

Review of Basic Statistical Concepts

The purpose of this review is to summarize the basic statistical concepts. Introductory statistics dealt with three main areas: descriptive statistics, probability, and inference.

Descriptive Statistics	Sample data may be summarized graphically or with summary statistics. Sample statistics include the mean , variance , standard deviation , and median. For the following definitions let x_1, x_2, \dots, x_n represent values of variable X obtained from a random sample of size n drawn from a population of interest.	
Sample Mean	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \dots + x_n}{n}$	The mean is just the average of the n values observed.
Sample Variance	$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$	The sample variance equals the mean squared deviation from \bar{x} . A small s^2 means that the observed values cluster around their average, while a large variance means that the x_i are more spread out. Thus, the variance is a measure of the “spread” in the sampled values.
Sample Standard Deviation	$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$	The sample standard deviation, s , is often a more useful measure of spread than the sample variance, s^2 , because s has the same units (inches, pounds, etc.) as the sample values (the x_i) and the sample mean \bar{x} .
StatGraphics	Common descriptive statistics can be obtained by following: Describe > Numeric Data > One-Variable Analysis > Tabular Options > <i>Summary Statistics</i>	
<u>Example</u>	The file LMF contains the three-year return for a random sample of 26 mutual funds. All of these funds involve a load (a type of sales charge). StatGraphics output is to the right.	<p>Summary Statistics for Return</p> <p>Count = 26 Average = 16.2346 Variance = 40.4208 Standard deviation = 6.35773 Minimum = 8.0 Maximum = 32.7 Range = 24.7 Std. skewness = 2.26003 Std. kurtosis = 1.1129</p>

Random Variables and their Probability Distributions

Random Variable	<p>A variable whose numerical value is determined by chance. The key elements here are that the variable assumes a number (sales volume, rate of return, test score, etc.) and that the sample selection process generates the numbers randomly, i.e., by a “random” selection.</p> <p>(In these notes, a random variable will be designated by a capital letter, such as X, to differentiate it from observed values x. For instance, X might represent the height of a man to be selected randomly. Once the man has been selected, his height is given by the value x, say $x = 68$ inches.)</p>	
Probability Distribution	Although the values of a random variable are subject to chance, some values are more likely to occur than others. For instance, the height of a randomly selected man is more likely to measure 6’ than 7’. It is the random variable’s probability distribution that determines the relative likelihood of possible values.	

Standardized Values

For the value x drawn from a population with mean μ and standard deviation σ , the **standardized** value $z = \frac{x - \mu}{\sigma}$ = the number of standard deviations above or below the mean that x is. For example, if incomes have a mean and standard deviation of \$48,000 and \$16,000, respectively, then someone making \$56,000 has a standardized income of $\frac{\$56,000 - \$48,000}{\$16,000} = \frac{\$8,000}{\$16,000} = \frac{1}{2}$ because their income is one-half standard deviation above the mean income. The advantage of standardizing is that it facilitates the comparison of values drawn from different populations. (Standardized values are one **measure of relative standing**, another is a value's **percentile**.)

Standardized Random Variables

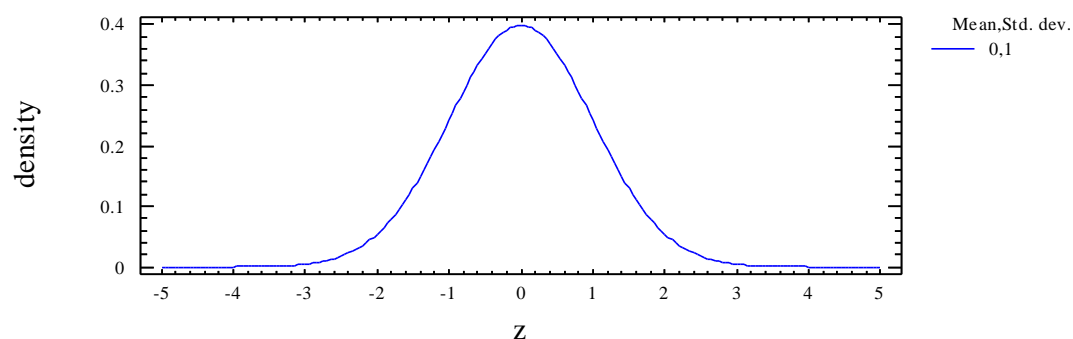
For the random variable X with mean μ and standard deviation σ , $Z = \frac{X - \mu}{\sigma}$ is the **Standardized** Random Variable. (Note: The Standardized Variable always has mean 0 and standard deviation 1.)

The Normal Distribution

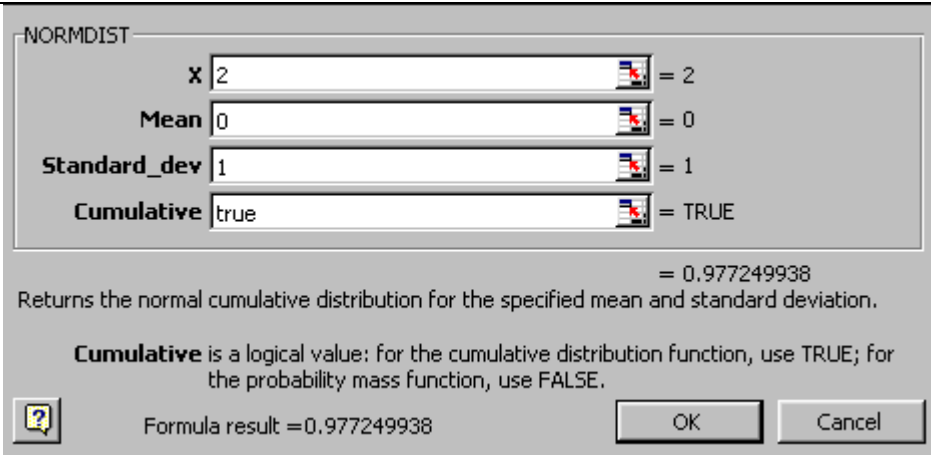
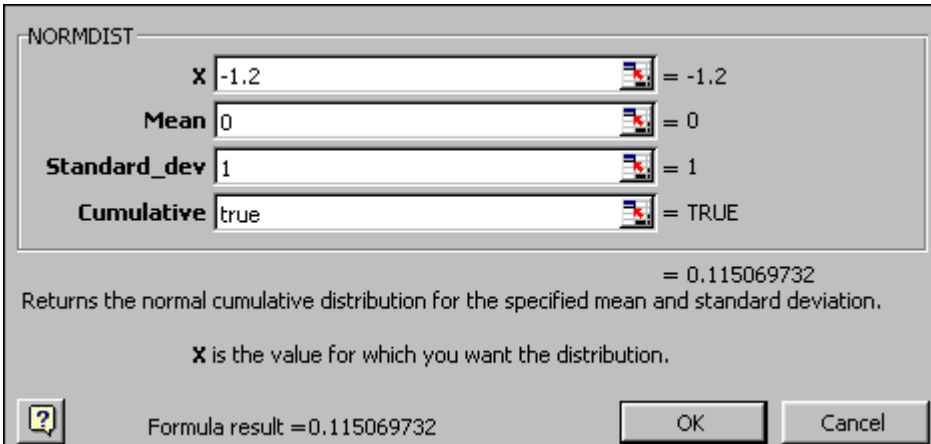
In this course we will make use of (at least) four distributions designed to model **continuous** data: the **Normal**, ***t***, ***F***, and ***Chi-Squared***. Of these, the normal distribution is by far the most important because of its role in **statistical inference**. Much of the logic behind what we do and why we do it is based upon an understanding of the properties of the normal distribution, and of the theorems involving it, particularly the **Central Limit Theorem**.

Properties	<ol style="list-style-type: none">1. Normal distributions are bell-shaped. (In fact, it is sometimes called the “Bell Curve”.)2. Normal distributions are symmetric about their mean.3. Normal distributions follow the 68-95-99.7 rule:<ul style="list-style-type: none">• (Approximately) 68% of the area under the curve is within <i>one</i> standard deviation σ of the mean μ• (Approximately) 95% of the area under the curve is within <i>two</i> standard deviations σ of the mean μ• (Approximately) 99.7% of the area under the curve is within <i>three</i> standard deviations σ of the mean μ4. If the random variable X is normal with mean μ and standard deviation σ, then the random variable $Z = \frac{X - \mu}{\sigma}$ is standard normal, i.e., is normal with mean equal 0 and standard deviation equal 1.
------------	--

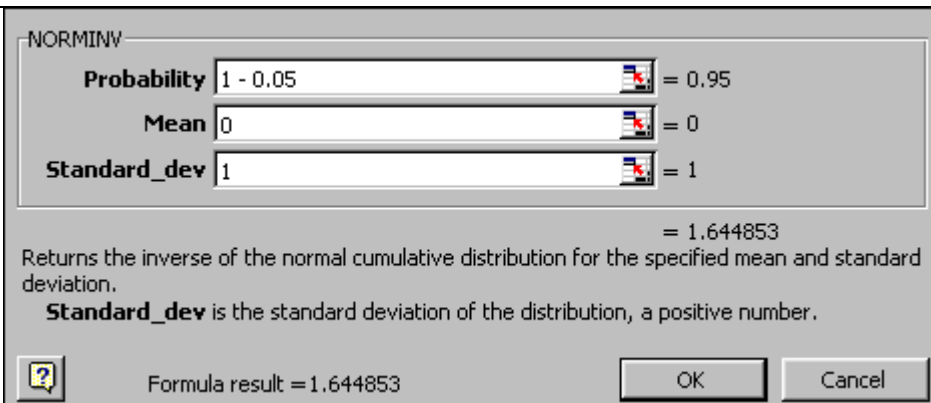
the Standard Normal Distribution, Z



Finding probabilities in Excel	Cumulative Probabilities for any normal random variable X , i.e., $P(X \leq x)$, are easy to find in Excel. Follow: $f_x > \text{Statistical} > \text{NORMDIST}$ and enter TRUE in the <i>Cumulative</i> field. Probabilities of the form $P(X > x)$ or $P(a < X < b)$ can be obtained by subtraction.
--------------------------------	---

Example	<p>To find $P(-1.2 < Z < 2)$, note that $P(-1.2 < Z < 2) = P(Z < 2) - P(Z \leq -1.2)$ and use the Excel output to the right.</p> <p>Answer = $0.9772 - 0.1151$ = 0.8621</p>	 <p>NORMDIST</p> <p>x 2 = 2</p> <p>Mean 0 = 0</p> <p>Standard_dev 1 = 1</p> <p>Cumulative true = TRUE</p> <p>= 0.977249938</p> <p>Returns the normal cumulative distribution for the specified mean and standard deviation.</p> <p>Cumulative is a logical value: for the cumulative distribution function, use TRUE; for the probability mass function, use FALSE.</p> <p>Formula result = 0.977249938</p> <p>OK Cancel</p>
		 <p>NORMDIST</p> <p>x -1.2 = -1.2</p> <p>Mean 0 = 0</p> <p>Standard_dev 1 = 1</p> <p>Cumulative true = TRUE</p> <p>= 0.115069732</p> <p>Returns the normal cumulative distribution for the specified mean and standard deviation.</p> <p>x is the value for which you want the distribution.</p> <p>Formula result = 0.115069732</p> <p>OK Cancel</p>

Critical Values	z_α is defined by $P(Z > z_\alpha) = \alpha$. Critical values are used in the construction of confidence intervals and (optionally) in hypotheses testing . To find the critical value associated with the significance level α , follow: $f_x > \text{Statistical} > \text{NORMINV}$ and enter $1 - \alpha$ in the <i>Probability</i> field.
-----------------	--

Example	<p>From the Excel output to the right we see that $z_{0.05} = 1.645$</p>	 <p>NORMINV</p> <p>Probability 1 - 0.05 = 0.95</p> <p>Mean 0 = 0</p> <p>Standard_dev 1 = 1</p> <p>= 1.644853</p> <p>Returns the inverse of the normal cumulative distribution for the specified mean and standard deviation.</p> <p>Standard_dev is the standard deviation of the distribution, a positive number.</p> <p>Formula result = 1.644853</p> <p>OK Cancel</p>
---------	---	--

The Distribution of the Sample Mean

Because, when we take a random sample, the values of a random variable are determined by chance, statistics such as the sample mean that are calculated from the values are themselves random

variables. Thus, the random variable $\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$ has a probability distribution of its

own. If we intend to use the sample mean $\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$ to estimate the mean μ of the

population from which the sample was drawn, then we need to know what values the random variable \bar{X} can assume and with what probability, i.e., we need to know the probability distribution of \bar{X} . It can be shown that \bar{X} has the following properties:

- The mean of \bar{X} equals the mean of X , i.e., $\mu_{\bar{x}} = \mu$. This just says that the sample mean \bar{x} is an **unbiased estimator** of the population mean μ .
- The variance of \bar{X} is less than that of X . In fact, $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$. This states that there is less variability in averaged values (and the variability *decreases* as the sample size *increases*) than there is in individual values. Hence, you might not be surprised if a randomly selected man measured 7', but you would be suspicious if someone claimed that 100 randomly chosen men *averaged* 7'!
- If the variable X is normally distributed, then \bar{X} will also be normal.

The properties above, however, don't describe the *shape* of the distribution of \bar{X} (needed for making inferences about μ) *except in the special case where X is normal!* They only contribute information about the mean and spread of the distribution. In general, the shape of the distribution of \bar{X} may be difficult to determine for small samples drawn from non-normal populations. However:

- For *large* samples the **Central Limit Theorem** states that \bar{X} will be at least approximately normal. (Most introductory statistics texts consider a sample large whenever $n \geq 30$.)

<u>Example</u>	The dean of a business school claims that the average weekly income of graduates of his school 1 year after graduation is \$600, with a standard deviation of \$100. Find the probability that a random sample of 36 graduates averages less than \$570.	<p>Solution: Let X = weekly income of a sampled graduate 1 year after graduation. We are asked to find $P(\bar{X} < \\$570)$ for 36 graduates.</p> $P(\bar{X} < \$570) = P\left(\frac{\bar{X} - \$600}{\$100/\sqrt{36}} < \frac{\$570 - \$600}{\$100/\sqrt{36}}\right) \cong P(Z < -1.8) = 0.0359$ <p>Note: Without the Central Limit Theorem we could not have approximated the probability that a sample of graduates average less than \$570 because the distribution of incomes is not usually normal.</p>
----------------	--	--

Statistical Inference: Estimation

Point Estimate	A single number used to estimate a parameter . For example, the sample mean \bar{x} is typically used to estimate the population mean μ .
Interval Estimate	A range of values used as an estimate of a population parameter. The width of the interval provides a sense of the accuracy of the point estimate.

Confidence Interval Estimates for μ

Confidence intervals for μ have a characteristic format: $\bar{x} \pm CV * \text{standard error}$, where CV stands for Critical Value and the standard error is the (usually estimated) standard deviation of \bar{X} .

Case I: X normal or $n \geq 30$, and σ is known	A $(1 - \alpha)*100\%$ confidence interval estimate for μ is given by $\bar{x} \pm z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$
Case II: $n \geq 30$ and σ is unknown	A $(1 - \alpha)*100\%$ confidence interval estimate for μ is given by $\bar{x} \pm t_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right), \text{ with } n-1 \text{ degrees of freedom}$
Case III: X is normal and σ is unknown	A $(1 - \alpha)*100\%$ confidence interval estimate for μ is given by $\bar{x} \pm t_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right), \text{ with } n-1 \text{ degrees of freedom}$

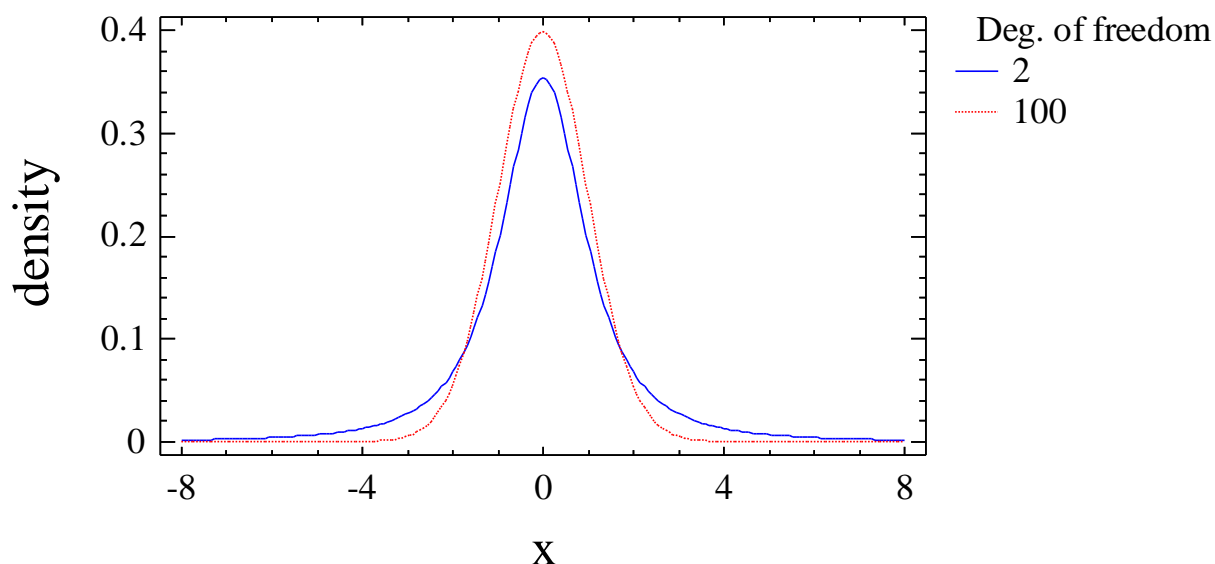
Case III requires some explanation. When X is normal, and we must use the sample standard deviation s to estimate the unknown population standard deviation σ , the **studentized** statistic

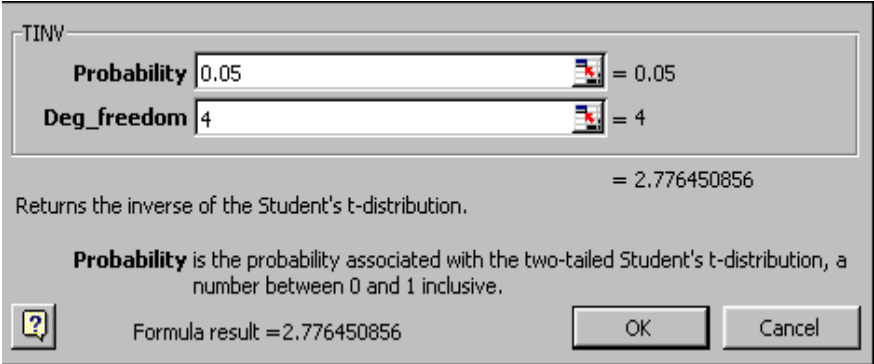
$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$ has a t distribution with $n-1$ degrees of freedom. Hence, we must use the critical value

$t_{\alpha/2}$ from the t distribution with $n-1$ degrees of freedom. The properties of the t distributions are similar to those for the **standard normal** distribution Z , except that the t has a larger spread to reflect the added uncertainty involved in estimating σ by s .

Note: For large samples, where $n \geq 30$, there is very little difference between the t distribution with $n-1$ degrees of freedom and the standard normal distribution Z . Therefore, for large samples (**Case II** in the table above) some texts replace $t_{\alpha/2}$ with $z_{\alpha/2}$ even when σ is unknown!

Student's t Distribution



<p><u>Example</u></p>	<p>A manufacturer wants to estimate the average life of an expensive component. Because the components are destroyed in the process, only 5 components are tested. The lifetimes (in hours) of the 5 randomly selected components are 92, 110, 115, 103, and 98. Assuming that component lifetimes are normal, construct a 95% confidence interval estimate of the component's life expectancy.</p>	<p>Solution: Using Excel, $\bar{x} = 103.6$ hours, and $s = 9.18$ hours. From the discussion above, the critical value is $t_{0.025} = 2.776$. (Note: In Excel, shown below, to find the critical value associated with the t distribution and significance level α, follow: $f_x > \text{Statistical} > \text{TINV}$ and enter α in the <i>Probability</i> field.)</p>  <p>Thus a 95% CIE for the mean lifetime of the components is given by</p> $103.6 \pm 2.776 \left(\frac{9.18}{\sqrt{5}} \right) \text{ or } (92.2, 115.0) \text{ hours}$
-----------------------	---	--

Statistical Inference: Decision Making

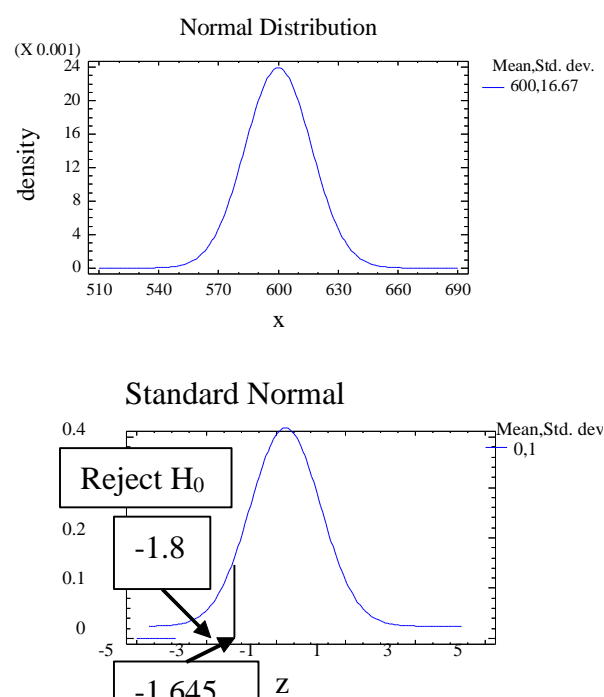
In hypothesis testing we are asked to evaluate a claim about something, such as a claim about a population mean. For instance, in a previous example a Business dean claimed that the average weekly income of graduates of his school one year after graduation is \$600. Suppose that you suspect the dean's claim may be exaggerated. Hypothesis testing provides a systematic framework, grounded in probability, for evaluating the dean's claim against your suspicions.

Although hypothesis testing uses probability distributions to arrive at a reasonable (and defensible) decision either to reject or "fail to reject" the claim associated with the null hypothesis of the test, H_0 , it does *not* guarantee that the decision is correct! The table below outlines the possible outcomes of a hypothesis test. (**Note:** We avoid "accepting" the null hypothesis for the same reason juries return verdicts of "not guilty" rather than "innocent")

TRUTH		
Decision	H_0 True	H_0 False
Reject H_0	Type I error	correct decision
Fail to Reject H_0	correct decision	Type II error

<p>Type I error</p>	<p>The error of incorrectly rejecting H_0 when, in fact, it's true. In a hypothesis test conducted at the significance level α, the probability of making a type I error, if H_0 is true, is at most α.</p>
<p>Type II error</p>	<p>The error of incorrectly failing to reject H_0 when, in fact, it's false. For a fixed sample size n, you cannot <i>simultaneously</i> reduce the probability of making a Type I error and the probability of making a Type II error. (This is the statistician's version of "there is no such thing as a free lunch.") However, if you can afford to take a larger sample, it is possible to reduce both probabilities.</p>

Decision Making: Hypothesis Testing

<u>Example</u>	Suppose that a sample of 36 graduates of the business school averaged \$570 per week one year after graduation. Test the dean's claim, against your suspicion, at the 5% level of significance.
	<p>Solution:</p> <ol style="list-style-type: none"> $H_0: \mu = \\$600$ (the dean's claim) $H_A: \mu < \\$600$ (your suspicion) $\alpha = 0.05$ (the probability of rejecting the dean's claim if she's right) Draw some pictures (see box to the right) Critical Value: $-z_{0.05} = -1.645$ From the sample - Standardized Test Statistic: $z = \frac{570 - 600}{100/\sqrt{36}} = -1.8$ Conclusion: There is sufficient evidence to reject the dean's claim at the 5% level of significance. <div style="text-align: right;">  </div>

the *P*-value Approach to Hypothesis Testing

<i>P</i> -value	<p>The smallest significance level at which you would reject H_0. The <i>p</i>-value is calculated from the test statistic, and is doubled for two-sided tests.</p> <p>Note: α and the <i>p</i>-value are the “before” and “after” significance levels for the test. We can reach a decision whether to reject H_0 by comparing the two significance levels.</p> <p>Rule: If the <i>p</i>-value $> \alpha$, then we "fail to reject" H_0 If the <i>p</i>-value $\leq \alpha$, then we reject H_0, i.e., we reject H_0 for <i>small</i> <i>p</i>-values</p>
-----------------	---

<u>Example</u>	Suppose that a sample of 36 graduates of the business school averaged \$570 per week one year after graduation. Use the <i>p</i> -value to test the dean's claim, against your suspicion, at the 5% level of significance.
	<p>Solution:</p> <p>Steps 1-3 are the same as before.</p> <p>4. Critical Values are not used in this approach.</p> <p>5. From the sample - Standardized Test Statistic: $z = \frac{570 - 600}{100/\sqrt{36}} = -1.8$ <i>p</i>-value = $P(Z < -1.8) = 0.0359 < 0.05 = \alpha$, where we have used the fact that the test is left-tailed!</p> <p>6. Conclusion: There is sufficient evidence to reject the dean's claim at the 5% level of significance.</p>

Notice that we rejected the Dean's claim under both the critical value and p -value approaches. This was not a coincidence: the two approaches *always* lead to the same decision. Since p -values are routinely computed by software such as Excel, we will usually use p -values to conduct significance tests.

Note: Many of the (hypothesis) tests conducted in this course are two-sided, and assume that we are sampling from a normal population with unknown variance. When this is the case, software will automatically return the correct p -value for the two-sided t test.