

## 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.**

<b>Descriptive Statistics</b>	Sample data may be summarized graphically or with summary statistics. Sample statistics include the <b>mean, variance, standard deviation</b> , and median. For the following definitions let $x_1, x_2, \dots, x_n$ represent the values obtaining from a random sample of size $n$ drawn from a <b>population</b> 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 the average, while a large variance means that they 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 sampled values and $\bar{x}$ .
---------------------------	---	---

<b>StatGraphics</b>	Common descriptive statistics can be obtained by following: <b>Describe</b> > Numeric Data > One-Variable Analysis > <b>Tabular Options</b> > <i>Summary Statistics</i>
---------------------	---

<u>Example 1</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 <b>number</b> (sales volume, rate of return, test score, etc.) and that the sample selection process generates the numbers randomly, i.e., by a “<b>random</b>” selection.</p> <p>(In these notes, a <b>random variable</b> will be designated by a capital letter, such as <math>X</math>, to differentiate it from observed values <math>x</math>. For instance, <math>X</math> might represent the height of a man to be selected randomly. Once the man has been selected, his height is given by the <b>value</b> <math>x</math>, say <math>x = 68</math> 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  $\mu$  and standard deviation  $\sigma$ , 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.

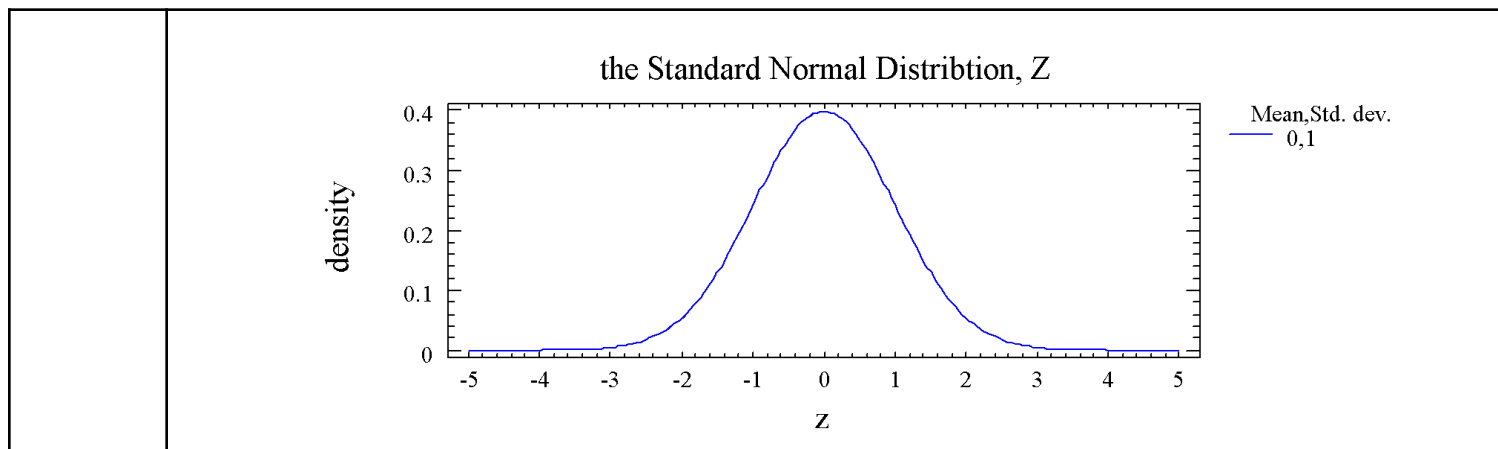
### Standardized Random Variables

For the random variable  $X$  with mean  $\mu$  and standard deviation  $\sigma$ ,  $Z = \frac{X - \mu}{\sigma}$  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-Square. 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**.

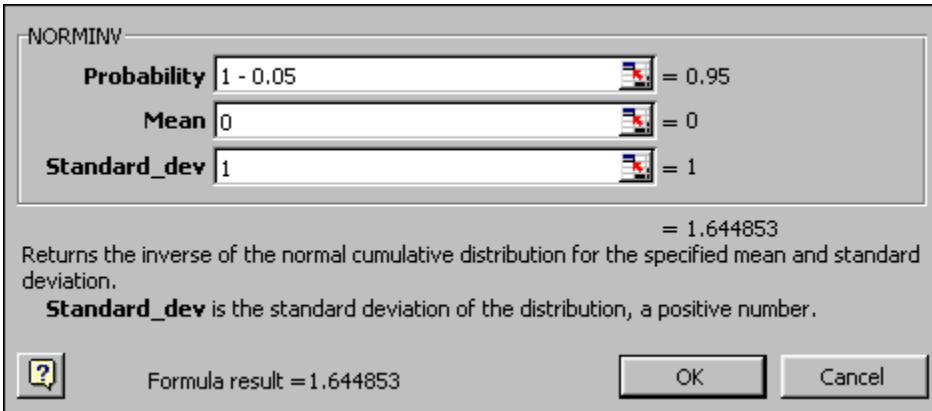
Properties	<ol style="list-style-type: none"> <li>Normal distributions are bell-shaped. (In fact, it is sometimes called the “Bell Curve”.)</li> <li>Normal distributions are symmetric about their mean.</li> <li>Normal distributions follow the 68-95-99.7 rule: <ul style="list-style-type: none"> <li>(Approximately)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



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: $f_x > Statistical > NORMDIST$ 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 <math>P(-1.2 &lt; Z &lt; 2)</math>, note that <math>P(-1.2 &lt; Z &lt; 2) = P(Z &lt; 2) - P(Z \leq -1.2)</math> and use the Excel output to the right.</p> <p>Answer = <math>0.9772 - 0.1151</math> = <b>0.8621</b></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	<p><math>z_\alpha</math> is defined by <math>P(Z &gt; z_\alpha) = \alpha</math>. Critical values are used in the construction of <b>confidence intervals</b> and (optionally) in <b>hypotheses testing</b>. To find the critical value associated with the significance level <math>\alpha</math>, follow: <math>f_x &gt; Statistical &gt; NORMINV</math> and enter <math>1 - \alpha</math> in the <i>Probability</i> field.</p>
-----------------	--

<u>Example</u>	From the Excel output to the right we see that $z_{0.05} = 1.645$	
----------------	---	--

### 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  $\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$  has a probability distribution of its

own. If we intend to use the sample mean  $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$  to estimate the mean  $\mu$  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 (using advanced calculus) 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  $\mu$ .
- 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 size of the sample *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  $\mu$ ) *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 non-normal populations and small samples. 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  $n > 30$ .)

<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 <math>X</math> = weekly income of a sampled graduate 1 year after graduation. We are asked to find <math>P(\bar{X} &lt; \\$570)</math> 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 <b>Central Limit Theorem</b> 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 <b>parameter</b> . For example, the sample mean $\bar{x}$ is typically used to estimate the population mean $\mu$ .
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 $\mu$

Confidence intervals for  $\mu$  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  $\bar{X}$ .

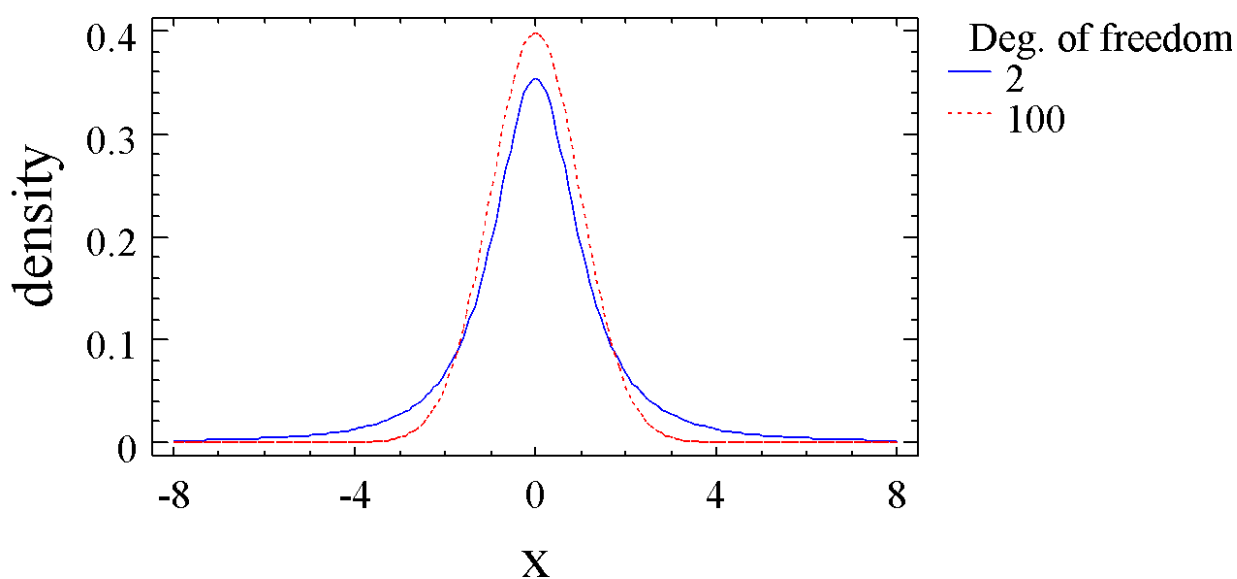
Case I: $X$ normal or $n > 30$ , and $\sigma$ is known	<p>A <math>(1 - \alpha) * 100\%</math> confidence interval estimate for <math>\mu</math> is given by</p> $\bar{x} \pm z_{\alpha/2} \left( \frac{\sigma}{\sqrt{n}} \right)$
Case II: $n \geq 30$ and $\sigma$ is unknown	<p>A <math>(1 - \alpha) * 100\%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

**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  $\sigma$ , the **studentized** statistic

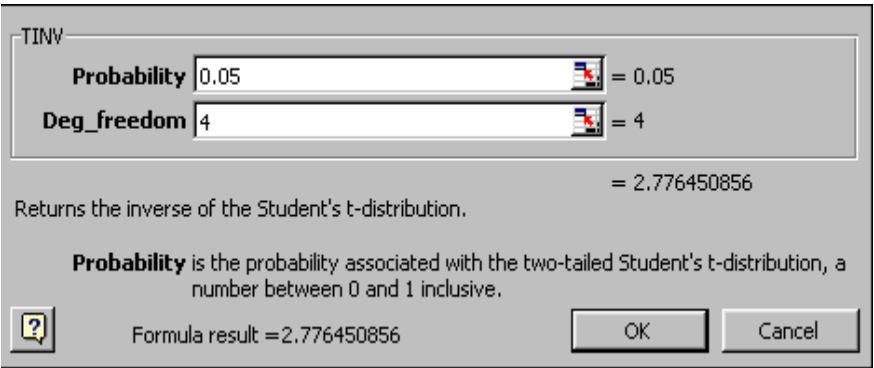
$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$  has a  $t$  distribution with  $n-1$  degrees of freedom. Hence, we must use the critical value  $t_{\alpha/2}$  from the  $t$  distribution with  $n-1$  degrees of freedom. The properties of the  $t$  distributions are similar to those for the **standard normal** distribution  $Z$ , except that the  $t$  has a larger spread to reflect the added uncertainty involved in estimating  $\sigma$  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  $X$  is normal and  $\sigma$  is unknown!

## Student's t Distribution



<u>Example</u>	<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</p>	<p>Solution: Using Excel, <math>\bar{x} = 103.6</math> hours, and <math>s = 9.18</math> hours. From the discussion above, the critical value is <math>t_{0.025} = 2.776</math>. (Note: In Excel, shown below, to find the critical value associated with the <math>t</math> distribution and significance level <math>\alpha</math>, follow: <math>f_x &gt; Statistical &gt; TINV</math> and enter <math>\alpha</math> in the <i>Probability</i> field.)</p>
----------------	---	--

	of the component's life expectancy.	 <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 of "innocent")

Decision:		
TRUTH	Accept $H_0$	Reject $H_0$
$H_0$ True	correct decision	Type I error
$H_0$ False	Type II error	correct decision

Type I error	The error of incorrectly rejecting $H_0$ when, in fact, it's true. In a hypothesis test conducted at the significance level $\alpha$ , the probability of making a type I error, if $H_0$ is true, is at most $\alpha$ .
Type II error	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.

### Decision Making: Hypothesis Testing

<b>Example</b>	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><b>Solution:</b></p> <ol style="list-style-type: none"> <li><math>H_0: \mu = \\$600</math> (the dean's claim)</li> <li><math>H_A: \mu &lt; \\$600</math> (your suspicion)</li> <li><math>\alpha = 0.05</math> (the probability of rejecting the dean's claim if she's right)</li> <li>Draw some pictures (see box to the right)</li> <li>Critical Value: <math>-z_{0.05} = -1.645</math></li> <li>From the sample - Standardized Test Statistic:  <math display="block">z = \frac{570 - 600}{100/\sqrt{36}} = -1.8</math> </li> <li>Conclusion: There is sufficient evidence to reject the dean's claim at the 5% level of significance.</li> </ol>	

### the P-value Approach to Hypothesis Testing

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

<b>Example</b>	Suppose that a sample of 36 graduates of the business school averaged \$570 per week one year after graduation. Use the p-value to test the dean's claim, against your suspicion, at the 5% level of significance.
<p><b>Solution:</b></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:  <math display="block">z = \frac{570 - 600}{100/\sqrt{36}} = -1.8</math> </p> <p>p-value = <math>P(Z &lt; -1.8) = 0.0359 &lt; 0.05 = \alpha</math>, 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 StatGraphics and 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, Statgraphics will automatically return the correct p-value for the two-sided *t* test.

---

### Example 1 (p.1)

Example 1. 提供了一組有關26個共同基金的三年回報率的統計數據。其逐項統計數據的內容與意義如下。

#### 1. 計數 (Count): 26

意義: 這表示樣本中共有26個共同基金的回報率數據。計數是進行統計分析的基礎, 樣本越大, 結果的可靠性通常越高。

#### 2. 平均值 (Average): 16.2346

意義: 這是所有26個共同基金三年回報率的算術平均值。它反映了這些基金在三年內的整體表現。平均值越高, 表示基金的回報潛力越好。

#### 3. 方差 (Variance): 40.4208

意義: 方差是用來衡量數據分散程度的指標。方差越大, 表示回報率的波動性越大。這意味著基金的回報率在不同基金之間差異較大。

#### 4. 標準差 (Standard Deviation): 6.35773

意義: 標準差是方差的平方根, 表示數據點與平均值之間的平均距離。這裡的標準差約為6.36, 顯示基金回報率的波動性。較高的標準差意味著回報率有較大的變化範圍。

#### 5. 最小值 (Minimum): 8.0

意義: 這是樣本中最低的回報率, 表示在這26個基金中, 最低的三年回報為8%。這可以幫助投資者了解最差的投資表現。

#### 6. 最大值 (Maximum): 32.7

意義: 這是樣本中最高的回報率, 表示在這26個基金中, 最高的三年回報為32.7%。這提供了投資者對於最佳投資表現的參考。

#### 7. 標準偏態 (Standard Skewness): 2.26003

意義: 標準偏態用來衡量數據分布的對稱性。正偏態(值大於0)表示數據的右側尾部較長, 意味著有較多的高回報率基金, 可能會吸引投資者的注意。

#### 8. 標準峰度 (Standard Kurtosis): 1.1129

意義: 標準峰度用來衡量數據分布的尖峭程度。值大於3表示數據分布比正態分布更尖峭, 值小於3則表示分布較平坦。這裡的值為1.1129, 顯示出這組數據的分布相對較平坦, 可能意味著回報率的極端值(非常高或非常低)較少。

#### 總結

這組統計數據提供了對26個共同基金回報率的全面了解。從平均值和標準差可以看出這些基金的整體表現和波動性, 而最小值和最大值則幫助投資者評估潛在的風險和回報。偏態和峰度的分析則進一步揭示了數據分布的特徵, 對於投資決策非常重要。

