

## Lecture Notes 7

Econ 29000, Principles of Statistics

Kevin R Foster, CCNY

Fall 2011

### Contents

- Hypothesis Testing
- Confidence Intervals
- p-values
- examples
- complications from a series of tests

### Learning Outcomes (from CFA exam Study Session 3, Quantitative Methods)

Students will be able to:

- construct and interpret a confidence interval for a normally distributed random variable, and determine the probability that a normally distributed random variable lies inside a given confidence interval;
- define the standard normal distribution, explain how to standardize a random variable, and calculate and interpret probabilities using the standard normal distribution;
- explain the construction of confidence intervals;
- define a hypothesis, describe the steps of hypothesis testing, interpret and discuss the choice of the null hypothesis and alternative hypothesis, and distinguish between one-tailed and two-tailed tests of hypotheses;
- define and interpret a test statistic, a Type I and a Type II error, and a significance level, and explain how significance levels are used in hypothesis testing;

### Hypothesis Testing

One of the principal tasks facing the statistician is to perform hypothesis tests. These are a formalization of the most basic questions that people ask and analyze every day – just contorted into odd shapes. But as long as you remember the basic common sense underneath them, you can look up the precise details of the formalization that lays on top.

The basic question is "How likely is it, that I'm being fooled?" Once we accept that the world is random (rather than a manifestation of some god's will), we must decide how to make our decisions, knowing that we cannot guarantee that we will always be right. There is some risk that the world will seem to be one way, when actually it is not. The stars are strewn randomly across the sky but some bright ones seem to line up into patterns. So too any data might sometimes line up into patterns.

A formal hypothesis sets a mathematical condition that I want to test. Often this condition takes the form of some parameter being zero for no relationship or no difference.

Statisticians tend to stand on their heads and ask: What if there were actually **no** relationship? (Usually they ask questions of the form, "suppose the conventional wisdom were true?") This statement, about "no relationship," is called the **Null Hypothesis**, sometimes abbreviated as  $H_0$ . The Null Hypothesis is tested against an **Alternative Hypothesis**,  $H_A$ .

Before we even begin looking at the data we can set down some rules for this test. We know that there is some probability that nature will fool me, that it will seem as though there is a relationship when actually there is none. The statistical test will create a model of a world where there is actually no relationship and then ask how likely it is that we could see what we actually see, "How likely is it, that I'm being fooled?"

The "likelihood that I'm being fooled" is the p-value.

For a scientific experiment we typically first choose the level of certainty that we desire. This is called the significance level. This answers, "How low does the p-value have to be, for me to accept the formal hypothesis?" To be fair, it is important that we set this value first because otherwise we might be biased in favor of an outcome that we want to see. By convention, economists typically use 10%, 5%, and 1%; 5% is the most common.

A five percent level of a test is conservative, it means that we want to see so much evidence that there is only a 5% chance that we could be fooled into thinking that there's something there, when nothing is actually there. Five percent is not perfect, though – it still means that of every 20 tests where I decide that there is a relationship there, it is likely that I'm being fooled in one of those – I'm seeing a relationship where there's nothing there.

To help ourselves to remember that we can never be truly certain of our judgment of a test, we have a peculiar language that we use for hypothesis testing. If the "likelihood that I'm being fooled" is less than 5% then we say that the data allow us to *reject* the null hypothesis. If the "likelihood that I'm being fooled" is more than 5% then the data *do not reject* the null hypothesis.

Note the formalism: we never "accept" the null hypothesis. Why not? Suppose I were doing something like measuring a piece of machinery, which is supposed to be a centimeter long. The null hypothesis is that it is not defective and so is one centimeter in length. If I measure with a ruler I might not find any difference to the eye. So I cannot reject the hypothesis that it is one centimeter. But if I looked with a microscope I might find that it is not quite one centimeter! The fact that, with my eye, I don't see any difference, does not imply that a better measurement could not find any difference. So I cannot say that it is truly exactly one centimeter; only that I can't tell that it isn't.

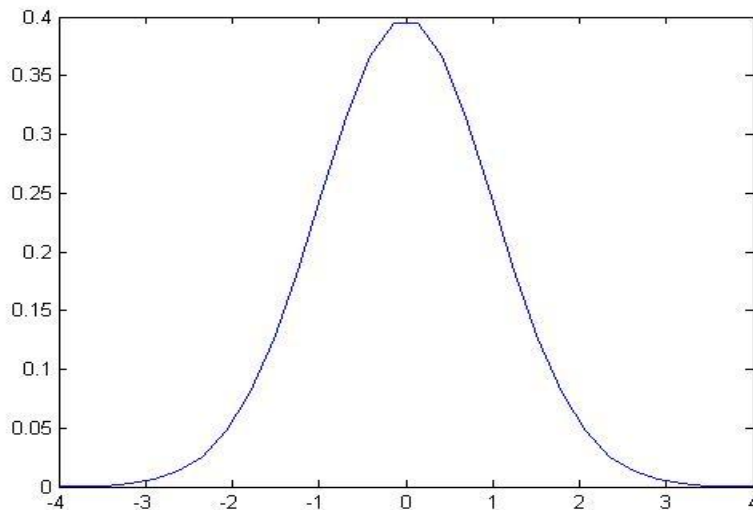
So too with statistics. If I'm looking to see if some portfolio strategy produces higher returns, then with one month of data I might not see any difference. So I would not reject the null hypothesis (that the new strategy is no improvement). But it is possible that the new strategy, if carried out for 100 months or 1000 months or more might show some tiny difference.

Not rejecting the null is saying that I'm not sure that I'm not being fooled. (Read that sentence again; it's not immediately clear but it's trying to make a subtle and important point.)

To summarize, Hypothesis Testing asks, "What is the chance that I would see the value that I've actually got, if there truly were no relationship?" If this p-value is lower than 5% then I reject the null hypothesis of "no relationship." If the p-value is greater than 5% then I do not reject the null hypothesis of "no relationship."

The rest is mechanics.

The null hypothesis would tell that a parameter has some particular value, say zero:  $H_0 : \mu = 0$ ; the alternative hypothesis is  $H_A : \mu \neq 0$ . Under the null hypothesis the parameter has some distribution (often normal), so  $H_0 : \mu \sim N(0, \sigma_{std\ err})$ . Generally we have an estimate for  $\sigma_{std\ err}$ , which is  $se$  (for small samples this inserts additional uncertainty). So I know that, under the null hypothesis,  $\frac{\mu}{se}$  has a standard normal distribution (mean of zero and standard deviation of one). I know exactly what this distribution looks like, it's the usual bell-shaped curve:



So from this I can calculate, "What is the chance that I would see the value that I've actually got, if there truly were no relationship?," by asking what is the area under the curve that is farther away from zero than the value that the data give. (I still don't know what value the data will give! I can do all of this calculation beforehand.)

Any particular estimate of  $\mu$  is generally going to be  $\bar{X}$ . So the test statistic is formed with  $\frac{\bar{X}}{se}$ .

Looking at the standard normal pdf, a value of the test statistic of 1.5 would not meet the 5% criterion (go back and calculate areas under the curve). A value of 2 would meet the 5% criterion, allowing us to reject the null hypothesis. For a 5% significance level, the standard normal **critical value** is 1.96: if the test statistic is larger than 1.96 (in absolute value) then its p-

value is less than 5%, and vice versa. (You can find critical values by looking them up in a table or using the computer.)

*Sidebar:* Sometimes you see people do a one-sided test, which is within the letter of the law but not necessarily the spirit of the law (particularly in regression formats). It allows for less restrictive testing, as long as we believe that we know that there is only one possible direction of deviation (so, for example, if the sample could be larger than zero but never smaller). But in this case maybe the normal distribution is inapplicable.

The test statistic can be transformed into measurements of  $\mu$  or into a confidence interval.

If I know that I will reject the null hypothesis of  $\mu = 0$  at a 5% level if the test statistic,  $\frac{\bar{X}}{se}$ , is greater than 1.96 (in absolute value), then I can change around this statement to be about  $\bar{X}$ . This says that if the estimated value of  $\bar{X}$  is less than 1.96 standard errors from zero, we cannot reject the null hypothesis. So cannot reject if:

$$\frac{|\bar{X}|}{se} < 1.96$$

$$|\bar{X}| < 1.96se$$

$$-1.96se < \bar{X} < 1.96se.$$

This range,  $(-1.96se, 1.96se)$ , is directly comparable to  $\bar{X}$ . If I divide  $\bar{X}$  by its standard error then this ratio has a normal distribution with mean zero and standard deviation of one. If I don't divide then  $\bar{X}$  has a normal distribution with mean zero and standard deviation,  $se$ .

If the null hypothesis is not zero but some other number,  $\mu_{null}$ , then under the null hypothesis the estimator would have a normal distribution with mean of  $\mu_{null}$  and standard error,  $se$ . To transform this to a standard normal would mean subtracting the mean and dividing by  $se$ , so

cannot reject if  $\frac{|\bar{X} - \mu_{null}|}{se} < 1.96$ , i.e. cannot reject if  $\bar{X}$  is within the range,

$$(\mu_{null} - 1.96se, \mu_{null} + 1.96se).$$

### Confidence Intervals

We can use the same critical values to construct a confidence interval for the estimator, usually expressed in the form  $\bar{X} \pm 1.96se$ . This shows that, for a given sample size (therefore  $se$ , which depends on the sample size) that there is a 95% likelihood that the interval formed around a given estimator contains the true value.

This relates to hypothesis testing because if the confidence interval includes the null hypothesis then we cannot reject the null; if the null hypothesis value is outside of the confidence interval then we can reject the null.

### Find p-values

We can also find p-values associated with a particular null hypothesis by turning around the process outlined above. If the null hypothesis is zero, then with a 5% significance level we reject the null if  $\frac{\bar{X}}{se}$  is greater than 1.96 in absolute value. What if the ratio  $\frac{\bar{X}}{se}$  were 2 – what is the smallest significance level that would still reject? (Check your understanding: is it more or less than 5%?)

We can compute the ratio  $\frac{\bar{X}}{se}$  and then convert this number to a p-value, which is the smallest significance level that would still reject the null hypothesis (and if the null is rejected at a low level then it would automatically be rejected at any higher levels).

### Type I and Type II Errors

Whenever we use statistics we must accept that there is a likelihood of errors. In fact we distinguish between two types of errors, called (unimaginatively) Type I and Type II. These errors arise because a null hypothesis could be either true or false and a particular value of a statistic could lead me to reject or not reject the null hypothesis,  $H_0$ . A table of the four outcomes is:

	$H_0$ is true	$H_0$ is false
Do not reject $H_0$	good!	oops – Type II
Reject $H_0$	oops – Type I	good!

Our chosen significance level (usually 5%) gives the probability of making an error of Type I. We cannot control the level of Type II error because we do not know just how far away  $H_0$  is from being true. If our null hypothesis is that there is a zero relationship between two variables, when actually there is a tiny, weak relationship of 0.0001%, then we could be very likely to make a Type II error. If there is a huge, strong relationship then we'd be much less likely to make a Type II error.

There is a tradeoff (as with so much else in economics!). If I push down the likelihood of making a Type I error (using 1% significance not 5%) then I must be increasing the likelihood of making a Type II error.

Edward Gibbon notes that the emperor Valens would "satisfy his anxious suspicions by the promiscuous execution of the innocent and the guilty" (chapter 26). This rejects the null hypothesis of "innocence"; so a great deal of Type I error was acceptable to avoid Type II error.

Every email system fights spam with some sort of test: what is the likelihood that a given message is spam? If it's spam, put it in the "Junk" folder; else put it in the inbox. A Type I error represents putting good mail into the "Junk" folder; Type II puts junk into your inbox.

### Examples

Let's do some examples.

Assume that the calculated average is 3, the sample standard deviation is 15, and there are 100 observations. The null hypothesis is that the average is zero. The standard error of the average is  $se = \frac{15}{\sqrt{100}} = 1.5$ . We can immediately see that the sample average is more than two standard errors away from zero so we can reject at a 95% confidence level.

Doing this step-by-step, the average over its standard error is  $\frac{\bar{X}}{se} = \frac{3}{1.5} = 2$ . Compare this to 1.96 and see that  $2 > 1.96$  so we can reject. Alternately we could calculate the interval,  $(-1.96s, 1.96s)$ , which is  $((-1.96 \cdot 1.5), (1.96 \cdot 1.5)) = (-2.94, 2.94)$ , outside of which we reject the null. And 3 is outside that interval. Or calculate a 95% confidence interval of  $3 \pm 2.94 = (0.06, 5.94)$ , which does not contain zero so we can reject the null. The critical value for the estimate of 3 is 4.55% (found from Excel either  $2*(1-NORMSDIST(2))$  if using the standard normal distribution or  $2*(1-NORMDIST(3,0,1.5,TRUE))$  if using the general normal distribution with a mean of zero and standard error of 1.5).

If the sample average were -3, with the same sample standard deviation and same 100 observations, then the conclusions would be exactly the same.

Or suppose you find that the average difference between two samples, X and Y, (i.e.

$\bar{X} - \bar{Y} = \frac{1}{n} \sum_{i=1}^n (X_i - Y_i)$ ) is -0.0378. The sample standard deviation is 0.357. The number of observations is 652. These three pieces of information are enough to find confidence intervals, do t-tests, and find p-values.

How?

First find the standard error of the average difference. This standard error is 0.357 divided by the square root of the number of observations, so  $\frac{.357}{\sqrt{652}} = 0.01398$ .

So we know (from the Central Limit Theorem) that the average has a normal distribution. Our best estimate of its true mean is the sample average, -0.0378. Our best estimate of its true standard error is the sample standard error, 0.01398. So we have a normal distribution with mean -0.0378 and standard error 0.01398.

We can make this into a standard normal distribution by adding 0.0378 and dividing by the sample standard error, so now the mean is zero and the standard error is one.

We want to see how likely it would be, if the true mean were actually zero, that we would see a value as extreme as -0.0378. (Remember: we're thinking like statisticians!)

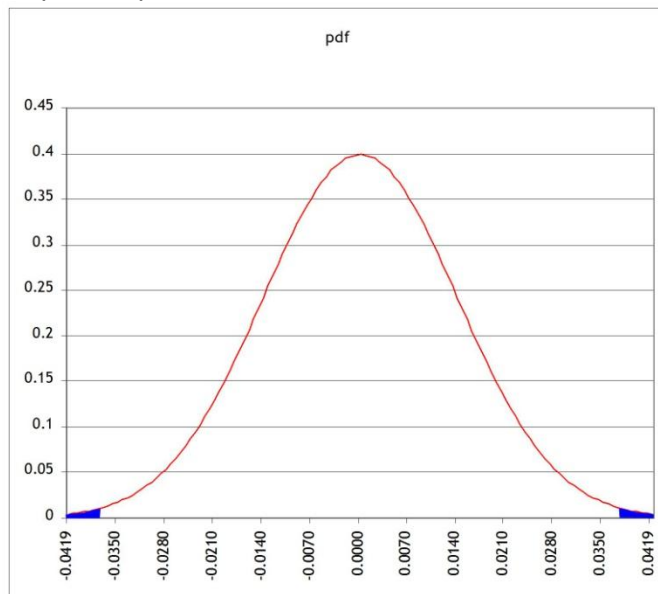
The value of -0.0378 is  $\frac{-0.0378}{0.01398} = -2.70$  standard deviations from zero.

From this we can either compare this against critical t-values or use it to get a p-value.

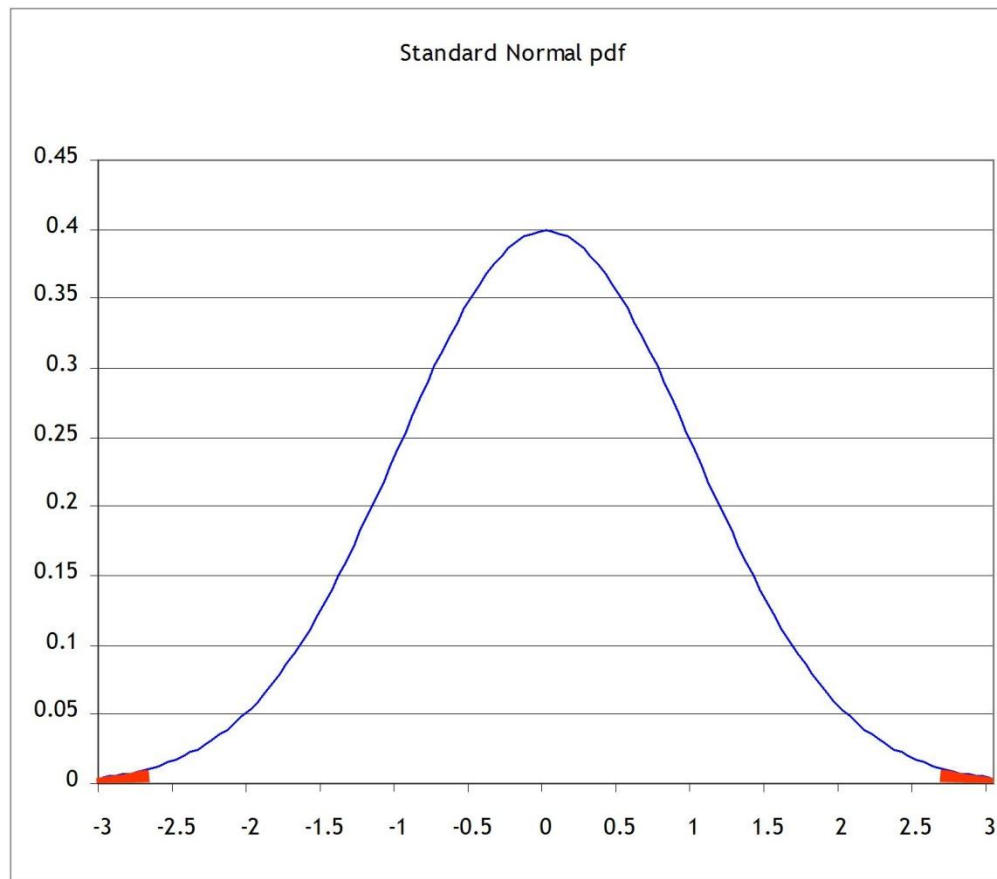
To find the p-value, we can use Excel just like in the homework assignment. If we have a standard normal distribution, what is the probability of finding a value as far from zero as -2.27, if the true mean were zero? This is  $2*(1-NORMSDIST(-2.27)) = 0.6\%$ . The p-value is 0.006 or 0.6%. If we are using a 5% level of significance then since 0.6% is less than 5%, we reject the null hypothesis of a zero mean. If we are using a 1% level of significance then we can reject the null hypothesis of a zero mean since 0.6% is less than 1%.

Or instead of standardizing we could have used Excel's other function to find the probability in the left tail, the area less than -0.0378, for a distribution with mean zero and standard error 0.01398, so  $2*NORMDIST(-0.0378,0,0.01398,TRUE) = 0.6\%$ .

Standardizing means (in this case) zooming in, moving from finding the area in the tail of a very small pdf, like this:



to moving to a standard normal, like this:



But since we're only changing the units on the x-axis, the two areas of probability are the same.

We could also work backwards. We know that if we find a standardized value greater (in absolute value) than 1.96, we would reject the null hypothesis of zero at the 5% level. (You can go back to your notes and/or HW1 to remind yourself of why 1.96 is so special.)

We found that for this particular case, each standard deviation is of size  $\frac{.357}{\sqrt{652}} = 0.01398$ . So we can multiply 1.96 times this value to see that if we get a value for the mean, which is farther from zero than  $0.01398 \times 1.96 = 0.0274$ , then we would reject the null. Sure enough, our value of -0.0378 is farther from zero, so we reject.

Alternately, use this 1.96 times the standard error to find a confidence interval. Choose 1.96 for a 95% confidence interval, so the confidence interval around -0.0378 is plus or minus 0.0274,  $-0.0378 \pm 0.0274$ , which is the interval (-0.0652, -0.0104). Since this interval does not include zero we can be 95% confident that we can reject the null hypothesis of zero.

### Complications from a Series of Hypothesis Tests

Often a modeler will make a series of hypothesis tests to attempt to understand the inter-relations of a dataset. However while this is often done, it is not usually done correctly. Recall



from our discussion of Type I and Type II errors that we are always at risk of making incorrect inferences about the world based on our limited data. If a test has a significance level of 5% then we will not reject a null hypothesis until there is just a 5% probability that we could be fooled into seeing a relationship where there is none. This is low but still is a 1-in-20 chance. If I do 20 hypothesis tests to find 20 variables that significantly impact some variable of interest, then it is likely that one of those variables is fooling me (I don't know which one, though). It is also likely that my high standard of proof meant that there are other variables which are more important but which didn't seem it.

Sometimes you see very stupid people who collect a large number of possible explanatory variables, run hundreds of regressions, and find the ones that give the "best-looking" test statistics – the ones that look good but are actually entirely fictitious. Many statistical programs have procedures that will help do this; help the user be as stupid as he wants.

Why is this stupid? It completely destroys the logical basis for the hypothesis tests and makes it impossible to determine whether or not the data are fooling me. In many cases this actually guarantees that, given a sufficiently rich collection of possible explanatory variables, I can run a regression and show that some variables have "good" test statistics – even though they are completely unconnected. Basically this is the infamous situation where a million monkeys randomly typing would eventually write Shakespeare's plays. A million earnest analysts, running random regressions, will eventually find a regression that looks great – where all of the proposed explanatory variables have test statistics that look great. But that's just due to persistence; it doesn't reflect anything about the larger world.

In finance, which throws out gigabytes of data, this phenomenon is common. For instance there used to be a relationship between which team won the Super Bowl (in January) and whether the stock market would have a good year. It seemed to be a solid result with decades of supporting evidence – but it was completely stupid and everybody knew it. Analysts still work to get slightly-less-implausible but still completely stupid results, which they use to sell their securities.

Consider the logical chain of making a number of hypothesis tests in order to find one supposedly-best model. When I make the first test, I have 5% chance of making a Type I error. Given the results of this test, I make the second test, again with a 5% chance of making a Type I error. The probability of not making an error on either test is  $(.95)(.95) = .9025$  so the significance level of the overall test procedure is  $1 - .9025 = 9.75\%$ . If I make three successive hypothesis tests, the probability of not making an error is  $.8574$  so the significance level is  $14.26\%$ . If I make 10 successive tests then the significance level is over 40%! This means that there is a 40% chance that the tester is being fooled, that there is not actually the relationship there that is hypothesized – and worse, the stupid tester believes that the significance level is just 5%.

## **Hypothesis Testing with two samples**

In our examples we have often come up with a question where we want to know if there is a difference in mean between two groups. From the PUMS data, we could ask if, say, people in NYC are paying too much of their income to rent (where "too much" is generally defined as one-third).

For the PUMS data, SPSS tells us (using "Analyze \ Descriptive Statistics \ Descriptives") that, of 74,793 households, the average "Gross Rent as percent of income" in the sample is 40, with a

standard deviation of 28.89. So the standard error of the average is  $\frac{s}{\sqrt{n}} = \frac{28.89}{\sqrt{74793}} = 0.1056$ ; so two standard errors from 40 is still only 39.7887 and 40.2113; the ideal amount of 33% of income is over 66 standard errors away! We don't need to look up NORMSDIST(-66) to figure that this is highly improbable. We might want to subdivide farther: to look if the burden of rent is worse in Manhattan, say, compared with the other boroughs.

Now we find that in Manhattan the fraction of income going to rent is 37.79 with standard deviation of 27.608, against 41.62 (with standard error of 29.213) for the outer boroughs (SPSS output in appendix below). Are these significantly different from each other? Could we just be observing a difference that is within the range of variation that we'd expect, given such a diverse group of households? Are they statistically significantly different from each other? Our formula that we learned last time has only one  $n$  – what do we do if we have two samples?

Basically we want to figure out how to use the two separate standard errors to estimate the joint standard error; once we get that we'll use the same basic strategy to get our estimate, subtract off the null hypothesis (usually zero), and divide by its standard error. We just need to know what is that new standard error.

To do this we use the sum of the two sample variances: if we are testing group 1 vs group 2, then a test of just group 1 would estimate its variance as  $\frac{s_1^2}{n_1}$ , a test of group 2 would use  $\frac{s_2^2}{n_2}$ ,

and a test of the group would estimate the standard error as  $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ .

We can make either of two assumptions: either the standard deviations are the same (even though we don't know them) or they're different (even though we don't know how different). It is more conservative to assume that they're different (i.e. don't assume that they're the same) – this makes the test less likely to reject the null.

## Appendix

SPSS output from "Analyze \ Descriptive Statistics \ Explore" where the "dependent" is "Gross Rent as percent of Income" and "factor" is "boro\_m" (a dummy variable that is one if the household is in Manhattan and zero else):

### Case Processing Summary

		Cases					
		Valid		Missing		Total	
		N	Percent	N	Percent	N	Percent
Gross Rent as percent of income	.00	53996	20.8%	205632	79.2%	259628	100.0%
	1.00	20797	37.0%	35346	63.0%	56143	100.0%

### Descriptives

boro_m			Statistic	Std. Error
Gross Rent as percent of income	.00	Mean	41.62	.126
		95% Confidence Interval for Mean	41.37	
		Lower Bound	41.87	
		Upper Bound	40.22	
		5% Trimmed Mean	31.00	
		Median	853.376	
		Variance	29.213	
		Std. Deviation	1	
		Minimum	101	
		Maximum	100	
		Range	36	
		Interquartile Range	.962	.011
		Skewness	-.339	.021
		Kurtosis		
	1.00	Mean	35.79	.191
		95% Confidence Interval for Mean	35.42	
		Lower Bound	36.17	
		Upper Bound	33.92	
		5% Trimmed Mean	27.00	
		Median	762.204	
		Variance	27.608	
		Std. Deviation		

Minimum	1	
Maximum	101	
Range	100	
Interquartile Range	30	
Skewness	1.227	.017
Kurtosis	.497	.034

Note that SPSS reports the "Standard error of the Statistic" so, for the Mean, it takes the standard deviation and divides by the square root of the number of observations to get the standard error of the mean. So for example, 0.126 is just  $29.213/\sqrt{53996}$ .