Chapter 2
Review of Basic Statistical Concepts

Terry Dielman

Applied Regression Analysis:
A Second Course in Business and Economic Statistics, fourth edition

2.1 Introduction

- Typically a problem in statistics is one of studying a particular *population*, perhaps all firms in an industry or the lifetime of a certain brand of tire.
- In most cases it is not possible to examine the entire population, so we work with a subset of the population called a *sample*.
- This study has two phases. In the first, *descriptive statistics* are used to explore the data.
- In the second, *inferential statistics* are used to generalize from the sample to the population.

Sampling and Statistics

- The most common type of sampling is *simple random sampling* where every item in the population is equally likely to be selected.
- Any numerical summary from a sample is a *statistic* and each one has a different *sampling distribution* that describes its theoretical behavior.
- This theoretical behavior provides the guidelines for the inference process.

2.2 Descriptive Statistics

- Textbook Table 2.1 shows the 5-year returns as of July 2002 for a random sample of 83 mutual funds.
- Just looking at a list of numbers like this provides little useful information.
- The field of descriptive statistics can provide several ways to meaningfully summarize such lists, even when there are far more data.

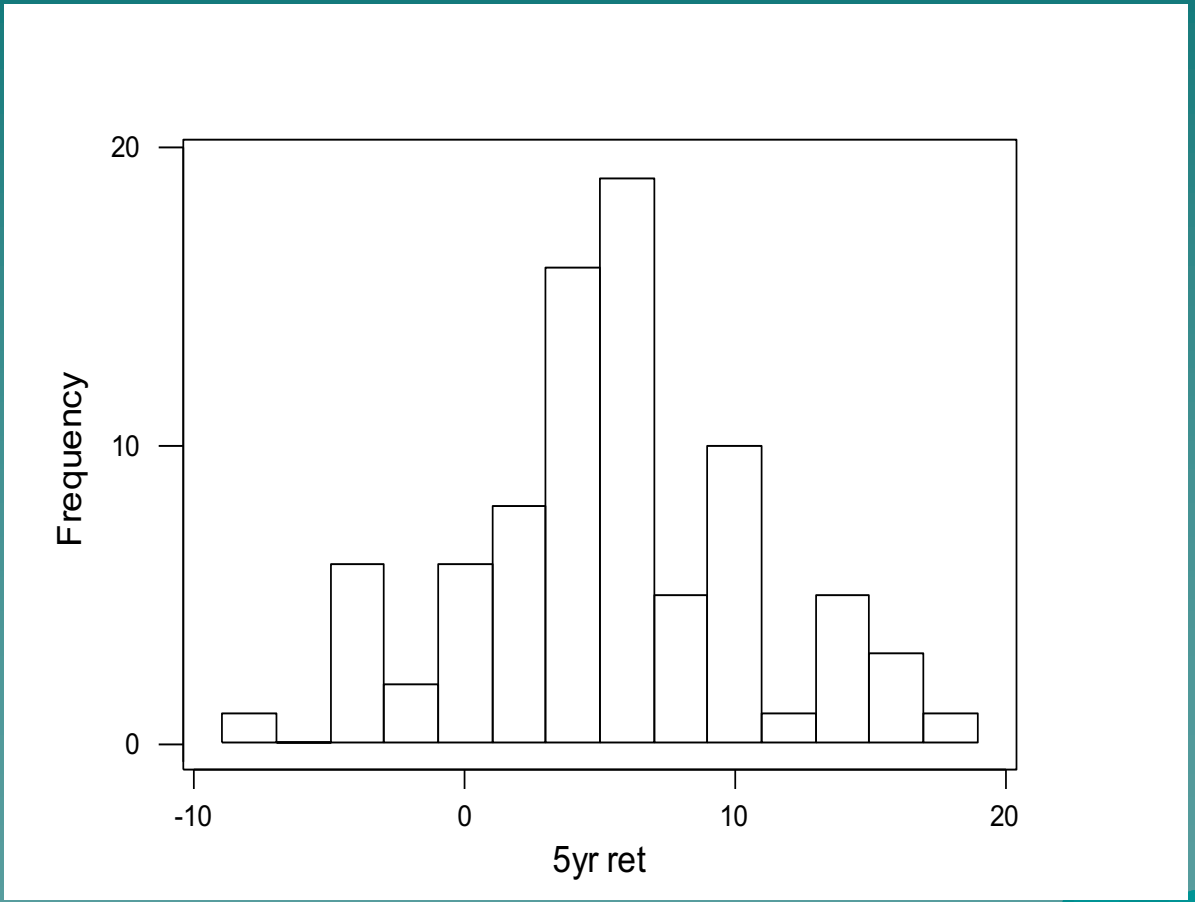
Frequency Distributions

- The table at right is constructed by breaking the return rates down into 7 categories of equal width.
- Each data value falls into a unique bin because rates are to nearest .1%

5-year rates of return	# of funds
-8% to -4.01%	5
-4% to -0.01%	6
0% to 3.99%	17
4% to 7.99%	34
8% to 11.99%	12
12% to 15.99%	8
16% to 19.99%	1

Histograms

- A graphical method to display a frequency distribution.
- You can get a quick look of the data's symmetry and spread.



Numerical Summaries

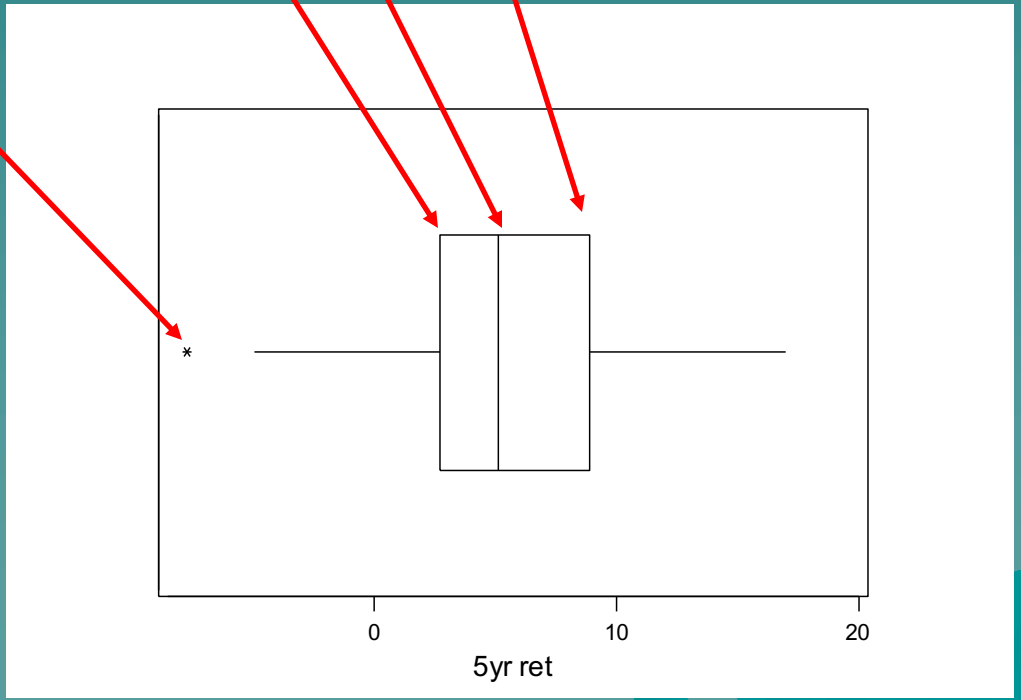
- These are single numbers computed from the sample to describe some characteristic of the data set.
- Measures of location include the mean, median and the first and third quartiles.
- Measures of spread include the standard deviation, range and midrange.

Descriptive Statistics in Minitab

Variable	N	Mean	Median	TrMean	StDev	SE Mean
5yr ret	83	5.371	5.100	5.391	5.229	0.574

Variable	Minimum	Maximum	Q1	Q3
5yr ret	-7.800	17.000	2.700	8.900

A boxplot is a good way to display many of the summary stats.



2.3 Discrete Random Variables and Probability Distributions

- A random variable is a rule that assigns a number to every possible outcome of an experiment.
- A discrete random variable is one with a definite distance between each of its possible values.
- For example, the number of heads in a single coin toss or the number of kings in two draws from a deck.

Probability Distributions

- In each experiment, the outcome is determined by chance.
- We can assign probabilities to these outcomes and thus to the values of the random variable.
- A probability distribution is a list of all the possible values and their associated probabilities.

Probability distribution for coin toss

Let X = # heads on a single coin toss

P(x) = the probability that the random variable X takes on value x

x	P(x)
0	.5
1	.5

Probability Distribution for 2-card Draw

Take Y = # kings in the 2-card draw

y	$P(y)$
0	$188/221 = .8507$
1	$32/221 = .1448$
2	$1/221 = .0045$

Where did these come from?

- With a 52-card deck, there are 52*51 = 2652 possible two card draws.
- There are 48 cards other than Kings, which is 48*47 = 2256 draws
- P(0 Kings) = 2256/2652 = 188/221

Laws of Probability

There are two conditions that probabilities must satisfy:

1. $0 \leq P(x) \leq 1$
2. Sum of all $P(x) = 1$

This says that each probability must be between 0 and 1, and the total probability in the distribution must add up to 1.

Numerical Summaries

- When we discussed a sample of observations, we discussed certain numerical summaries.
- The ones most often used were the sample mean and standard deviation.
- We have similar measures here that describe certain features of the distribution.

Profit From Two Investments

x	P(x)
-2000	.05
-1000	.10
1000	.10
2000	.25
5000	.50

y	P(y)
0	.40
1000	.20
2000	.20
3000	.10
4000	.10

Expected Value

- The expected value of a discrete random variable is:

$$E(X) = \mu_X = \sum xP(x)$$

- This is a weighted average of the values of X, where the weights are the probabilities.

Expected Profit From Two Investments

x	P(x)	$xP(x)$
-2000	.05	-100
-1000	.10	-100
1000	.10	100
2000	.25	500
5000	.50	2500
	E(X)=	2900

y	P(y)	$yP(y)$
0	.40	0
1000	.20	200
2000	.20	400
3000	.10	300
4000	.10	400
	E(Y)=	1300

An Interpretation

- The expected return of investment X is 2900 but this value will never occur. The probability of getting this value is 0, so we can't "expect" it to happen.
- A better interpretation of this number is that if we could do this investment many times, this would be the average of all those returns.

Investment Risk

- Although X clearly has the higher expected return, note that 15% of the time it loses money.
- Some people may thus prefer Y because it never loses.
- The variance of the returns is sometimes used as a measure of risk.

Variance

- The variance of a discrete random variable is:

$$Var(X) = \sigma_X^2 = \left(x - \right)^2 P(x)$$

- Because this is in squared units ($ squared here), we often work with the square root, σ_X.

Variance of X

x	$P(x)$	$x-\mu_x$	$(x-\mu_x)^2$	$(x-\mu_x)^2P(x)$
-2000	.05	-4900	24,010,000	1,200,500
-1000	.10	-3900	15,210,000	1,521,000
1000	.10	-1900	3,610,000	361,000
2000	.25	-900	810,000	202,500
5000	.50	2100	4,410,000	2,205,000
			$\sigma_X^2 =$	5,490,000

X Has More Risk

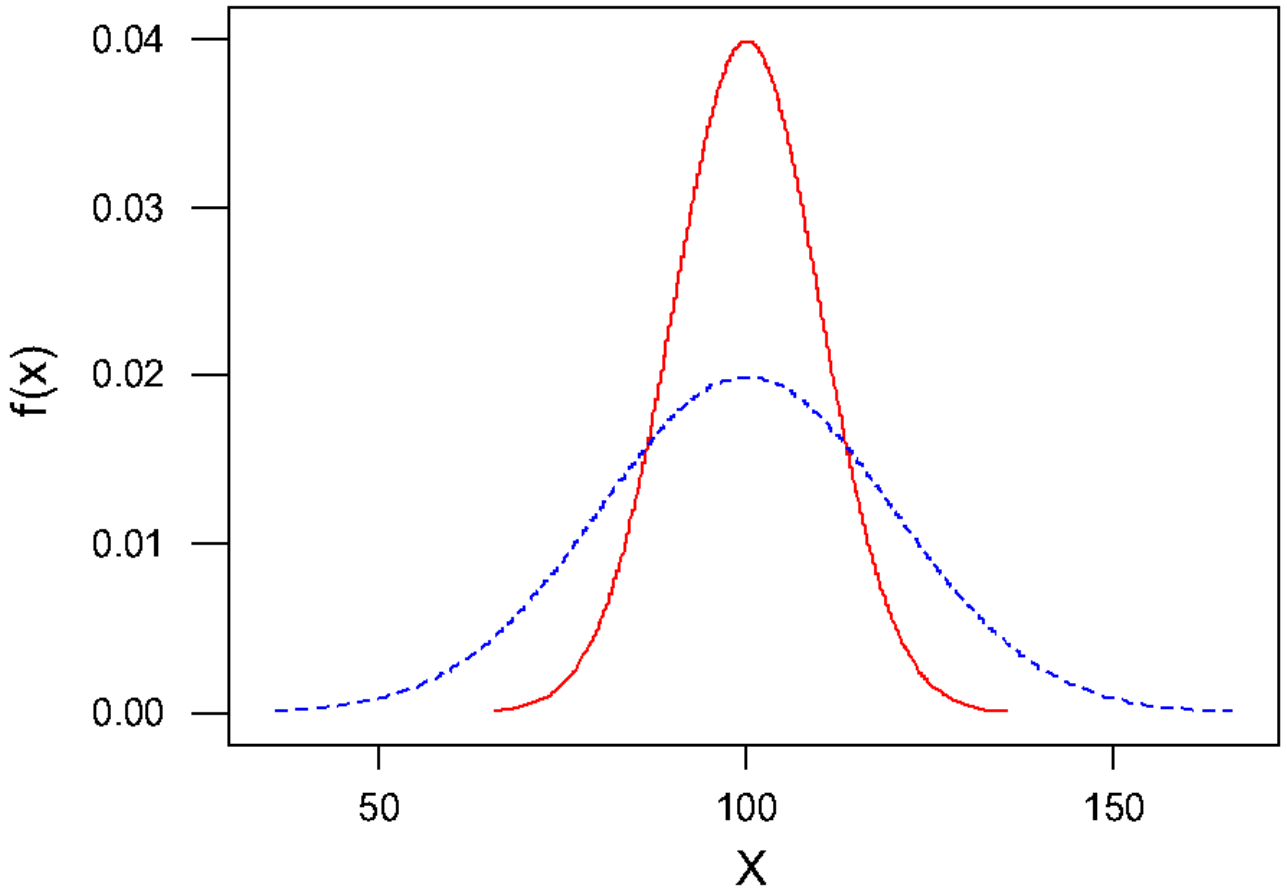
- We calculate that $\sigma_X^2 = 5{,}490{,}000$.
- A similar calculation for the returns on *Y* yields $\sigma_Y^2 = 1{,}810{,}000$ so *X* both returns and varies more.
- The two standard deviations are $\sigma_X = 2343.07$ and $\sigma_Y = 1345.36$

2.4 The Normal Distribution

- A continuous random variable can take any value over a given range.
- The most important one is no doubt the normal random variable whose probability distribution is often depicted as bell-shaped.
- It is centered at μ and most of its probability is within 3σ of the mean.

Two Normal Distributions:

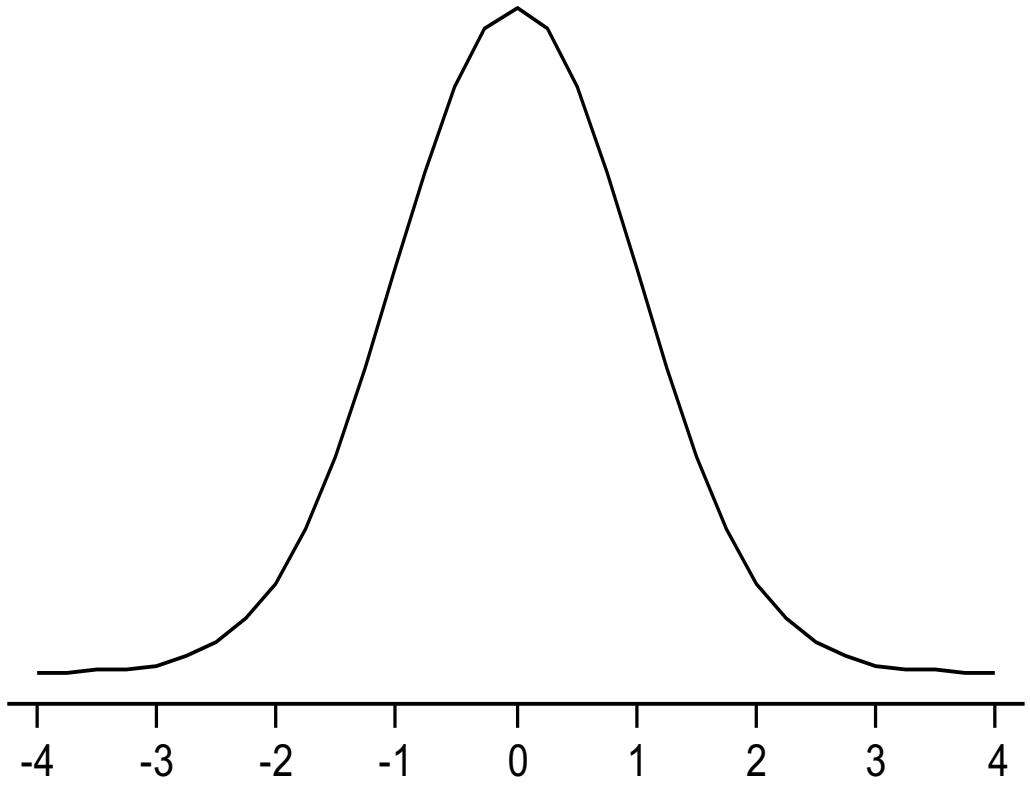
— $\mu = 100, \sigma = 10$ ---- $\mu = 100, \sigma = 20$



Calculating Normal Probabilities

- For any continuous variable, probabilities are areas under the probability curve.
- For the normal, these computations have been tabulated (Table B.1).
- Because each combination of μ and σ defines a different distribution, the table shows the "standard normal" with mean 0 and standard deviation 1.

The Standard Normal Distribution



Units of measurement are standard deviations above/below the mean

Standardizing

- To use the tables with any normally-distributed variable, we convert with the standardizing transformation:

$$Z = (X - \mu_X)/\sigma_X$$

- For example, if X has mean 10 and standard deviation 2:

$$Z = (X - 10)/2$$

Reverse Standardizing

- Sometimes we have the reverse problem, where we know a probability and want to find the *X* value associated with it.
- The probability yields the Z-value, then we get *X* by:

$$X = \mu_X + Z\ \sigma_X$$

Example 2.2

- A large retail firm has accounts receivable that are assumed to be normal with mean $\mu = \$281$ and standard deviation $\sigma = \$35$.

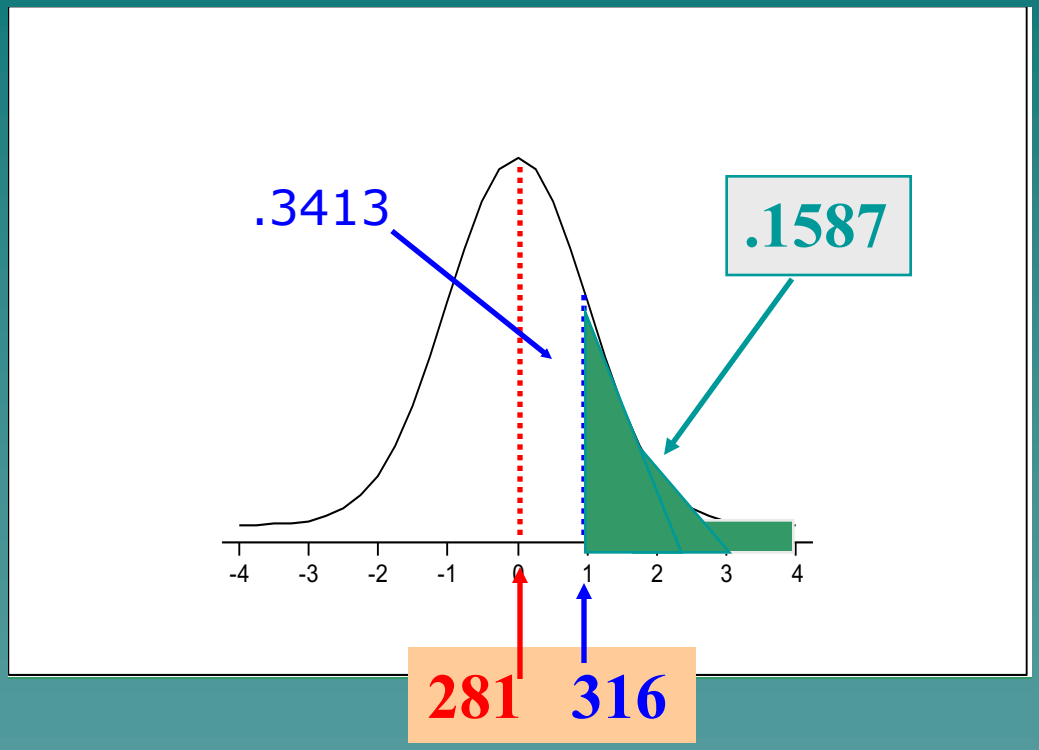
- What proportion of accounts are above $316?
- Above what value do 13.57% of the accounts lie?

Computations for Above 316

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

$$= (316 - 281)/35$$

$$= 35/35 = 1.00$$



$$P(Z > 1.00) = .5 - .3413 = .1587$$

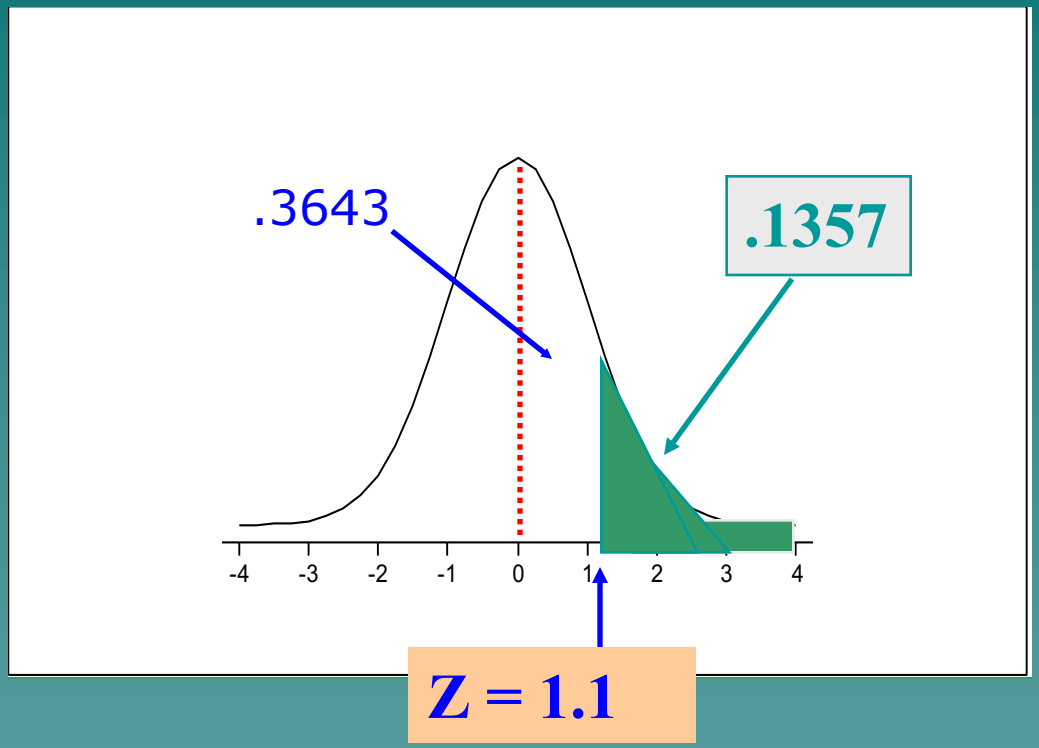
Where are the upper 13.57%?

$$X = \mu + Z\sigma$$

$$= 281 + 1.1(35)$$

$$= 281 + 38.5$$

$$= 319.50$$



2.5 Populations, Samples and Sampling Distributions

- A statistic uses its own sampling distribution to make inference about the population.
- For example, $\bar{y}$ is the statistic most often used to make inference about the population mean μ.
- The sample mean is a random variable thus follows a probability distribution.

The Sampling Distribution of $\bar{y}$

- Assume the random variable *Y* has mean μ and standard deviation σ.

- We observe a random sample of size *n* on *Y* and compute the sample average. The expected value of $\bar{y}$ is μ and the standard deviation of $\bar{y}$ is:

$$_{\bar{y}} = \quad _{Y} / \sqrt{n}$$

Distributional Form

- If the observations come from a normal distribution, the sampling distribution is also normal.
- The Central Limit Theorem states that the distribution is approximately normal as long as the sample size *n* is large (usually 30 or more).

Example 2.3

- In a manufacturing process, the diameter of a certain part averages 40 cm. The variation appears to be normally distributed with standard deviation .2 cm.
- If a sample of 16 parts is chosen, what is the probability that the average diameter is greater than 40.1 cm?

Distribution of average diameter

- For *n=16*, the standard error is

$$\sigma_{\bar{y}} = \sigma_Y / \sqrt{n} = 0.2 / \sqrt{16} = .05$$

- Thus,

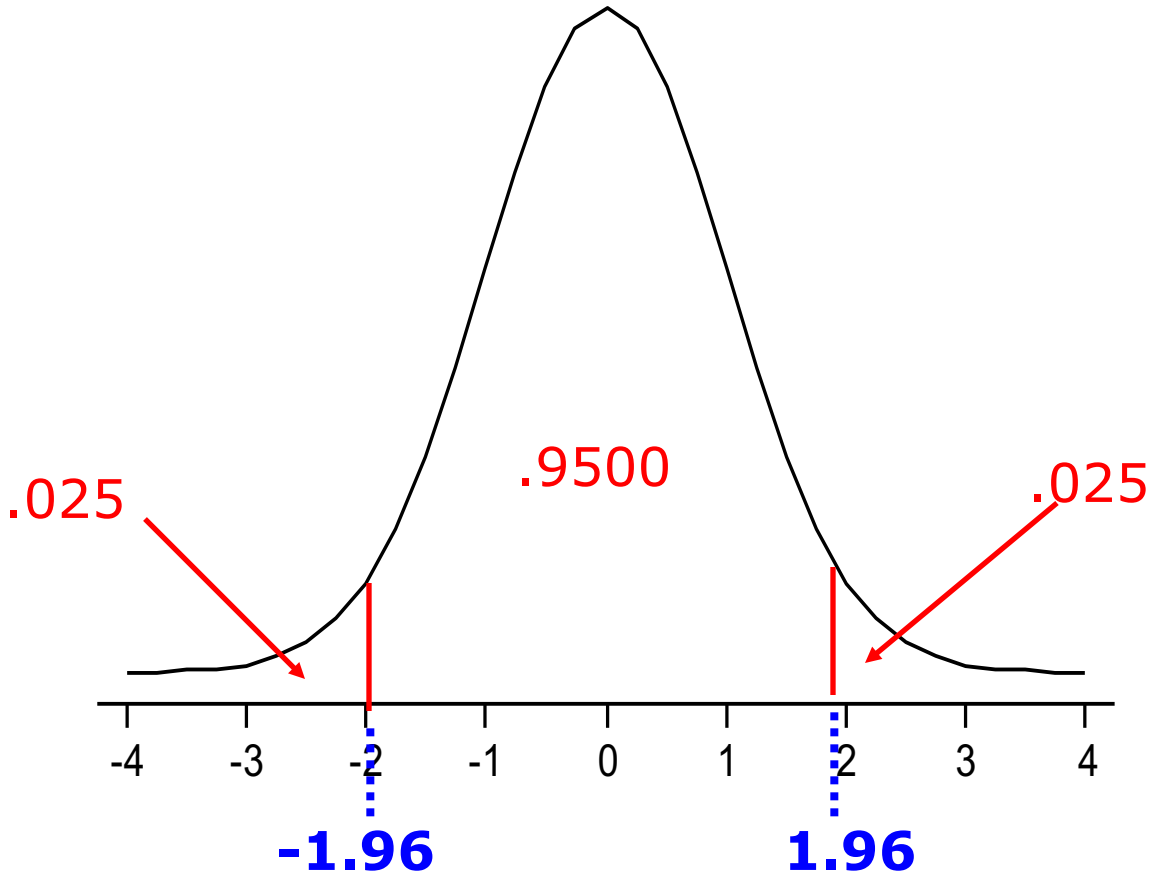
$$P(\bar{y} > 40.1) = P\left(Z > \frac{40.1 - 40}{.05}\right) =$$

$$P(Z > 2) = .5 - .4772 = .0228$$

2.6 Estimating a Population Mean

- Point estimates are single numbers used as an estimate of a population parameter.
- In general, these will never be exactly right so we use them as the basis for an interval estimate.
- Because we base the interval on the sampling distribution, we know the probability content of the interval.

The 95% Interval



The General Interval

In general, let α denote the total tail probability. A 100(1- α)% confidence interval estimate of μ is:

$$\bar{y} \quad z_{\alpha/2}\frac{}{\sqrt{n}}, \bar{y} + z_{\alpha/2}\frac{}{\sqrt{n}}$$

where $z_{\alpha/2}$ is the standard normal value that has tail probability $\alpha/2$.

Example 2.5

- In a department store, the charge account balances of customers has a standard deviation of $45.
- They will take a sample of 100 accounts and want to make a 90% interval estimate for all accounts.
- If the sample mean is $245, what is the interval estimate?

Interval estimate

For a 90% interval, the multiplier is *z=1.65.* The interval is:

$$\bar{y} \quad 1.65\frac{}{\sqrt{n}}, \bar{y} + 1.65\frac{}{\sqrt{n}}$$

$$= \quad 245 \quad 1.65\frac{45}{\sqrt{100}}, 245 + 1.65\frac{45}{\sqrt{100}}$$

or 245 ± 7.42 = ($237.58 to $252.42)

When σ Is Unknown

- The previous result required us to know the population standard deviation σ.
- In most cases, this is not known and we estimate it by the sample standard deviation, thus have an estimated standard error:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

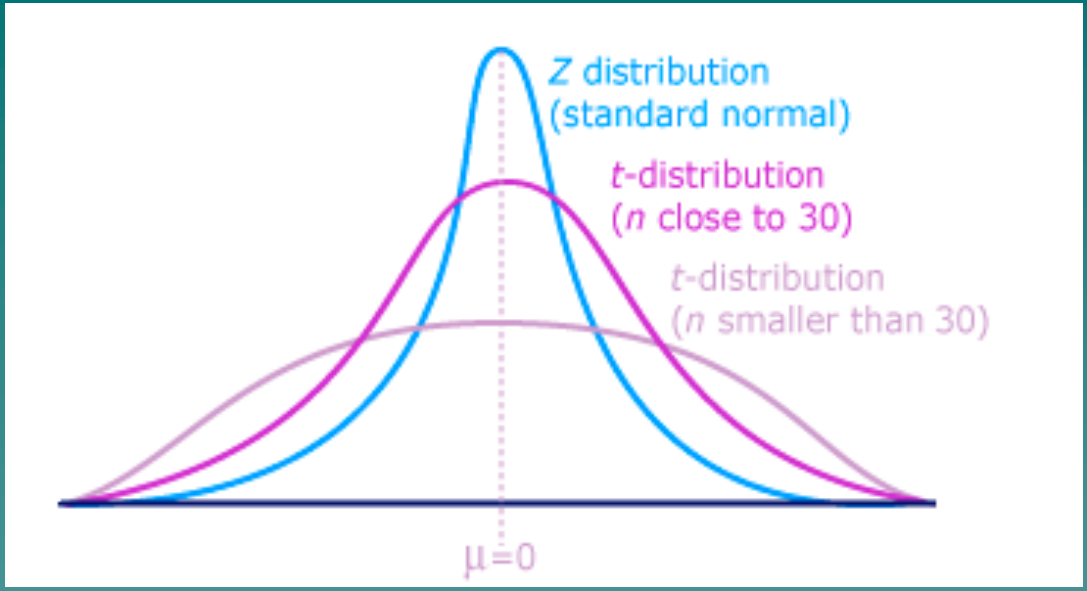
t Distribution Interval

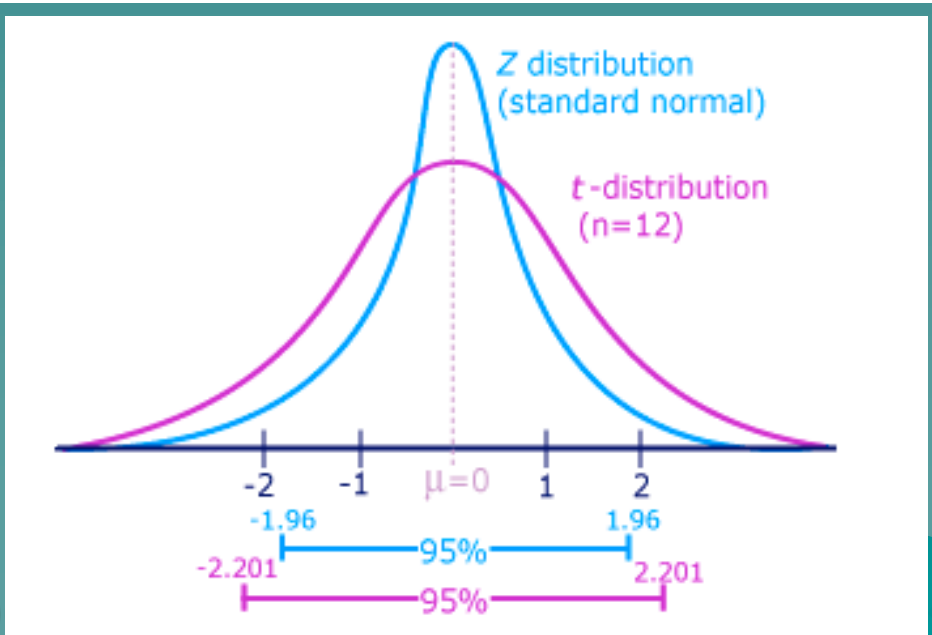
- Statistical theory tells us that our interval estimator is now:

$$\bar{y} \pm t_{/2,n\ 1} \frac{s}{\sqrt{n}}$$

- where $t_{\alpha/2,n-1}$ is a multiplier from the *t* distribution with *n-1* degrees of freedom.

t-distribution vs. Normal distribution





Example 2.6

- A manufacturer wants to estimate the average life on an expensive electrical component.
- Because the test destroys the component, a small sample is used.
- If the test results are 92, 110, 115, 103 and 98, find a 95% interval estimate of the average lifetime.

Computations

- We get $\bar{y} = 103.6$ and $s = 9.18$
- For $n=5$, we use the t distribution with 4 degrees of freedom. The multiplier is $t_{.025,4} = 2.776$.
- The interval is:

  $103.6 \pm 2.776(9.18/\sqrt{5})$

  103.6 ± 11.40

  92.2 to 115.0 hours

2.7 Hypothesis Tests About a Population Mean

- In the previous section we discussed estimating an unknown parameter.
- Here we use the sample data to test a preconceived belief about the value of a parameter.
- We state this belief in a hypothesis, thus the procedure is called hypothesis testing.

An Example

Is the population average 10?

$H : = 10$ (On average, mean is 10)

$H_a : \quad 10$ (No it isn't)

Notation and Terminology

H_0 is called the Null Hypothesis

H_a is called the Alternative or Research Hypothesis

We will set up a decision rule to determine whether we accept or reject the null hypothesis. This rule is constructed after we choose the test's level of significance.

A test statistic will be computed from the sample information to make the decision.

Level of Significance

- If we perform the test and the null hypothesis really is correct, there is a chance we will say it is false because we happened to get some fairly extreme values in our sample.
- We control for this by setting up the decision rule so there is a small probability of this happening.
- This probability is the level of significance.

Continuing

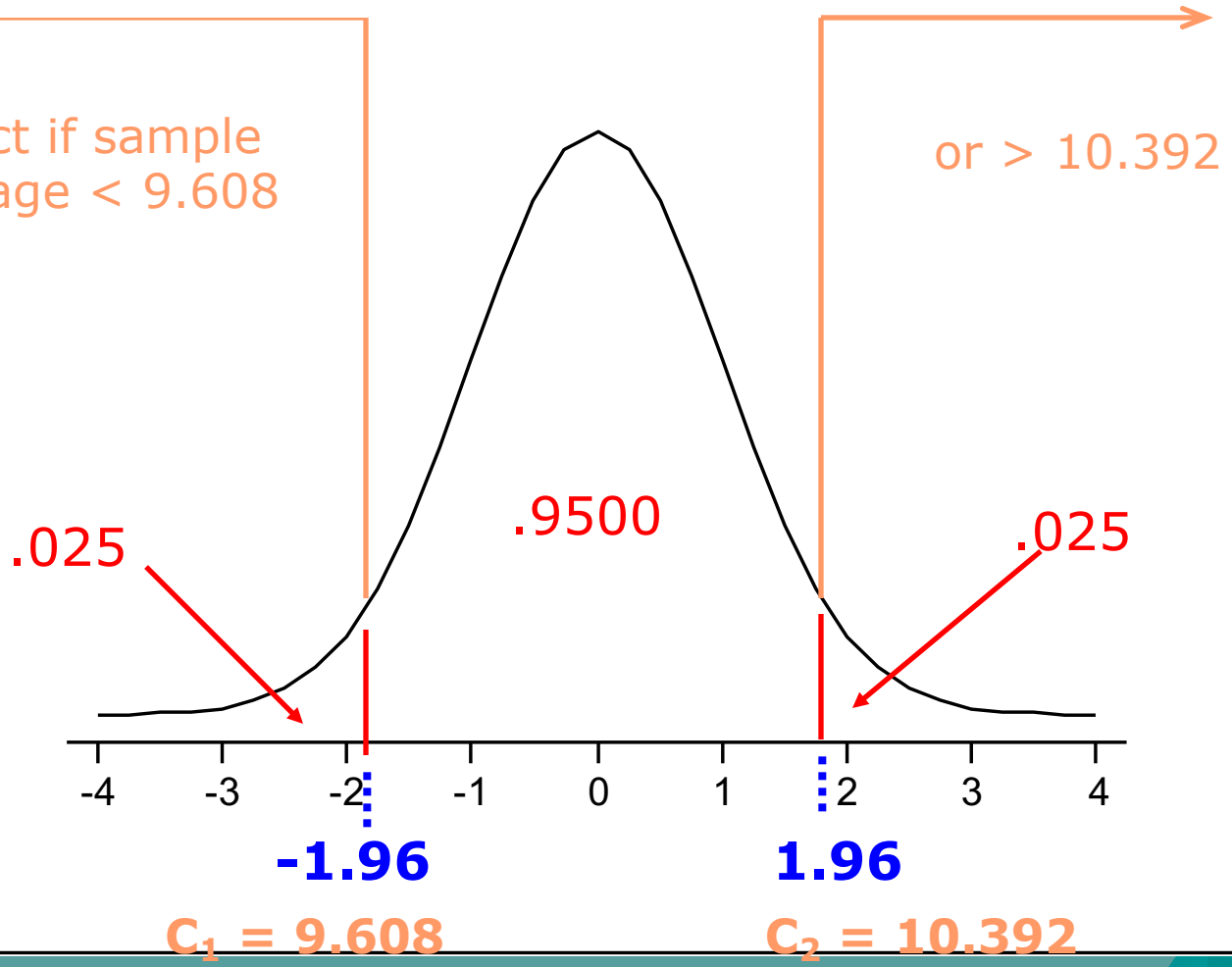
Suppose we know that the population standard deviation is 2, and we wish to use a 5% level of significance.

$$H: \quad = 10$$

$$H_a: \quad 10$$

We will reject H_0 if we get either a very large or very small value of the sample average.

The Decision Rule for a 5% Level of Significance



Sample Average as Test Statistic

- In this example we were able to use the sample average as our test statistic (reject if average < 9.608 or > 10.392).
- Because we knew the standard deviation σ we were able to figure out what points corresponded to the two *z* critical values.
- We could also have worked with a standardized test statistic, which is what we will have to use if σ is unknown.

Standardized Test Statistic

- We could have worked with:

$$z = \frac{\bar{y} - 0}{/\sqrt{n}}$$

- If so, the decision rule is:

  At a 5% level of significance, reject H_0 if $z > 1.96$ or $z < -1.96$

Tests When σ Is Unknown

- As in the estimation problem, use the sample standard deviation.
- The standardized test statistic is now:

$$t = \frac{\bar{y} \quad 0}{s/\sqrt{n}}$$

- The critical values now come from the *t* distribution with *n-1* degrees of freedom.

Example 2.8

- A company that manufactures rulers wants to insure the average length is correct (12 inches).
- From each production run, a sample of 25 rulers is selected and checked with accurate equipment.
- One particular sample had an average of 12.02 inches with a standard deviation .02 inch.
- Using a 1% level of significance, test to see if production is on target.

Our example

H_0: $\mu = 12$ (everything OK)

H_a: $\mu \neq 12$ (something wrong)

Decision process: compute $t = \dfrac{\bar{y} \quad 12}{s/\sqrt{25}}$

Decision rule:

Reject H_0 if $t > t_{.005,25} = 2.797$

or if $t < -2.797$

Results

The sample averaged 12.02 with a standard deviation of .02.

$$t = \frac{\bar{y} \quad 12}{s/\sqrt{25}} = \frac{12.02 \quad 12}{.02/\sqrt{25}} = \frac{.02}{.004} = 5$$

Because this is beyond 2.797, we reject H_0 and conclude we are producing rulers longer than 12 inches.

One-Sided Tests (Lower Tail)

$$H_0: \quad = \quad_0 \quad (\textbf{or} \quad)$$

$$H_a: \quad < \quad_0$$

Reject H_0 if $\bar{y}$ is too small

One-Sided Tests (Upper Tail)

$$H_0: \quad = \quad_0 \quad (\textbf{or} \quad)$$

$$H_a: \quad > \quad_0$$

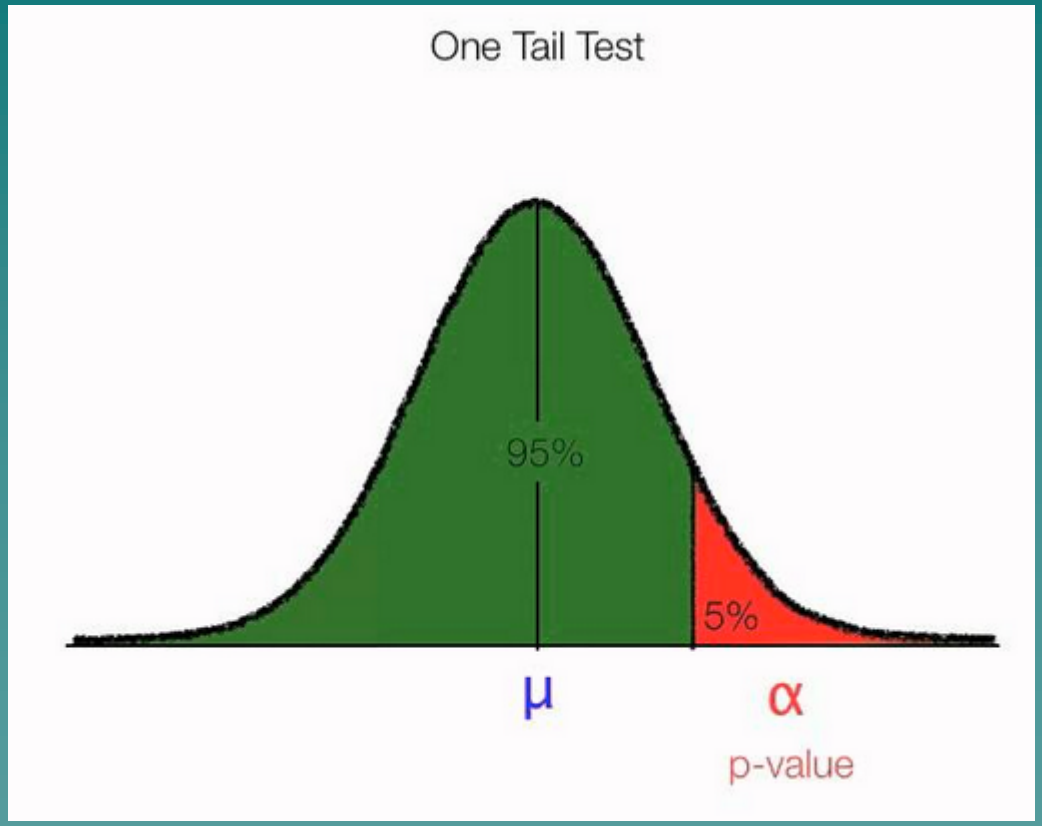
This time reject H_0 if $\bar{y}$ is too large

P-Value

- Most software packages report the results of a hypothesis test computing the p-value of the test.
- This is just a probability that says how far out in the tail the test statistic fell.
- An equivalent decision rule is this:

  Reject H_0 if the p-value < α

Alpha & P-value



If p-value < alpha , then we could reject Ho : mu = mu_o

Example

- We have a one–sided test:

$$H_0: \mu \leq 10$$
$$H_a: \mu > 10$$

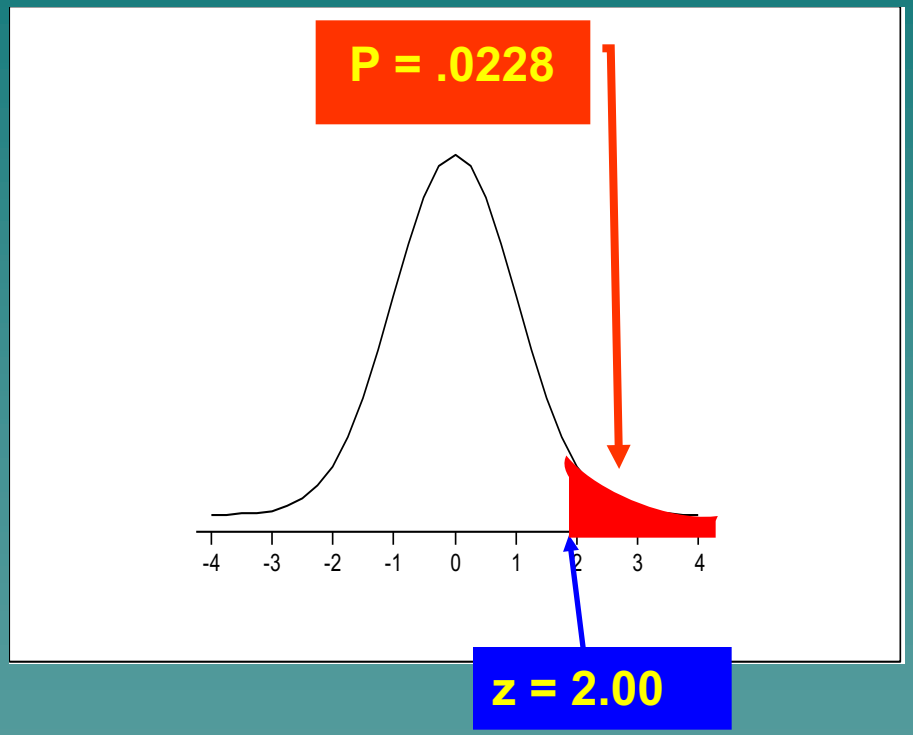
- If we knew that σ was 10, sampled $n=100$ and obtained a sample average of 12, what is the p-value?

P Value

- **First find the value of the test statistic.**

$$z = \frac{12 \quad 10}{10/\sqrt{100}} = 2.0$$

- **Now find tail probability beyond the computed value.**



2.8 Estimating the Difference Between Two Population Means

Here we have two samples and two sets of statistics:

Sample 1:	n_1	$\bar{y}_1$	s_1
Sample 2:	n_2	$\bar{y}_2$	s_2

and want to use them to estimate the difference between the two population means, μ_1 and μ_2

Estimate and Standard Error

- A good estimate of the difference in means, $(\mu_1 - \mu_2)$ is the difference in sample means, $\bar{y}_1 \quad \bar{y}_2$.
- If we know the standard deviations, the standard error of $\bar{y}_1 \quad \bar{y}_2$ is:

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Interval Estimate

- If we are sampling from two normal populations, an interval estimate is:

$$(\bar{y}_1 \quad \bar{y}_2) \pm z_{/2} \sqrt{\frac{^2_1}{n_1} + \frac{^2_2}{n_2}}$$

- We can also use this as a good approximate interval if both sample sizes are large ($n_1 \geq 30$ and $n_2 \geq 30$).

Unknown σ_1 and σ_2

- We can use this formula only if the population standard deviations are known.
- If they are not, we can use the sample standard deviations and get:

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

The Approximate Interval

- As before, use of the sample standard deviations means we use a *t* distribution for the multiplier.
- In this case, the results are only approximate and the *t* distribution has Δ degrees of freedom (see the text for how Δ is computed.)

The Pooled Variance Estimate

- In some cases, it may be reasonable to assume that σ_1 and σ_2 are approximately equal, in which case we need only estimate their common value.
- For this purpose, we "pool" the two sample variances and get S_p^2 which is a weighted average of the two sample variances.

The Exact (pooled sample) Interval

If this is the situation, we can compute an exact interval:

$$(\bar{y}_1 \quad \bar{y}_2) \pm t_{/2,n_1+n_2} \; 2\sqrt{S_p^2 \; \frac{1}{n_1} + \frac{1}{n_2}}$$

Note that the pooling allows us to combine degrees of freedom:

$$df = (n_1\text{-}1)+(n_2\text{-}1) = n_1 + n_2 \text{ -}2$$

What Should We Use?

- If we know the two population variances are about equal, use the exact procedure.
- If we think they differ a lot, we should use the approximate result.
- If we do not really know, the approximate approach is probably best.

Example 2.10

- For the 83 mutual funds we discussed earlier, we want to compare the five-year returns for load funds versus no-load funds.
- The Minitab output for both procedures is on the next slide. The exact procedure output is on the lower half.

Minitab Two-Sample Output

```
Two-sample T for 5yr ret

LoadNoLo     N      Mean     StDev   SE Mean
0           32      5.95      5.88       1.0
1           51      5.01      4.80      0.67

Difference = mu (0) - mu (1)
Estimate for difference:  0.94
95% CI for difference: (-1.54, 3.42)
T-Test of difference = 0 (vs not =): T-Value = 0.76  P-Value = 0.450 DF=56
```

Approximate

```
Two-sample T for 5yr ret

LoadNoLo     N      Mean     StDev   SE Mean
0           32      5.95      5.88       1.0
1           51      5.01      4.80      0.67

Difference = mu (0) - mu (1)
Estimate for difference:  0.94
95% CI for difference: (-1.41, 3.29)
T-Test of difference = 0 (vs not =): T-Value = 0.80  P-Value = 0.428 DF=81
Both use Pooled StDev = 5.24
```

Exact
(uses pooled SD)

SAS Two-Sample Output

The TTEST Procedure

Variable: return

type	N	Mean	Std Dev	Std Err	Minimum	Maximum
load	51	5.0078	4.8001	0.6721	-5.0000	15.9000
noload	32	5.9500	5.8833	1.0400	-7.8000	17.0000
Diff (1-2)		-0.9422	5.2411	1.1820		

type	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
load		5.0078	3.6578	6.3579	4.8001	4.0163	5.9669
noload		5.9500	3.8289	8.0711	5.8833	4.7166	7.8217
Diff (1-2)	Pooled	-0.9422	-3.2939	1.4096	5.2411	4.5435	6.1938
Diff (1-2)	Satterthwaite	-0.9422	-3.4226	1.5383			

Exact

Approximate

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	81	-0.80	0.4277
Satterthwaite	Unequal	56.223	-0.76	0.4499

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	31	50	1.50	0.1963

HW #1 (due to Sept. 8th)

- write the SAS source code for

1) Make a library "library" at "c:\sastemp" directory
2) Copy "loadno2.xpt" to "c:\temp" directory
3) Import "loadno2.xpt" file at c:\temp as a SAS data set at "library" (c:\sastemp)
   (see the manual at e-class)
4) Create a data set "final" which has
   'Type' and 'Return' variables only
   Type = "noload5" or "loan5"
   Return = return value
5) Use PROC TTEST for two-sample t-test
   which will produce the previous page output

Interpretation

- Since we do not have information that the population variances are equal, it is best to use the approximate procedure.
- The degrees of freedom are $\Delta=56$ and the interval estimate of $(\mu_{NoLoad} - \mu_{Load})$ is -1.538 to 3.423.
- Because this interval contains zero, we can conclude the return rates are not that different.

2.9 Hypothesis Tests About the Difference Between Two Population Means

Our test is of the form:

H_0: $\mu_1 = \mu_2$ (No difference)
H_a: $\mu_1 \neq \mu_2$ (One is higher)

which has an equivalent form:

H_0: $\mu_1 - \mu_2 = 0$ (Difference is zero)
H_a: $\mu_1 - \mu_2 \neq 0$ (Difference not zero)

Test Statistic

- For the hypothesis of zero difference, the test statistic is just:

$$t = \frac{\bar{y}_1 \quad \bar{y}_2}{SE}$$

- The standard error (*SE*) is either:

$$\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \quad \text{or} \quad \sqrt{S_p^2 \quad \frac{1}{n_1} + \frac{1}{n_2}}$$

Choice of Procedure

- As before, we use the approximate procedure with Δ degrees of freedom if we cannot assume σ_1 and σ_2 are equal to some common value.
- If that is a reasonable assumption, we compute the pooled standard error and use the exact procedure with (n_1+n_2-2) degrees of freedom.

Example

To test the hypothesis that load and no load funds have the same return, we write:

$$H_0: \mu_N - \mu_L = 0$$
$$H_a: \mu_N - \mu_L \neq 0$$

We do not know that the variances are equal, so we use the approximate procedure which has $\Delta = 56$ degrees of freedom.

Results

At a 5% level of significance,

Reject H_0 if $t > t_{.025,56} \approx 1.96^{**}$

or $t < -1.96$

Minitab gives us $t = 0.76$ so we accept H_0 and will conclude there is no difference in average return.

**The correct value for a t_{56} is 2.003.