

Chapter 2

Two aspects of statistics

- **Descriptive Statistics** Summarizing and organizing information. Usually the information comes from a sample of some population.
- **Inferential Statistics** Making inferences about the characteristics of a population based on the characteristics of a sample from the population.

Descriptive Statistics

I. Tabular Summaries of Data

(a) **Frequency Distributions** Data is grouped into bins, or classes, and the number of data values in each class is given.

(b) **Relative Frequency Distributions** Similar to frequency tables, except that proportions or percentages of data values in each class are given.

(c) **Cumulative Frequency Distributions** In this case the number of data values *at or below* each class limit is given.

- Be careful how the lower and upper endpoints of a bin or class are defined.

Example (a) Frequency distribution for total scores on the first three Calc exams

Total Score	Frequency
$180 \leq x < 190$	3
$190 \leq x < 200$	3
$200 \leq x < 210$	4
$210 \leq x < 220$	5
$220 \leq x < 230$	5
$230 \leq x < 240$	4
$240 \leq x < 250$	3
$250 \leq x < 260$	2
$260 \leq x < 270$	1

Example (b) Relative frequency distribution for total scores on the first three Calc exams

Total Score	Relative Frequency
190	10%
200	10%
210	13.33%
220	16.67%
230	16.67%
240	13.33%
250	10%
260	6.67%
270	3.33%

Example (c) Cumulative frequency distribution for total scores on the first three Calc exams

Total Score	Cumulative Frequency
$180 \leq x < 190$	10%
$190 \leq x < 200$	20%
$200 \leq x < 210$	33.33%
$210 \leq x < 220$	50%
$220 \leq x < 230$	66.67%
$230 \leq x < 240$	80%
$240 \leq x < 250$	90%
$250 \leq x < 260$	96.67%
$260 \leq x < 270$	100%

II. Graphical Summaries of Data

(a) Frequency Histograms Data is grouped into bins, or classes, and the number of data values in each class is given.

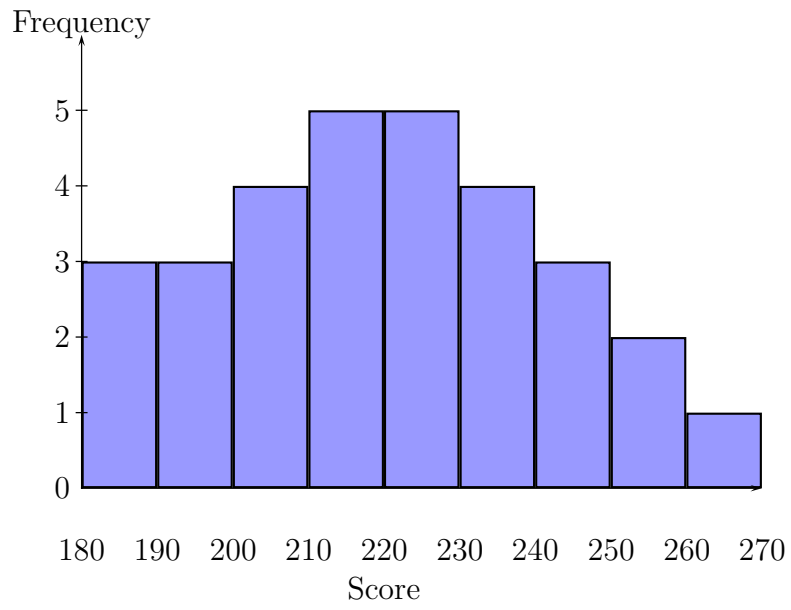
(b) Relative Frequency Histograms Similar to frequency tables, except that proportions or percentages of data values in each class are given.

(c) Cumulative Frequency Histograms In this case the number of data values *at or below* each class limit is given.

- In each case, the horizontal axis represents the value of the data, the bars cover the width of a bin
- The height of a bar represents the frequency, relative frequency, or cumulative frequency of that bin (*assuming all bins have the same width*)

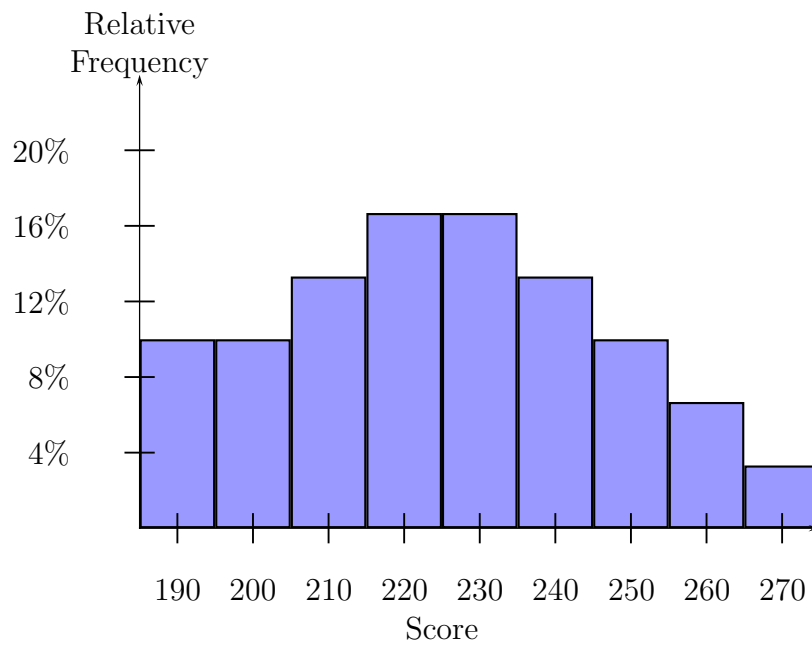
Example (a)

Frequency Histogram for Calc Scores



Example (b)

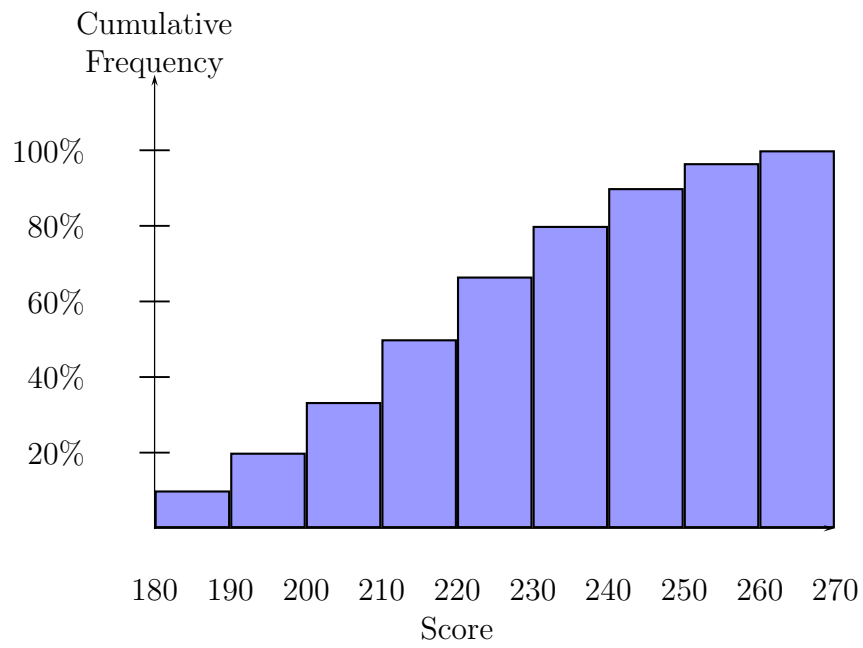
Relative Frequency Histogram for Calc Scores



- Notice that the bar of the histogram is centered over the value of the data set.
- Still, in this case, the bar represents a range of data.

Example (c)

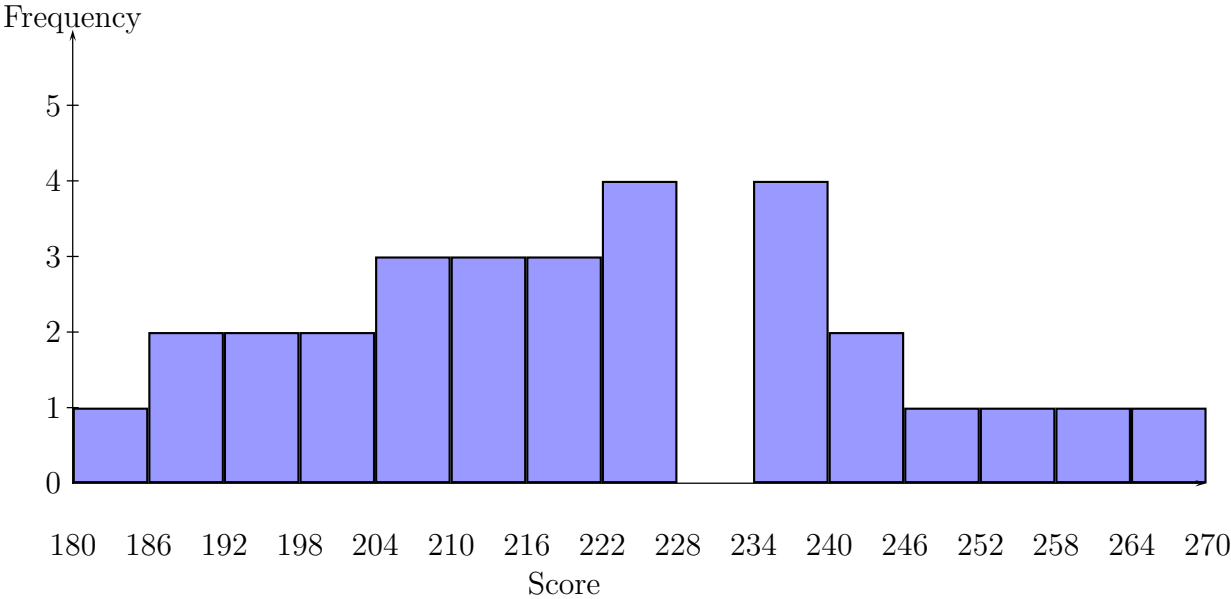
Cumulative Histogram for Calc Scores



Remark The bin widths are up to you. Notice that different bin widths can have a large impact on the histogram.

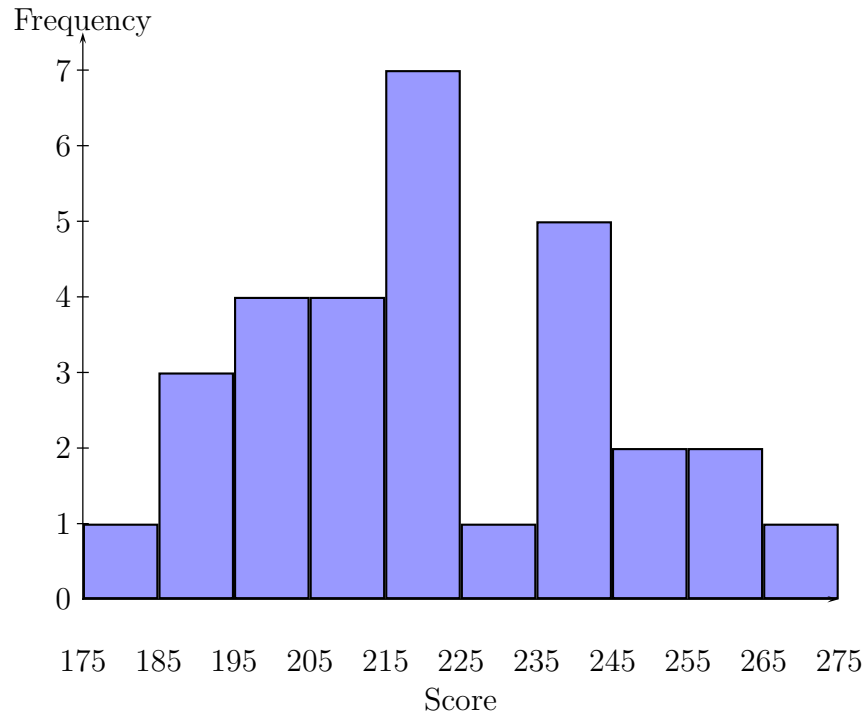
- Bin Width = 6

Frequency Histogram for Calc Scores



- Bin Width = 10, with a different minimum and maximum

Frequency Histogram for Calc Scores



Numerical Summaries of Data Single numbers that summarize one aspect of a data set

I. Measures of Central Tendency

- **Sample Mean**

$$\bar{x} = \frac{\sum x}{n}$$

where n = sample size.

- **Sample Median** The number that divides the bottom 50% of the data from the top 50%.

- If there is an even number of observations, take the average of the middle two observations.

- If there is an odd number of observations, the median is the middle observation.

- **Comparison** The mean is sensitive to extreme observations, the median is not. ie, the mean is the center in terms of value of the numbers in the data set, the median is the center in terms of the position of the numbers.

Example Find the median of

5	8	19	31	22
16	11	24	17	

Example Find the median of 3, 5, 5, 6, 8, 9

II. Measures of Dispersion

- **Sample Variance**

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

the average of the squared difference between each data point from the mean

- **Sample Standard Deviation**

$$s = \sqrt{s^2} = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

roughly the average deviation from the mean

- **Why divide by $n - 1$ rather than n ?** This is a theoretical question, but in practice we use the *sample* variance and standard deviation to estimate the *population* variance and standard deviation. When we divide by n in the sample variance formula, we tend to consistently get an underestimate of the population variance. When we divide by $n - 1$, we tend to get a more accurate, or unbiased, estimate of the true population variance.

SPSS Example Exercise 2, p. 12

- In SPSS, open the file U:MT Student File Area/Stat378SS/Stat378 Data/Dielman, 4ed Data/Chapter 2 Data/SPSS/GRADRATE2.sav
- To obtain a frequency table, run **Analyze/Descriptive Statistics/Frequencies**

- To obtain measures of central tendency and dispersion, run **Analyze/Descriptive Statistics/Explore**
- To obtain a histogram, run **Graphs/Histogram**
- To put a title on the histogram, label axis, rescale axes, etc,
 - Right click on the graph
 - Open the chart editor

Random Variable: A rule that assigns a number to every possible outcome of an experiment.

Two types of Random Variables

- **Discrete Random Variables** The variable can take on only finitely many values, or else the values of the variable can be “listed”
- **Continuous Random Variables** The values of the variable come from an interval of numbers.

Probability Distribution of a Random Variable: A table, graph, or formula that provides

1. the possible values of the random variable, and
2. their corresponding probabilities.

Probability Histogram of a Random Variable: Same thing as a relative frequency histogram

I. Discrete Random Variables

Example

Suppose you toss a coin 3 times. Let X be the number of heads observed in three tosses.

- What are the possible values of the random variable X ?

- Determine $P(X = 2)$.

- Find the probability distribution of X .

- Find $P(X \leq 2)$.

Sum of Probabilities of a Discrete Random Variable must equal 1

ie, if X is a random variable, then

$$\sum P(X = x) = 1$$

The Mean of a Discrete Random Variable

$$\mu_x = \sum xP(X = x)$$

Remarks

- The term **expected value** is also used for the mean

Example

You toss a coin 3 times, let X denote the number of times Heads appears. Find the mean of the random variable X .

Interpretation of the mean If an experiment is repeated a large number of times, and the value of the random variable is recorded each time, then the mean of the recorded values of the variable will approximately be the mean of the random variable. The more the experiment is repeated, the closer the mean of the recorded values will tend to be to the mean of the random variable.

Standard Deviation of a Discrete Random Variable

$$\sigma_x = \sqrt{\sum (x - \mu)^2 P(X = x)}$$

alternatively, we could use the computing formula

$$\sigma_x = \sqrt{\sum x^2 P(X = x) - \mu^2}$$

Let X denote the age of a randomly selected student. A probability distribution for X is as follows:

Age x	Probability $P(X = x)$
19	0.250
20	0.375
21	0.250
27	0.125

Find the standard deviation of the ages of the students.

- In a normally distributed random variable, a *Normal Curve* approximates the probability histogram. Areas under this normal curve approximate areas of the bars of the histogram.

Standard Normal Distribution A normally distributed random variable with a mean of 0 and a standard deviation of 1.

Standardized Random Variable For a random variable x , the standardized version of x is

$$z = \frac{x - \mu}{\sigma}$$

- z is the number of units of standard deviation that x is from the mean.
- If x is normally distributed, then z is normally distributed
- If x is normally distributed, then the mean of z is always 0.
- If x is normally distributed, then the standard deviation of z is always 1.
- The percentage of all possible observations of x that lie between a and b is the same as the percentage of all possible observations of z that lie between

$$\frac{a - \mu}{\sigma} \text{ and } \frac{b - \mu}{\sigma},$$

which is just the area under the curve for z between $\frac{a - \mu}{\sigma}$ and $\frac{b - \mu}{\sigma}$.

Empirical Rule: If the data are approximately bell-shaped, then

- At least 68.26% of the observations lie within 1 standard deviation of the mean.
- At least 95.44% of the observations lie within 2 standard deviations of the mean.
- At least 99.74% of the observations lie within 3 standard deviations of the mean.

Example The annual wages, excluding board, of U.S. farm laborers in 1926 were normally distributed with a mean of \$586 and a standard deviation of \$97. In 1926 what percentage of U.S. farm laborers had an annual wage of

- between \$500 and \$700?

– Find the area to the left of $z = 1.18$

– Use Table II.

z	.0007	.08	.09	...
.				.		
.				.		
.				.		
1.0				0.8599		
1.1	0.8643	...	0.8790	0.8810	0.8830	...
1.2				0.8997		
.				.		
.				.		
.				.		

– Find the area to the left of $z = -0.89$ under the standard normal curve.

– Use Table II.

z	.0008	.09
.				.
.				.
.				.
-0.9				0.1611
-0.8	0.2119	...	0.2177	0.1867
-0.7				0.2148
.				.
.				.
.				.

– Subtract the two areas

- What percentage had an annual wage of at least \$400?

Example

The length of the western rattlesnake is normally distributed with a mean of 42 inches and a standard deviation of 2 inches. Let X denote the length of one of these snakes selected at random.

- Find $P(41 < X < 45)$ (Use table B.1 in the appendix)

- Find $P(X < 38)$.

Sampling Distribution of the Mean The probability distribution of the random variable \bar{x} .

- In other words, the sampling distribution of the mean is the distribution of **all possible sample means** for samples of a given size.
- **The mean of the random variable \bar{x} :**

$$\mu_{\bar{x}} = \mu$$

ie, the mean of all the possible sample means for samples of a given size is the same as the population mean.

- **The Standard Deviation of the Variable \bar{x} .**

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

ie, the standard deviation of all the possible sample means is the population standard deviation divided by \sqrt{n} .

Important Fact Suppose a random variable x is normally distributed with mean μ and standard deviation σ . For samples of size n , the variable \bar{x} is normally distributed and has mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

Example Suppose the population mean weekly salary for migrant workers in California is

$$\mu = 250$$

and the population standard deviation is

$$\sigma = 80$$

Suppose also that the weekly salaries for migrant workers are normally distributed. If we take a random sample of 36 migrant workers and determine the mean salary, what percentage of the time will the sample mean be between 238 and 262?

Central Limit Theorem

For a relatively large sample size, the variable \bar{x} is approximately normally distributed, regardless of the variable under consideration. The approximation becomes better and better with increasing sample size.

- The farther the variable is from being normally distributed, the larger the sample size must be for the distribution of \bar{x} to be normal.
- Usually, a sample size of $n = 30$ is enough.

Example An economist samples 500 families and determines the mean weekly food cost, \bar{x} , and uses this as an estimate of the true mean weekly food cost, μ .

Suppose the population standard deviation is known to be $\sigma = \$17.20$.

What is the probability that the sampling error made in estimating μ by \bar{x} will be at most \$1?

Confidence Intervals for the Population Mean μ An interval of numbers centered at \bar{x} which we are confident to a certain level contains μ .

Notation z_α = the z score having an area of α to its right under the standard normal curve.

- Suppose we want to know where the range for the middle 95% of z scores

- Then we need to know the z scores that separate the lower and upper 2.5% of all z scores

- ie, we need to know $-z_{0.025}$ and $z_{0.025}$.

- Suppose we want a confidence interval for the mean μ .
- We find a sample of size n , and compute \bar{x} .
- From the sampling distribution of \bar{x} , there is some interval centered at μ in which \bar{x} should fall, say, 95% of the time.

- Thus, 95% of the time, the z score for \bar{x} should fall between $-z_{0.025}$ and $z_{0.025}$.

- Thus, 95% of the time, \bar{x} should fall between $\mu - z_{0.025} \cdot \sigma_{\bar{x}}$ and $\mu + z_{0.025} \cdot \sigma_{\bar{x}}$.
- However, we don't know μ . Solving for \bar{x} , we see that 95% of the time, we should have that

$$\bar{x} - z_{0.025} \cdot \sigma_{\bar{x}} \leq \mu \leq \bar{x} + z_{0.025} \cdot \sigma_{\bar{x}}.$$

Formula for a $(1 - \alpha) \cdot 100\%$ Confidence Interval for μ , σ known

$$\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \text{ to } \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Example Twenty randomly selected science books have the prices, to the nearest dollar, shown in the table.

91	61	99	88	93
73	100	54	100	75
85	78	70	80	87
52	97	102	80	75

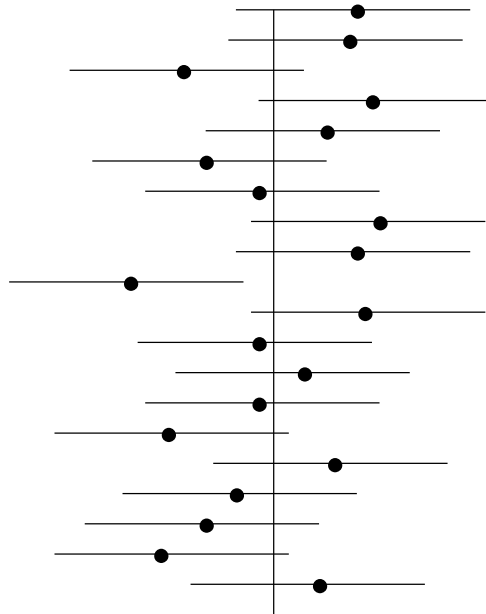
Assume that science book prices are normally distributed, and that $\sigma = \$16$. Determine a 95.44% confidence interval for the mean price of a science book.

Interpretation For a confidence interval with a 95.44% confidence level, in theory 95.44% of the samples will result in an interval which contains μ . So we have a 95.44% chance that the method we have just used will result in an interval which contains μ . Therefore we say we are 95.44% confident that the true mean lies in our confidence interval.

- Suppose, in the last example, that the mean price of all textbooks is actually $\mu = \$86$.
- We can do a computer simulation of 20 random samples, each containing 20 textbooks.
- The sample mean, along with the confidence interval for each sample, is given in the table.

Sample	\bar{x}	Confidence Interval
1	\$89.80	\$82.64 to \$96.96
2	\$89.47	\$82.31 to \$96.63
3	\$82.01	\$74.85 to \$89.17
4	\$92.30	\$85.14 to \$99.46
5	\$87.72	\$80.56 to \$94.88
6	\$83.62	\$76.46 to \$90.78
7	\$85.60	\$78.44 to \$92.76
8	\$90.07	\$82.91 to \$97.23
9	\$89.71	\$82.55 to \$96.87
10	\$78.39	\$71.23 to \$85.55
11	\$92.12	\$84.96 to \$99.28
12	\$84.41	\$77.25 to \$91.57
13	\$87.66	\$80.50 to \$94.82
14	\$85.71	\$78.55 to \$92.87
15	\$80.97	\$73.81 to \$88.13
16	\$89.12	\$81.96 to \$96.28
17	\$84.30	\$77.14 to \$91.46
18	\$82.76	\$75.6 to \$89.92
19	\$78.88	\$71.72 to \$86.04
20	\$88.73	\$81.57 to \$95.89

- Notice that 19 out of the 20 confidence intervals contains μ .



Finding Confidence Intervals when σ is not known

- We must estimate σ by the sample standard deviation s
- The approximation to $\sigma_{\bar{x}}$ is then s/\sqrt{n} .
- Then we transform the variable \bar{x} using

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}},$$

- This no longer has the standard normal distribution.
- It has the **Student's t-Distribution** with $n - 1$ **Degrees of Freedom**
- **Notation** $df = n - 1$ means $n - 1$ degrees of freedom.

Facts About the t distribution

- A t curve looks like a normal curve, but actually has more variation.
- As the number of degrees of freedom increases, t curves look more and more normal
- We compute areas under the t curves using Table B.2.

Confidence Interval for μ when σ is unknown

$$\bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \text{ to } \bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

Example

The heights, in inches, of 20 randomly selected 6 year old girls are as follows:

44	44	47	46	38
42	46	41	50	43
40	51	47	43	47
48	48	45	41	46

- Obtain a 90% confidence interval for the mean height of the 6 year old girls. (*Note:* $\bar{x} = 44.85$ inches, and $s = 3.392$ inches)

- Interpret your answer.

SPSS Example Problem 26, p. 31

- In SPSS, open the file U:MT Student File Area/Stat378SS/Stat378 Data/Dielman, 4ed Data/Chapter 2 Data/SPSS/ONERET2.sav
- Run the command **Analyze/Descriptive Statistics/Explore**
- Click on the **Options** box to specify the confidence level.

Hypothesis Testing Test whether or not there is support for a given hypothesis

- $H_0 =$ **Null Hypothesis** The hypothesis to be tested
- $H_a =$ **Alternative Hypothesis** The hypothesis that is considered as an alternative to the null hypothesis

The Logic Behind Hypothesis Testing

- Assume that the null hypothesis is true
- From a sample of the population, compute a statistic, called the **Test Statistic**, from the data in the sample

- Find the probability of getting such a value for the test statistic in the event that the null hypothesis is true.
- If this probability is too small, then there was something wrong with our original assumption, and the null hypothesis is rejected in favor of the alternative hypothesis.
- If the probability is not small, then the null hypothesis is not necessarily accepted as true. However, there is not sufficient evidence to reject it.

Example

- A coin-operated soft drink machine was designed to discharge, on the average, 7 ounces of beverage per cup.
- We want to check and see if the machine is working properly.
 - ie, we want to test the hypothesis that $\mu = 7$.
- From past experience with the machine, we know the standard deviation σ is 0.12 ounces per cup.
- Suppose we randomly sample 10 cupfuls of beverage from the machine, and record the following amounts per cup.

7.09	6.98	7.2	6.97	7.14
7.15	7.05	7.18	7.08	7.16

- Compute the sample mean \bar{x} for this sample.
- **Questions**
 - Can we conclude, based on this sample, that the mean amount of beverage discharged from the machine is different from 7 ounces?
 - Or was this sample just due to chance?
- Null and Alternative Hypotheses

- Decide how confident we want to be in our results. If we want to be 95% confident, then our sample mean should fall where 95% of all the sample means fall, assuming that the true mean is 7.
- Use the standard normal curve and find the z -scores corresponding to this range.
- Compute the z score of our sample mean. This is the **Test Statistic**.
- Does this z score fall where we would expect it too, 95% of the time?
- Decide whether or not to reject the null hypothesis that $\mu = 7$.

- Notice that it is possible that $\mu = 7$, and our sample fell outside the expected range. If $\mu = 7$, the probability of this is less than 5%. We say that our **significance level** is 0.05.

Type I and Type II Errors

	Do not reject H_0	Reject H_0
H_0 True	Correct Decision	Type I Error
H_0 False	Type II Error	Correct Decision

Significance Level Two interpretations:

- The probability of making a type I error. ie, the probability of rejecting a true null hypothesis
- The area of the rejection region for the test statistic

If σ is **unknown** use the sample standard deviation and use t scores for the test statistic

Example In 1998, a certain car manufacturer reported that a particular model equipped with a four-speed manual transmission average 29 miles per gallon on the highway. Suppose the EPA tested 15 of the cars and obtained the following gas mileages:

27.3	31.2	29.4	31.6	28.6
30.9	29.7	28.5	27.8	27.3
25.9	28.8	28.9	27.8	27.6

The distribution of gas mileages is approximately normal. Is there sufficient evidence to support the manufacturers claims? Perform a hypothesis test at the 5% significance level. (Notice: $\bar{x} = 28.753$ and $s = 1.595$)

Example One of the strongest arguments in favor of halfway houses for nonviolent criminals is the cost factor. Government officials claim that the mean cost per house is only \$1.50 per day per resident. A neighborhood association opposed to the construction of a new halfway house

in its area counters the argument by surveying 14 such houses. They find that $\bar{x} = 1.65$ and $s = .35$. Can we conclude that the mean cost is actually greater than \$1.50? Use $\alpha = 0.05$.

Hypothesis Testing, P value approach

- The P-value is the area of the rejection region if we take our *test statistic* as the critical value.
- The P-value is the probability that a sample will result in a sample mean such as the one we obtained if the null hypothesis is true.
- The P-value could be viewed as the significance level of the test statistic.
- P-value of a hypothesis test is the *smallest significance level at which the null hypothesis can be rejected*

A Criteria For Rejecting H_0 in Terms of P-values

- If the P-value is less than or equal to the specified significance level, then reject the null hypothesis. Otherwise, do not reject the null hypothesis.
- In general, the smaller the p-value the more significant the result is

Example A company that produces snack foods uses a machine to package 454-gram bags of pretzels. We will assume that the net weights are normally distributed and that the population standard deviation of all such weights is 7.8 grams. A random sample of 25 bags of pretzels has the net weights in the table:

465	456	438	454	447
449	442	449	446	447
468	433	454	463	450
446	447	456	452	444
447	456	456	435	450

Do the data provide sufficient evidence to conclude that the packaging machine is not working properly?

Example Ten years ago, the mean age at first marriage for men in a certain region was 24 years of age. A sample of 10 first time, newlywed men in the region had the following ages:

23	26	27	19	32
29	38	28	32	30

Assume that the ages are normally distributed, and that $\sigma = 5.2$ years. Is this sufficient evidence to conclude that the mean age of men at first marriage has gone up in this community at the 5% level?

Using SPSS to find p values

- You need to have data entered in the spreadsheet
- Enter your test statistic in a variable cell. Give it a name like *teststat*.
- Run the program **Transform/Compute**
- Assign a variable name, like *pvalue*, in the **Target Variable** box
- In the **Numeric Expression** box, use the **CDF.T(teststat,df)** function.

- For a left tailed test, use **CDF.T(teststat,df)**
- For a right tailed test, use **1-CDF.T(teststat,df)**
- For a two tailed test, use either **2*CDF.T(teststat,df)** or **2(1-CDF.T(teststat,df))**

Example The average retail price for bananas in 1994 was 46 cents per pound. Recently, a random sample of 15 bananas had a mean price of 48.4 cents per pound, and a sample standard deviation of 3.5 cents per pound. Can we conclude that the mean price of bananas has increased since 1994?

Using SPSS to perform a hypothesis test

- Run the program **Analyze/Compare Means/One Sample T Test**
- Put the hypothesized mean in the **Test Value** Box
- This is a two tailed test.

Example The mean consumption of beef per person in 1990 was 64 lb. A sample of 40 people taken this year yielded data on last year's beef consumptions. This data is stored on the U-drive under the file **U: MT Student File Area/dgarth/stat190/beef.sav**

Two Tailed Test

Left Tailed Test

Hypothesis Testing for the difference between two population means, equal variances

ASSUMPTIONS

1. The samples are independent
2. Normal populations or large samples
3. The populations have the same variances, ie, $\sigma_1^2 = \sigma_2^2$

- The **Test Statistic** is

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

- This test statistic has the t -distribution with $n_1 + n_2 - 2$ degrees of freedom

Example A study was conducted by the Florida Game and fish commission to assess the amounts of chemical residues found in the brain tissue of brown pelicans. In a test for DDT, random samples of $n_1 = 10$ and $n_2 = 13$ nestlings gave the results shown in the table.

Juveniles	Nestlings
$n_1 = 10$	$n_2 = 13$
$\bar{x}_1 = 0.041$	$\bar{x}_2 = 0.026$
$s_1 = 0.017$	$s_2 = 0.006$

Test the hypothesis that there is no difference between mean amounts of DDT found in juveniles and nestlings versus the alternative that the juveniles have a larger mean. Use $\alpha = 0.05$.

Confidence interval for the difference in two population means, $\sigma_1^2 = \sigma_2^2$

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

where $df = n_1 + n_2 - 2$.

Example Construct a 95% confidence interval for $\mu_1 - \mu_2$ in the previous example.

Hypothesis Testing for the difference between two population means, *unequal variances*

ASSUMPTIONS

1. The samples are independent
2. Normal populations or large samples
3. The populations have the same variances, ie, $\sigma_1^2 \neq \sigma_2^2$

- The **Test Statistic** is

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}$$

- This test statistic has the t -distribution, and the degrees of freedom are estimated by

$$\Delta = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{[(s_1^2/n_1)^2/(n_1 - 1)] + [(s_2^2/n_2)^2/(n_2 - 1)]}$$

Confidence interval for $\mu_1 - \mu_2$ when $\sigma_1^2 \neq \sigma_2^2$

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2} \sqrt{(s_1^2/n_1) + (s_2^2/n_2)}$$

where df is approximately Δ

When in Doubt Assume unequal standard deviations

SPSS Example Problem 40, p. 48.

- Run the program **Analyze/Compare Means/Independent Samples T Test**
- Put the variable, in this case, *Loans*, in the **Test Variable** box
- Put the variable *Type* in the **Grouping Variable** box
- Define the groups as 0 and 1