheses, one positive and one negative. The critical value of t, t_{crit}, is written with both a plus and minus sign (±). For example, the critical value of t when there are ten degrees of freedom (df=10) and a is set to .05, is t_{crit}= ± 2.228. The sampling distribution model used in a two-tailed t-test is illustrated below:



One-Tailed Tests

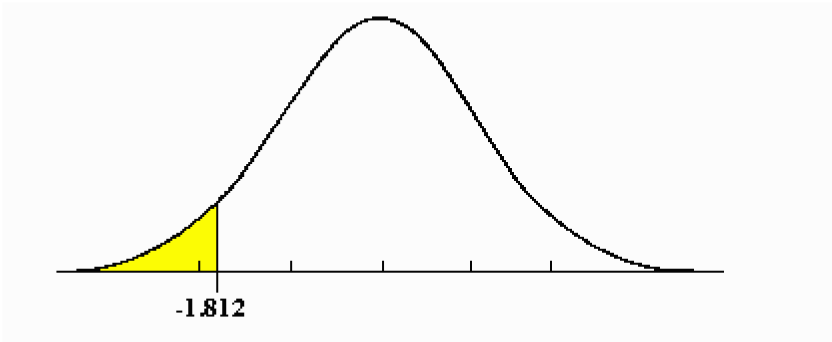
There are really two different one-tailed t-tests, one for each tail. In a one-tailed t-test, all the area associated with a is placed in either one tail or the other. Selection of the tail depends upon which direction t_{obs} would be (+ or -) if the results of the experiment came out as expected. The selection of the tail must be made before the experiment is conducted and analyzed.

A one-tailed t-test in the positive direction is illustrated below:



The value t_{crit} would be positive. For example when a is set to .05 with ten degrees of freedom (df=10), t_{crit} would be equal to +1.812.

A one-tailed t-test in the negative direction is illustrated below:



The value t_{crit} would be negative. For example, when a is set to .05 with ten degrees of freedom (df=10), t_{crit} would be equal to -1.812.

What Does This All Mean… an example!

Consider we have two instruments that sit side by side and measure a DO (dissolved oxygen) level in a stream where our company discharges wastewater upstream. Due to random fluctuations, they never read exactly the same value (6.22283 ppm vs 6.14223 ppm) but the long term behavior expected is that both machines will have the higher value 50% of the time.

Suppose we record the data from the machines once each 15 minutes and look back at the last 5 hours of data (20 values). We find that the "A" machine is high 15 times and the "B" machine is high 5 times. Will we judge that the machines are "broken" (not in agreement) or will we judge the machines are operating "as expected" and not needing repair?

<u>Major Issues</u>

Two types of error are possible: If we judge the machines are "ok" and they are, in fact "broken" will we needlessly try to "fix" machines that aren't really broken. If we judge the machines are "broken" and they are, in fact "ok" we will allow data to be recorded and acted on that is in fact "erroreous".

This problem is similar to flipping a coin, that is, there should be a 50% probability of seeing the "A" or "B" higher. Thus we can express this situation in terms of flipping coins (since we can more easily visualize that). Just how surprising is it to see 15 "heads" when we expect p*n=10

What is the probability of seeing 15 heads?

What is the probability of seeing at least 15 heads?

Notice that what would surprise us is not seeing 15 H's but seeing as 15 or more heads.

If you were applying for a job and you were expecting a starting salary of around $55,000 and you were offered $42,000 or $88,888 are you surprised because of the "specific" (exact) number or the "size" of the number (range of values)?

Therefore, we aren't interested in

P(x=15) = BINOMDIST(15,20,0.5,FALSE) = 0.014785767

but rather

P(x>=15) = 1 – P(x<14) = 1-BINOMDIST(14,20,0.5,TRUE) = 0.020694733

Thus, we are rather surprised to see 15 or more H's (or A's>B's) because by chance we should only see this happening 2% of the time. Thus, because we DID see it happen, we will believe we are seeing something "real" because it happens less often than the 5% of the time it will happen by chance.

In other words, suppose someone has just flipped a "claimed" fair coin 20 times and they got 19 heads. Do you think you just witnessed a once-in-a-lifetime occurance that happened by random chance or do you think this is evidence that the coin is not fair and you aren't seeing anything particularly rare at all?

heads	p(x>=H)
0	1.00000
1	1.00000
2	0.99998
3	0.99980
4	0.99871
5	0.99409
6	0.97931
7	0.94234
8	0.86841
9	0.74828
10	0.58810
11	0.41190
12	0.25172
13	0.13159
14	0.05766
15	0.02069
16	0.00591
17	0.00129
18	0.00020
19	0.00002
20	0.00000

To be correct 95% of the time we will be wrong 5% of the time. Hence, when the item we are concerned with is occurring "in the tails" of the expected behavior, we will be "Rejecting Null" and when we are in the "main" (between the tails) we will be "FTR Null" (failing to reject the Null).

In this case, our evidence strongly suggests that one or both of the two machines need to be repaired. If one was ALWAYS higher, of course, the case would be "obvious". Also, if we found A>B 9 or 10 or 11 we probably wouldn't have thought we should repair them (because the behavior was as near expected for n=20 with p=0.5).

One Sample t-Tests for Mean

When we deal with "small samples" (say, n<30) the distribution against which we compare "expected behavior" is not "normal" (gaussian) but rather a related function (called the student-t distribution). The t-distribution function contains a correction for small "n" (number of degrees of freedom).

A worksheet (t-one_mean.xls) has been provided to simplify making t-tests on one sample.

Two Tailed Examples

1. We have a machine that (when working properly) puts 90g of candy in each bag. We have sampled 10 bags and find xbar=89.5g with sd=5g. Is the machine properly filling the bags? Use 95% CL.

This is a two-tail case since we are interested in "putting in 90" vs "not putting in 90".

Two-Tailed Test (u=uo)

Hypotheses		
Ho: μ	=	90
	<	
H1: μ	>	90
α	=	0.05
Sample Evidence		
Sample Mean is	=	89.5
Sample SD is	=	5
Sample Size is	=	10

Calculations

t-statistic	-0.316227766
p-value	0.759040654
Decision	FTR Null
CI Lower Bound	85.25462509
CI Upper Bound	93.74537491

The test statistic for this case was -0.316 (no where near -2). The "decision" we should make is "FTR Null" that is, I fail to be able to reject the Null. In other words, I will "accept the Null" (although no statistician ever says this!).

Wording 1: At a 95% CL (confidence level) I cannot reject that the machine is working properly.

Wording 2: At a 95% CL I accept that the machine is working properly. (That is, no one needs to "fix it").

2. We have a machine that (when working properly) puts 90g of candy in each bag. We have sampled 10 bags and find xbar=87.0g with sd=5g. Is the machine properly filling the bags? Use 95% CL.

Sample Evidence		
Sample Mean is	=	87
Sample SD is	=	5
Sample Size is	=	10

Calculations

t-statistic	-1.897366596
p-value	0.090267331
Decision	FTR Null

Wording 2: At a 95% CL I accept that the machine is working properly. (That is, no one needs to "fix it").

3. We have a machine that (when working properly) puts 90g of candy in each bag. We have sampled 10 bags and find xbar=85.0g with sd=5g. Is the machine properly filling the bags? Use 95% CL.

Sample Evidence		
Sample Mean is	=	85
Sample SD is	=	5
Sample Size is	=	10

Calculations

t-statistic	-3.16227766
p-value	0.011507985
Decision	Reject Null

Wording 2: At a 95% CL I reject that the machine is working properly. (That is, I believe H1, this behavior is not expected of samples coming from a population with a mean of 90, and I believe someone needs to "fix the machine").

4. We have a machine that (when working properly) puts 90g of candy in each bag. We have sampled 10 bags and find xbar=85.0g with sd=10g. Is the machine properly filling the bags? Use 95% CL.

Sample Evidence		
Sample Mean is	=	85
Sample SD is	=	10
Sample Size is	=	10

Calculations

t-statistic	-1.58113883
p-value	0.148304704
Decision	FTR Null

Wording 2: At a 95% CL I accept that the machine is working properly. (That is, no one needs to "fix it").

One Tailed Examples

5. We have readjusted our machine to produce bags labeled "contains 95g". We have sampled 10 bags and find xbar=92.0g with sd=5g. Is the machine putting in at least 95g? Use 95% CL.

This is a one-tail case since we are interested in "putting in at least 95" vs "not putting in at least 95" (that is, putting in less than 95).

One-Tailed (Right Tail)

Hypotheses		
	<	
Ho: μ	=	95
H1: μ	>	95
α	=	0.05
Sample Evidence		
Sample Mean is	=	92
Sample SD is	=	5
Sample Size is	=	10

Calculations

t-statistic	-1.897366596
p-value	0.954866335
Decision	FTR Null
CI Lower Bound	87.75462509
CI Upper Bound	96.24537491

The decision "FTR Null" means we accept the null, that is, accept that filling less than 95. We are accepting Ho and not H1 and so (at a 95% CL) we do NOT think the machine is putting in at least 95. We should adjust the machine.

6. We have again adjusted our machine and now find a sample of 10 bags contain 99g with sd=5g. Is the machine putting in at least 95g? Use 95% CL.

One-Tailed (Right Tail)

Hypotheses		
	<	
Ho: μ	=	95
H1: μ	>	95
α	=	0.05
Sample Evidence		
Sample Mean is	=	99
Sample SD is	=	5
Sample Size is	=	10

Calculations

t-statistic	2.529822128
p-value	0.01612239
Decision	Reject Null
CI Lower Bound	94.75462509

The decision is "Reject Null" that is, reject Ho (and accept H1). Therefore, we can accept that the machine is putting in at least 95 and no further adjustment is necessary.

7. We need to cut back on expenses and therefore we have labels printed saying "contains 85g". We want to make sure we are putting in "no more than 87g" and now find a random sample of 10 bags contain 88 with sd=5g. Is the machine putting in no more than 87g? Use 95% CL.

One-Tailed (Left Tail)

Hypotheses		
	>	
Ho: μ	=	87
H1: μ	<	87

α	=	0.05
Sample Evidence		
Sample Mean is	=	88
Sample SD is	=	5
Sample Size is	=	10

Calculations

t-statistic	0.632455532
p-value	0.728589521
Decision	FTR Null
CI Lower Bound	83.75462509
CI Upper Bound	92.24537491

Our decision is "FTR Null" means accepting Ho means accepting "more than 87". Hence, we are not able to say we are putting in "less than" 87 grams. The machine will need to be adjusted to achieve that.

Two Sample t-Tests for Mean

Often times we have available "before" and "after" data that represents a "treatment" or a hoped-for change in state (increased yield or decreased rate of product rejects). In this case, we are not comparing our sample to a "known" population but rather against another sample.

A worksheet (t-two_means.xls) has been provided to simplify making t-tests on one sample.

LowFlow Shower Head		
User	Before	After
1	5.21	4.66
2	4.33	2.02
3	2.09	2.17
4	5.72	4.98
5	7.31	4.23
6	3.96	1.55
7	4.88	5
8	5.6	2.75
9	7.35	3.68
10	10.95	7.71
mean	5.74	3.875
stdev	2.397235	1.855725

Suppose we wanted to see if there was an actual difference when using the new shower heads (flowrate during shower). Obviously there is a "difference" but we wish to investigate the issue "statistically" using the t-two_means spreadsheet.

If there is no difference then $\mu1-\mu2 = 0$

Two-Tailed Test (u=uo)

Hypotheses				
	Ho:	μ1-μ2	=	0
			<	
	H1:	μ1-μ2	>	0
		α	=	0.05
Sample Evidence				
		Sample 1		Sample 2
Sample Mean		5.7400		3.8750
Sample SD		2.3972		1.8557
Sample Size		10		10

Calculations

t-statistic	1.945406 9	
p-value	0.069513 5	
Decision	FTR Null	
CI Lower Bound	- 0.505670 2	
CI Upper Bound	4.235670 2	

The decision "FTR Null" indicates there IS NOT a significant difference between the two types of shower heads. Even though there appears to be a large difference, we need to appreciate that there is a large standard deviation (hence, the data is of low quality).

Let's check if there the data allows us to claim a difference with a confidence level of 90%.

Two-Tailed Test (u=uo)

Hypotheses				
	Ho:	μ1-μ2	=	0
			<	
	H1:	μ1-μ2	>	0

	α	=	0.1
Sample Evidence			
	Sample 1		Sample 2
Sample Mean	5.7400		3.8750
Sample SD	2.3972		1.8557
Sample Size	10		10

Calculations

t-statistic	1.945406 9
p-value	0.069513 5
Decision	Reject Null
CI Lower Bound	-0.167286 1
CI Upper Bound	3.897286 1

The decision "Reject Null" indicates that at a 90% CL we cannot reject the null hypothesis, therefore, the data supports a difference in the flow rate. Notice that we are less "confident" in the statement and hence risk being wrong more often (10 percent of the time).

The two other tables in the spreadsheet allow one to assess cases where the difference is "great than" or "less than" some specified value. We could use this to investigate if the difference was at least 1 gpm, etc.

Be sure to carefully assign (1) and (2) when determining the difference, $\mu 1-\mu 2$.

One Sample z-Tests for Proportion

When we deal with "success vs failure" issues we are attempting to establish the "probability" that is associated with a success. For example, a fair coin has a probability p=0.5 (success=heads) and a die has a probability p=0.1666667 (success=three dots).

We would now like to test samples where we have counted the number of successes and trials against particular probability statements.

For example, we might have a "trick coin" which the manufacturer claims comes up heads 70% of the time; p=0.7 (heads). In this case we might have flipped the coin in question 20 times and only observed 10 heads. Is this reason to reject the claim of the manufacturer?

A worksheet (z-one_prop.xls) has been provided to simplify making z-tests on one sample.

Two Tailed Examples

1. We have a trick coin which is claimed to come up heads 70% of the time. We have made 20 "random" flips and saw only 10 heads. Should we reject the claim of the toy manufacturer? Use 95% CL.

This is a two-tail case since we are interested in "p=0.70" vs "p <>0.70".

Two-Tailed Test (u=uo)

Hypotheses			
	Ho: p	=	0.7
	H1: p	< >	0.7
	α	=	0.05
Sample Evidence			
	Sample Prop.	=	0.5
	Sample Size	=	20

Calculations

z-statistic	-1.9518
p-value	0.05096 2
Decision	FTR Null

Our decision is "FTR Null", that is, we "accept Ho", that is, we accept p=0.7.

Wording: At a 95% confidence level, the sample supports the manufacturer claim that the coin comes up heads 70% of the time.

2. We have a trick coin which is claimed to come up heads 70% of the time. We have made 50 "random" flips and saw only 25 heads. Should we reject the claim of the toy manufacturer? Use 95% CL.

Two-Tailed Test (u=uo)

Hypotheses			
	Ho: p	=	0.7
		<	
	H1: p	>	0.7
	α	=	0.05
Sample Evidence			
	Sample Prop.	=	0.5
	Sample Size	=	50

Calculations

z-statistic	-3.08607
p-value	0.002028
Decision	Reject Null

Our decision is “Reject Null”, that is, we “Reject Ho”, that is, we reject p=0.7 and accept p<>0.7.

Wording: At a 95% confidence level, the sample does not support the manufacturer claim that the coin comes up heads 70% of the time.

3. We have devised an ESP experiment in which a subject is asked to state the color (red or blue) of a card chosen from a deck of 50 well-shuffled cards by an individual in another room. The subject does not know how many red or blue cards are in the deck. Suppose a test subject correctly identifies the color of 32 cards. Does this test support the claim that the subject has ESP?

Ho: p=0.5 (the subject is simply guessing and the successes are due to chance.

H1: p>0.5 (the subject has being able to get more right than mere guessing can explain)

p*=32/50=0.64

One-Tailed (Right Tail)

Hypotheses			
		<	
	Ho: p	=	0.5
	H1: p	>	0.5
	α	=	0.05
Sample Evidence			

Sample Prop.	=	0.64
Sample Size	=	50

Calculations

z-statistic	1.979899
p-value	0.023857 4
Decision	Reject Null

Our decision is "Reject Null", that is, reject p=0.5 and accept p>0.5.

Wording: At a 95% confidence level, the experiment supports the claim that the subject has a degree of ESP. This outcome (32 out of 50) happens less than 5% of the time DUE TO CHANCE when someone is merely guessing.

4. A drug manufacturer claims that their cold remedy was 90% effective in relieving allergy symptoms for 8 hours. In a sample of 200 people who had the allergy, the medicine provided 8 hour relief for 160 people. Is the manufacturer's claim legitimate at a confidence level of 0.01?

Ho: p=0.9

H1: p<0.9

This is a "left-tail" test. Problem solution left for student.

Two Sample z-Tests for Proportion

When we deal two samples we are usually testing "before" and "after" situations or "with and without" treatment.

A worksheet (z-two_prop.xls) has been provided to simplify making z-tests on two samples.

5. Two groups of people, A and B, each consist of 100 people who have a particular disease. A serum (medicine) is given to the Group A people and a placebo (fake) is given to group B (the control group). It was found that in Group A and B, 75 and 65 people recovered from the disease. Test the hypothesis that the drug administered "helps" to cure the disease using a significance level of 0.01.

Ho: p1=p2

H1: p1>p2

This is a "right-tail" test for two samples.

One-Tailed (Right Tail)

Hypotheses				
			<	
	Ho:	p1-p2	=	0
	H1:	p1-p2	>	0
		α	=	0.01
Sample Evidence				
		Sample 1		Sample 2
Sample Prop.		0.75		0.65
Sample Size		100		100
		p combo	=	0.7
Calculations				
z-score		1.552301		1.543033
p-value		0.060295		0.061411
Decision		FTR Null		

Our decision is to FTR Null, that is, fail to reject Null, that is, fail to reject Ho. We therefore reject H1.

Wording: At a 99% confidence level, the samples provided (experimental data) do not support the claim of effectiveness of the medicine. Random chance could make the data appear as it does more than 1% of the time.

6. A sample poll of 300 voters from district A and 200 voters from district B showed that 56% and 48% respectively favor the candidate Joe Schmo. (a) At a level of 95%, test the hypothesis that there is a difference in preference between the two districts. (b) At a level of 95%, test the hypothesis that the candidate is preferred in district A.

Case (6a) Ho: p1=p2 H1: p1<>p2

Two-Tailed Test

Hypotheses				
	Ho:	p1-p2	=	0
			<	
	H1:	p1-p2	>	0
		α	=	0.05

Sample Evidence			
	Sample 1		Sample 2
Sample Prop.	0.56		0.48
Sample Size	300		200
	p combo	=	0.528
Calculations			
z-score	1.758631		1.755467
p-value	0.07864		0.079179
Decision	FTR Null		

The decision is FTR Null, that is, we accept Ho and reject H1. At a 95% confidence level, the sampled data supports that there is no difference between candidate preference in the two districts.

Case (6b) Ho: p1<=p2 H1: p1>p2

One-Tailed (Right Tail)			
Hypotheses			
Ho:	p1-p2	<=	0
H1:	p1-p2	>	0
	α	=	0.05
Sample Evidence			
	Sample 1		Sample 2
Sample Prop.	0.56		0.48
Sample Size	300		200
	p combo	=	0.528
Calculations			
z-score	1.758631		1.755467
p-value	0.03932		0.03959
Decision	Reject Null		

The decision is "Reject Null", that is, reject Ho, that is, accept H1.

Wording: At a 95% confidence level, the data available does not support the statement that Joe Schmo is less popular in District A. (The data does support the statement that Joe Schmo is the preferred candidate in district A.)

Note that this is not a contradiction of 6a. Each case is considered on its own.