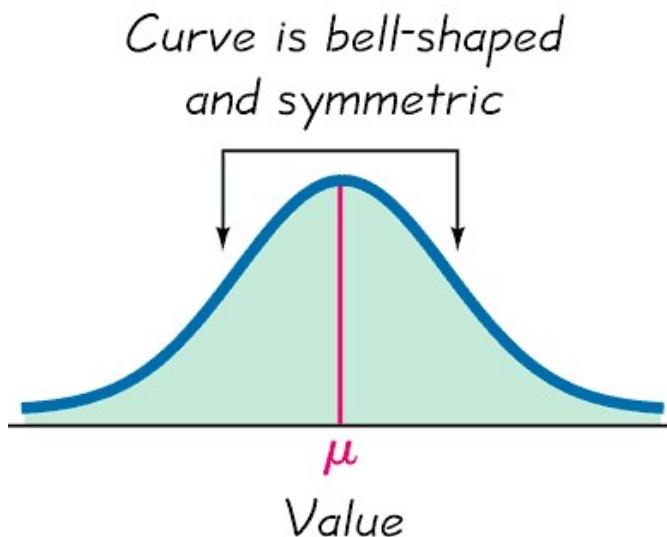


Chapter 6. Continuous Probability Distributions

- ❖ Continuous random variables
- ❖ Normal distributions



$$f(x) = \frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}}$$

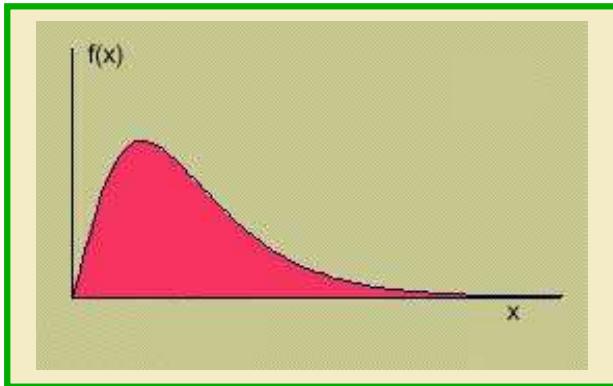
Distribution determined by fixed values of mean and standard deviation

Continuous Random Variables

- ❖ **Continuous Random Variable**
if it can assume infinitely many values corresponding to the points on a **real line interval** (without gaps or interruptions).
- Examples:
 - The **time** it takes to complete a task.
 - **Something measured:** Weight, Height, Temperature, Liters of water, etc.

Continuous Random Variables

- **Density Curve:** A smooth curve (with function $f(x) \geq 0$, called density function) describes the probability distribution of a continuous random variable.

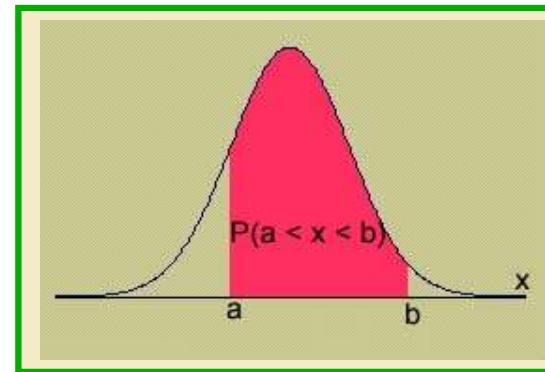


1. Every point on the curve must have a vertical height that is 0 or greater.
2. The total **area** under the curve must equal **1**

Area and Probability

Because the total area under the density curve is equal to 1, there is a correspondence between **area** and **probability**.

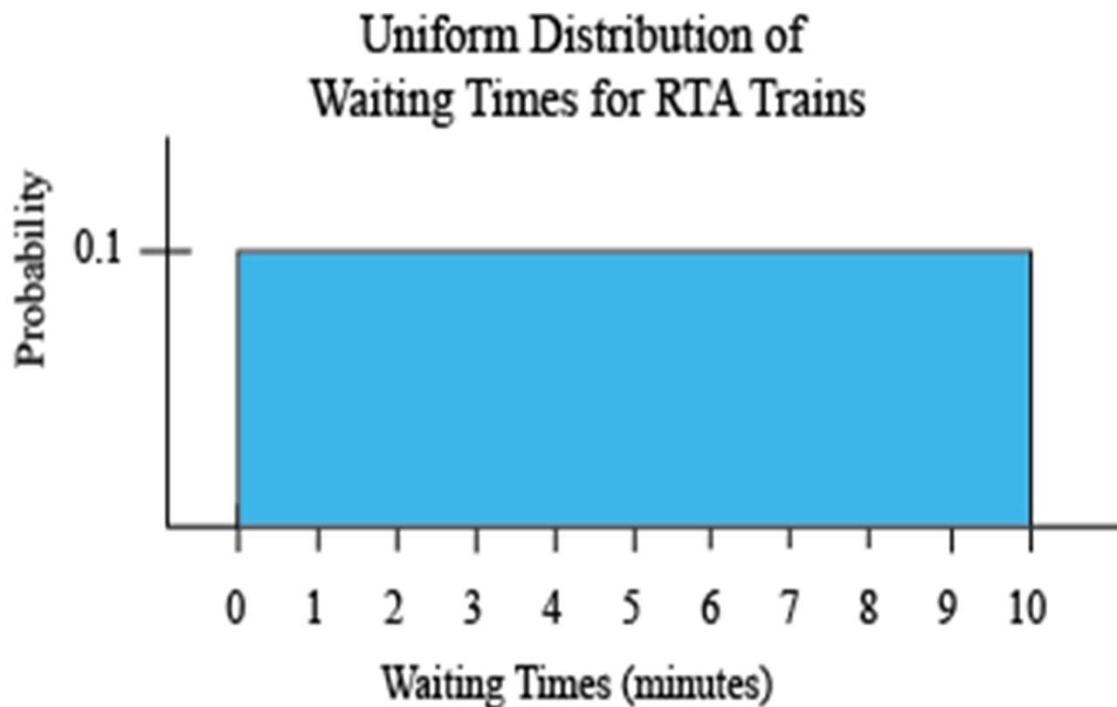
1. $P(a \leq x \leq b) = \text{area under the curve between } a \text{ and } b.$
That is,



2. There is no probability attached to any single value of x . That is, $P(x = a) = 0$.

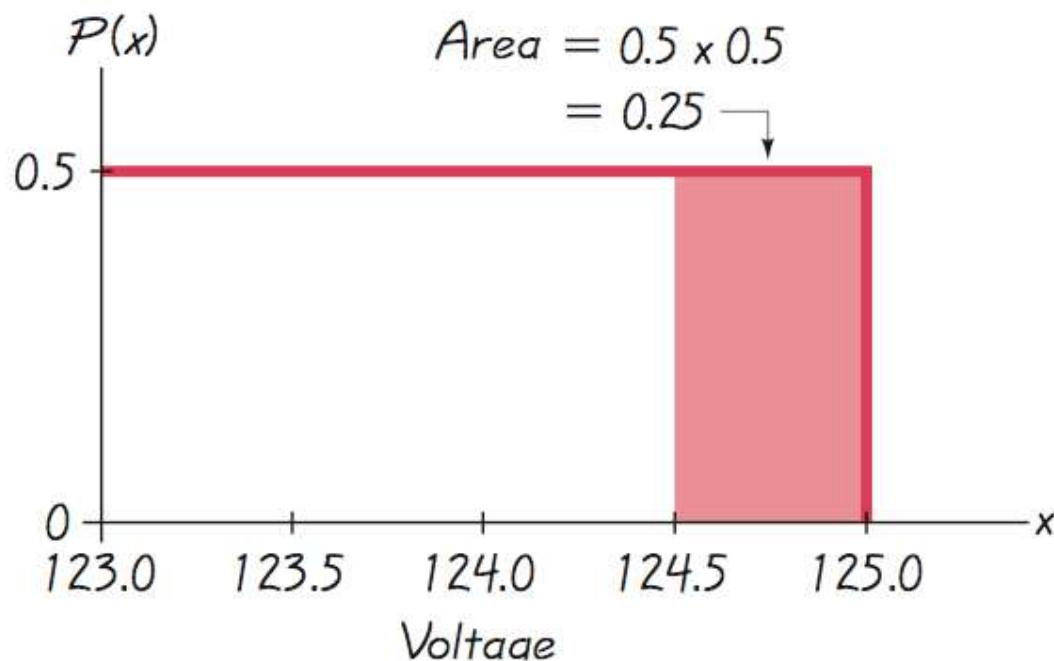
6.1 Uniform Distribution

- A continuous random variable has a **uniform distribution** if its values are spread **evenly** over the range of probabilities. The graph of a uniform distribution results in a rectangular shape.



Using Area to Find Probability

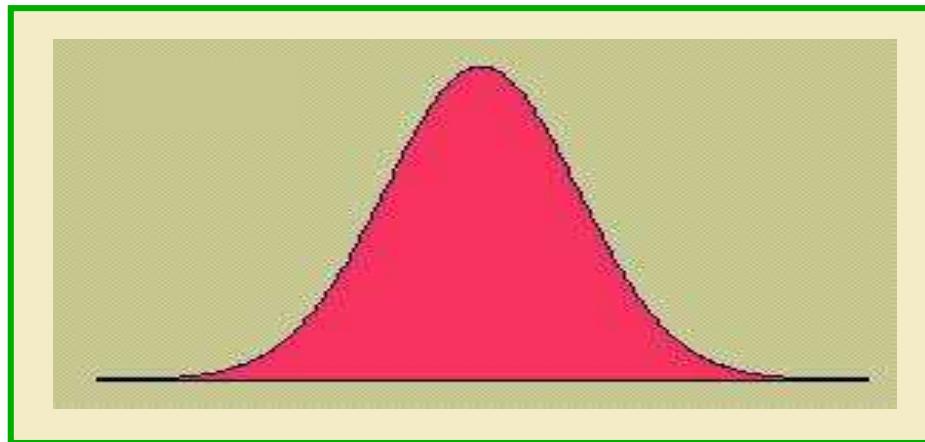
Example. Given the uniform distribution illustrated, find the probability that a randomly selected voltage level is greater than 124.5 volts.



Shaded area
represents voltage
levels greater than
124.5 volts.

6.2 Normal Probability Distribution

- The density function that generates the Normal probability distribution is:



$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \text{ for } -\infty < x < \infty$$

$$e \approx 2.71828 \quad \pi \approx 3.1416$$

μ and σ are the population mean and standard deviation.

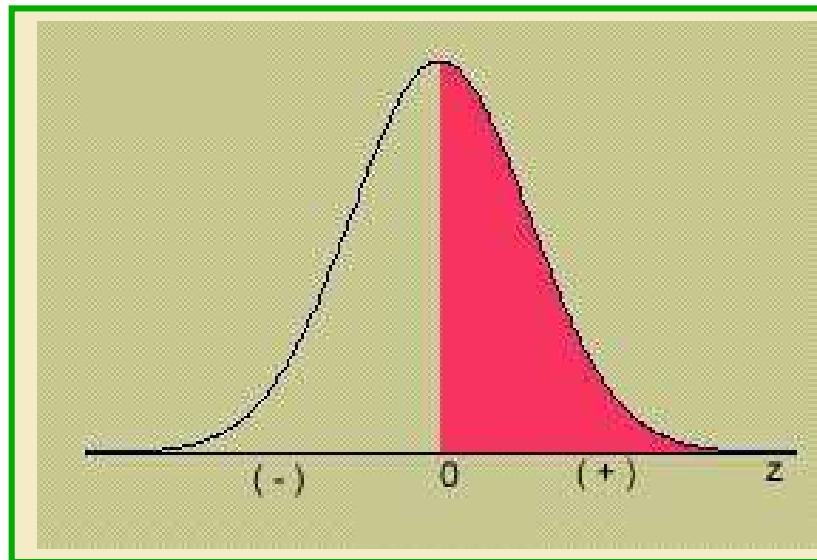
Normal Probability Distribution

- Bell Shaped: **unimodal, symmetric**
- A Normal model with mean μ and standard deviation σ .
 - ✓ μ (read “mew”) represents the population mean.
 - ✓ σ (read “sigma”) represents the population standard deviation.
 - ✓ $N(\mu, \sigma)$ represents a Normal model with mean μ and standard deviation σ .

Parameters and Statistics

- **Parameters:** Numbers that help specify the model
 - μ, σ
- **Statistics:** Numbers that summarize the data
 - ✓ $\bar{x}, s,$
- ❖ The Normal model should only be used if the sample **data** is approximately **symmetric** and **unimodal**.
- ❖ $N(0, 1)$ is called the **standard Normal model**, or the **standard Normal distribution**.

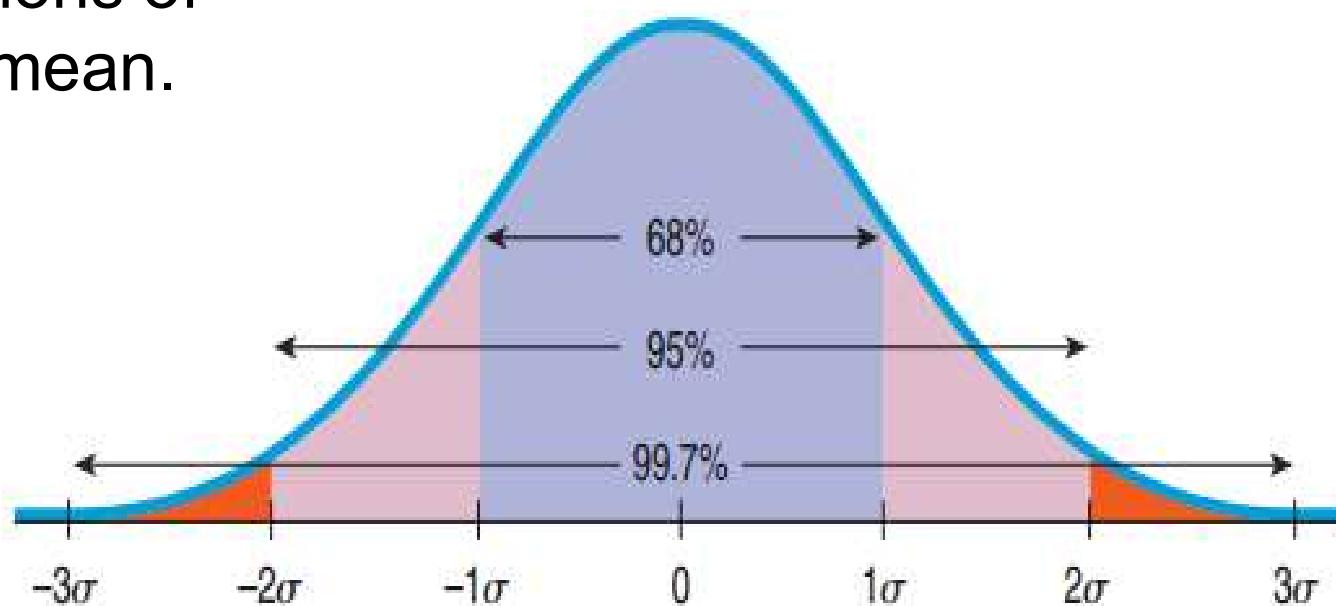
Standard Normal Distribution



- Mean = 0; Standard deviation = 1
- Symmetric about $z = 0$
- Values of z to the left of center are negative
- Values of z to the right of center are positive
- Total area under the curve is 1.

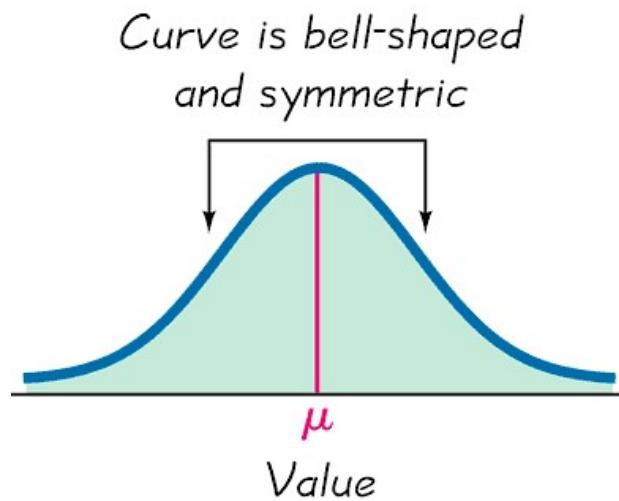
Recall: The 68-95-99.7 Rule

- 68% of the values fall within 1 standard deviation of the mean.
- 95% of the values fall within ≈ 2 (1.96) standard deviations of the mean.
- 99.7% of the values fall within ≈ 3 (2.97) standard deviations of the mean.



General Normal Distributions

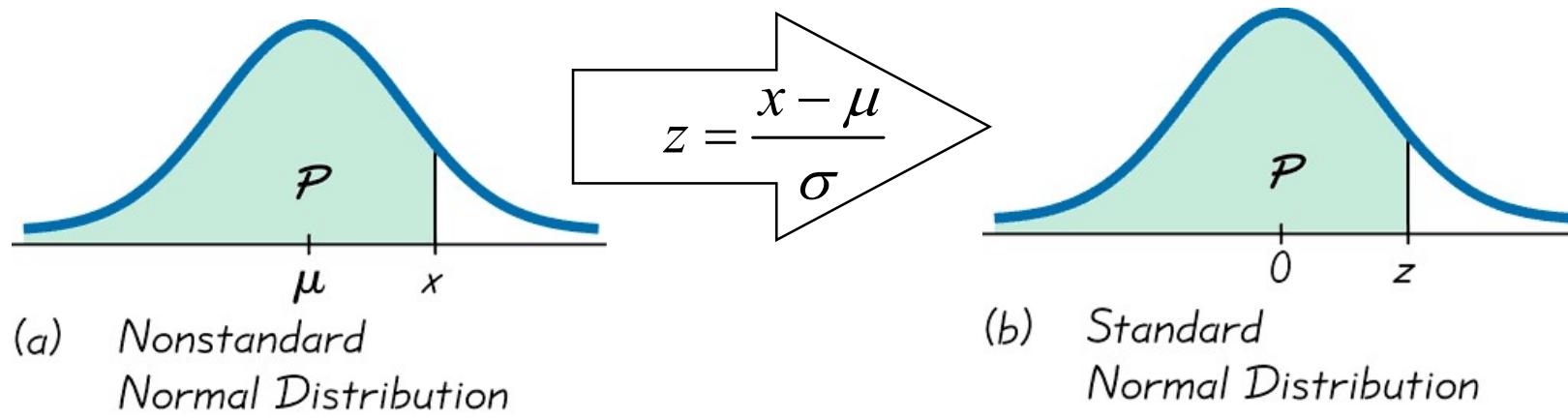
General Normal distributions



$$f(x) = \frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}}, \quad -\infty < x < \infty$$

Distribution determined by fixed values of mean μ and standard deviation σ

Converting to a Standard Normal Distribution



Activity: Center and Spread of Normal Distributions

<https://istats.shinyapps.io/NormalDist/>

Key Concepts

- Continuous distributions
- Uniform distributions
- (Standard) Normal distributions

6.3 Probabilities and Quantiles for Normal Distributions

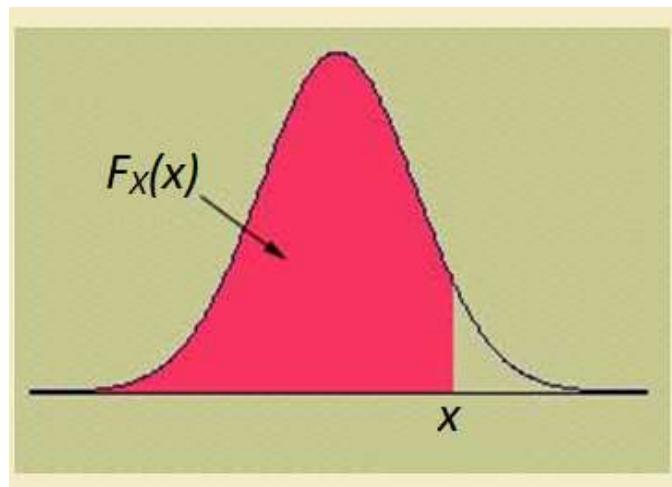
Objectives

- Develop the skill to find **areas** (or probabilities) corresponding to various regions under the graph of the normal distributions.
- Find normal **quantiles** corresponding to a left/right area under the normal density curve.

Finding Probabilities for the Normal Distribution

Definition. The **cumulative distribution function** or cdf of a random variable X , denoted by $F_X(x)$, is defined by

$$F_X(x) = P(X \leq x) \text{ for any } x.$$



Find Normal Probabilities

Let X be a normal random variable with mean μ and standard deviation σ . Then (using cdf),

$$P(X < U) = ?$$

$$P(X > L) = 1 - P(X \leq L)$$

$$P(L < X < U) = P(X \leq U) - P(X \leq L)$$

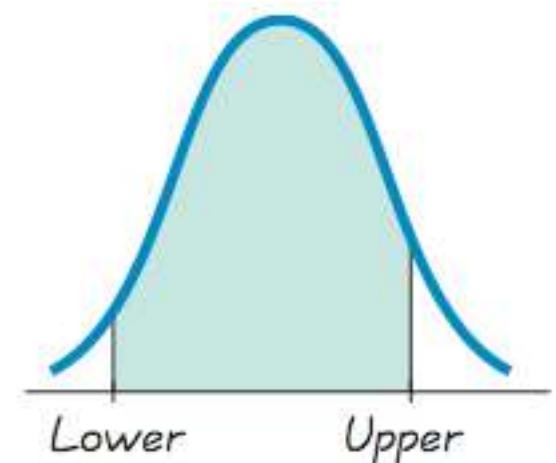
We can get the results directly using R.

Find Normal Probabilities

We can get the result

$$P(L < X < U)$$

directly using R without
using cdf.



Find Normal Probabilities

<https://esumath.shinyapps.io/Rstats>

<https://istats.shinyapps.io/NormalDist/>

Converting to a Standard Normal Distribution(optional)

- To find $P(L < X < U)$ using the z-Table, we need to standardize each value of x by expressing it as a z-score, the number of standard deviations it lies from the mean μ .

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

Round the z scores to 2 decimal places.

- Then

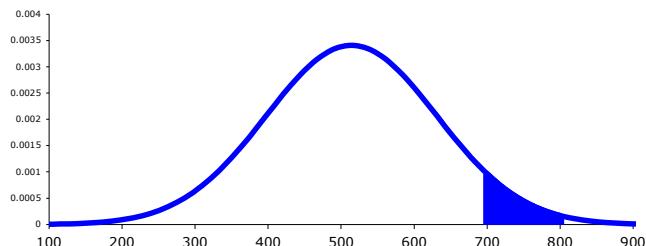
$$P(L < X < U) = P[(L-\mu)/\sigma < Z < (U-\mu)/\sigma]$$

Example

The mean mathematics SAT score in 2012 was 514 with a standard deviation of 117 ("Total group profile," 2012).

Assume the mathematics SAT score is normally distributed.

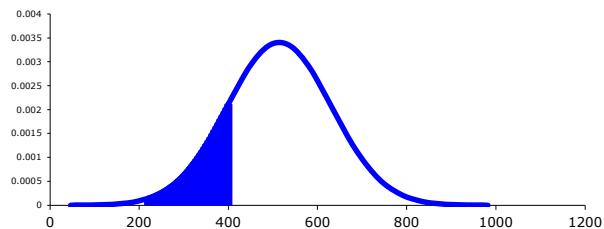
- State the random variable.
- Find the probability that a person has a mathematics SAT score over 700.



R: `1-pnorm(700, 514, 117)`
`pnorm(700, 514, 117, lower.tail = FALSE)`

Example

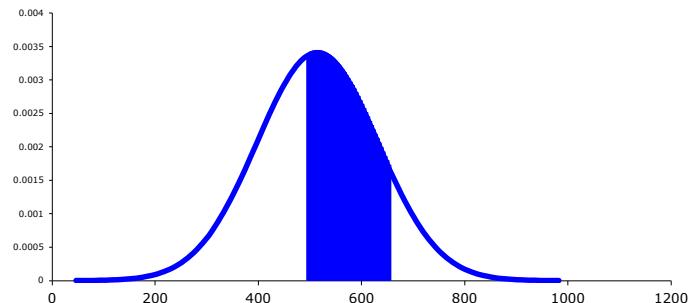
(c) Find the probability that a person has a mathematics SAT score of less than 400.



R: `pnorm(400, 514, 117)`

Example

(d) Find the probability that a person has a mathematics SAT score between a 500 and a 650.



R: `pnorm(650, 514, 117)-pnorm(500, 514, 117)`

Calculation Question

If Z is a **standard** normal variable, find the probability that Z lies between 0.7 and 1.98.

- A. 0.2175
- B. -0.2181
- C. 1.7341
- D. 0.2181

R: `pnorm(1.98,0,1)-pnorm(0.7,0,1)`

From Probabilities to Quantiles: X in Reverse

- Definition. Let X be a continuous random variable and p be a probability. Then the **p th quantile** of X , denoted by Φ_p , is the value such that

$$P(X \leq \Phi_p) = p$$

- Percentiles are special quantiles.

From Probabilities to Quantiles: X in Reverse

Let X be a normal random variable with mean μ and variance σ^2 . Let p_1 and p_2 be two probabilities. There are two typical problems

- $P(X < ?) = p_1$
- $P(X > ?) = p_2$ (**Note:** $P(X \leq ?) = 1 - p_2$)

See Handout using R.

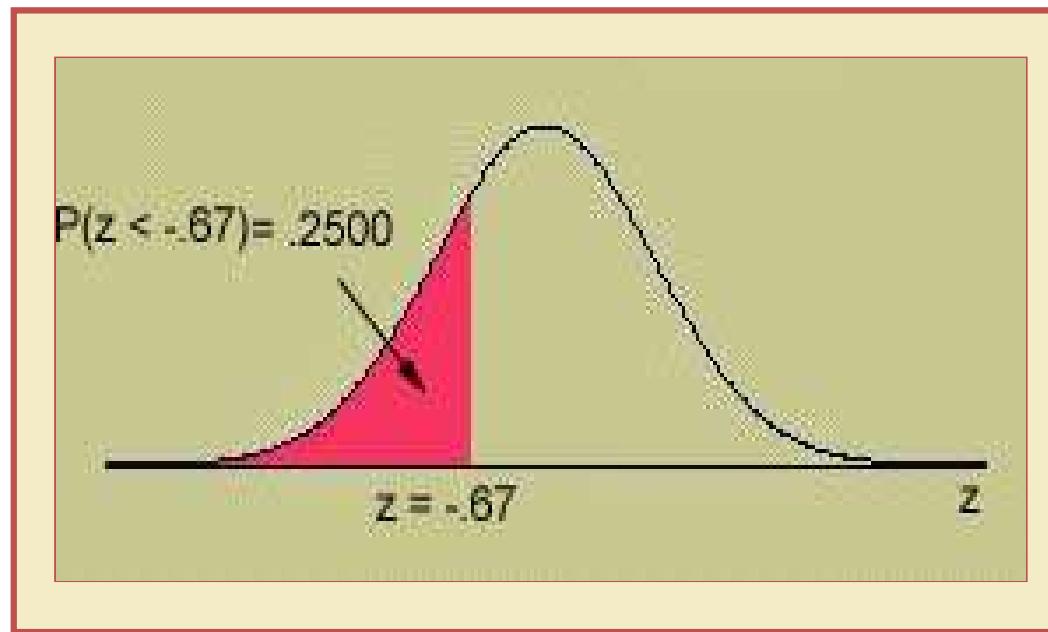
From Probabilities to Quantiles: X in Reverse

<https://esumath.shinyapps.io/Rstats>

<https://istats.shinyapps.io/NormalDist/>

Example 1

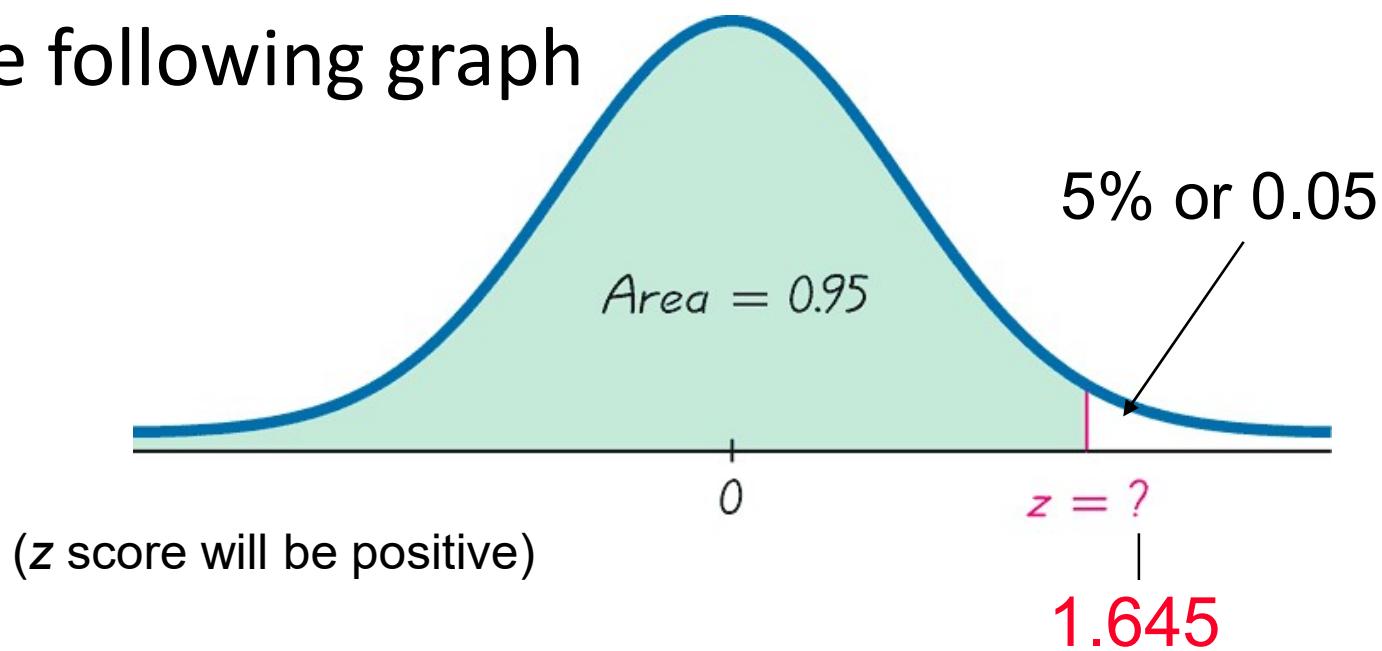
Find the value of Z that has area 0.25 to its left.



R: `qnorm(0.25,0,1)`

Example 2

Find the Z-score in
the following graph



It is the 95th Percentile

R: `qnorm(0.95,0,1)`

Critical Value

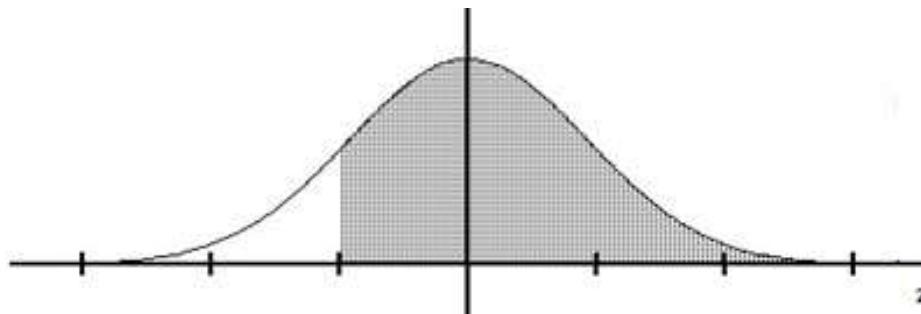
Definition. For the standard normal distribution, a **critical value, denoted by z_α** , is a z score such that the area of α (*a small probability*) is to its **right**.

Some special z-scores:

- $z_{0.05} = 1.645$.
- $z_{0.025} = 1.96$.
- $z_{0.005} = 2.576$.

Calculation Question

Find the indicated z-score. Shaded area is 0.8599.



- A. 0.8051
- B. 0.5557
- C. -1.08
- D. 1.08

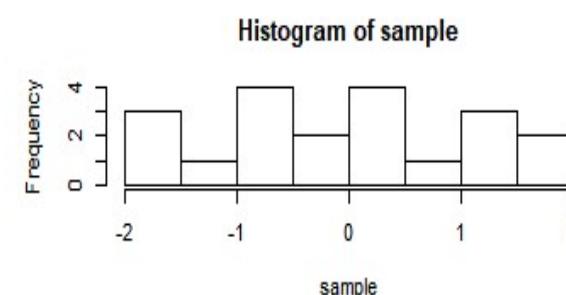
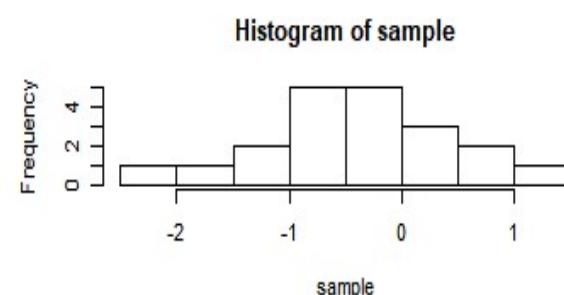
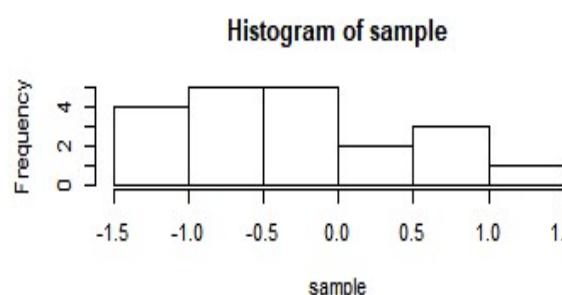
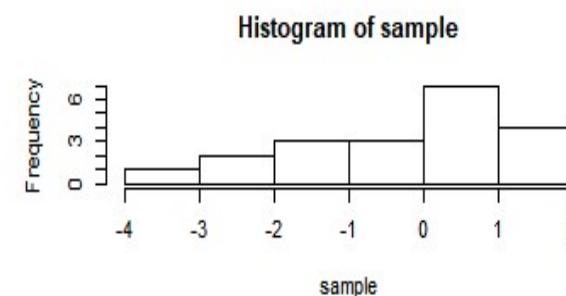
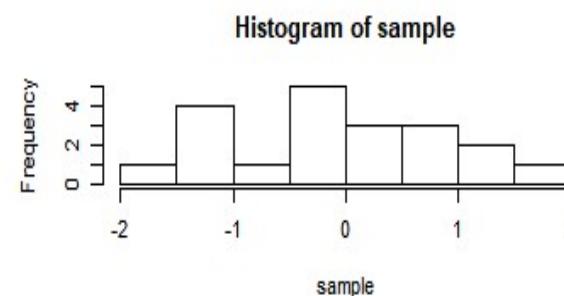
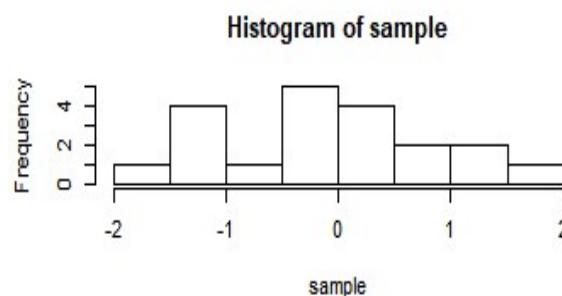
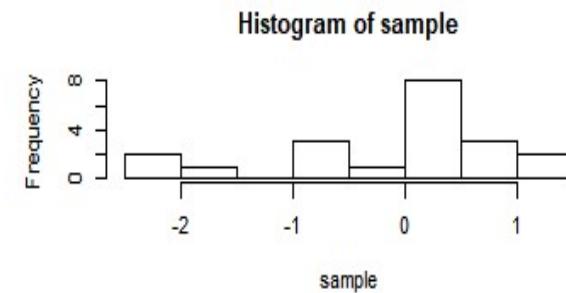
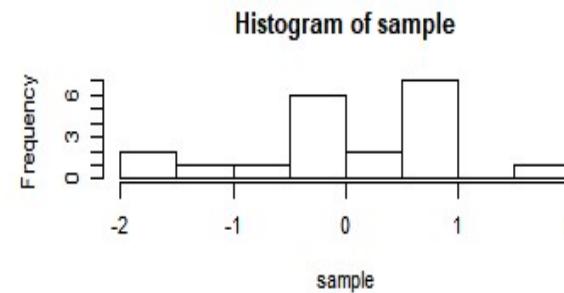
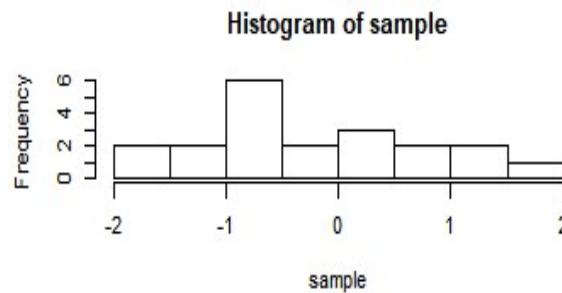
6.4 Assessing Normality

- This section presents criteria for determining whether the requirement of a normal distribution is satisfied.
- The criteria involves visual inspection of a histogram to see if it is roughly bell shaped, identifying any outliers, and constructing a graph called a **normal quantile plot**.

Procedure for Determining Whether It Is Reasonable to Assume that Sample Data Are from a Normally Distributed Population

1. **Histogram:** Construct a histogram. Reject normality if the histogram departs **dramatically** from a bell shape.
2. **Outliers:** Identify outliers. Reject normality if there is more than one outlier present.
3. **Normal Quantile Plot:** If the histogram is basically symmetric and there is at most one outlier, use technology to generate a **normal quantile plot**.

Histograms of 9 samples of size 20 from standard normal



Histograms of samples of size 20 from standard normal

R code:

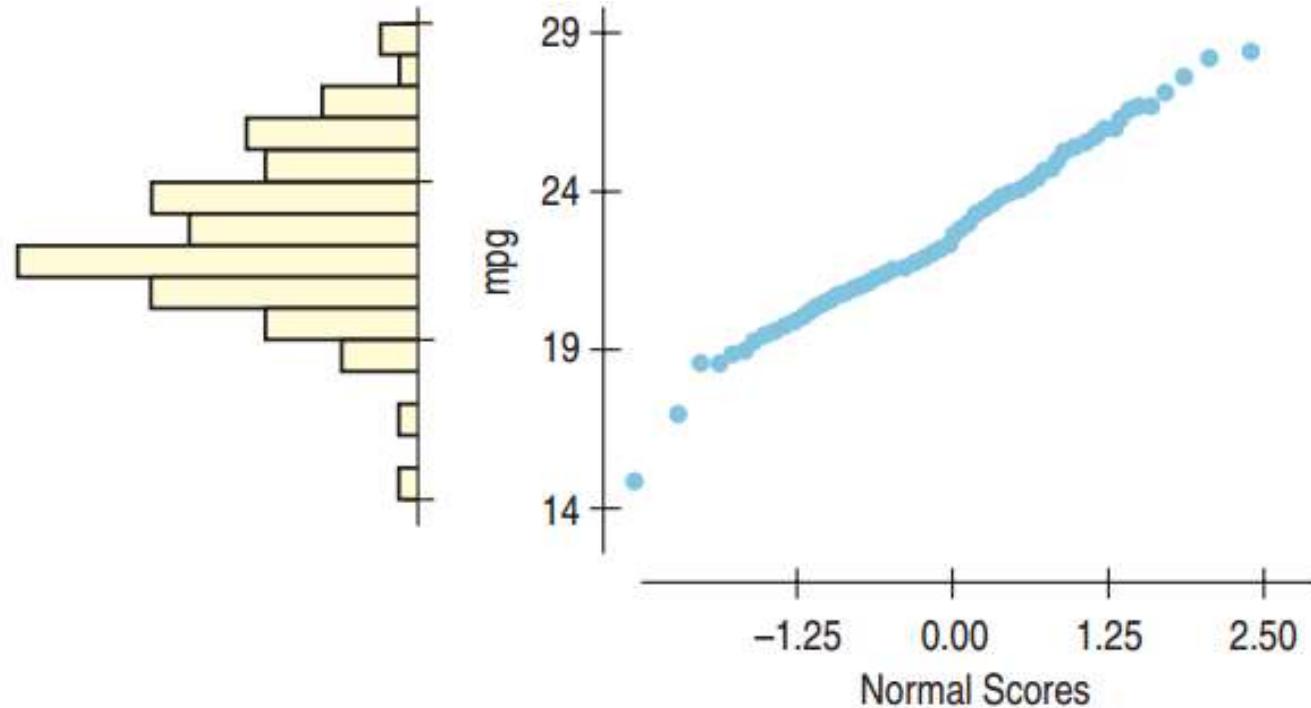
```
par(mfrow=c(3,3));
for (i in 1:9)
{
  sample=rnorm(n=20, mean = 0, sd = 1);
  #a sample of size 20 from standard normal
  hist(sample);
}
```

Checking Normality Using Normal Quantile Plot

Use the following criteria to determine whether or not the distribution is normal.

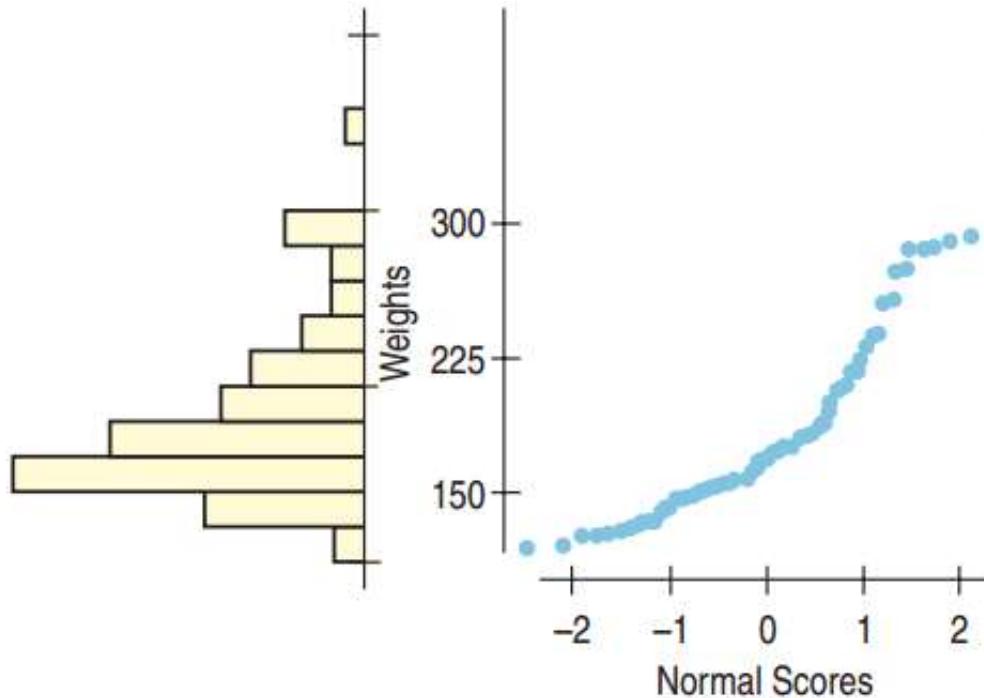
- **Normal Distribution:** The population distribution is normal if the pattern of the points is reasonably close to a straight line and the points do not show some systematic pattern that is not a straight-line pattern.
- **Not a Normal Distribution:** The population distribution is not normal if either or both of these two conditions applies:
 - The points do not lie reasonably close to a straight line.
 - The points show some systematic pattern that is not a straight-line pattern.

The Normal Model Applies



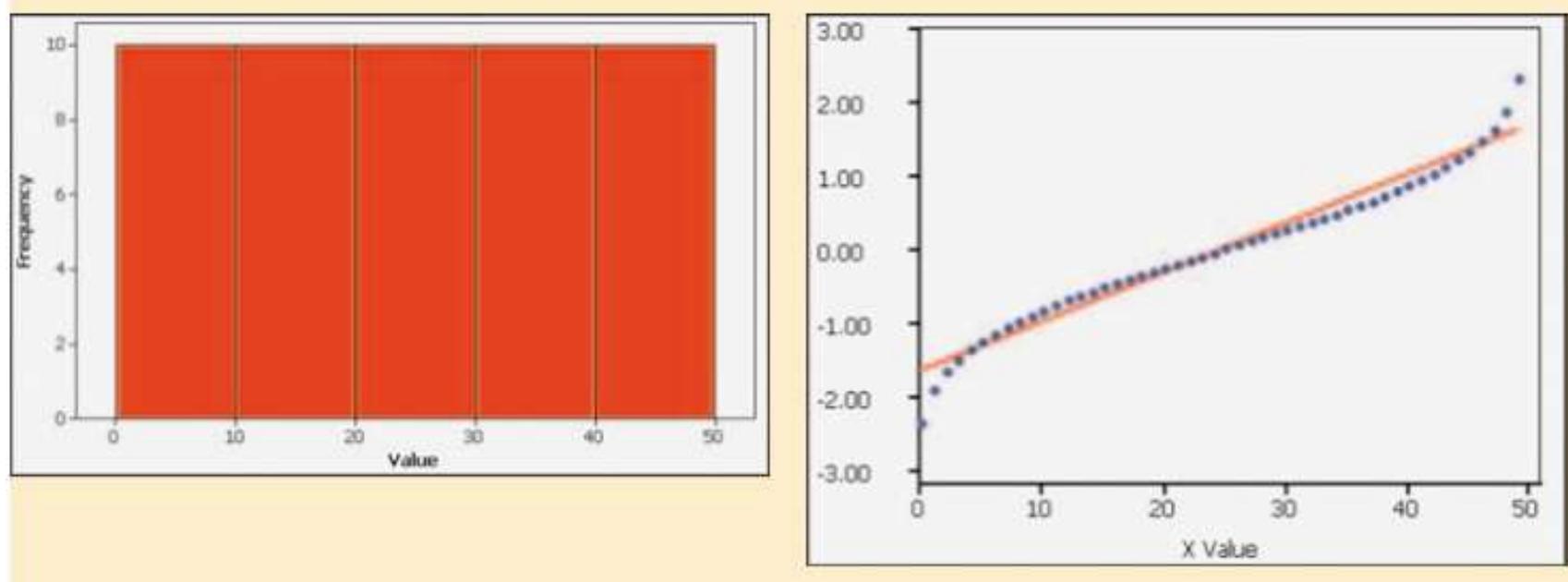
- The Normal probability plot is **nearly straight**, so the Normal model applies. Note that the histogram is unimodal and somewhat symmetric.

The Normal Model Does Not Apply



- The Normal probability **plot is not straight**, so the Normal model does not apply applies. Note that the histogram is skewed to the right.

The Normal Model Does Not Apply



Uniform: Histogram of data having a uniform distribution. The corresponding normal quantile plot suggests that the points are not normally distributed because the points show a systematic pattern that is not a straight-line pattern. These sample values are not from a population having a normal distribution.

Normal Quantile Plot Using Technology

R: `qqnorm(data)`

Combining Random Variables

Let c be a constant and X a random variable

(1) $E(X + c) = E(X) + c, E(X - c) = E(X) - c$

$$\text{Var}(X+c) = \text{Var}(X), \text{SD}(X+c) = \text{SD}(X).$$

(2) $E(cX) = cE(X), \text{Var}(cX) = c^2\text{Var}(X), \text{SD}(cX) = |c|\text{SD}(X)$

(3) $E(X \pm Y) = E(X) \pm E(Y)$

(4) If X and Y are **independent**, then

$$\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y)$$

- These properties will be used in **statistical inference**.

Key Concepts

- Calculation of Normal probabilities
- Find the normal quantile for given left/right tail probabilities
- Assessing Normality using Histogram and normal QQ-plot

6.5 Sampling Distribution and the Central Limit Theorem

Objectives

- Understand the concept of a **sampling distribution of a statistic**, which is the distribution of all values of that statistic when all possible samples of the same size are taken from the same population.
- The **Central Limit Theorem** tells us that for a population with **any** distribution, the distribution of the sample means approaches a normal distribution as the sample size increases.

Sampling Distributions

- Recall: **Parameters** are numerical descriptive measures for populations.
 - For the normal distribution, the location and shape are described by μ and σ .
 - For a **binary** population, the parameter is the population proportion p .
- Often the **values of parameters** that specify the exact form of a distribution are **unknown**.
- We must rely on the **sample** to learn about these parameters.

Sampling Distributions

- Numerical descriptive measures calculated from the sample are called **statistics**.
- Statistics **vary** from sample to sample and hence are random variables.
- The probability distributions for statistics are called **sampling distributions**. In repeated sampling, they tell us **what values** of the statistics can occur and **how often** each value occurs.
- **Definition.** The **sampling distribution of a statistic** is the distribution of all values of the statistic when all possible samples of the **same size n** are taken from the **same population**.

Definitions

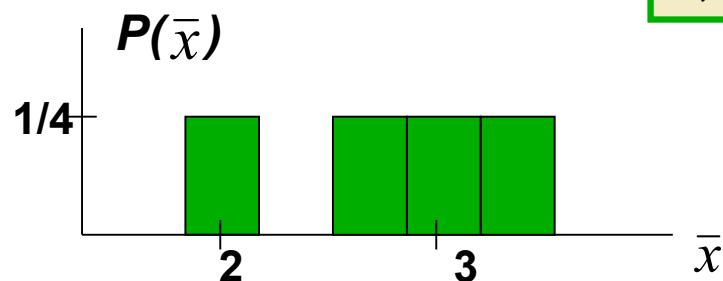
- ✓ The **sampling distribution of the sample mean** is the distribution of **all possible** sample means, with all samples having the **same sample size n** taken from the **same** population.
- ✓ The **sampling distribution of the sample proportion** is the distribution of **all possible** sample proportions, with all samples having the **same sample size n** taken from the **same** population.
- ✓ The **sampling distribution of the sample variance** is the distribution of **all possible** sample variances, with all samples having the **same sample size n** taken from the **same** population (not to be discussed).

Example 1

Assume that we have a population of small size $N=4$.

Population: 3, 5, 2, 1
Draw samples of size $n = 3$
without replacement

<u>Possible samples</u>	\bar{x}
3, 5, 2	$10/3 = 3.33$
3, 5, 1	$9/3 = 3$
3, 2, 1	$6/3 = 2$
5, 2, 1	$8/3 = 2.67$



Each value of x -bar
is equally likely, with
probability $1/4$

Example 2

Assume that we have a binary population of small size $N=5$: $x_1=0, x_2=0, x_3=1, x_4=1, x_5=1$. Thus, $p=3/5=0.6$.

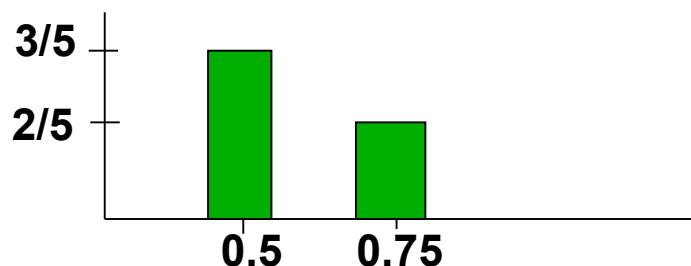
Population: 0, 0, 1, 1, 1

Draw samples of size $n = 4$
without replacement

Possible samples \hat{p}

0, 0, 1, 1	$2 / 4 = 0.5$
0, 0, 1, 1	$2 / 4 = 0.5$
0, 0, 1, 1	$2 / 4 = 0.5$
0, 1, 1, 1	$3 / 4 = 0.75$
0, 1, 1, 1	$3 / 4 = 0.75$

$P(\hat{p})$



A Simulation Study - Sampling Distributions of Sample Means

Consider repeating this process: **Roll a balanced die 5 times.** Find the mean of the five numbers, \bar{x} .

What do we know about the behavior of **all** sample means that are generated as this process **continues indefinitely?**

Note.

- The **population** is the results of “rolling the die infinitely many times”.
- We have 1,000 samples of size 5 if we continue the process 1,000 times.

A Simulation Study - Sampling Distributions of Sample Means (discrete population)

- Let X be the random variable of the results of rolling a balanced die.

x	$P(x)$
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

$$\mu = \sum xP(x) = 21/6 = 3.5$$

$$\sigma \approx 1.708$$

See also https://istats.shinyapps.io/SampDist_discrete/

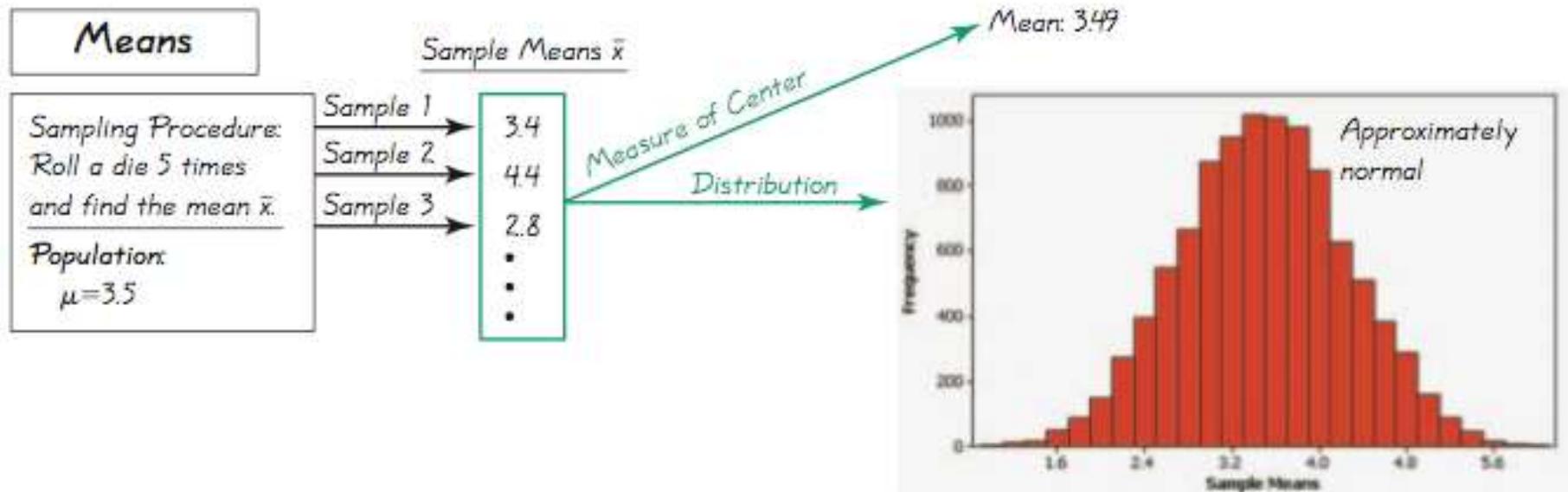
A Simulation Study - Sampling Distributions of Sample Means (discrete population)

- R code

```
iter = 10000;
n=50;    #sample size
means=numeric(); #a vector to store sample means
for (i in 1:iter)
{
  x=sample(1:6, n, replace=T); #roll the die 5 times
  means[i]=mean(x); #sample mean of each
}
hist(means, breaks=20); #histogram of the 10000 sample means
mean(means);    #average of the 10000 sample means
sd(means);      #standard deviations of the 10000 sample means
1.708/sqrt(n);
```

A Simulation Study - Sampling Distributions of Sample Means (discrete population)

Specific results from 10,000 trials



The population mean is 3.5; the mean of the 10,000 trials is 3.49. If continued **indefinitely**, the sample mean will be 3.5. Also, notice the distribution is “normal.”

A Simulation Study - Sampling Distributions of Sample Means (continuous population)

[https://istats.shinyapps.io/sampdist cont/](https://istats.shinyapps.io/sampdist_cont/)

(try after class)

Unbiased Estimators

- A statistic is an **unbiased** estimator of the population parameter if the **mean** (expected value) of the statistic is equal to the true value of the population parameter.
- ❖ Sample means, variances and proportions are **unbiased estimators**.
These statistics are better compared with other estimators in estimating the population parameters.

Summary: The Central Limit Theorem

Given:

1. The population has a distribution (which usually is NOT normal) with mean μ and standard deviation σ .
2. Simple random samples **all of size n** are selected from the population.

Conclusions:

The distribution of sample mean \bar{x} will, if the sample size n is large or the population is close to Normal, approach a **normal** distribution with

- mean μ
- standard deviation $\frac{\sigma}{\sqrt{n}}$

Summary: The Central Limit Theorem

- Requirements
 - Independent
 - Randomly collected sample
- The sampling distribution of the means is close to Normal if either:
 - Large sample size or
 - Population close to Normal

Summary: The Central Limit Theorem

- What does “central limit” mean?
 - **Central:** all sample means (of size n) have the population mean μ as the center; or the sample mean is an unbiased statistic for μ .
 - **Limit:** If the population is not normal (which is generally true), the sample means have a normal distribution only if the sample size n is **large**.

How Large is Large?

1. If the sample is from a Normal population, then the sampling distribution of \bar{x} will also be exactly normal, no matter what the sample size is.
2. If the population is unimodal and symmetric, a fairly small sample is okay ($15 \leq n \leq 30$).
3. For samples of size n larger than 30, the distribution of the sample (from any population) means can be approximated reasonably well by a normal distribution. The approximation becomes closer to a normal distribution as the sample size n becomes larger.

How Large is Large?

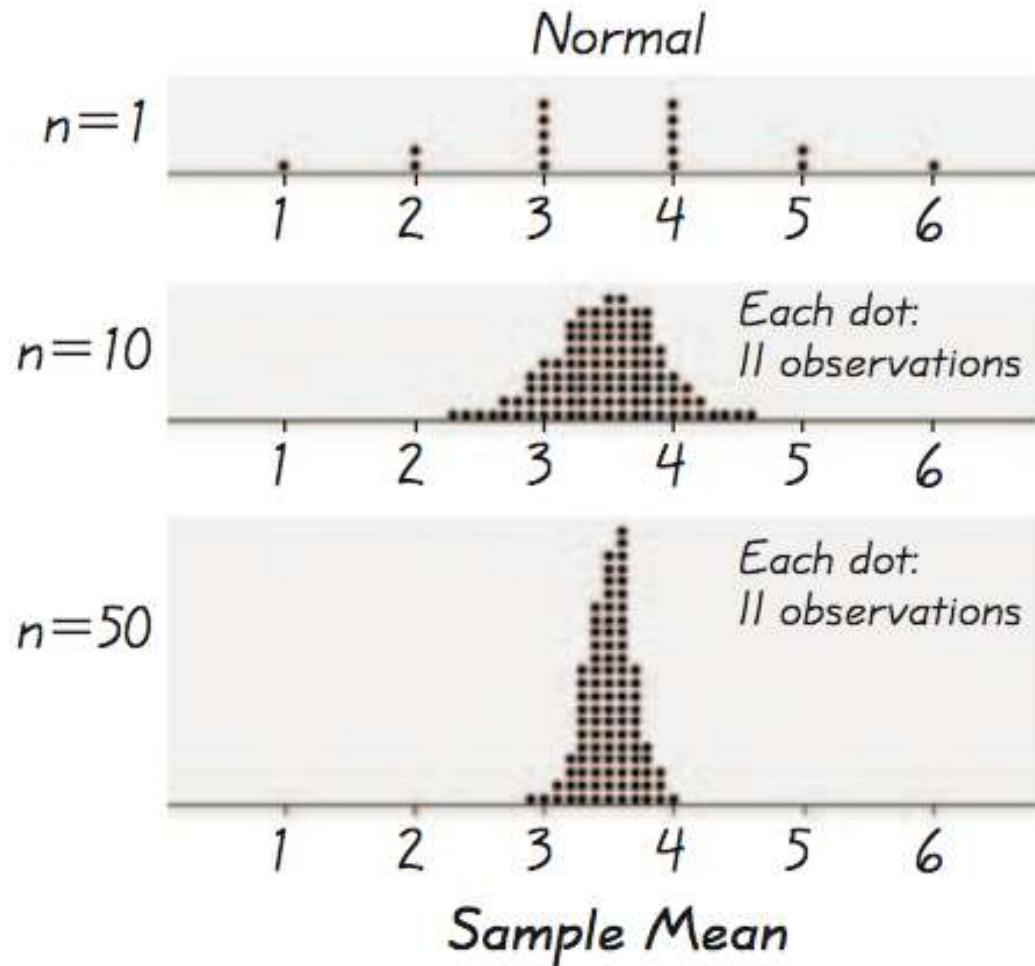
- Population Distribution

Sampling Distribution for the Means

- Normal → Normal (**any sample size**)
- Uniform → Normal (large sample size)
- Bimodal → Normal (larger sample size)
- Skewed → Normal (larger sample size)

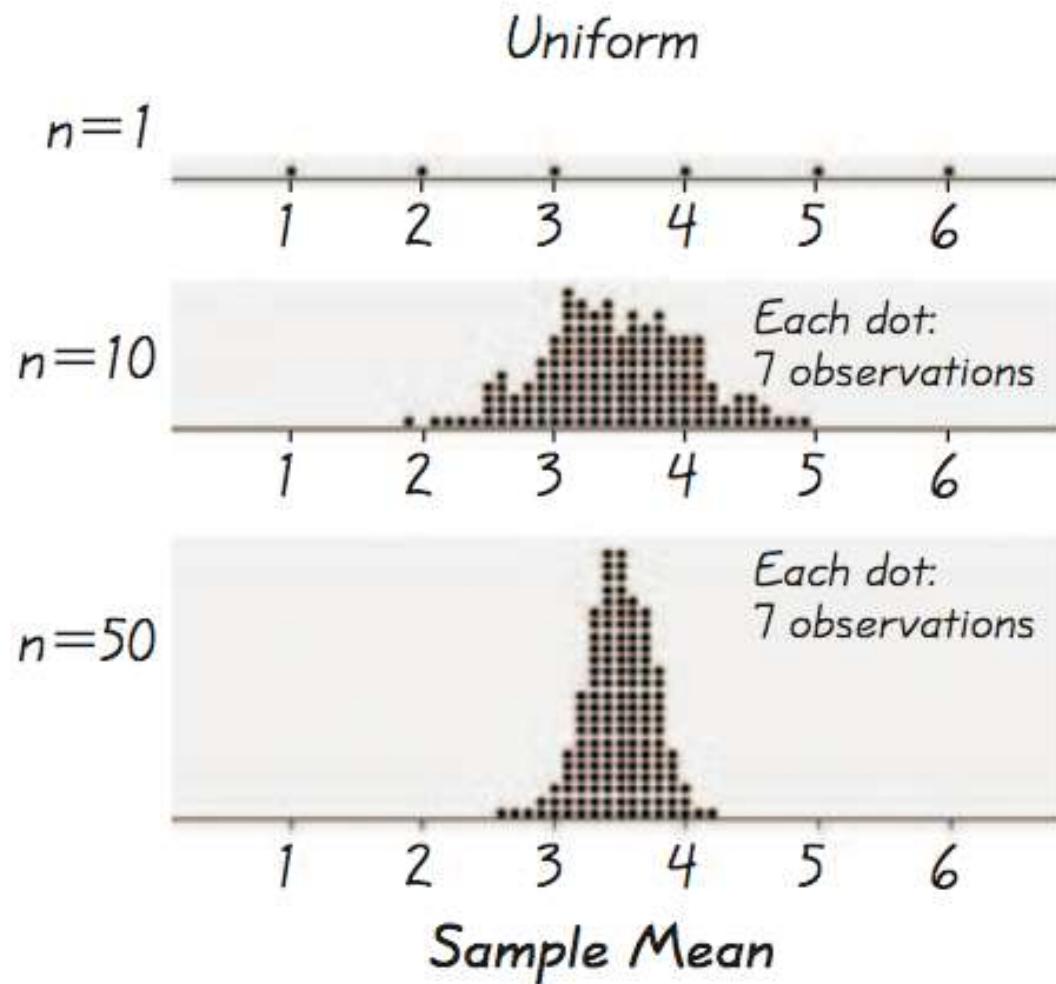
Example 1- Normal Distribution

As we proceed from $n = 1$ to $n = 50$, we see that the distribution of sample means is approaching the shape of a normal distribution.



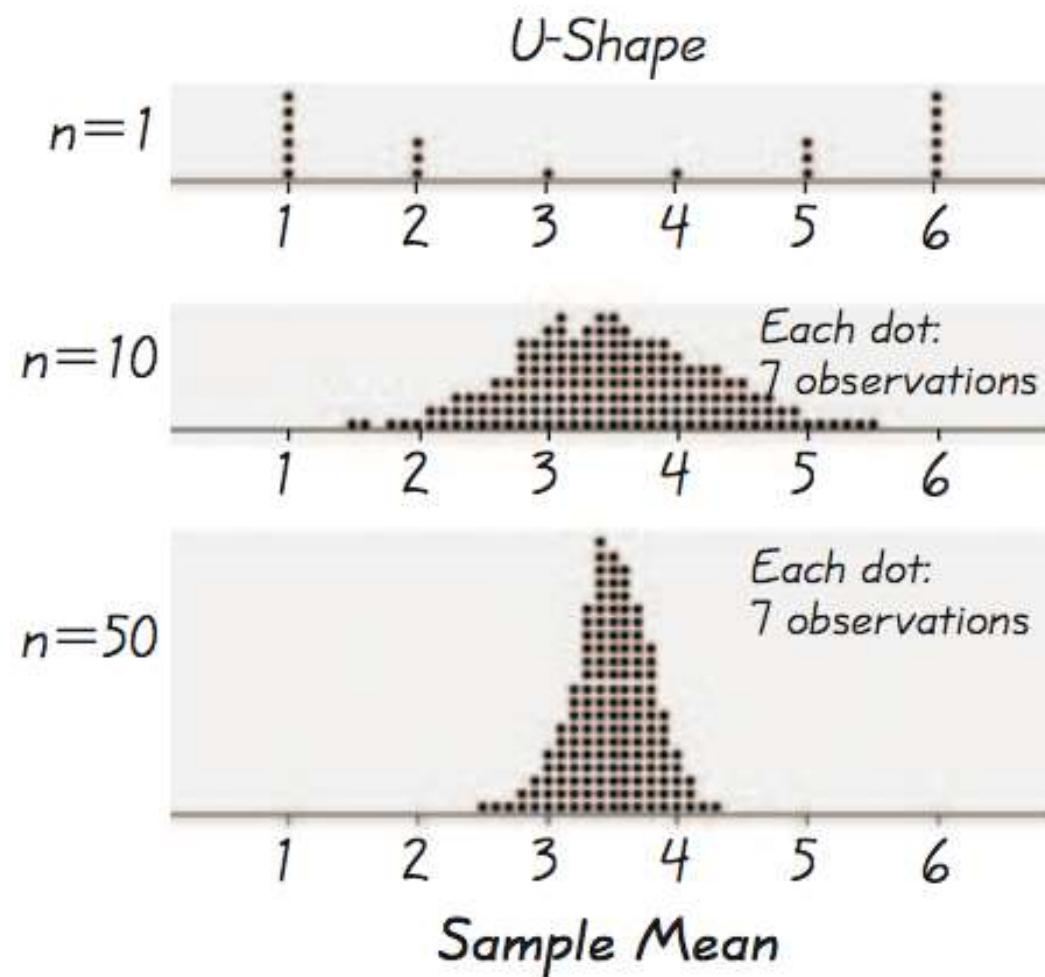
Example 2 - Uniform Distribution

As we proceed from $n = 1$ to $n = 50$, we see that the distribution of sample means is approaching the shape of a normal distribution.

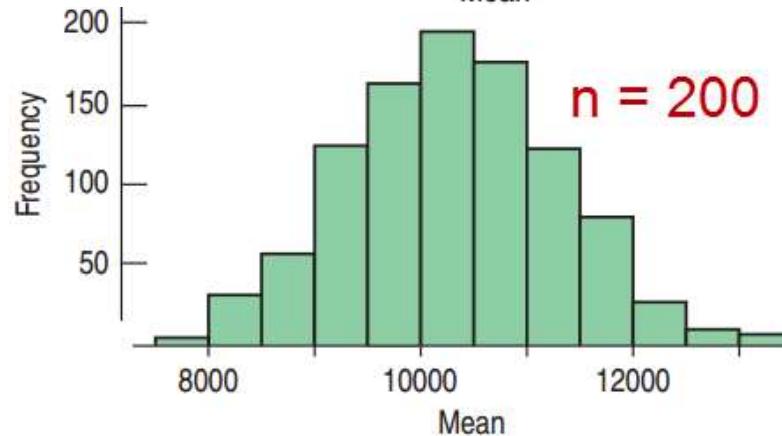
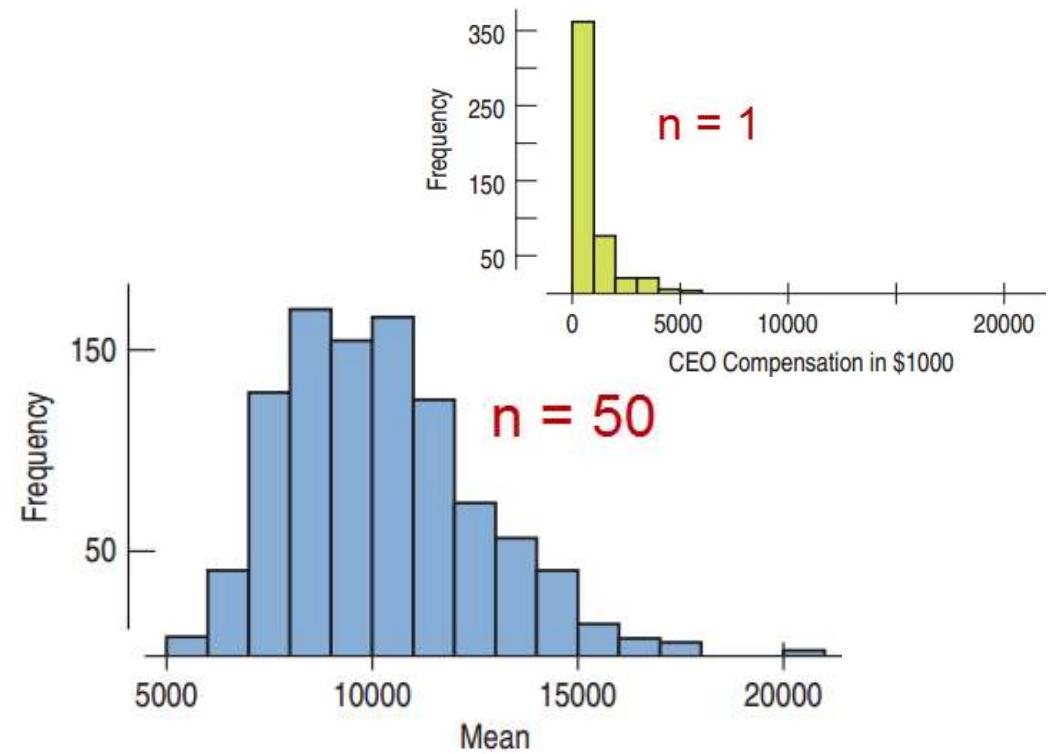
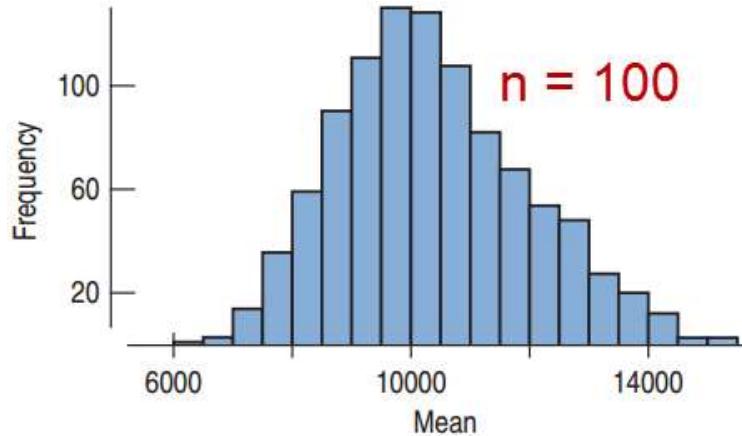
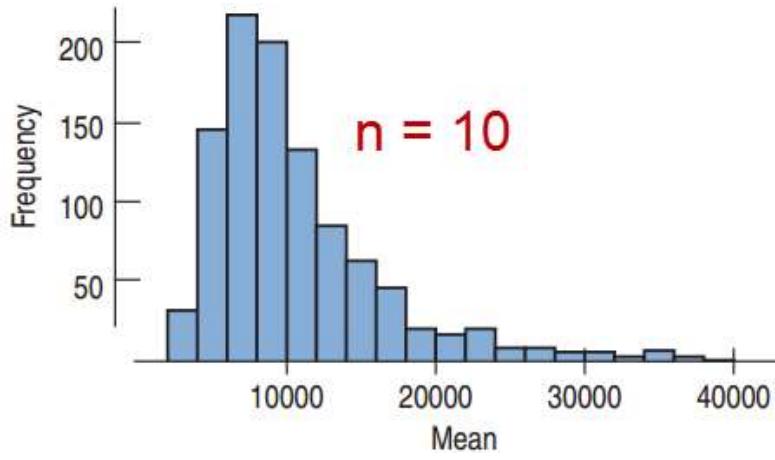


Example 3 - U-Shaped Distribution

As we proceed from $n = 1$ to $n = 50$, we see that the distribution of sample means is approaching the shape of a normal distribution.



Example 4- skewed Distribution



Calculation Question

One truck from Lakeland Trucking, Inc. can carry a load of 2959.2 lb. Records show that the weights of boxes that it carries have a mean of 78 lb and a standard deviation of 12 lb. For samples of size 36, find the mean and standard deviation of the **sample mean**, respectively.

- A. 78; 12
- B. 78; 2

Example

- 14. Groceries** A grocery store's receipts show that Sunday customer purchases have a skewed distribution with a mean of \$32 and a standard deviation of \$20.
- Explain why you cannot determine the probability that the next Sunday customer will spend at least \$40.
 - Can you estimate the probability that the next 10 Sunday customers will spend an average of at least \$40? Explain.
 - Is it likely that the next 50 Sunday customers will spend an average of at least \$40? Explain.
 - Suppose the store had 312 customers this Sunday. Estimate the probability that the store's revenues were at least \$10,000.

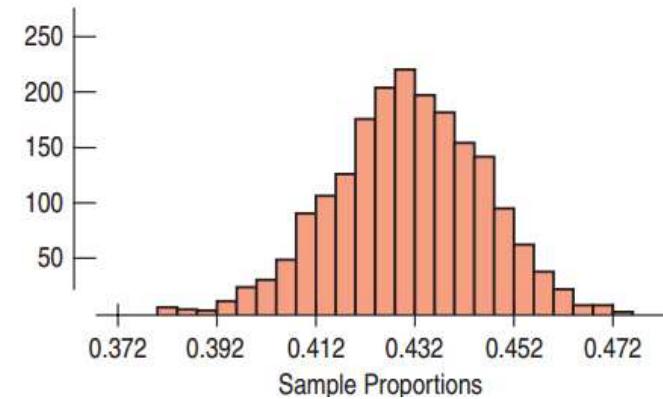
What Can Go Wrong?

- Don't confuse the **sampling distribution** with the distribution of the sample data.
 - A histogram of the data shows the sample's distribution. The sampling distribution is more theoretical.
- The sample must be random. Beware of observations that are **not independent**.
 - The CLT fails for dependent samples. A good survey design can ensure independence.
- Watch out for **small samples** from skewed or bimodal populations.
 - The CLT requires large samples or a Normal population or both.

Introductory Example: Sampling About Evolution



- According to a Gallup poll, 43% believe in evolution. Assume this is true of all Americans.
 - If many surveys were done of 1007 Americans, we could calculate the sample proportion for each.
 - The histogram shows the distribution of a simulation of 1000 sample proportions.
 - It seems that the sampling distribution of the sample proportions is bell-shaped.



Simulation Study without coding - Sampling Distributions of Sample Proportions

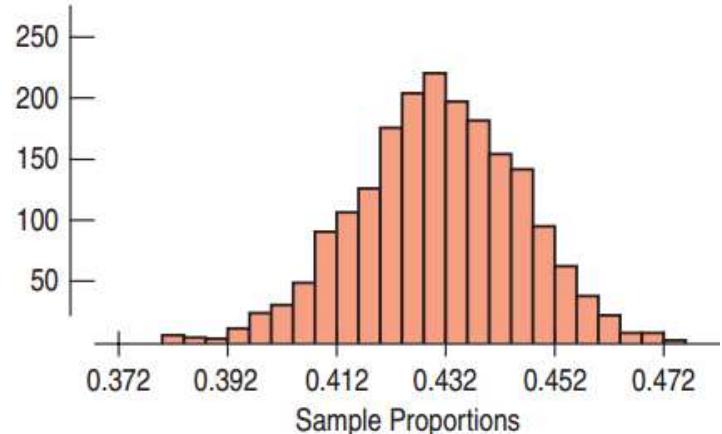
https://istats.shinyapps.io/SampDist_Prop/

Simulation Study with coding - Sampling Distributions of Sample Proportions

- R code

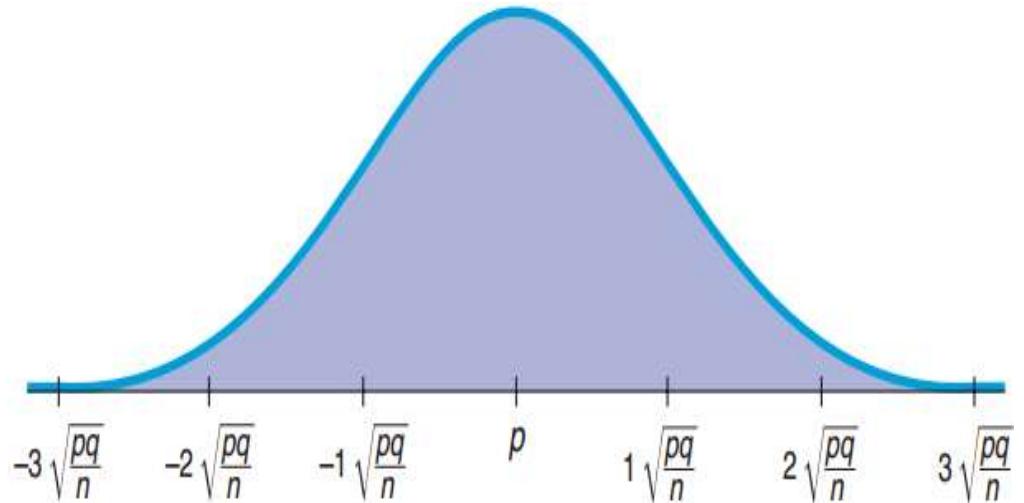
```
p=0.43;          #population proportion  
iter = 1000;  
n=1007;          #sample size  
ps=numeric();    #a vector to store sample proportions  
for (i in 1:iter)  
{  
  x=rbinom(n, size=1, prob=p); # survey 1007 people  
  ps[i]=mean(x);            #sample proportion of each sample  
  print(ps[i]);  
}  
hist(ps, breaks=20); #histogram of the 1000 sample means  
mean(ps);           #average of the 1000 sample means  
sd(ps);            #standard deviation of the 1000 sample means
```

Sampling Distribution for Proportions



- For categorical data (two categories) – Yes/No, etc.
- Symmetric
- Unimodal
- Centered at p
- The **sampling distribution** follows the **Normal model**

The CLT for the Sample Proportion



✓ The sampling distribution of $\hat{p} = x / n$ is approximately Normal when $np \geq 10$, $n(1-p) \geq 10$, with

- mean p and
- standard deviation $\sqrt{p(1-p)/n}$

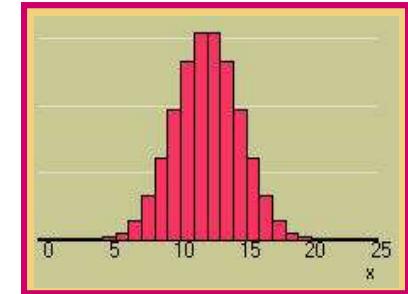
- In summary, $\hat{p} \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$, $q=1-p$

Calculation Question

State of Pennsylvania has an unemployment rate of 6% in July 2015. One survey select a sample of size 200 from PA. What is the value of the standard deviation of the sample proportion of unemployment?

- A. 0.017
- B. 0.237
- C. 0.0002

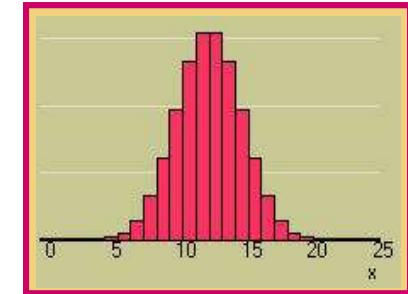
Remark: The CLT for the Sample Proportion is Derived from the CLT for Sample Mean



- Consider a random sample selected from a binary population such that

$$P(x_i=1)=p \text{ and } P(x_i=0)=1-p, \quad i=1, 2, \dots, n.$$

Remark: The CLT for the Sample Proportion is Derived from the CLT for Sample Mean



- If n is large ($np \geq 10$ and $n(1-p) \geq 10$ or $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$ if p is unknown, are required), we have
 - ✓ By the Central Limit Theorem, the sample proportion ($= (x_1+x_2+\dots+x_n)/n$) is approximately Normal, with mean p and standard deviation $\sqrt{p(1-p)/n}$.

Example

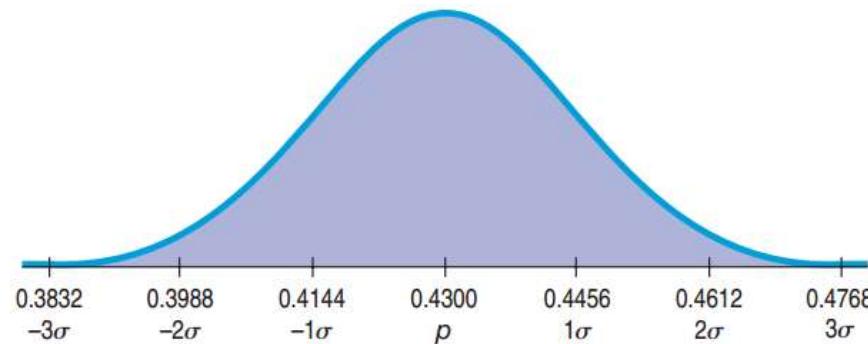
2. **Marketing** The proportion of adult women in the United States is approximately 51%. A marketing survey telephones 400 people at random.
- What proportion of the sample of 400 would you expect to be women?
 - What would the standard deviation of the sampling distribution be?
 - How many women, on average, would you expect to find in a sample of that size?

Example: The Normal Model for Evolution

- Population: $p = 0.43$, $n = 1007$. Sampling Distribution:

- Mean = 0.43

- Standard deviation = $\sigma(\hat{p}) = \sqrt{\frac{(0.43)(0.57)}{1007}} \approx 0.0156$



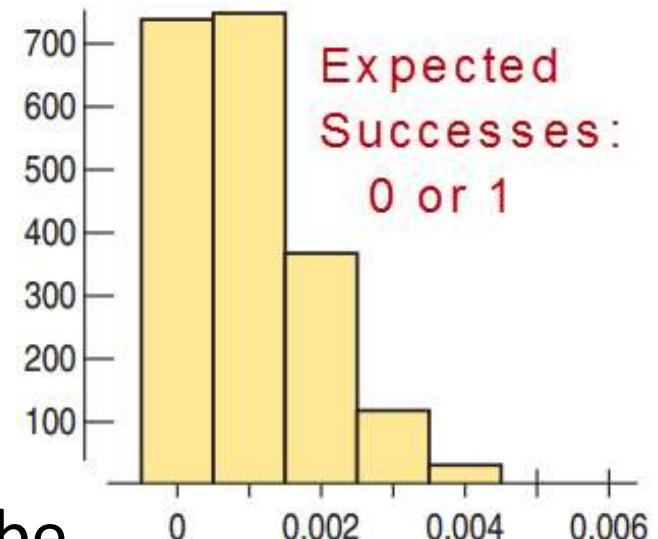
When Does the Normal Model Work?

- Success Failure Condition:

$np \geq 10, nq \geq 10$ There must be at least 10 expected successes and failures, where $q=1-p$.

- Independent trials: Check for the Randomization Condition.

- 10% Condition: Sample size must be less than 10% of the population size.



Example

6. **Campus sample** For her final project, Stacy plans on surveying a random sample of 50 students on whether they plan to go to Florida for Spring Break. From past years, she guesses that about 10% of the class goes. Is it reasonable for her to use a Normal model for the sampling distribution of the sample proportion? Why or why not?

Example

- 28. Contacts** Assume that 30% of students at a university wear contact lenses.
- We randomly pick 100 students. Let \hat{p} represent the proportion of students in this sample who wear contacts. What's the appropriate model for the distribution of \hat{p} ? Specify the name of the distribution, the mean, and the standard deviation. Be sure to verify that the conditions are met.
 - What's the approximate probability that more than one third of this sample wear contacts?

Key Concepts

- Sampling distribution: statistics are random variables in repeated samplings
- Unbiased estimators
- Central Limit Theorem for sample means
- Central Limit Theorem for sample proportions

Correlation and Regression

The material is covered by Section 10.1 and 10.2 of the book “Statistics Using Technology”.

- ✓ Correlation
- ✓ Regression

Bivariate Data

- When two variables are measured (not always but usually on a single unit), the resulting data are called bivariate data. We denote one variable by X , and the second variable as Y .
- Suppose both X and Y are quantitative or numerical.
- We consider a sample data set of size n

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

which is taken from a **bivariate population**.

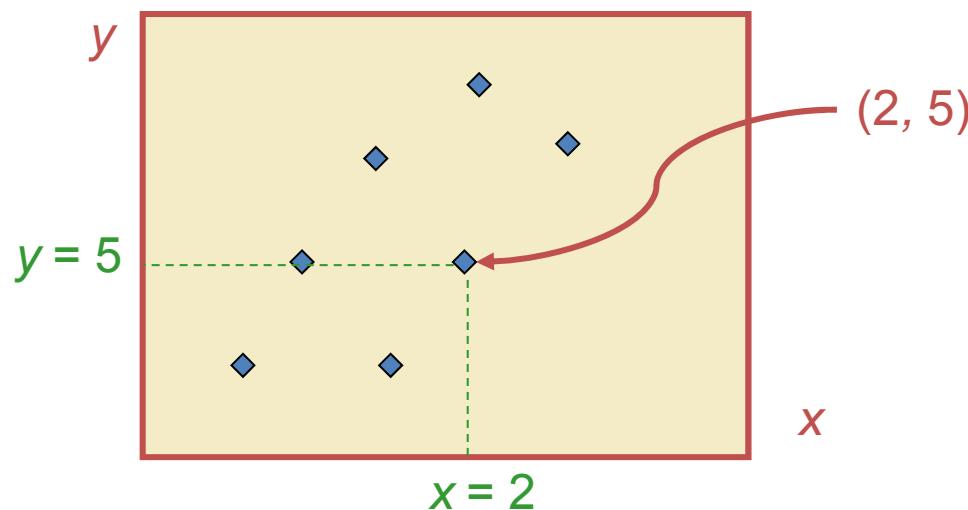
- We are interested in making inferences about the relationship between X and Y based on the sample information.
- We describe the sample data only in this course.

Section 1: Scatterplots, Association, and Correlation

- When two variables are measured (not always but usually on a single unit), the resulting data are called bivariate data (or Paired data).
- We can describe each variable individually, and we can also explore the relationship between the two variables.
- Bivariate data can be described with
 - Graphs
 - Numerical Measures

Scatterplot of Two Quantitative Variables

When both of the variables are **quantitative**, denote one variable x (independent or explanatory variable) and the other y (dependent or response variable.). A single measurement is a pair of numbers (x, y) that can be plotted using a two-dimensional graph called a **scatterplot**.



What can we see from a scatter plot?

https://istats.shinyapps.io/Association_Quantitative/

- Scatterplots exhibit the relationship between two variables.
- Used for detecting patterns, trends, relationships, and extraordinary values

The Direction of the Association

- Negative Direction: As one goes up, the other goes down.



- Positive Direction: As one goes up, the other goes up also.



- No Direction:

Form

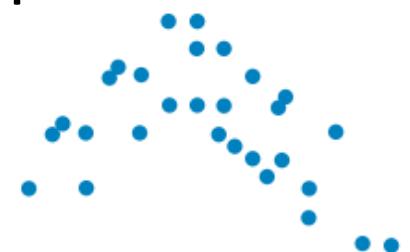
- Linear: The points cluster near a line.



- Gently curves in a direction. May be able to straighten with a transformation.



Curves up and down. Difficult to straighten



Strength of the Relationship

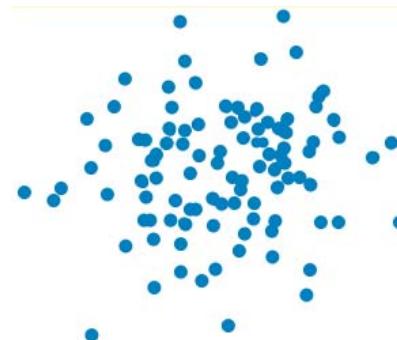
- **Strong Linear Relationship:**



- Moderate Linear Relationship:

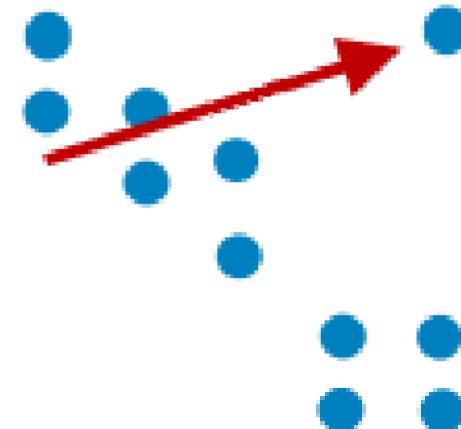


- No Linear Relationship:



Influential Points

- An **influential point** is a point on a scatterplot that stands away from the overall pattern of the scatterplot.
- **Influential points** are **special outliers** which are almost always interesting and always deserves special attention.



Influential Points

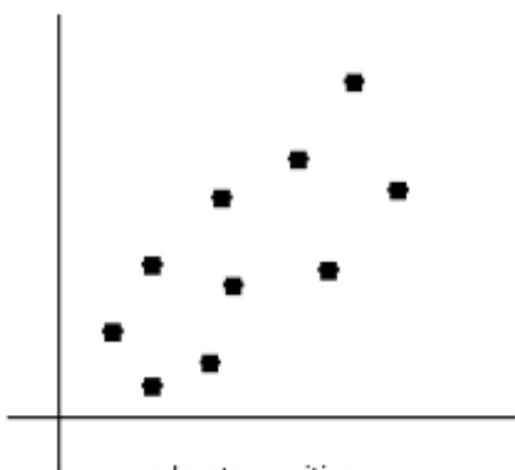
- Activity: draw an influential point

<https://istats.shinyapps.io/ExploreLinReg/>

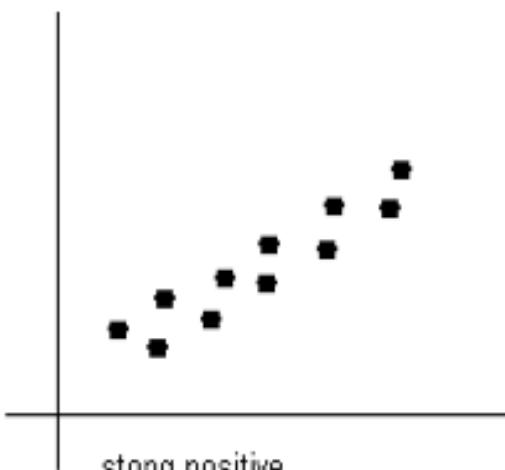
Summary: Describing the Scatterplot

When analyzing a scatterplot, we should look for:

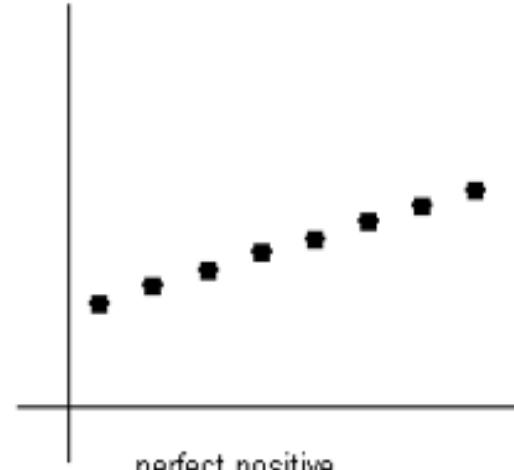
- Association. This is a pattern in the scatterplot.
 - Linear
 - Nonlinear
- Direction of Association
- Strength of Association?
 - Strong, moderate, or weak
- Outliers and influential points?



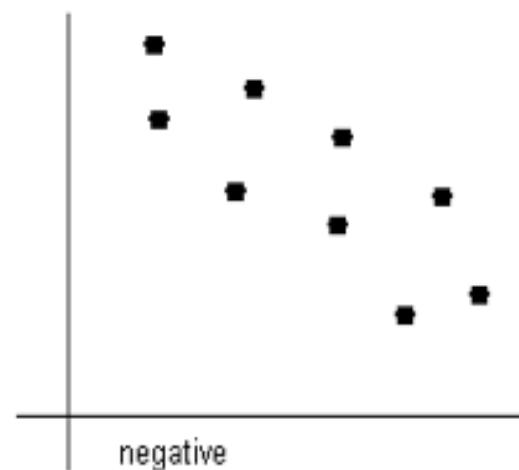
moderate positive



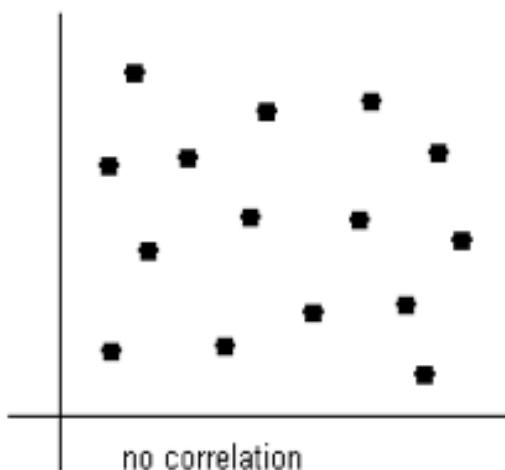
strong positive



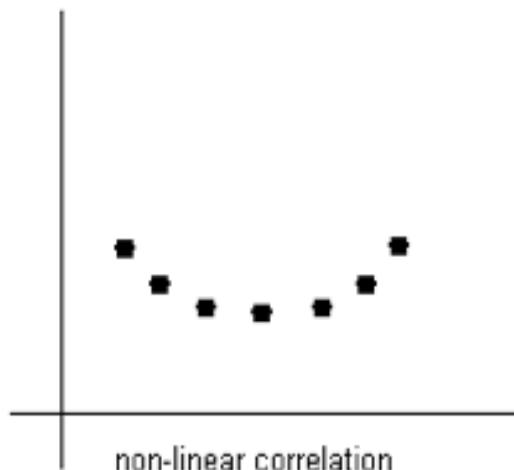
perfect positive
correlation



negative



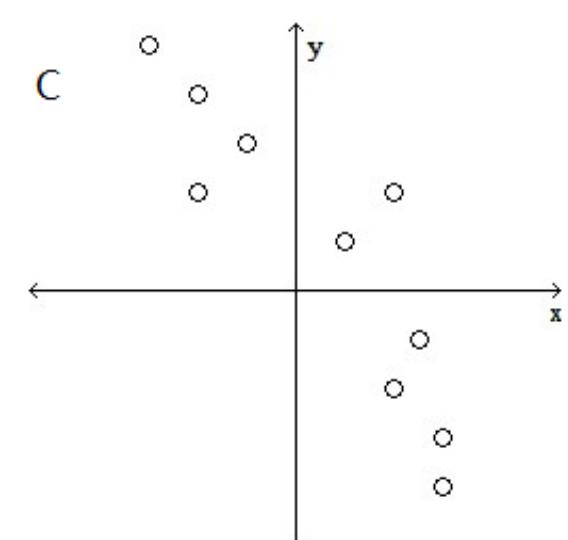
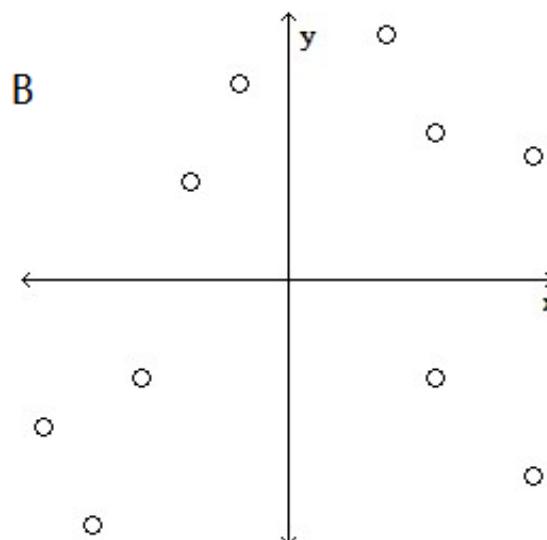
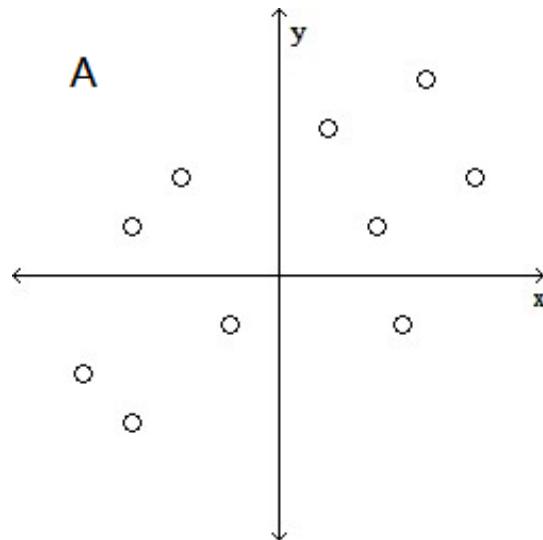
no correlation



non-linear correlation

Concept Question

- Which shows the strongest linear correlation?

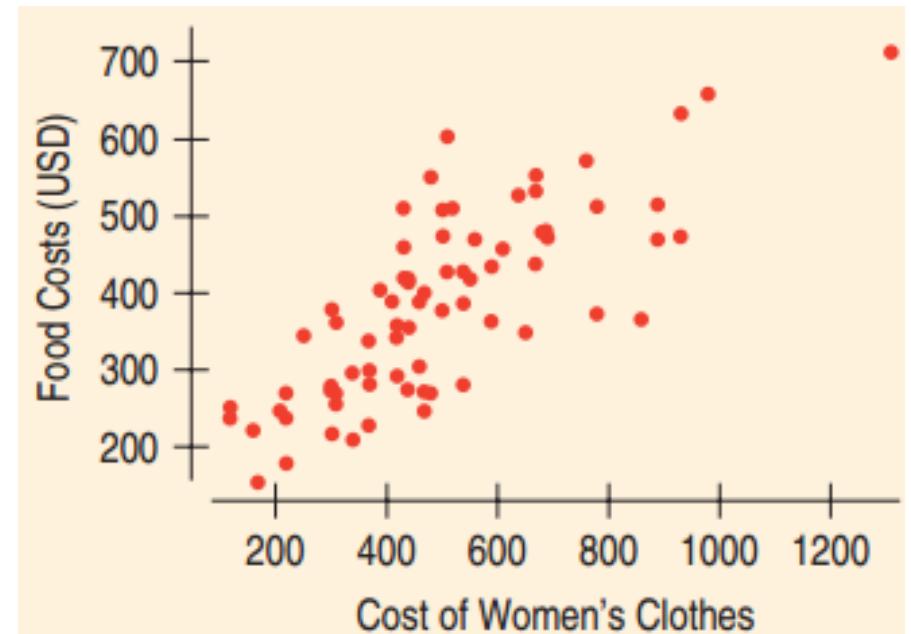


Notation: Roles of Variables

- **Dependent or Response Variable (Y):** The variable of interest. It is what we want to **predict**.
- **Independent, Explanatory or Predictor Variable (X):** The variable that we use to provide information for a prediction of the response variable.
- ❖ Choosing the response variable and the explanatory variable depends on how we think about the problem.

Example 1: Comparing Prices Worldwide

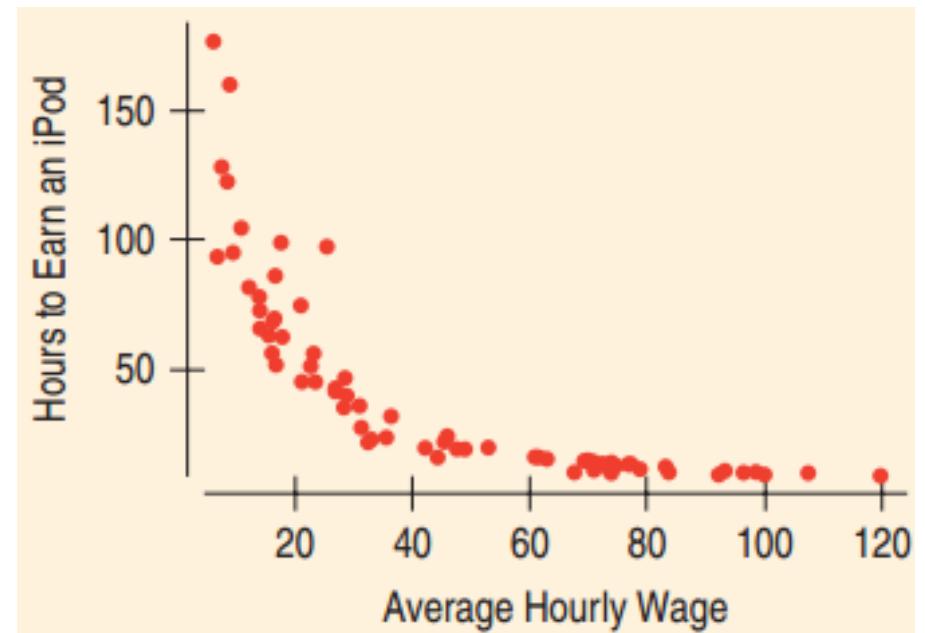
- Describe the patterns.
 - *Cost of Women's Clothes* and *Food Costs* are positively associated. The association is straight.
 - Higher clothes costs correspond to higher food costs.



Data from 73 International Cities

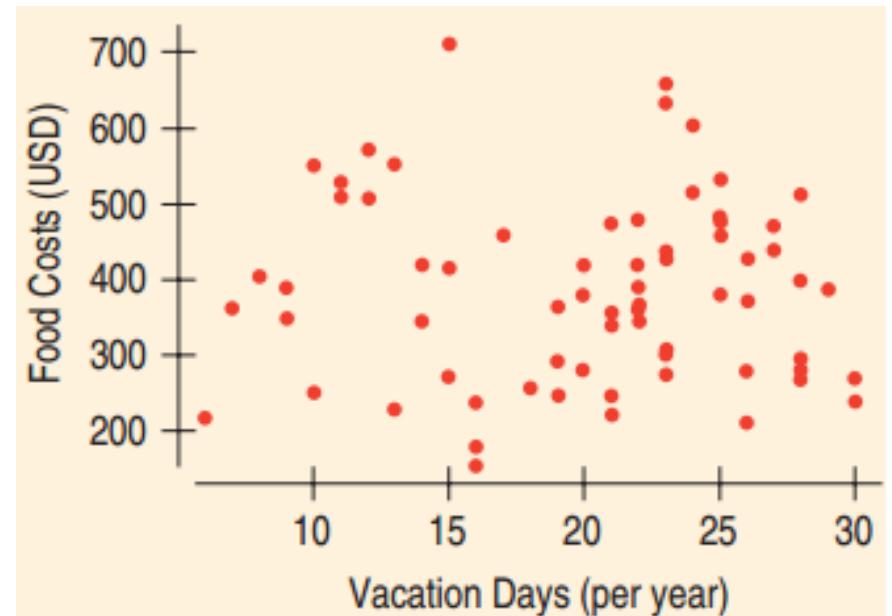
Example 2: Comparing Prices Worldwide

- Describe the patterns.
 - *Average Hourly Wage* and *Hours to Earn an iPod* are negatively associated. The association is **not straight**.
 - **Higher** average hourly wages correspond to **fewer** hours to earn an iPod.



Example 3: Comparing Prices Worldwide

- Describe the patterns.
 - There seems to be no association between *Vacation Days* and *Food Costs*.
 - Knowing the vacation days per year tell us nothing about what food will cost.



Example

4. **Disk drives** Disk drives have been getting larger. Their capacity is now often given in *terabytes* (TB) where 1 TB = 1000 gigabytes, or about a trillion bytes. A survey of prices for external disk drives found the following data:

- a) Prepare a scatterplot of *Price* against *Capacity*.
- b) What can you say about the direction of the association?
- c) What can you say about the form of the relationship?
- d) What can you say about the strength of the relationship?
- e) Does the scatterplot show any outliers?

See **handout** for scatter-plot using R.

R:

```
x=c(0.08,0.12,0.2,0.25,0.32,1,2,4);  
y=c(29.95, 35, 299, 49.95, 69.95, 99, 205, 449);  
plot(x, y);
```

Capacity (in TB)	Price (in \$)
0.080	29.95
0.120	35.00
0.200	299.00
0.250	49.95
0.320	69.95
1.0	99.00
2.0	205.00
4.0	449.00

Linear Correlation

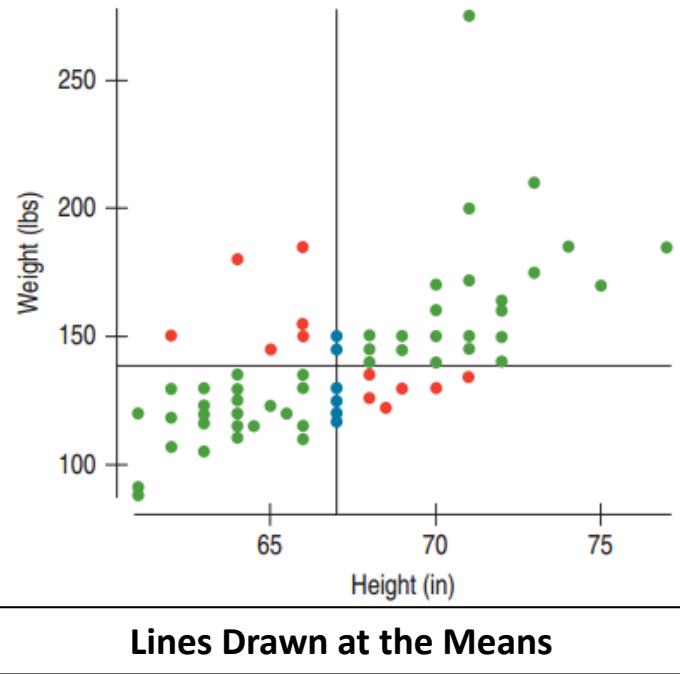
- An **association** exists between two variables when the values of one are somehow **associated** with the values of the other in some way.
- A (**linear**) **correlation** exists between two variables when there is an association and the plotted points of paired data result in a pattern that can be approximated by a **straight line**.
- ❖ We focus on the study of the **linear correlation** between two variables in this course.

Numerical Measures for Two Quantitative Variables

- Assume that the two variables x and y exhibit a linear pattern or form.
- There are two numerical measures to describe:
 - **Linear Correlation Coefficient:** the strength and direction of the relationship between x and y .
 - **Linear Regression Model:** the form of the relationship.

Example: Height and Weight

- It seems that there is an association between height and weight?
- It looks positive.
- How strong is the positive association?
- We use **linear correlation coefficient** to measure the direction and strength of the association.



Requirements for Linear Correlation

1. The sample of paired (x, y) data is a simple random sample.
2. Visual examination of the scatterplot must confirm that the points approximate a straight-line pattern.
3. **Outliers and Influential points**
 - The outliers/influential points must be removed if they are known to be errors;
 - The effects of any other influential points should be considered by calculating the linear correlation coefficient **with** and **without** the influential points included.

Notations for the Linear Correlation Coefficient

n number of pairs of sample data

\sum denotes the addition of the items indicated

$\sum x$ sum of all x -values

$\sum x^2$ indicates that each x -value should be squared and then those squares added

$(\sum x)^2$ indicates that each x -value should be added and the total then squared

$\sum xy$ indicates each x -value is multiplied by its corresponding y -value. Then sum those up.

r linear correlation coefficient for sample data

ρ linear correlation coefficient for a population of paired data

Linear Correlation Coefficient

- The **sample** of bivariate (x, y) data is a simple random sample from a **population** of bivariate data.
- ❖ The **population linear correlation coefficient** is denoted by ρ . And we make statistical inferences about ρ using the sample information $(x_i, y_i), i=1, 2, \dots, n$.

Linear Correlation Coefficient

The **linear correlation coefficient** r measures the strength and direction of a **linear relationship** between the paired values in a **sample**.

$$r = \frac{s_{xy}}{\sqrt{s_{xx}} \sqrt{s_{yy}}}$$

where

$$s_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

$$s_{xx} = \sum_i (x_i - \bar{x})^2, \quad s_{yy} = \sum_i (y_i - \bar{y})^2$$

Linear Correlation Coefficient

A second formula of linear correlation coefficient r :

$$r = \frac{\sum(z_x z_y)}{n-1}$$

where

$$z_{x_i} = \frac{x_i - \bar{x}}{s_x}, \quad z_{y_i} = \frac{y_i - \bar{y}}{s_y}$$

are z-scores

If $s_x = 0$ and/or $s_y = 0$, we define r to be 0. This is because

- The formula doesn't work in this case (division of 0 by 0)
- The standard deviation equals 0 if and only if all data values are equal, so there can be **no association** since there is no variation.

Properties of the Linear Correlation Coefficient r

1. $-1 \leq r \leq 1$
2. If all values of either variable are converted to a different **scale**, the value of r does not change.
Therefore, changing the units of x or y does not affect r .
 - Measuring in dollars, cents, or Euros will all produce the same correlation.
3. The value of r is **not affected by the order** of x and y . Interchange all x and y -values and the value of r will not change.

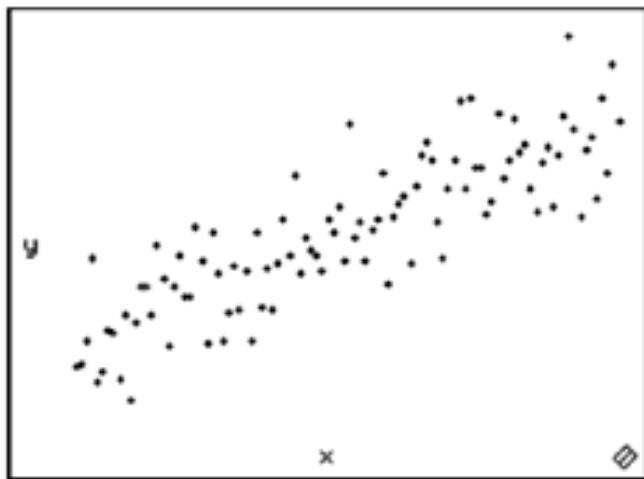
Properties of the Linear Correlation Coefficient r

4. r measures strength and direction of a **linear relationship**.
5. r is very **sensitive to outliers**, which can dramatically affect the value of r .
 - The adjectives *weak*, *moderate*, and *strong* can describe correlation, but there are no agreed upon boundaries.
 - If data has a strong association but is not linear, **don't use r** since it measures **linear correlation only**.

Interpreting r

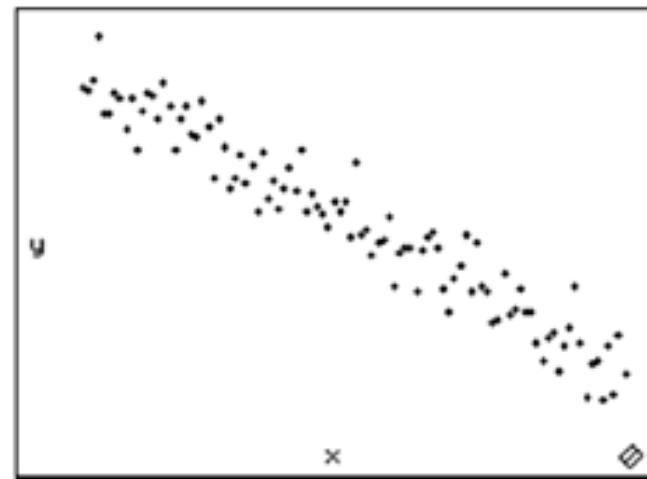
- $-1 \leq r \leq 1$ Sign of r indicates direction of the linear relationship.
- $r \approx 0$ Weak relationship; random scatter of points
- $r \approx 1$ or -1 Strong relationship; either positive or negative
- $r = 1$ or -1 All points fall exactly on a straight line.

Scatterplot and Linear Correlation Coefficient



(a) Positive correlation:

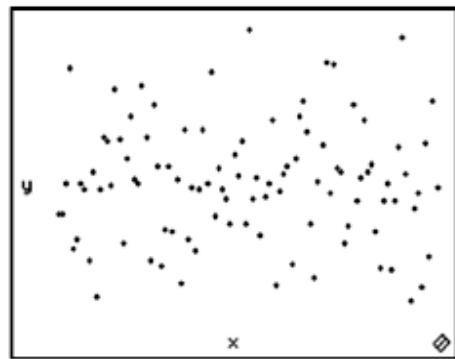
$$r = 0.851$$



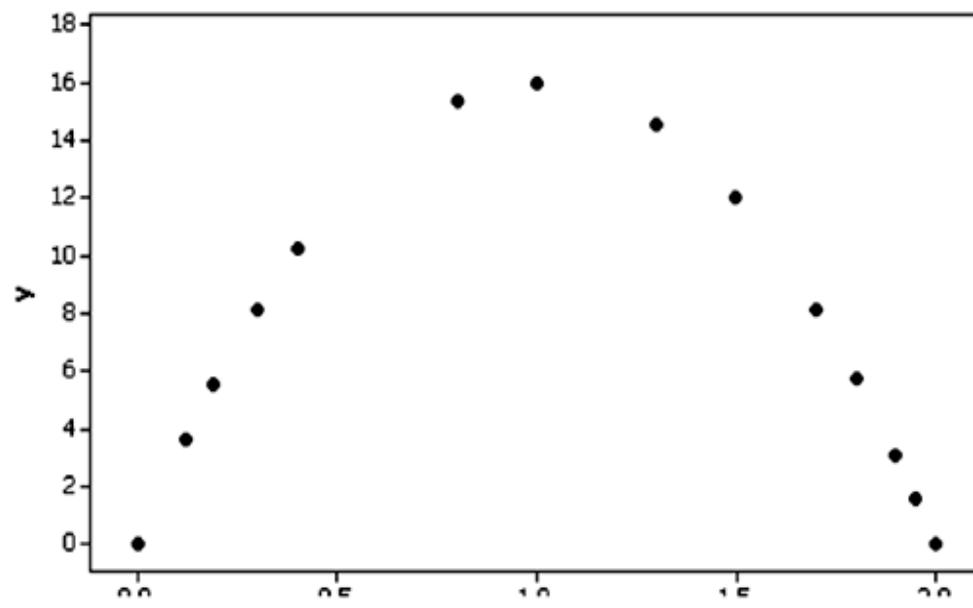
(b) Negative correlation:

$$r = -0.965$$

Scatterplot and Linear Correlation Coefficient



(c) No correlation: $r = 0$



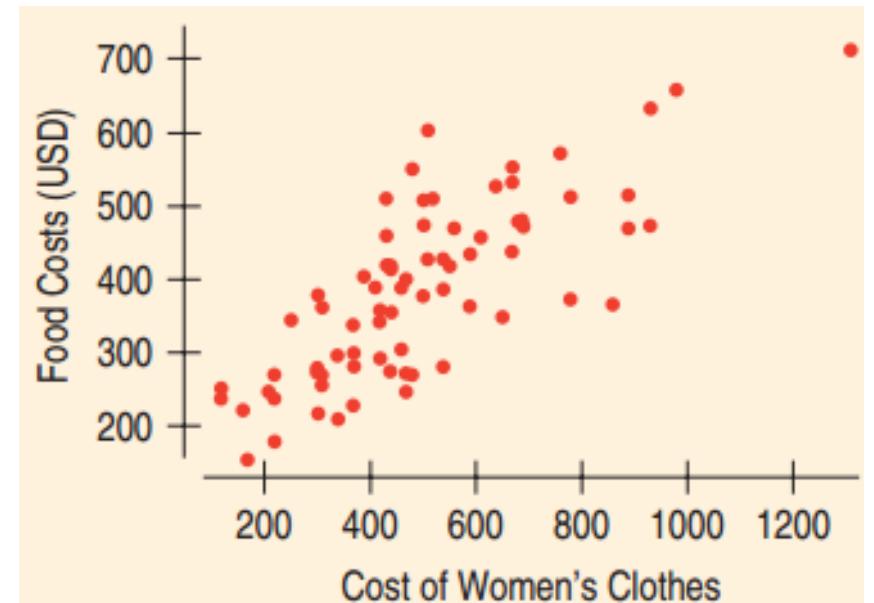
(d) Nonlinear relationship: $r = -0.087$

Activity: Guess the Correlation

<https://istats.shinyapps.io/guesscorr/>

Example 1: Clothes and Food Revisited

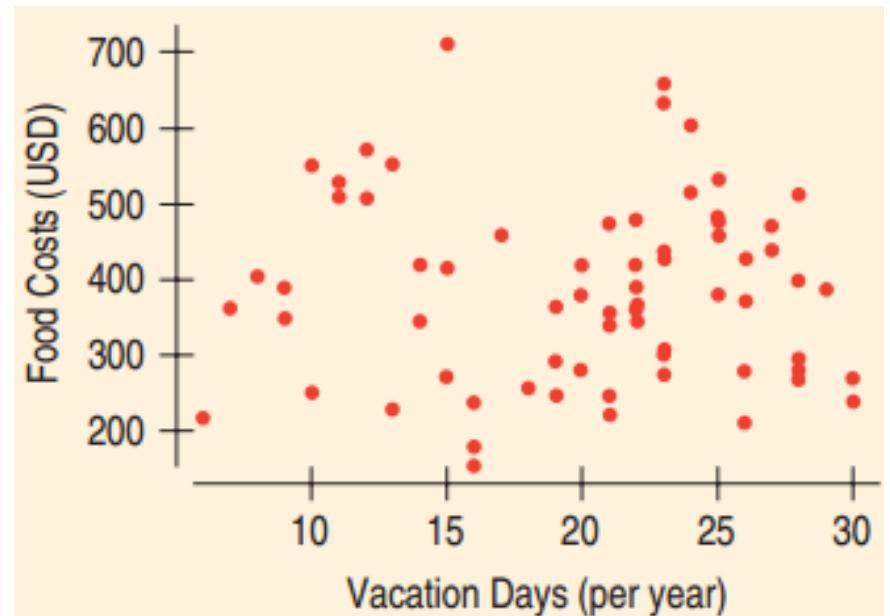
- The scatterplot indicates a straight-line pattern. The variables are both quantitative (\$), and there are no strong outliers away from the linear pattern.



- The correlation of $r = 0.774$ represents a strong positive association.

Example 2: Vacation and Food Revisited

- The scatterplot indicates that there may be no underlying linear relationship between vacation days and food costs.



- The correlation is $r = -0.022$ supports that there may be no linear association between the two.

More examples

https://istats.shinyapps.io/Association_Quantitative/

Caution

- The methods of this section apply to a ***linear correlation*** only.
- If $|r|$ is close to zero and you conclude that there does not appear to be linear correlation, it is possible that there might be some **other associations** that are not linear. So we have to study both r and the scatterplot.

Example

- We use R programming language to calculate r . See handout.

Capacity (in TB)	Price (in \$)
0.080	29.95
0.120	35.00
0.200	299.00
0.250	49.95
0.320	69.95
1.0	99.00
2.0	205.00
4.0	449.00

```
R:x=c(0.08,0.12,0.2,0.25,0.32,1,2,4);  
y=c(29.95, 35, 299, 49.95, 69.95, 99, 205, 449);  
plot(x, y);  
cor(x, y);
```

Calculation Question

- Find the value of the linear correlation coefficient r .

x	47.0	46.6	27.4	33.2	40.9
y	8	10	10	5	10

- A. 0.156
- B. 0.175
- C. 0
- D. -0.175

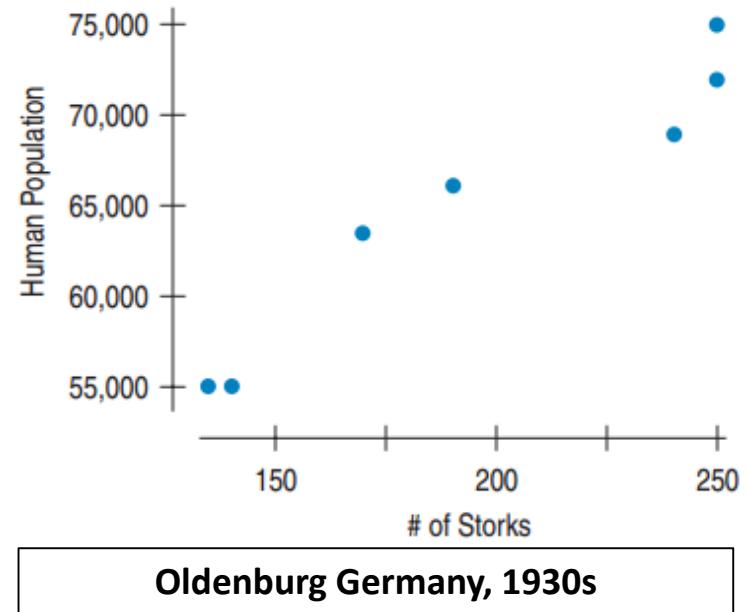
Caution: Correlation \neq Causation

- Causation is a possibility, but more must be done to prove causation.
 - The causation could be in reverse (y causes x)
 - Furthermore, a **lurking variable** may cause both.
 - **Example:** Number of gray hairs and number of wrinkles are strongly correlated, but dyeing hair black does not undo wrinkles. Age is the lurking variable that causes both to increase.

Causation may be in reverse

- **Example: Storks and Babies**

There is a clear positive association between the number of storks and the population.



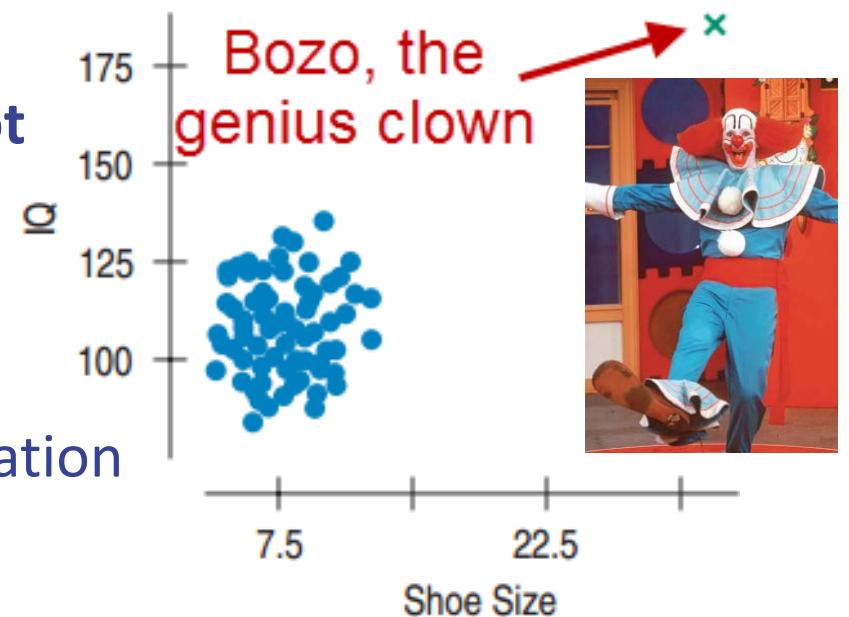
- This does not prove that an increase in storks has caused an increase in babies being born.
- Here, **causation is in reverse**. Storks nest on house chimneys, so the increased population has increased nesting sites.

What Can Go Wrong?

- Don't say "correlation" when you mean "association."
 - **Correlation** in general implies a **linear** relationship.
Association means any relationship.
- Don't correlate categorical variables.
 - It makes no sense to say *car model* and *personality type* are correlated.
- Don't confuse correlation with causation.
 - Correlation only implies general tendencies.

What Can Go Wrong?

- Make sure the association is **linear**.
 - Always look at the scatterplot to check.
- Don't assume the association is linear just because the **correlation coefficient** is high.
 - Always look at the scatterplot to check.
- Beware of influential points!
 - $r = 0.5$, but there is no correlation between shoe size and IQ.



How to Report Correlation

- **Bad:** Raising salaries increases productivity.
- **Good:** Employees with higher salaries tend to be more productive.
- **Bad:** $r = -0.99$. This proves that drinking more red wine lowers cholesterol.
- **Good:** There is a strong negative association between red wine consumption and cholesterol level.

DIY Example

42. Drug abuse A survey was conducted in the United States and 10 countries of Western Europe to determine the percentage of teenagers who had used marijuana and other drugs. The results are summarized in the following table.

- a) Create a scatterplot.
- b) What is the correlation between the percent of teens who have used marijuana and the percent who have used other drugs?
- c) Write a brief description of the association.
- d) Do these results confirm that marijuana is a “gateway drug,” that is, that marijuana use leads to the use of other drugs? Explain.

Country	Percent Who Have Used	
	Marijuana	Other Drugs
Czech Rep.	22	4
Denmark	17	3
England	40	21
Finland	5	1
Ireland	37	16
Italy	19	8
No. Ireland	23	14
Norway	6	3
Portugal	7	3
Scotland	53	31
United States	34	24

Example Solution

(b) $r = 0.934$.

(c) The association between the percent of teens who have used marijuana and the percent of teens who have used other drugs is linear, positive, and strong. Countries with higher percentage of teens who have used marijuana tend to have higher percentage of teens that have used other drugs.

(d) These results do not confirm that marijuana is “gateway drug”. An association exists between the percent of teens who have used marijuana and the percent of teens who have used other drugs. However, it does not mean that one caused the other.

Key Concept

- Scatter plot of bivariate data.
- Calculation of r (linear correlation coefficient), a measure of direction and strength of the linear correlation between the two variables.

Section 2 Simple Linear Regression

- ❖ The **simple linear regression model** studies the **form** of the relationship between two variables:

x - the **explanatory variable, independent variable**

and

- y - the **response variable or dependent variable**).

Simple Linear Regression

- If we want to describe the relationship between y and x for the **whole population**, there are two models we can choose

- Deterministic Model: $y = \beta_0 + \beta_1 x$
- Probabilistic Model:
 - y = deterministic model + random error
 - $y = \beta_0 + \beta_1 x + \varepsilon$
 - where ε is Normally distributed with mean 0 and variance σ^2

Simple Linear Regression

A random sample of paired data is of the form
(with sample size n)

y	x
y_1	x_1
y_2	x_2
\vdots	\vdots
y_n	x_n

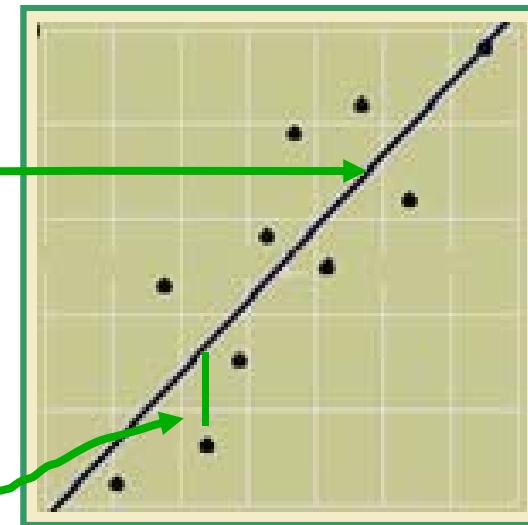
Simple Linear Regression

- Since the bivariate measurements that we observe do not generally fall exactly on a straight line, we choose to use:
- Probabilistic Model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

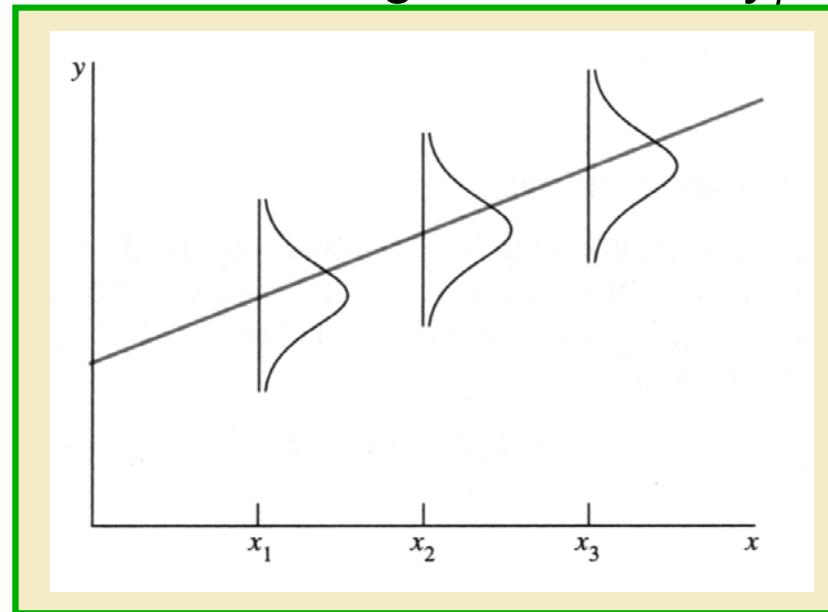
$$E(y_i) = \beta_0 + \beta_1 x_i, i=1, \dots, n$$

Points y_i deviate from the line of **means** by an amount ε_i where ε_i are **independent normal** random variables with mean 0 and **common variance** σ^2 .



The Random Error

- The line of means, $E(y_i) = \beta_0 + \beta_1 x_i$, describes average value of y_i for any **fixed value of x_i** , $i=1,2, \dots, n$.
- The line of means, $E(y) = \beta_0 + \beta_1 x$ (*we usually write $y = \beta_0 + \beta_1 x$*) is called the **regression line**.
- The population of measurements is generated as y_i deviates from the population line of means by ε_i . We estimate β_0 and β_1 using sample information.



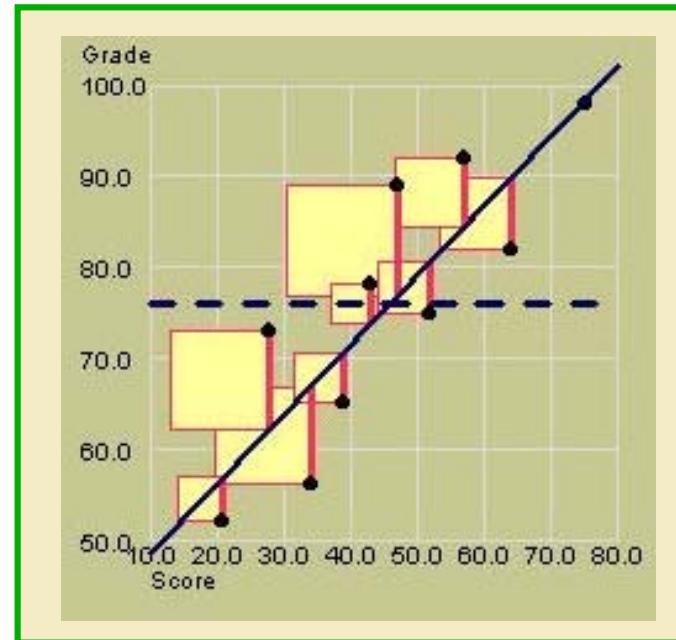
The Method of Least Squares

- The equation of the best-fitting line is calculated using a set of n pairs (x_i, y_i) , $i=1, \dots, n$.
- We choose our estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ to estimate β_0 and β_1 so that the vertical distances of the points from the line, are minimized.

Best fitting line : $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize

$$SSE = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$



Notations

The fitted regression line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

is called the **least squares line**. \hat{y} is called **fitted value** or **predicted value**.

	Population Parameter	Sample Statistic
y -Intercept of regression equation	β_0	$\hat{\beta}_0$
Slope of regression equation	β_1	$\hat{\beta}_1$
Equation of the regression line	$y = \beta_0 + \beta_1 x$	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Explore Linear Regression Parameter and its LS estimate

<https://istats.shinyapps.io/ExploreLinReg/>

Formulas for $\hat{\beta}_1$ and $\hat{\beta}_0$

(1) Slope. $\hat{\beta}_1 = r \frac{s_y}{s_x}$ or

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Note: $\hat{\beta}_1$ and r have the **same sign**.

(2) Intercept $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Note: $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$. That is, (\bar{x}, \bar{y}) is always on the fitted regression line.

Concept Question

Use the given data to find the equation of the fitted regression line.

x	2	4	5	6
y	7	11	13	20

- A. $\hat{y} = 3.0x$
- B. $\hat{y} = 0.15 + 3.0x$
- C. $\hat{y} = 2.8x$
- D. $\hat{y} = 0.15 + 2.8x$

How to Find LS Line

- **Method 1 – Usually for Large Data Sets With Summary Info**
 - Formulas (not used in practical data analysis)
- **Method 2 – Using R coding**
- **Method 3 – web app**

<http://esumath.shinyapps.io/rstats>

<https://istats.shinyapps.io/LinearRegression/>

Example

4. Disk drives Disk drives have been getting larger. Their capacity is now often given in *terabytes* (TB) where 1 TB = 1000 gigabytes, or about a trillion bytes. A survey of prices for external disk drives found the following data:

Capacity (in TB)	Price (in \$)
0.080	29.95
0.120	35.00
0.250	49.95
0.320	69.95
1.0	99.00
2.0	205.00
4.0	449.00

$\bar{x} = 1.110$ $\bar{y} = 133.98$
 $SD(x) = 1.4469$ $SD(y) = 151.26$
 $r = 0.994$

Fit an SLR model.

R:

```
x=c(0.08,0.12,0.25,0.32,1,2,4);  
y=c(29.95, 35, 49.95, 69.95, 99, 205, 449);  
fit=lm(y~x);  
summary(fit);
```

Conditions for Using Regression

- Only use the regression line to make predictions if:
 - **Quantitative Variable Condition:** Regression analysis can only be used for quantitative variables.
 - **Linear Condition:** The scatterplot should indicate a relatively straight pattern.
 - **Outlier Condition:** Outliers dramatically influence the fit of the least squares line. Don't use regression with outliers.

Interpreting the Line of Best Fit

- Protein and Fat

- $\hat{Fat} = 8.4 + 0.91 \text{ Protein}$
- Slope = 0.91: A Burger King item with one more gram of protein is **expected** to have 0.91 additional grams of fat.
- **y-intercept** = 8.4: A Burger King item with no grams of protein is expected to have 8.4 grams of fat. In reality the two items with no protein also have no fat.
 - The intercept does not always have a meaningful interpretation!

Examples

16. Horsepower

We previously examined

the relationship between the fuel economy (mpg) and horsepower for 15 models of cars. Further analysis produces the regression model $\widehat{mpg} = 43.45 - 0.070 HP$. If the car you are thinking of buying has a 200-horsepower engine, what does this model suggest your gas mileage would be?

20. More horsepower

In Exercise 16, the regression model $\widehat{mpg} = 43.45 - 0.070 HP$ relates cars' horsepower to their fuel economy (in mpg). Explain what the slope means.

Example:

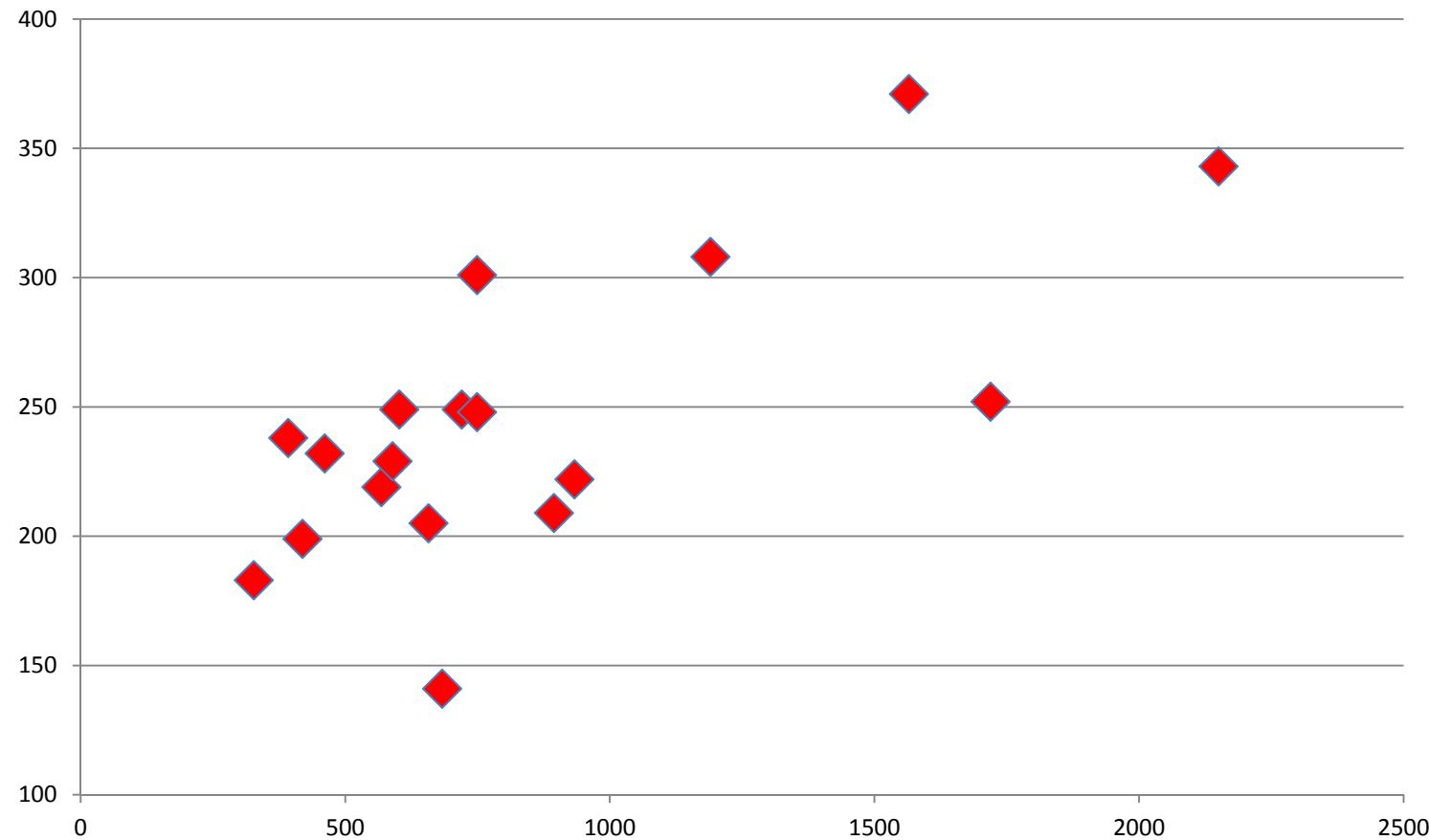
Flights From Atlanta (DIY)

Is there an association
between the distance of the
flight and its cost?

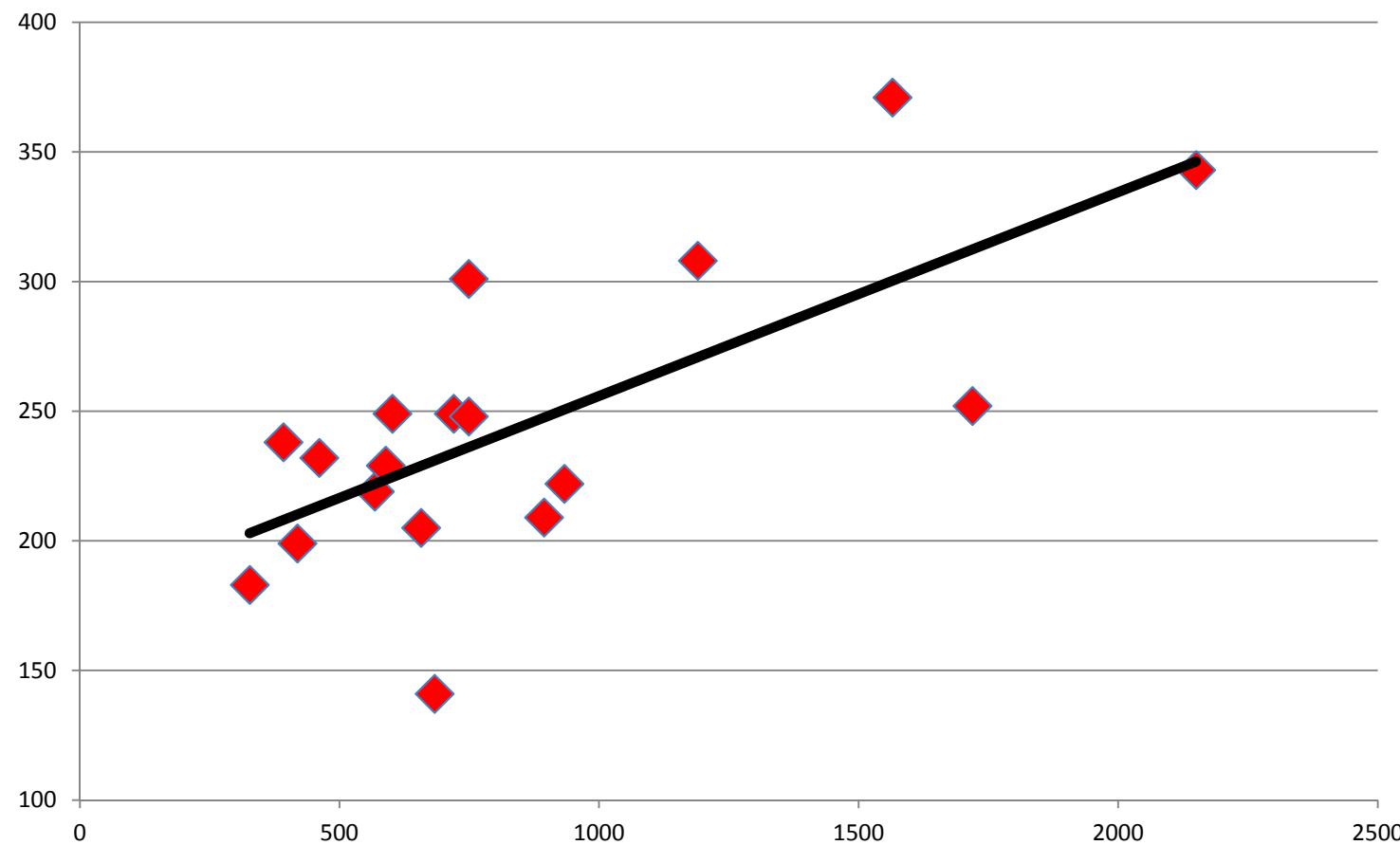
If there is a linear correlation,
fit the simple linear regression
model.

Atlanta to:	Distance (mi)	Fare (\$)
Baltimore	568	219
Boston	933	222
Dallas	720	249
Denver	1190	308
Detroit	602	249
Kansas City	683	141
Las Vegas	1719	252
Miami	589	229
Memphis	327	183
Minneapolis	894	209
New Orleans	419	199
New York City	749	248
Oklahoma City	749	301
Orlando	392	238
Philadelphia	657	205
St. Louis	461	232
Salt lake City	1565	371
Seattle	2150	343
Mean	853.7222222	244.3333333
Std Dev	497.7907434	56.36957878
Correlation	0.69427207	

Example: Flights From Atlanta



Example: Flights From Atlanta with LSR



Residuals

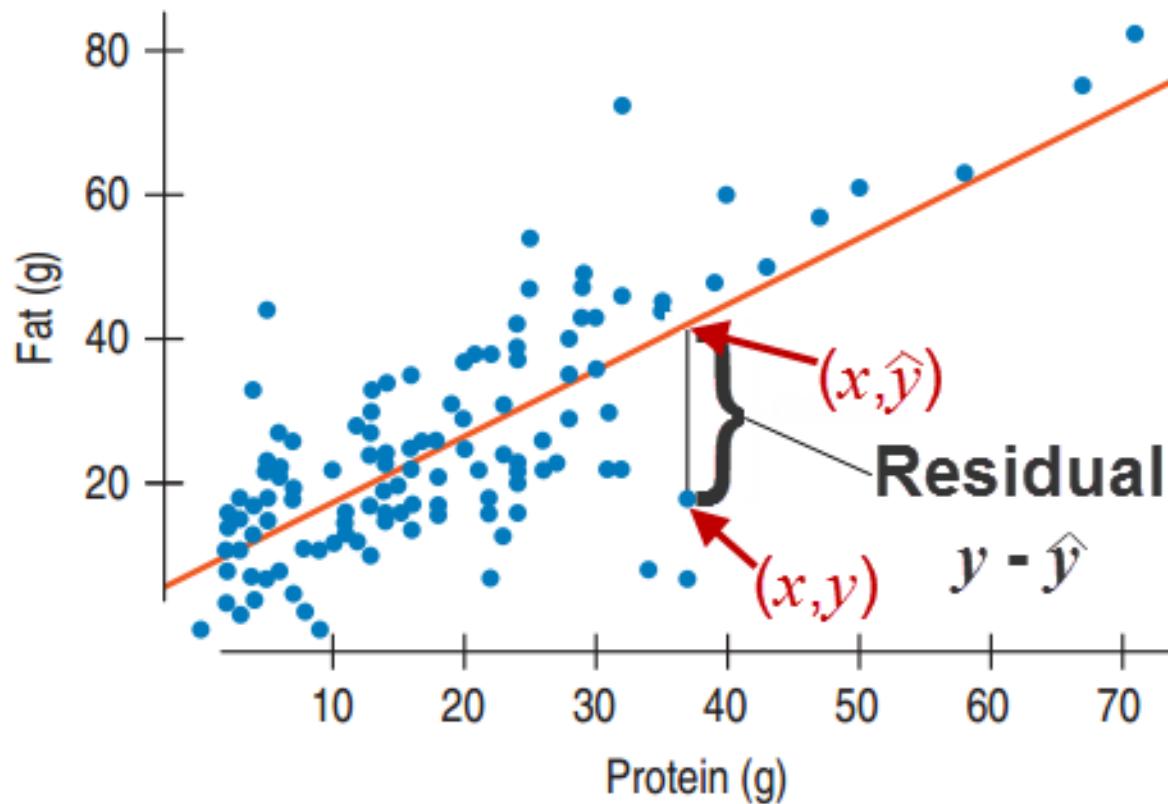
Least Squares line : $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

- ❖ \hat{y} is called the predicted or fitted value.
- ❖ For a pair of sample x and y values, the **residual** is the difference between the **observed** sample value of y and the **predicted** y -value.

That is:

$$\text{residual} = \text{observed } y - \text{predicted } y = y - \hat{y}$$

- ❖ For observation i : the residual $e_i = y_i - \hat{y}_i, i=1, \dots, n.$
- ❖ A residual is an observed random error ε in the model.

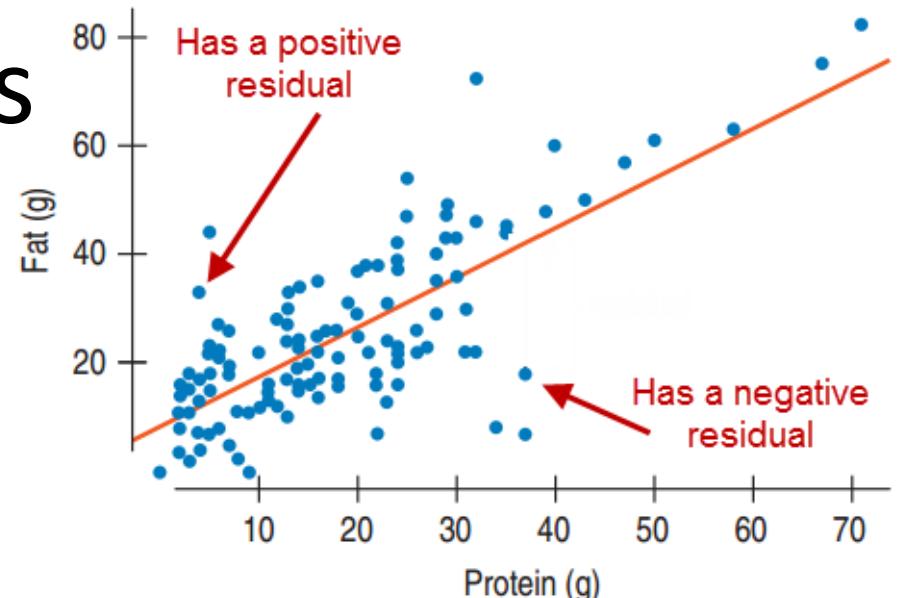


Protein and Fat in Burger King Foods

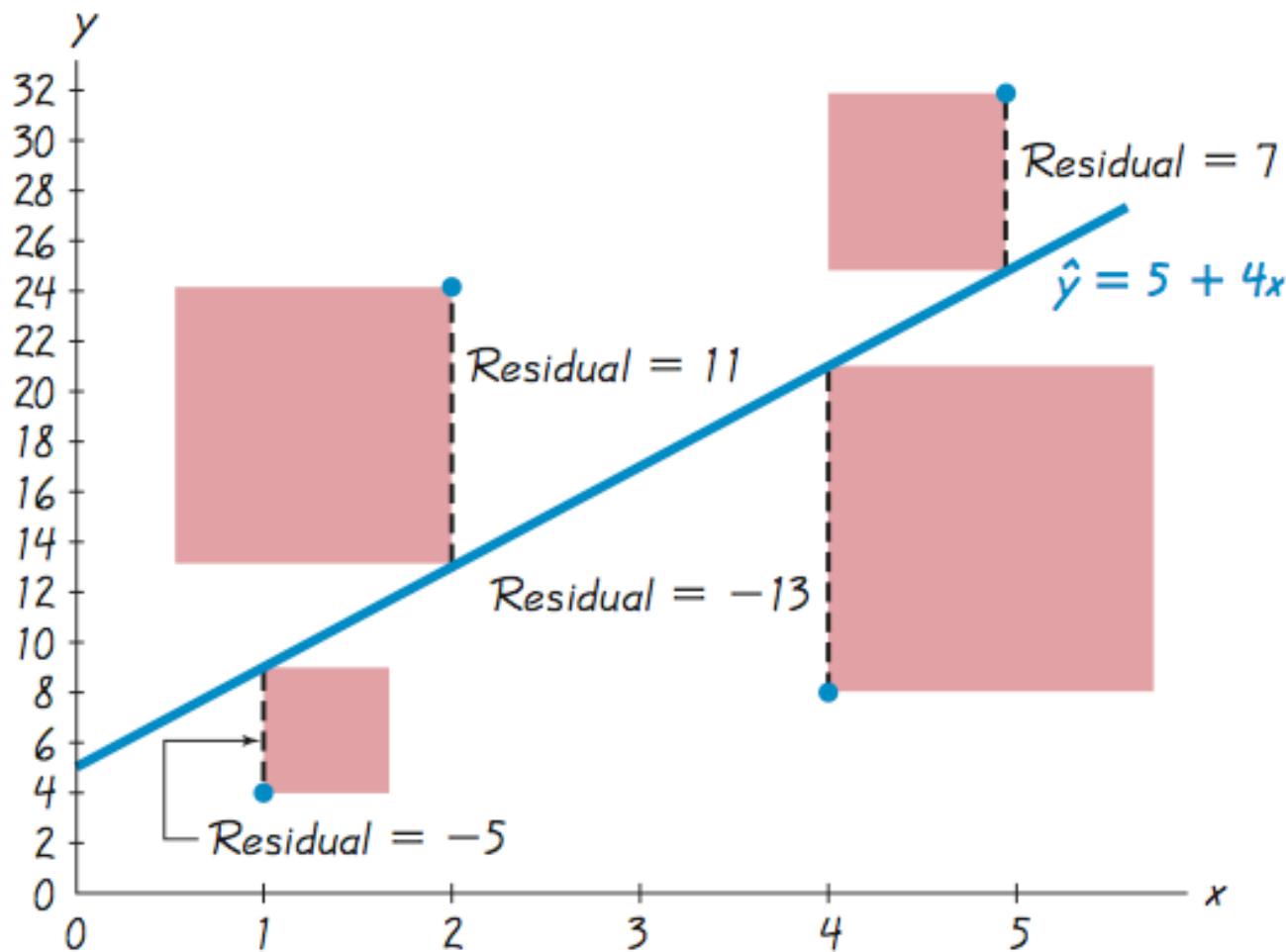
More on Residuals

Residual:

- Observed – Predicted
- Points *above* the line have *positive* residuals
- Points *below* the line have *negative* residuals.
- This line gives the average fat content expected for a given amount of protein.



Example: Residuals



Residual Plot

Definition: A **residual plot** is a scatterplot of the (x, y) or (\hat{y}, y) values after each of the y -coordinate values has been replaced by the residual value $y - \hat{y}$ (where \hat{y} denotes the predicted value of y).

That is, a residual plot is a graph of the points

$$(x_i, y_i - \hat{y}_i), \quad i=1, \dots, n.$$

or

$$(\hat{y}_i, y_i - \hat{y}_i), \quad i=1, \dots, n.$$

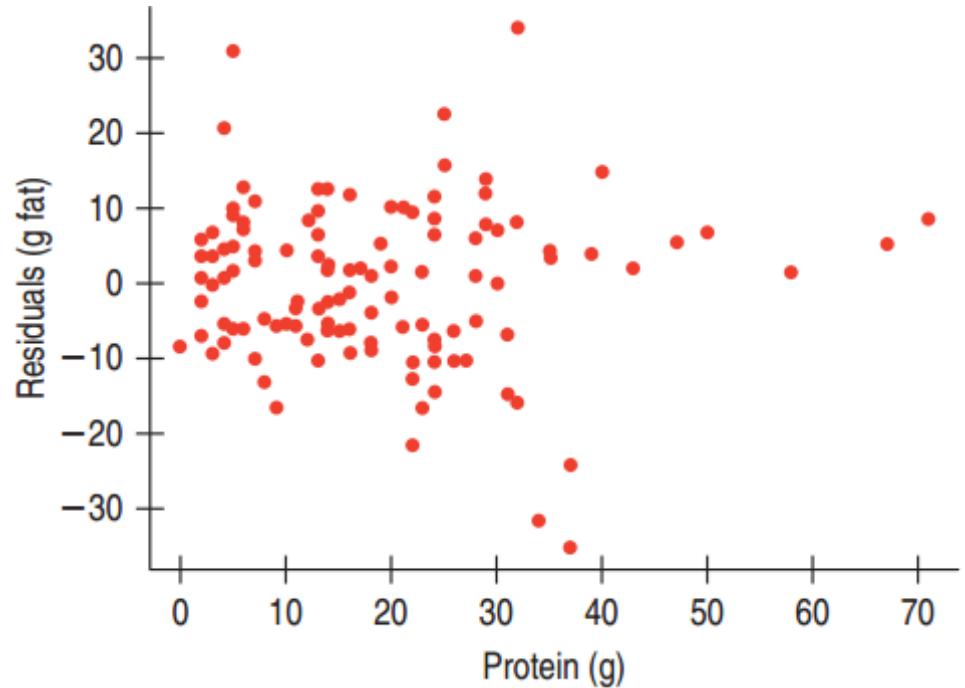
Residual Plot Analysis

When analyzing a residual plot, look for a **pattern** in the way the points are configured, and use these criteria:

- The residual plot **should not have any obvious patterns** (not even a straight line pattern) because the responses y_i or random errors ε_i are **independent**. This confirms that the scatterplot of the sample data is a straight-line pattern.
- The residual plot should not become thicker (or thinner) when viewed from left to right. This confirms the requirement that for different fixed values of x , the distributions of the corresponding y values or random errors all have the **constant variance σ^2** .

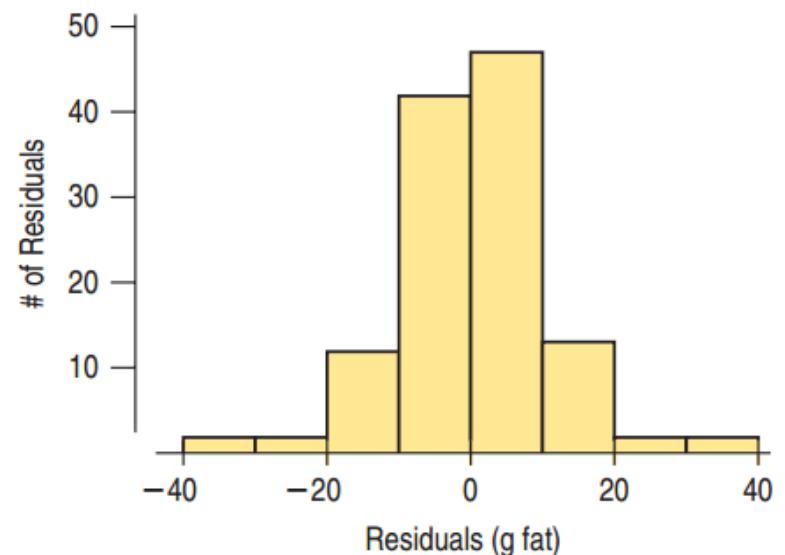
A Good Regression Model

- The regression model is a good model if the residual scatterplot has no interesting features.
 - No direction
 - No shape
 - No bends
 - No outliers
 - No identifiable pattern
 - Does not become thicker or thinner



The Residual Standard Deviation

- The mean of the residuals is **0**
- The standard deviation of the residuals shows how small the residuals vary. Written s_e
- **Equal Variance Assumption:** A good model will have the **spread of the residuals consistent** and small.
- A histogram of the residuals helps us understand the residuals' distribution.



Residual Analysis with Web App

<http://esumath.shinyapps.io/rstats>

<https://istats.shinyapps.io/LinearRegression/>

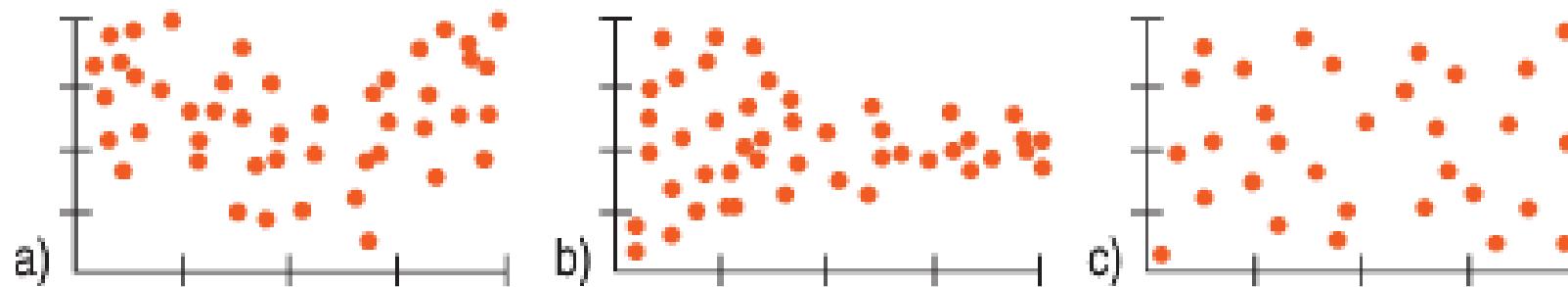
Concept Question

In simple linear regression, the plot of residuals versus fitted values \hat{y} can be used to check for which of the following?

- A. Normality
- B. Independence
- C. Constant variance
- D. Both independence and constant variance

Residual Plot Example

28. **Residuals** Tell what each of the residual plots below indicates about the appropriateness of the linear model that was fit to the data.



Analysis of Variance (ANOVA)

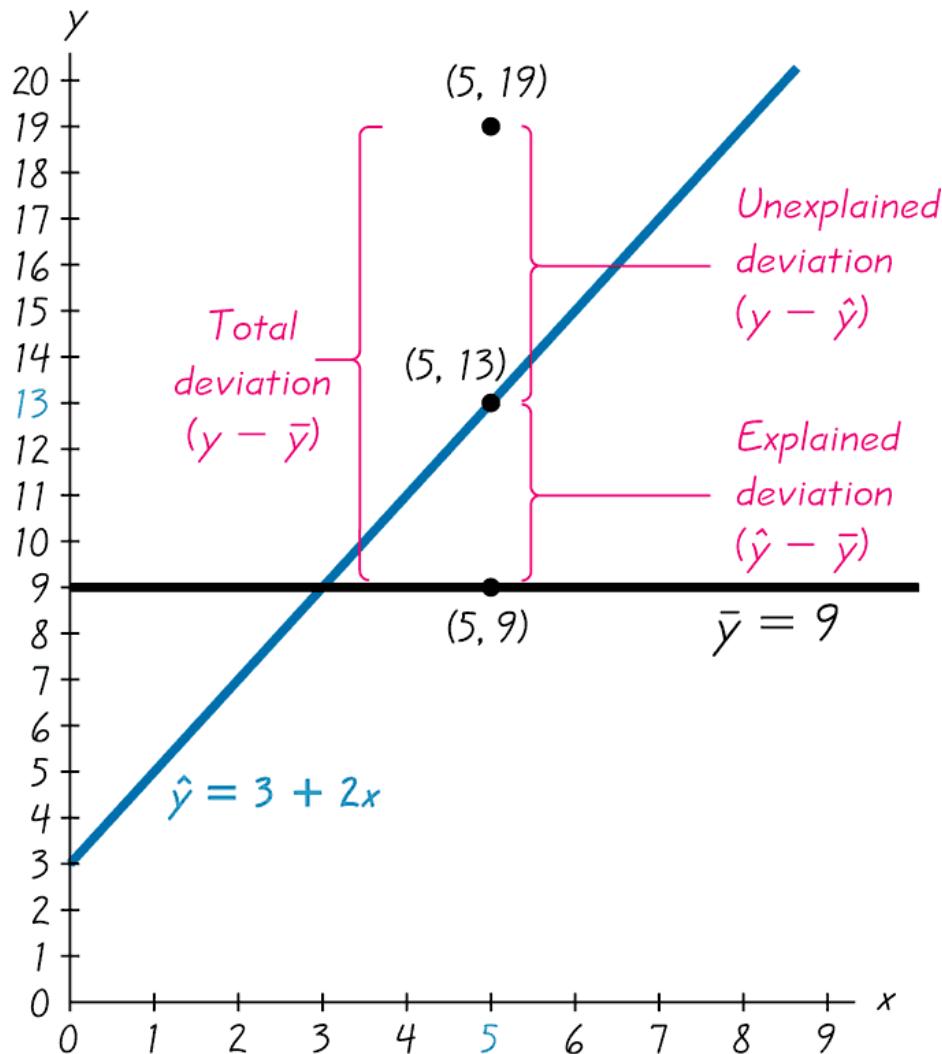
Consider the **Total Sum of Squares (SS_{total})**, a measure of the variation in the response values ignoring the regression model:

$$SS_{\text{total}} = \sum(y_i - \bar{y})^2$$

Now compare this with the **Sum of Squares for Error (SSE)**, a measure of the variation remaining in the response values after predicting them using the fitted regression equation:

$$SSE = \sum(y_i - \hat{y}_i)^2$$

Explained and Unexplained Variation



Example:

- There is sufficient evidence of a linear correlation.
- The equation of the line is $\hat{y} = 3 + 2x$
- The sample mean of the y -values is 9.
- One of the pairs of sample data is $x = 5$ and $y = 19$.
- The point **(5,13)** is on the fitted regression line.

Explained and Unexplained Variation

The figure shows $(5, 13)$ lies on the regression line, but $(5, 19)$ does not.

We arrive at:

Total Deviation (from $\bar{y} = 9$) of the point $(5, 19) = y - \bar{y} = 19 - 9 = 10$.

Explained Deviation (from $\bar{y} = 9$) of the point $(5, 19) = \hat{y} - \bar{y} = 13 - 9 = 4$.

Unexplained Deviation (from $\bar{y} = 9$) of the point $(5, 19) = y - \hat{y} = 19 - 13 = 6$.

ANOVA(Analysis of Variance)

(total deviation) = (explained deviation) + (unexplained deviation)

$$(y - \bar{y}) = (\hat{y} - \bar{y}) + (y - \hat{y})$$

(total variation) = (explained variation) + (unexplained variation)

$$\sum(y - \bar{y})^2 = \sum(\hat{y} - \bar{y})^2 + \sum(y - \hat{y})^2$$

That is, $\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2$

$$SS_{\text{total}} = SSR + SSE$$

✓ SSR (sum of squares for regression): measures the variation explained by using x in the regression model.

ANOVA Table

Total $df =$	$n - 1$	Mean Squares
Regression $df =$	1	$MSR = SSR/(1)$
Error $df =$	$n - 1 - 1 = n - 2$	$MSE = SSE/(n-2)$

Source	df	SS	MS	F	p-value
Regression	1	SSR	$MSR = SSR/1$	MSR/MSE	?
Error	$n - 2$	SSE	$MSE = SSE/(n-2)$		
Total	$n - 1$	Total SS			

MSE is the best estimate of the common variance σ^2

Estimation of σ

σ^2 is estimated by $MSE = SSE/(n-2)$. That is

$$s_e^2 = MSE = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$

MSE is unbiased for σ^2 .

σ is estimated by

$$s_e = \sqrt{MSE} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

Coefficient of Determination

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

$$R^2 = \frac{SSR}{SS_{\text{total}}} = 1 - \frac{SSE}{SS_{\text{total}}}$$

- It can be shown that

$$R^2 = (r)^2$$

Coefficient of Determination

- The coefficient of determination, R^2 (or r^2), is the proportion of the variation in y that is explained (accounted for) by the regression line

$$r^2 = \frac{\text{explained variation}}{\text{total variation}}$$

- That is, it is a measure of
 - How much of the variation in the response is “explained” by the regression (the linear relationship between x and y)

or

- How much of the variation in the response is reduced by predicting it using the regression

Concept Question

Find the coefficient of determination, given that the value of the linear correlation coefficient, r , is -0.721 .

- A. 0.721**
- B. 0.520**
- C. 0.480**
- D. 0.279**

Coefficient of Determination

If we know the value of R^2 , what is the value of r ?

$$r = \sqrt{R^2} ? \quad \text{or}$$

$$r = -\sqrt{R^2} ?$$

- ❖ The sign of r is the same as that of β_1 – the sign of the association between x and y

ANOVA with Web App

<http://esumath.shinyapps.io/rstats>

<https://istats.shinyapps.io/LinearRegression/>

Example

6. Disk drives again In Chapter 6, Exercise 4, we saw some data on hard drives. After correcting for an outlier, these data look like this: we want to predict *Price* from *Capacity*.

Capacity (in TB)	Price (in \$)
0.080	29.95
0.120	35.00
0.250	49.95
0.320	69.95
1.0	99.00
2.0	205.00
4.0	449.00

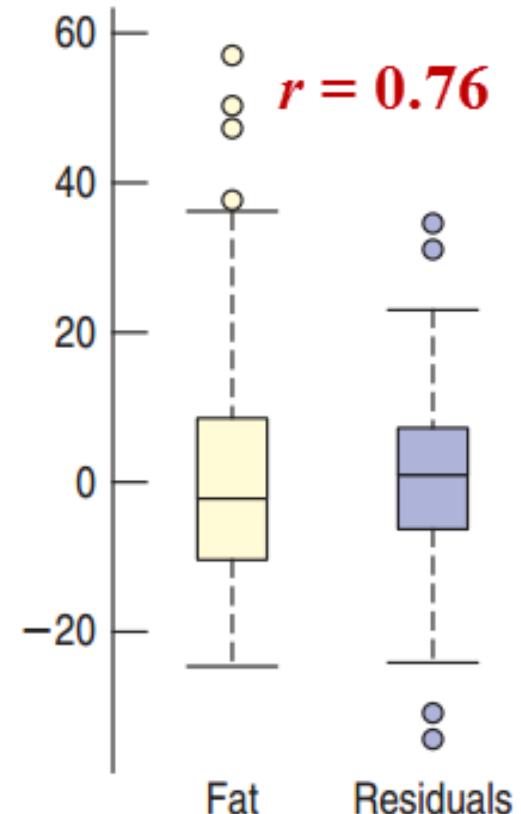
$\bar{x} = 1.110$	$\bar{y} = 133.98$
$SD(x) = 1.4469$	$SD(y) = 151.26$
$r = 0.994$	R:

- (1) Find R^2 when you fit an SLR model.
- (2) Construct the ANOVA table

```
x=c(0.08,0.12,0.2,0.25,0.32,1,2,4)
y=c(29.95, 35, 299, 49.95, 69.95, 99, 205, 449)
fit=lm(y~x)
summary(fit)
anova(fit)
```

Variation of y and the Variation of the Residuals (Continued)

- $R^2 = 0.76^2 = 0.58$
 - 58% of the variability in fat content in Burger King's menu items is accounted for by the variation in the protein content.
 - 42% of the variability in fat content is left in the residuals.
 - Other factors such as how the food is prepared account for this remaining variability.



Example

- 22. Another car** The correlation between a car's horsepower and its fuel economy (in mpg) is $r = -0.909$. What fraction of the variability in fuel economy is accounted for by the horsepower?

When is R^2 Big Enough

- R^2 provides us with a measure of how useful the regression line is as a prediction tool.
- If R^2 is close to 1, then the regression line is useful.
- If R^2 is close to 0, then the regression line is not useful.
- What “close to” means depends on who is using it.
 - **Good Practice:** Always report R^2 and let the researcher decide.

Causation and Regression

- Never report out a cause and effect relationship based solely on regression analysis.
 - Even though the correlation was high and the model was reasonably linear for pressure vs. wind in the hurricane data, we would need a scientific explanation to conclude cause and effect. Regression analysis alone can never prove cause and effect.

What Can Go Wrong?

- Don’t fit a straight line to a nonlinear relationship.
 - If there are curves and bends in the scatterplot, don’t use regression analysis.
- Don’t ignore outliers.
 - Instead report them out and think twice before using regression analysis.
- **Don’t invert** the regression.
 - Switching x and y does not mean just solving for x in the least squares line. You must start over.

Key Concepts

Simple Linear Regression Model

- Understand the model
- Fit a SLR model
- Analysis of Variance
- Coefficient of Determination

Estimation

This chapter presents the beginning of inferential statistics.

- After we take a random sample of size n we can make two types of inference about the population parameters:
 - (i) **Estimation**
 - (a) Point estimation
 - (b) Interval estimation
 - (ii) **Hypothesis Testing** (to be discussed)

Estimation

- ✓ Confidence intervals of the population proportion
- ✓ Confidence intervals of the population mean

Point Estimation

Definition. A **point estimate** is a single value (or point) used to approximate a population parameter.

Examples:

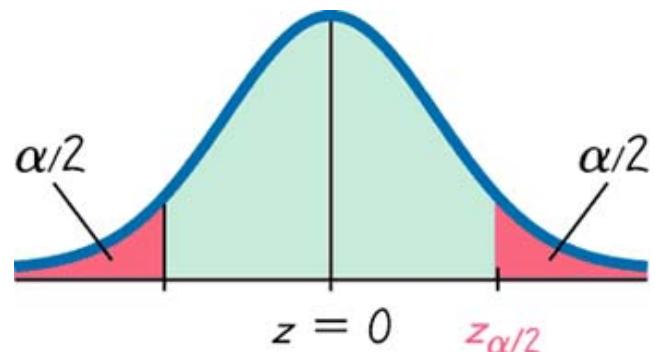
- (1) \bar{y} is a point estimate of the unknown population parameter μ
- (2) $\hat{p} = y/n$ is the point estimate of the unknown population parameter p , where y is the number of successes in the sample.

Interval Estimation

- A **confidence interval** (or **interval estimate**) is a an interval (a, b) of values so that you are **fairly sure (with high probability)** that the population parameter lies between these two values. A confidence interval is abbreviated as CI.
- A **confidence level** is the probability $1 - \alpha$ (often expressed as $100(1-\alpha)\%$) that the confidence interval actually does contain the population parameter, assuming that the estimation process is repeated a large number of times in **repeated sampling**.
 - The confidence level is also called **confidence coefficient**. Some common choices of $1 - \alpha$ are **90%** ($\alpha=0.1$), **95%** ($\alpha=0.05$), or **99%** ($\alpha=0.01$).

Recall: Z Critical Values

The number $z_{\alpha/2}$ is a critical value that is a z score with the property that it separates an area of $\alpha/2$ in the right tail of the standard normal distribution



Examples.

$$Z_{0.05} = 1.645, \quad Z_{0.025} = 1.96, \quad Z_{0.005} = 2.575$$

R: Find $Z_{0.01} = ?$

$$\text{qnorm}(0.99, 0, 1) \approx 2.326.$$

8.2 Interval Estimation of a Population Proportion

- Recall by C.L.T., if n is large ($np \geq 10$ and $n(1-p) \geq 10$), \hat{p} is approximately Normal with mean p and standard deviation $\sqrt{p(1-p)/n}$. That is

$$Z = \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}}$$

is approximately standard Normal.

Note. We estimate $\sqrt{p(1-p)/n}$ by $\sqrt{\hat{p}(1-\hat{p})/n}$ since p is unknown.

Standard Error for a Sample Proportion

- After taking a sample, we only know the sample proportion \hat{p} , which we use as an approximation.
- \hat{p} is regarded as a random variable in repeated sampling.
- The **standard deviation** of \hat{p} is called **standard error** of \hat{p} which is given by

$$\text{SE}(\hat{p}) = \sqrt{\hat{p}(1 - \hat{p})/n}$$

Estimating a Population Proportion

- To find the $1-\alpha$ confidence interval of p , we need to find the lower limit L and upper limit U such that

$$P(L < p < U) = 1 - \alpha$$

- We start the derivation from

$$P\left(-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{\hat{p}(1 - \hat{p})/n}} < z_{\alpha/2}\right) = 1 - \alpha$$

Estimating a Population Proportion

- Multiply through by $\sqrt{\hat{p}(1-\hat{p})/n}$

$$P\left(-z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} < \hat{p} - p < z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\right) = 1 - \alpha$$

- Subtract \hat{p} from each term

$$P\left(-\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} < -p < -\hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\right) = 1 - \alpha$$

Estimating a Population Proportion

- Multiply through by -1 to eliminate the minus sign in front of p

$$P\left(\hat{p} - z_{\alpha/2} \sqrt{\hat{p}(1-\hat{p})/n} < p < \hat{p} + z_{\alpha/2} \sqrt{\hat{p}(1-\hat{p})/n}\right) = 1 - \alpha$$

- That is

$$P\left(\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1 - \alpha$$

Estimating a Population Proportion

- A $100(1-\alpha)\%$ confidence interval of p is

$$\left(\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \quad \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

- ❖ The interval is a **random interval** since \hat{p} is a random variable.
- ❖ This interval is **not random** any more if the statistic \hat{p} is replaced by a fixed sample proportion calculated from a specific sample.

Calculation Question

Find the critical value that corresponds to a 98% confidence level of p using R.

- A. 2.05
- B. 2.33
- C. 2.575
- D. 1.75

Example: Facebook Daily Status Updates

- A recent survey found that 48 of 156 (or 30.8%) update their Facebook status daily.



$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} = \sqrt{\frac{(0.308)(0.692)}{156}} \approx 0.037$$

- So a 95% confidence interval of p is given by
 - $30.8\% \pm 1.96 \times 0.037 = (0.235, 0.381)$
 - See **handout** to get the interval using R!

Calculation Question

Use the degree of confidence 95% and sample data: $n=56$, $x=30$ to construct a confidence interval for the population proportion p .

- A. $0.405 < p < 0.667$
- B. $0.426 < p < 0.646$

CI calculation with Technology

R: prop.test(x=30, n=56, conf.level=0.95, correct =FALSE)

Inferences->Statistics in

<https://esumath.shinyapps.io/Rstats/>

https://istats.shinyapps.io/Inference_prop/

Naming the Confidence Interval

- This confidence interval is a **one-proportion z-interval.**
 - “**One**” since there is a single survey question (one sample).
 - “**Proportion**” since we are interested in the *proportion* of Facebook users who update their status daily.
 - “**z-interval**” since the distribution is approximately **normal**.

Capturing a Proportion

- The confidence interval **may or may not contain** the true population proportion. We do not know!
- Consider **repeating the sampling over and over again**, each time with the same sample size.
 - Each time we would get a different \hat{p} .
 - From each \hat{p} , a different confidence interval could be computed.
 - About 95% of **these confidence intervals** will capture the true proportion.
 - 5% will be duds.

Simulating Confidence Intervals

- Population proportion $p=0.4$
- Sample size $n=80$
- Take 100 samples from the population
- Construct a 95% symmetric two-tailed confidence interval of p
- Count the number of confidence intervals containing p
- Repeat the process for 1000, 10,000 ... samples.

Simulating Confidence Intervals

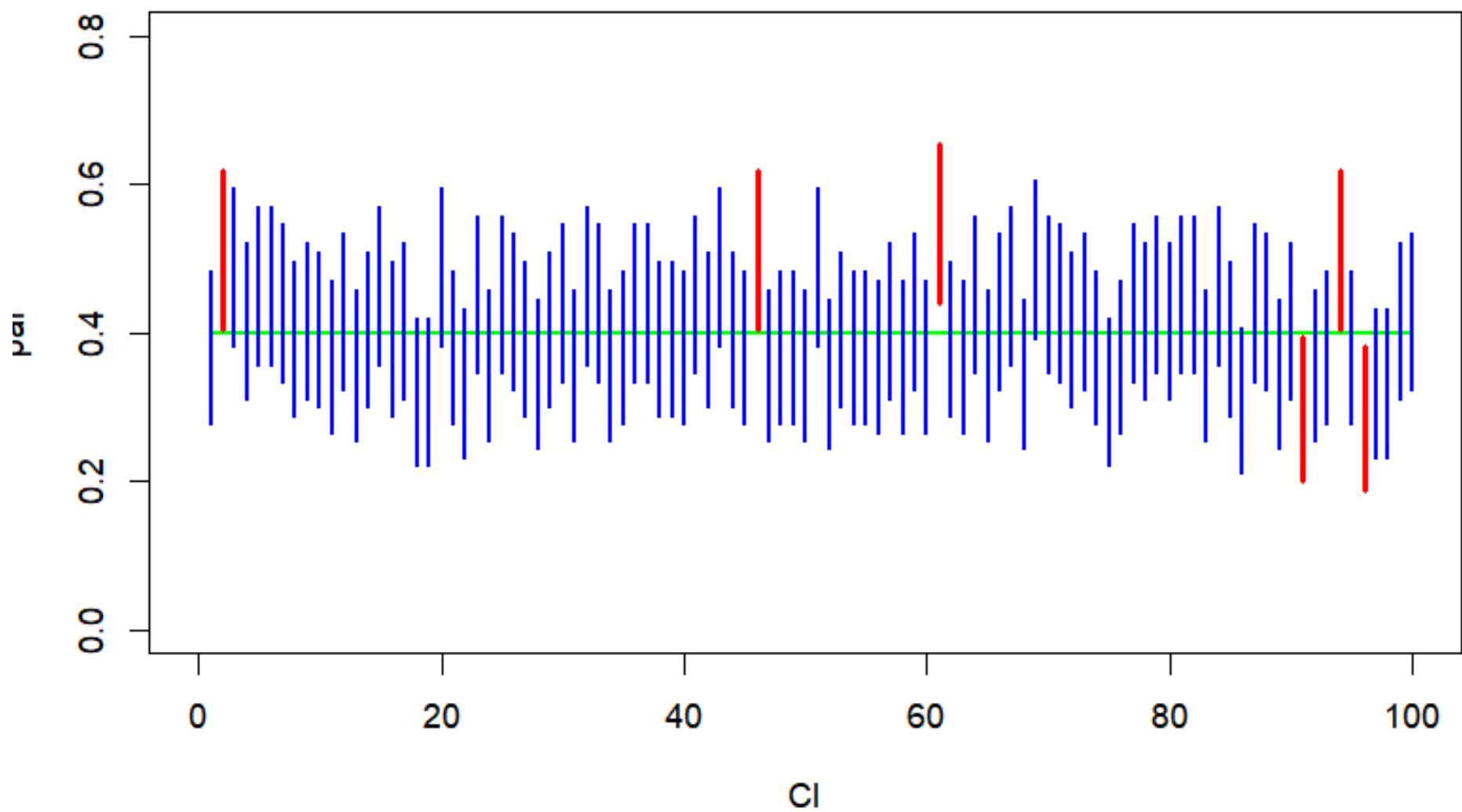
- <https://www.statcrunch.com/applets/type3&ciprop>
- Or use R coding in the next slide

Simulating Confidence Intervals

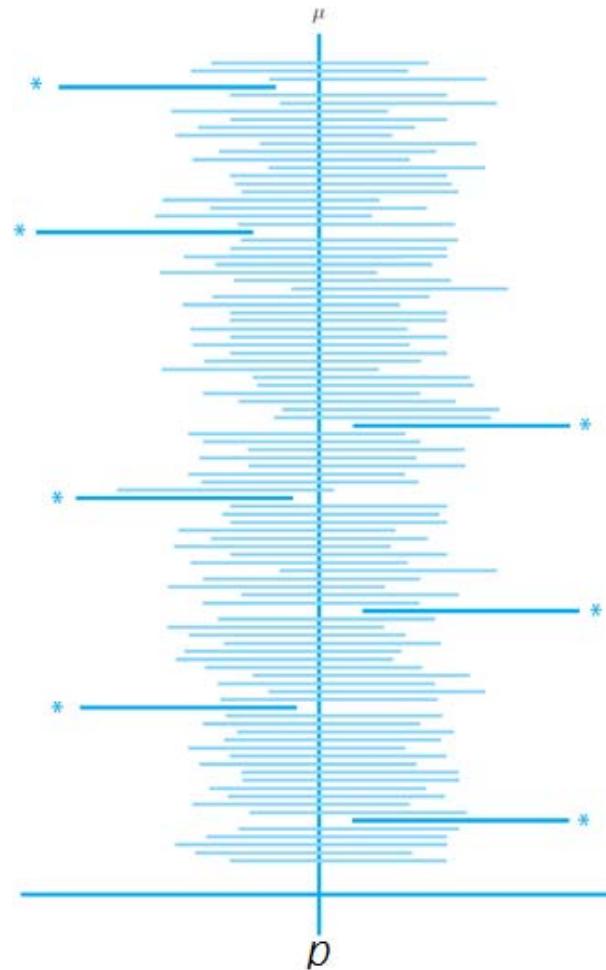
R code

```
p=0.4;                      #population proportion
n = 80;                      #sample size
conlevel=0.95;                #confidence level
iter=1000000;
count=0;                      #number of intervals containing p
L=numeric(); U=numeric(); CI=1:iter; par=rep(p,iter);
plot(CI,par,type="l",xlim=c(0,iter), ylim=c(0,0.8),lwd=2, col="green");
for (i in 1:iter)
{#Sys.sleep(0.05);          #suspend 0.5 second
x0=rbinom(n=1, size=n, prob=p);  #number of success in a sample
interval=prop.test(x=x0, n, conf.level = conlevel, correct = FALSE)$conf.int;    #The limits of the
C.I.
L[i]= interval[1];  U[i]=interval[2];
if (L[i]<=p && U[i]>=p) count=count+1;
time=c(i,i); bds=c(L[i], U[i]);
if (L[i]>p | U[i]<p)
{lines(time,bds,type="l", col="red", lwd=3);
} else { lines(time,bds,type="l", col="blue", lwd=2);}
}
count;
count/iter;
```

Simulating Confidence Intervals



Interpreting a Confidence Interval



One hundred 95% CIs (asterisks identify intervals that do not include p).

Interpreting a Confidence Interval

- There are a huge number of confidence intervals that could be drawn.
 - In theory, all the confidence intervals could be listed.
 - **95% (in the long run)** will “work” (capture the true proportion).
 - **5% (in the long run)** will be “duds” (not capture the true proportion).

Interpreting a Confidence Interval

- ❖ If we repeat the sampling and compute the $100(1-\alpha)\%$ confidence interval of p from each sample following the formula

$$\left(\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \quad \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

- ❖ Then in the long run, $100(1-\alpha)\%$ such confidence intervals will actually contain the true population proportion p .

Interpreting a Confidence Interval

- Recall: Facebook Daily Status Updates
A recent survey found that 48 of 156 (or 30.8%) update their Facebook status daily.



$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} = \sqrt{\frac{(0.308)(0.692)}{156}} \approx 0.037$$

- The 95% confidence interval of p is given by
 $30.8\% \pm 1.96 \times 0.037 = (0.235, 0.380)$

Interpreting a Confidence Interval

The 95% confidence interval (0.235, 0.380)

We will never know whether it captures the population proportion.

- Technically Correct

- We are **95% confident** that the interval from 23.5% to 38.0% captures the true proportion of Facebook users who update daily.

- More Casual But Fine

- We are **95% confident** that between 23.5% and 38.0% of Facebook users update daily.

- Do NOT use probability language

- There is a **95% chance** that the interval (0.235, 0.380) contains the true proportion of Facebook users who update daily. This is wrong!
Do **not** use this interpretation.

Margin of Error

Half length of the confidence interval, is called the margin of error, denoted by ME . For $100(1-\alpha)\%$ confidence interval of p , we have

$$ME = Z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Z-Confidence Interval for Estimating a Population Proportion

$$\hat{p} - ME < p < \hat{p} + ME, \text{ where } ME = Z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

or

$$(\hat{p} - ME, \hat{p} + ME)$$

Finding the Point Estimate and ME from a Confidence Interval

Point estimate of \hat{p} (*midpoint of the CI*) :

$$\hat{p} = \frac{\text{(upper confidence limit)} + \text{(lower confidence limit)}}{2}$$

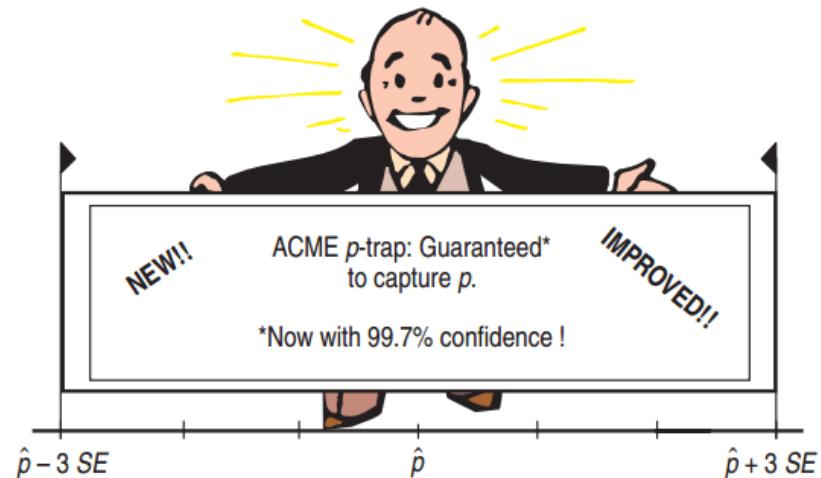
Margin of Error (half length of the CI):

$$ME = \frac{\text{(upper confidence limit)} - \text{(lower confidence limit)}}{2}$$

Example. For the “Facebook Daily Status Updates” example, a 95% confidence interval of the p - population proportion of people who update their Facebook status daily is given by (0.235, 0.380). Find the sample proportion and the margin of error.

Certainty vs. Precision

- Instead of a **95%** confidence interval, any percent can be used.
 - Increasing the confidence (e.g. **99%**) increases the margin of error.
 - Decreasing the confidence (e.g. **90%**) decreases the margin of error.



Certainty vs. Precision

- Assume that the confidence level and the value of the sample proportion does not change.
 - Increasing the **sample size** decreases the margin of error.
 - Decreasing the **sample size** increases the margin of error.

Critical Values

- For a **95%** confidence interval, the margin of error was **1.96SE**.
 - The critical value $Z_{0.025}=1.96$ comes from the normal curve
 - 95% of the area is within about $1.96SE$ from the proportion.

Example: Finding the Margin of Error

- Yale/George Mason Poll: 1010 US adults, 40% think scientists disagree about global warming ($\hat{p} = 0.4$). Find the margin of error at 95% and 90% confidence, respectively .

$$SE(\hat{p}) = \sqrt{\frac{(0.4)(0.6)}{1010}} \approx 0.0154$$

- For 95%, $z_{0.025} \approx 1.96$: $ME = (1.96)(0.0154) = 0.030$.
- For 90%, $z_{0.05} \approx 1.645$: $ME = (1.645)(0.0154) = 0.025$
which gives a smaller margin of error which is *good*.
- **Drawback:** lower level of confidence which is *bad*

Concept Question

Find the **margin of error** for the 95% confidence interval used to estimate the population proportion with $n = 163$ and $x = 96$.

- A. 0.0680
- B. 0.0755
- C. 0.132
- D. 0.00291

Assumptions and Conditions: Independence and Sample Size

– Independence Condition

- If data is collected using **SRS** or a randomized experiment
→ Randomization Condition
- Some data values do not influence others.
- Check for the **10% Condition**: The sample size is less than 10% of the population size.

– Success/Failure Condition

- There must be at least 10 successes: $n\hat{p} \geq 10$.
- There must be at least 10 failures: $n(1 - \hat{p}) \geq 10$.

Summary: One-Proportion Z-Interval

- First check for randomization, independence, **10% rule** and conditions on sample size.
- Confidence level **$1-\alpha$** , sample size **n** , sample proportion **\hat{p}** .
- Confidence interval:

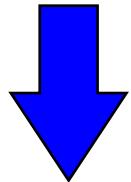
$$\hat{p} \pm Z_{\alpha/2} \cdot SE(\hat{p})$$

- where $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$
- $Z_{\alpha/2}$: the **critical value** corresponding to the confidence level **$1-\alpha$**

Choosing Sample Size

$$z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq ME,$$

where ME is the desired margin of error



(solve for n by algebra)

$$n \geq \frac{(z_{\alpha/2})^2 \hat{p}(1 - \hat{p})}{ME^2}$$

and round n up.

Choosing Sample Size

(1) When an estimate of \hat{p} is known:

$$n = \frac{(z_{\alpha/2})^2 \hat{p}(1 - \hat{p})}{ME^2}$$

(2) When no estimate of \hat{p} is known, let $\hat{p} = 0.5$

because 0.5 maximizes the $\hat{p}(1 - \hat{p})$

$$n = \frac{(z_{\alpha/2})^2 0.25}{ME^2}$$

Thoughts on Sample Size and ME

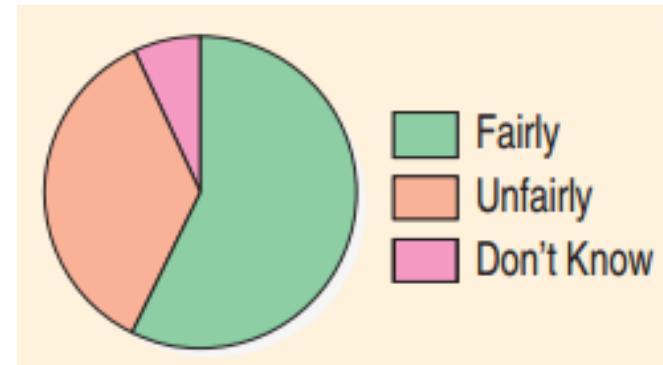
- Obtaining a large sample size can be expensive and/or take a long time.
- For a pilot study, $ME = 10\%$ can be acceptable.
- For full studies, $ME \leq 5\%$ is better.
- Public opinion polls typically use $ME = 3\%$, $n = 1000$.
- If p is expected to be very small such as 0.005 , then much smaller ME such as 0.1% is required. See Example 3.

Example 1: Do You Believe the Death Penalty is Applied Fairly?

– Sample size: **510**

– Answers:

- 58% “Fairly”
- 36% “Unfairly”
- 7% “Don’t Know”
- Construct a confidence interval for the population proportion that would reply “**Fairly**.”



Example 1: Do You Believe the Death Penalty is Applied Fairly?

- $n = 510$, $\hat{p} = 0.58$
- $SE(\hat{p}) = \sqrt{\frac{(0.58)(0.42)}{510}} \approx 0.022$
- $Z_{0.025} \approx 1.96$
- $ME \approx (1.96)(0.022) \approx 0.043$
- The 95% Confidence Interval is
 0.58 ± 0.043 or $(0.537, 0.623)$

Example 1: Do You Believe the Death Penalty is Applied Fairly?

- **Conclusion:** We are **95%** confident that between **57.3%** and **62.3%** of all US adults think that the death penalty is applied fairly.

Example 2: The Yale/George Mason Survey and Sample Size

- Poll: 40% believe scientists disagree on global warming.
 - For a follow-up survey, what sample size is needed to obtain a 95% confidence interval with $ME \leq 3\%$?
$$n = \frac{(1.96)^2 0.4 \times 0.6}{0.03^2}$$
 - $n \approx 1024.43$ and then round up.
 - The group will need at least 1025 respondents.

Example 2: The Yale/George Mason Survey and Sample Size

- What Sample Size if **no estimate** of p is available?
 - For **95%**, $Z_{0.025} = 1.96$
 - Let $\hat{p} = 0.5$
 - For example, to ensure a **$ME \leq 3\%$** :
$$n = \frac{(1.96)^2 0.25}{0.03^2}$$
 - The formula gives $n \approx 1067.1$ and then round up.
 - We need to survey at least **1068** to ensure a **ME** less than **0.03** for the **95%** confidence interval.

Example 3: Credit Cards and Sample Size

- A pilot study showed that 0.5% of credit card offers in the mail end up with the person signing up.
 - To be within 0.1% of the true rate with 95% confidence, how big does the test mailing have to be?

$$n = \frac{(1.96)^2 0.005 \times 0.995}{0.001^2}$$

- $n \approx 19,111.96$ and then **round up**
- The test mailing should include at least 19,112 offers.

Calculation Question

Margin of error 0.07; confidence level 95%; from a prior study, sample proportion is estimated by the decimal equivalent of 92%. **Find the minimum sample size required to estimate the population proportion.**

- A. 4
- B. 51
- C. 58
- D. 174

What Can Go Wrong?

- Don't claim other samples will agree with yours.
 - **Wrong:** In 95% of samples, between 43% and 51% agree with decriminalization of marijuana.
- Don't be certain about the parameter.
 - **Wrong:** Between 23% and 38% of Facebook users update daily. Don't forget to include the **confidence**.
- Don't forget that it's about the parameter.
 - **Wrong:** I'm 95% confident that \hat{p} is between 23% and 38%. You know for sure exactly what \hat{p} is.
- Do treat the whole interval equally.
 - The middle of the interval is not necessarily more plausible than the edges.

What Can Go Wrong?

- Beware of margins of error that are too large to be useful.
 - Between 10% and 90% update daily is not useful. Use a larger sample size to shrink the *ME*.
- Don't claim to know too much (what is your population?).
 - **Wrong:** I'm 95% confident that between 23% and 38% of all Facebook users in the world update daily. The survey was just about **US residents** between 18 and 22.
- Don't use probability language to interpret a CI.
 - **Wrong:** There is a 95% chance that the interval (23%, 38%) contains the true parameter.
- Don't suggest that the parameter varies.
 - **Wrong:** There is a 95% chance that the true parameter is between 23% and 38%.

Key Concepts

Interval estimation of population proportion p

- Understand the confidence interval formula
 - ✓ Point estimate and Margin of error
 - ✓ Certainty versus Precision
 - ✓ Interpretation of confidence intervals
- Find a confidence interval using technology
- Sample size calculation

8.3 Interval Estimation about Means

Conditions:

- (1) The population standard deviation σ is unknown.
- (2) Take a random sample (of any size, large or small) from a Normal population **or**
Take a random sample of size n ($n \geq 30$) from a population of any distribution type
- We estimate σ by the sample standard deviation s , but we do not use the Normal distribution in estimation.
- The new type of distribution to be used is **Student t Distribution**

Recall: The Central Limit Theorem (CLT)

- When a random sample is drawn from a population with mean μ and standard deviation σ , the sampling distribution has:
 - Mean: μ
 - Standard deviation: $\frac{\sigma}{\sqrt{n}}$
 - **Approximately Normal** distribution as long as the sample size is large.
 - The larger the distribution, the closer to Normal.

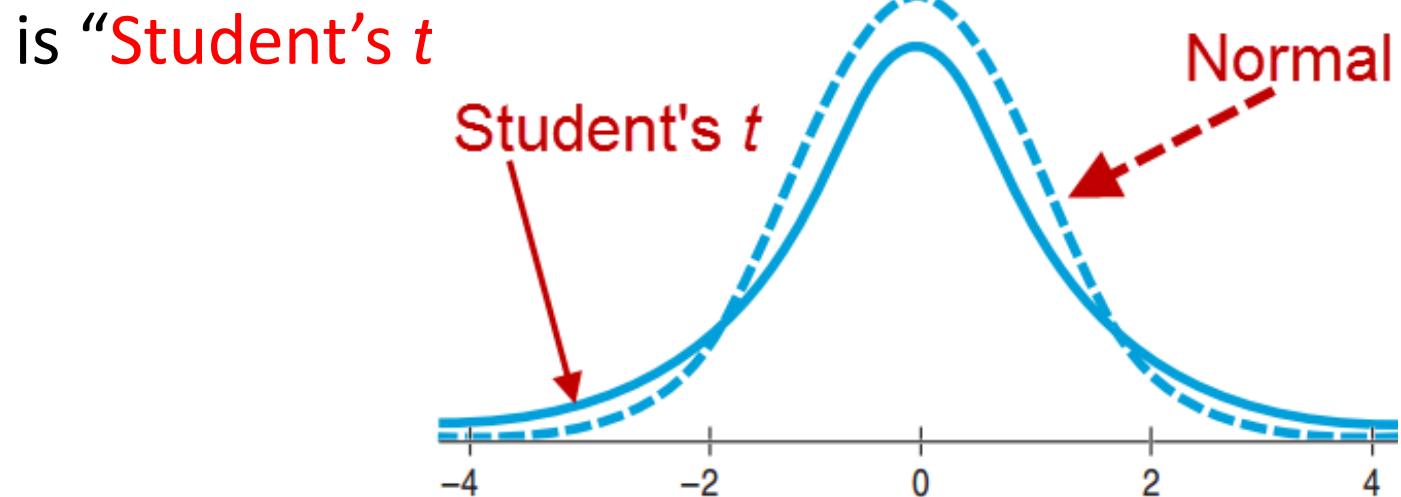
The Challenge of the CLT

- CLT tells us $SD(\bar{y}) = \frac{\sigma}{\sqrt{n}}$
- We would like to use this for Confidence Intervals and Hypothesis Testing.
- Unfortunately, we almost never know σ .
- Using s almost works: $SE(\bar{y}) = \frac{s}{\sqrt{n}}$, but not quite.
- When using s , the Normal model has some error.
- **William Gosset** came up with new models, one for each n that works better.

Student *t* Distribution

- If the distribution of a population is **approximately Normal**, or the sample size is large, then the correct model of

$$\frac{\bar{y} - \mu}{s / \sqrt{n}}$$

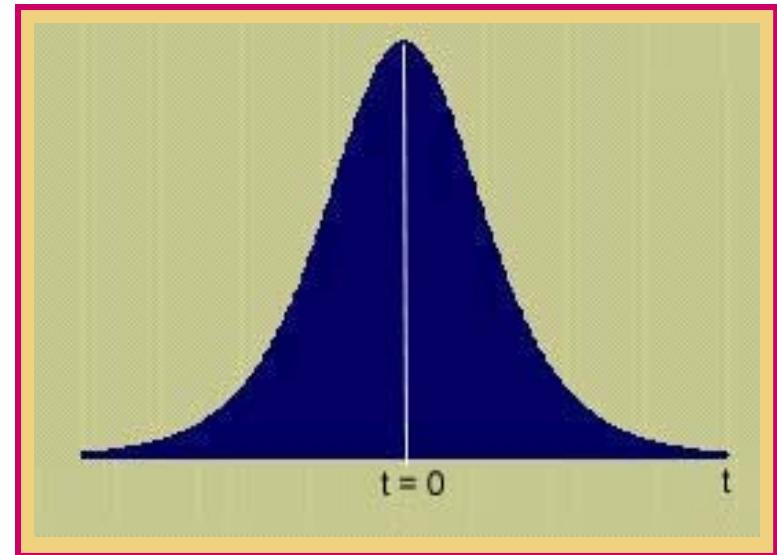


Student t Distribution

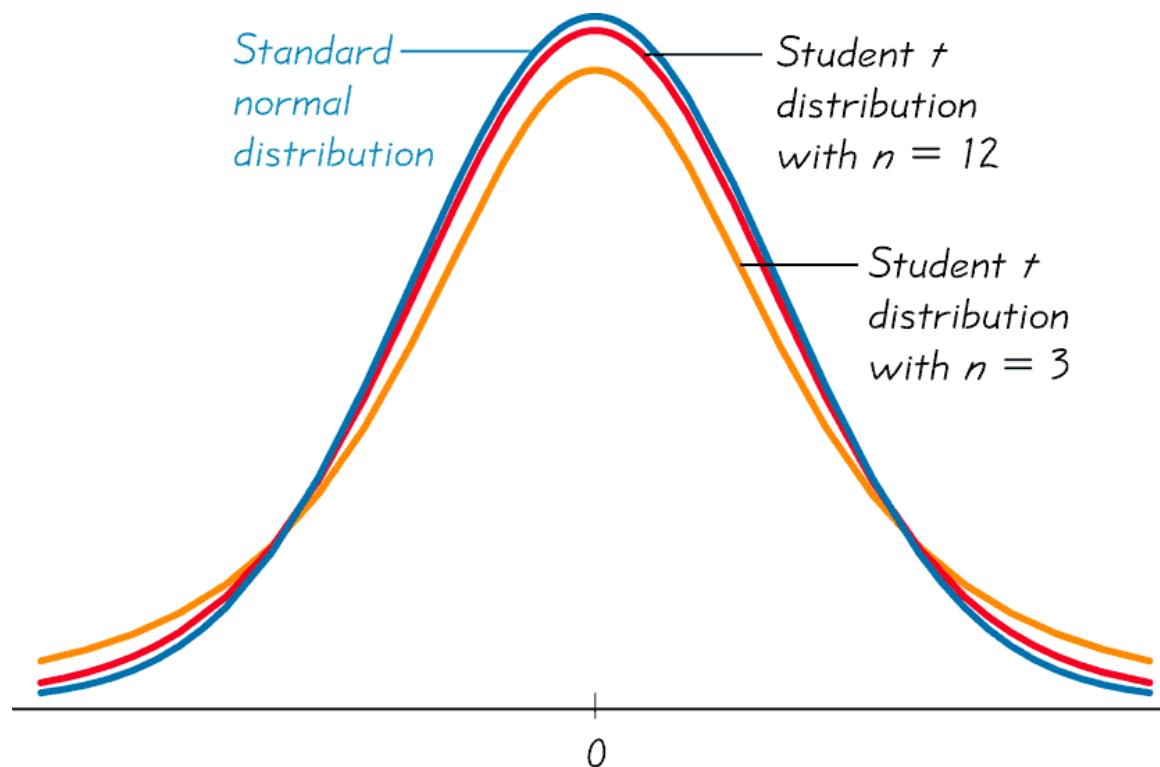
- For every sample size n there is a different Student's t distribution (the shape depends on $n-1$)
- Degrees of freedom: $df = n - 1$.
- Similar to the “ $n - 1$ ” in the formula for sample standard deviation
- It is the number of independent quantities left after we've estimated the parameters.

Properties of the Student t Distribution

1. The Student t distribution has the same general symmetric bell shape as the z standard normal distribution and symmetric about 0 (has a mean 0).
2. It reflects the **greater variability** (with heavier tails and the standard deviation **greater than 1**) compared to z , the standard normal distribution.
3. Shape depends on the sample size n or the degrees of freedom, $n-1$.
4. As n increases the shapes of the t and z distributions become almost identical.



Student t Distributions for $n = 3$ and $n = 12$



Thoughts about z and t

- The Student's t distribution:
 - Is very close to Normal for large n .
 - Is needed because we are using s as an estimate for σ .
- If you happen to know σ , which almost never happens, use the **Normal model** and not Student's t .

Summary: Sampling Distribution Model for Sample Means

- With certain conditions (seen later), the standardized sample mean follows the Student's t model with $n - 1$ degrees of freedom.

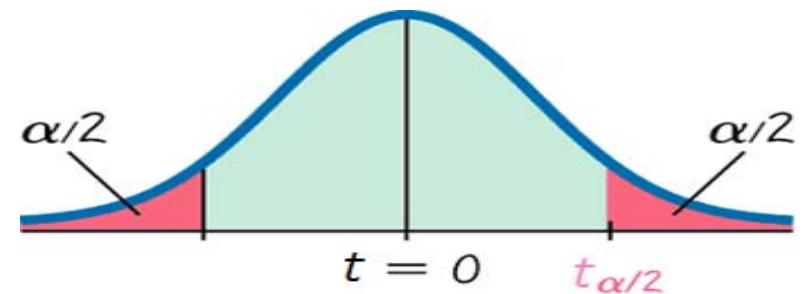
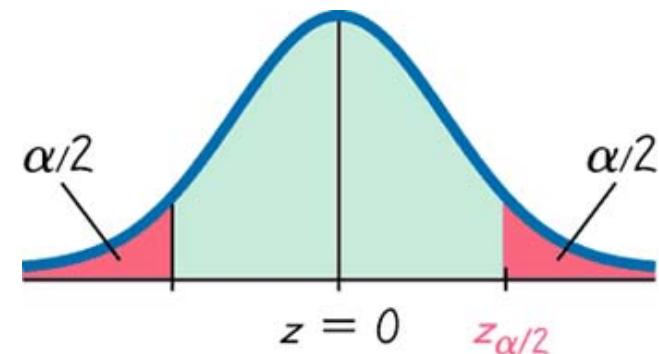
$$t = \frac{\bar{y} - \mu}{SE(\bar{y})}$$

where we estimate the standard deviation of the sample mean with

$$SE(\bar{y}) = \frac{s}{\sqrt{n}}$$

Critical Values: Z and t distribution

- Recall the number $z_{\alpha/2}$ is a critical value that is a z score with the property that it separates an area of $\alpha/2$ in the right tail of the standard normal distribution
- The number $t_{\alpha/2}$ is a critical value that is a t score with the property that it separates an area of $\alpha/2$ in the right tail of the t -distribution



Construction of a Confidence Interval of μ

- To construct a $100(1-\alpha)\%$ confidence interval of μ from a random sample of size n , we need

$$P\left(-t_{\alpha/2} < \frac{\bar{y} - \mu}{s / \sqrt{n}} < t_{\alpha/2}\right) = 1 - \alpha$$

Now let's manipulate the inequalities inside the parentheses in so that they appear in the form of $L < \mu < U$ such that

$$P(L < \mu < U) = 1 - \alpha$$

Construction of a Confidence Interval of μ

- Multiply through by $\frac{s}{\sqrt{n}}$

$$P\left(-t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} < \bar{y} - \mu < t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

- Subtract \bar{y} from each term

$$P\left(-\bar{y} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} < -\mu < -\bar{y} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

Construction of a Confidence Interval of μ

- Multiply through by -1 to eliminate the minus sign in front of μ

$$P\left(\bar{y} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} > \mu > \bar{y} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

- That is

$$P\left(\bar{y} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{y} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

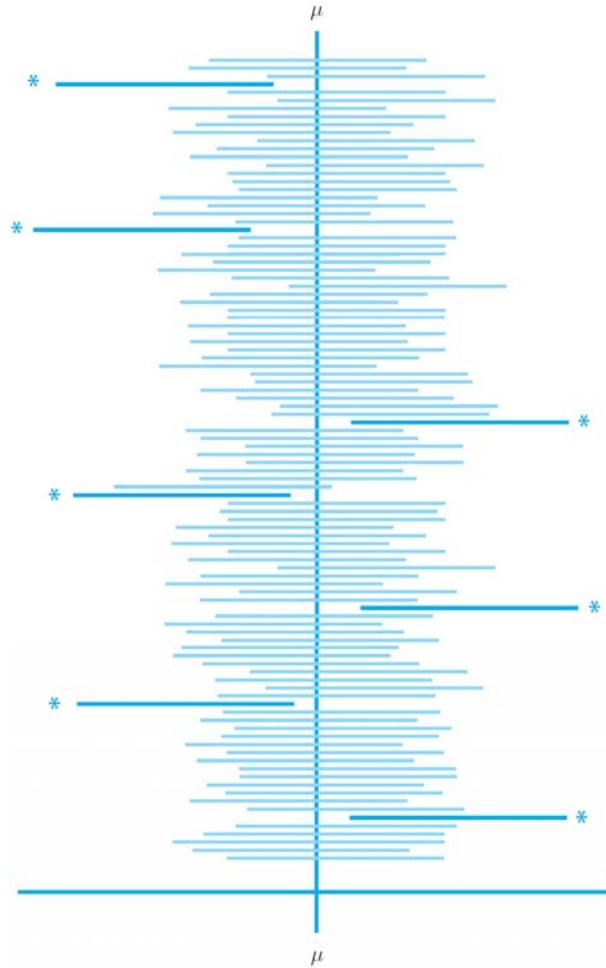
Construction of a Confidence Interval of μ

- A $100(1-\alpha)\%$ confidence interval of μ is

$$\left(\bar{y} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}, \quad \bar{y} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \right)$$

- ❖ **Note.** This interval is not **random** any more if the statistic \bar{y} is replaced by a sample mean calculated from a specific sample.

Interpreting a Confidence Interval



One hundred 95% CIs (asterisks identify intervals that do not include μ).

Simulating Confidence Intervals

- Population (standard normal) mean $\mu=0$
- Sample size $n=15$
- Take 100 samples from the population
- Construct a 95% symmetric two-tailed confidence interval of μ
- Count the number of confidence intervals containing μ
- Repeat the process for 1000, 10,000 ... samples.

Simulating Confidence Intervals

- <https://www.statcrunch.com/applets/type3&comean>
- Or use R coding in the next slide

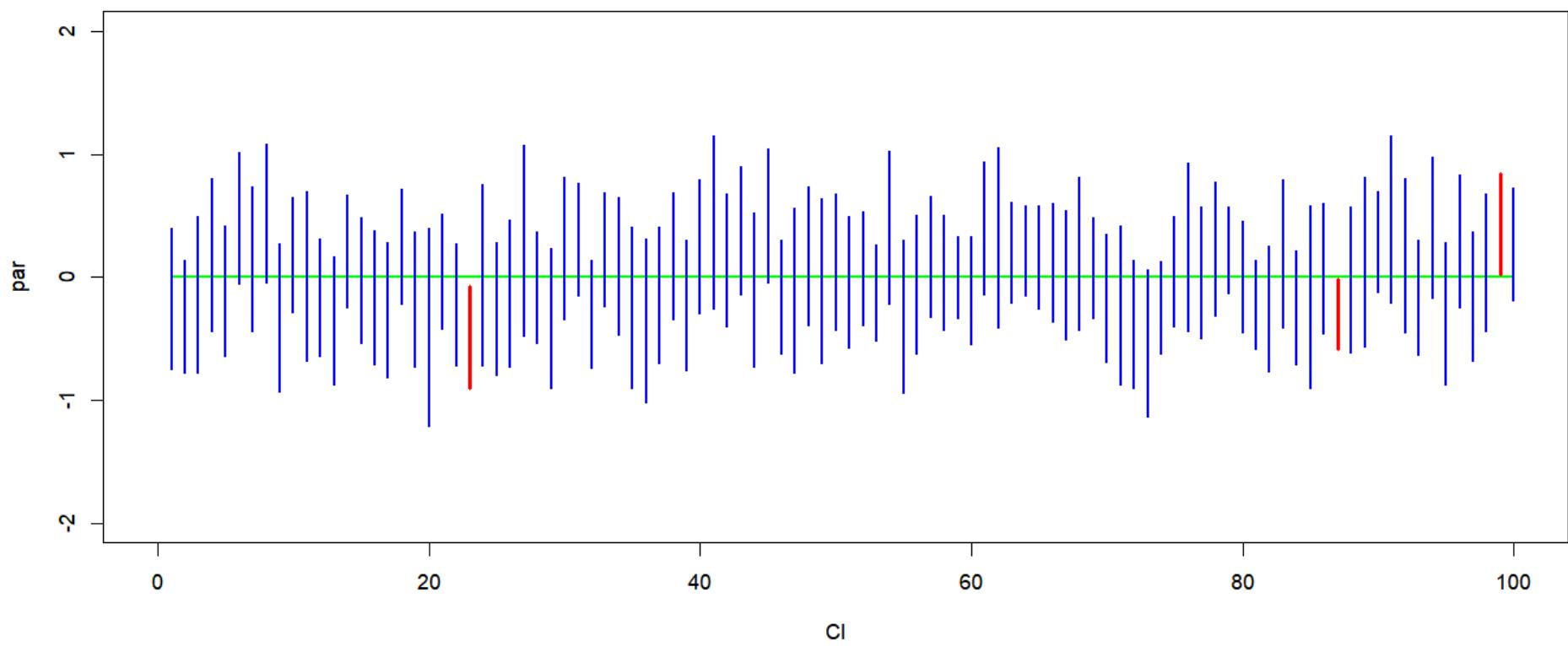
Simulating Confidence Intervals

R code

```
n = 15;                      #sample size
mu=0;   conlevel=0.95;      #confidence level
iter=100;
count=0;          #number of intervals containing mu
L=numeric(); U=numeric(); CI=1:iter; par=rep(mu,iter);
plot(CI,par,type="l",xlim=c(0,iter), ylim=c(-2,2),lwd=2, col="green");
for (i in 1:iter)
{Sys.sleep(0.5);           #suspend 0.5 second
x=rnorm(n, mean=0, sd=1);    #Generation of samples
interval = t.test(x, conf.level=conlevel)$conf.int;  #C.I.
L[i]= interval[1]; U[i]=interval[2];
if (L[i]<=mu && U[i]>=mu) count=count+1;
time=c(i,i); bds=c(L[i], U[i]);
if (L[i]>mu | U[i]<mu)
{lines(time,bds,type="l", col="red", lwd=3);
} else {
lines(time,bds,type="l", col="blue", lwd=2);
}
}
count;
count/iter;
```

Simulating Confidence Intervals

One simulation result



Interpreting a Confidence Interval

- ❖ If we repeat the sampling and compute the $100(1-\alpha)\%$ confidence interval of μ from each sample following the formula

$$\left(\bar{y} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}, \quad \bar{y} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \right)$$

- ❖ Then in the long run, $100(1-\alpha)\%$ such confidence intervals will actually contain the true mean μ .

Interpreting a Confidence Interval

For a specific of 95% CI of μ : (0.828, 0.872) which is constructed from a single sample such that the sample mean is a **number** instead of a statistic. **We will never know whether it captures the population mean.**

We can simply say

- ❖ “We are 95% **confident** that the interval from 0.828 to 0.872 actually does contain the true value of the population mean μ .”
- ❖ But we **cannot** say
 - (1) “There is a 95% chance/probability that the true value of μ will fall between 0.828 and 0.872”;
 - (2) “95% of the sample means will fall between 0.828 and 0.872”.

Summary: One Sample t -Interval for the Mean

- When the assumptions are met (seen later), the confidence interval for the mean is

$$\left(\bar{y} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}, \quad \bar{y} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \right)$$

- The critical value $t_{\alpha/2}$ depends on the confidence level $1 - \alpha$, and the degrees of freedom $n - 1$.

CI Calculation with Technology

- (1) Inferences->Statistics if summary statistics are given
or
Inferences->Data if detailed data are given in
<https://esumath.shinyapps.io/Rstats/>

- (2) https://istats.shinyapps.io/Inference_mean/

Example 1: Contaminated Salmon

- A study of mirex concentrations in salmon found
 - $n = 150$, $\bar{y} = 0.0913$ ppm, $s = 0.0495$ ppm
 - Find a 95% confidence interval for mirex concentrations in salmon.
 - $df = 150 - 1 = 149$
 - $SE(\bar{y}) = \frac{0.0495}{\sqrt{150}} \approx 0.0040$
 - Critical value $t_{0.025=1.976}$
(R: $qt(0.975, 149)$)

Example 1: Contaminated Salmon

- Confidence Interval for μ (or use R! See handout!)

$$\bar{y} \pm t_{0.025} \times SE(\bar{y}) = 0.0913 \pm 1.976 \times 0.0040$$

$$= 0.0913 \pm 0.0079$$

$$= (0.0834, 0.0992)$$

- We are 95% confident that the mean level of mirex concentration in farm-raised salmon is between 0.0834 and 0.0992 parts per million.

Calculation Question

Thirty randomly selected students took the calculus final. If the sample mean was 95 and the standard deviation was 6.6, construct a 99% confidence interval for the mean score of all students.

- A. $92.03 < \mu < 97.97$
- B. $91.68 < \mu < 98.32$

Example 2: Paint Drying Time

The Labels on 1-gallon cans of paint usually indicate the drying time and the area that can be covered in one coat. A random sample of ten 1-gallon cans of one brand white paint were used to paint ten identical areas using the same kind of equipment. The actual areas covered by these 10 gallons of paint are given here:

310, 311, 412, 368, 447, 376, 303, 410, 365, 350

Assume the distribution of the relevant **population is normal**. Find a 98% confidence interval of μ , the average coverage area for this white paint.

Example 2: Paint Drying Time

- Confidence Interval for μ (using R)

- $(322.002, \quad 408.398)$
 - We are 98% confident that the confidence interval $(322.002, 408.398)$ contains the true average coverage area for this white paint.

Assumptions and Conditions (Recall CLT)

- Independence Condition

- Randomization Condition: The data should arise from a suitably randomized experiment.
- Sample size $< 10\%$ of the population size.

- Nearly Normal

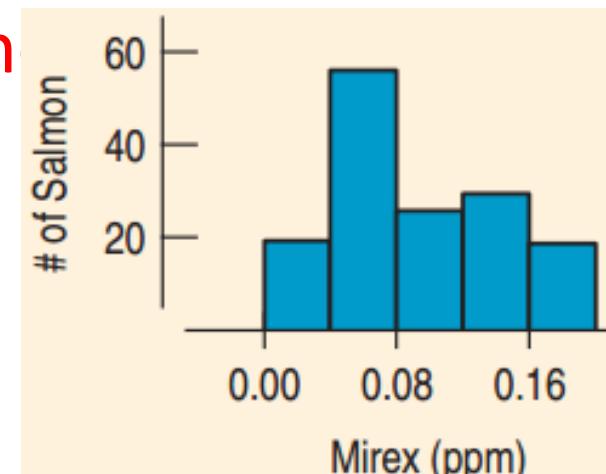
- For large sample sizes ($n > 30$), not severely skewed.
- ($15 \leq n \leq 30$): Need unimodal and symmetric.
- ($n < 15$): Need almost perfectly normal.
- Check with a histogram.

Assumptions for Salmon Contamination Study

- Researchers tested 150 salmon from 51 farms in eight regions in six countries. Are the assumptions and conditions for inference satisfied?

✓ **Independence Assumption:**
Not a random sample, but likely independently selected.

✓ **Nearly Normal Condition:**
The histogram is unimodal and not highly skewed. OK since sample size is 51.



Make a Picture, Make a Picture, Make a Picture

- **Always Test the Normality Assumption**

- Create a **histogram** to check for near normality.
 - Good for seeing the nature of the deviations: outliers, not symmetric, skewed
- Also create a **normal quantile plot** to see that it is reasonably straight.
 - Good for checking for normality
- With technology at hand, there is no excuse not to make these two plots.

Margin of Error

Recall: Half length of the confidence interval, is called the **margin of error**, denoted by ***ME***. We have

$$ME = t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

t-Confidence Interval for Estimating a Population Mean

$$\bar{y} - ME < \mu < \bar{y} + ME, \text{ where } ME = t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

or

$$(\bar{y} - ME, \bar{y} + ME)$$

Finding the Point Estimate and ME from a Confidence Interval

Point estimate of μ (*midpoint of the CI*):

$$\bar{y} = \frac{\text{(upper confidence limit)} + \text{(lower confidence limit)}}{2}$$

Margin of Error (*half length of the CI*):

$$ME = \frac{\text{(upper confidence limit)} - \text{(lower confidence limit)}}{2}$$

Concept Question

A 98% confidence interval of population mean is given by (0.434, 0.548). Find the margin of error ME of this confidence interval. Round the margin of error to four decimal places.

- A. 0.114
- B. 0.057
- C. 0.0285

Certainty vs. Precision

- Instead of a **95%** confidence interval, any percent can be used.
 - Increasing the confidence (e.g. **99%**) increases the margin of error.
 - Decreasing the confidence (e.g. **90%**) decreases the margin of error.

Certainty vs. Precision

- Assume that the confidence level and the value of the sample mean does not change.
 - Increasing the **sample size** decreases the margin of error.
 - Decreasing the **sample size** increases the margin of error.

Choosing Sample Size

(reading material; not in any tests)

1. Determine the size of the margin of error, E , that you are willing to tolerate.
2. Choose the sample size by solving for n in the equality:

$$ME = t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

However, $t_{\alpha/2}$ depends on n .

The Challenge of Finding the Sample Size

$$ME = t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

- To find the necessary sample size in order to have a small enough margin of error:
 - Decide on acceptable ME .
 - Determine s : Use a pilot to estimate s .
 - Determine $t_{\alpha/2}$: Use *z-critical value* as an estimate. For example, by the 68-95-99.7 Rule, use 2 (≈ 1.96) for 95% confidence.

Finding the Sample Size

- Example.** Should we buy a movie download accelerator?
With a free trial we can test several download times in
order to obtain a 95% CI with a $ME < 8$ minutes.
Assume $\sigma = 10$ minutes.

- With $z_{0.025} \approx 2$, $8 = 2 \times \frac{10}{\sqrt{n}}$, $\sqrt{n} = \frac{20}{8}$, $n = 6.25$
- Note. With the small sample size, z and t *critical values* differ.
With $n=6$ and thus $df=6-1=5$,

$$t_{0.025} = 2.571 \text{ (qt}(0.975, 5) \text{)}$$

Finding the Sample Size

Example continued.

– Estimate n again: $8 = 2.571 \times \frac{10}{\sqrt{n}}$

$$\sqrt{n} = \frac{2.571 \times 10}{8} \approx 3.214, \quad n \approx 10.33$$

- To make sure ME is no larger than 8 minutes, **round up**.
- We'll find the downloading times for 11 movies.

Finding the Sample Size

- Sample size calculations are **never** exact. But it is always a good idea to know whether your sample size is large enough to give you a good chance of being able to tell you what you want to know before you collect your data.

Key Concepts

Interval estimation of population mean μ

- Understand the confidence formula
 - ✓ t distribution
 - ✓ Point estimate and Margin of error
 - ✓ Certainty versus Precision
 - ✓ Interpretation of confidence intervals
- Find a confidence interval using technology

Hypothesis Testing

The material is covered by Chapter 7 of the book “Statistics Using Technology”.

- ✓ Hypothesis test of population proportion
- ✓ Hypothesis test of population mean

Section 1

Basics of Hypothesis Testing

Inferential statistics:

- (1) **Estimation** of a population parameter
- (2) **Test a hypothesis** or claim about a population parameter.

Example

- **Health:** It is often claimed that the **mean** body temperature is 98.6 degrees. We can test this claim using a **sample** of 106 body temperatures.

Introduction



(Antibiotic Potency) Suppose that a drug manufacturer is concerned if the mean potency μ of an antibiotic meets the minimum government potency standards μ_0 . They need to decide between two possibilities:

- The mean potency μ does not exceed the minimum allowable potency μ_0 ;
- The mean potency μ exceeds the minimum allowable potency μ_0 .

This is an example of **hypothesis**.

Any statement regarding the value of a **population parameter** is called a **hypothesis**.

Null Hypothesis

Null Hypothesis

- The **null hypothesis** (denoted by H_0) is a statement that the value of a population parameter (such as mean, or proportion) is **equal to** some claimed value of a parameter. (Only Equal sign is used in H_0 , to be explained).
- Note the parameter describes the **population** not the sample.

Null Hypothesis

Null Hypothesis

- We test the null hypothesis directly in the sense that we **assume it is true** and reach a conclusion to either **reject H_0** or **fail to reject H_0** .

Alternative Hypothesis

- The **alternative hypothesis** (denoted by H_1 or H_A) is the statement that is **contradictory** to the null hypothesis.
- The symbolic form of the alternative hypothesis must use one of these symbols: $>$, $<$, \neq . For example,
 - $H_A: \mu > \mu_0$ if we decide whether a population mean μ is greater than a specified value μ_0 .
 - $H_A: \mu < \mu_0$ if we decide whether a population mean μ is less than a specified value μ_0 .
 - $H_A: \mu \neq \mu_0$ if we decide whether a population mean μ is different from a specified value μ_0 .

Alternative Hypothesis

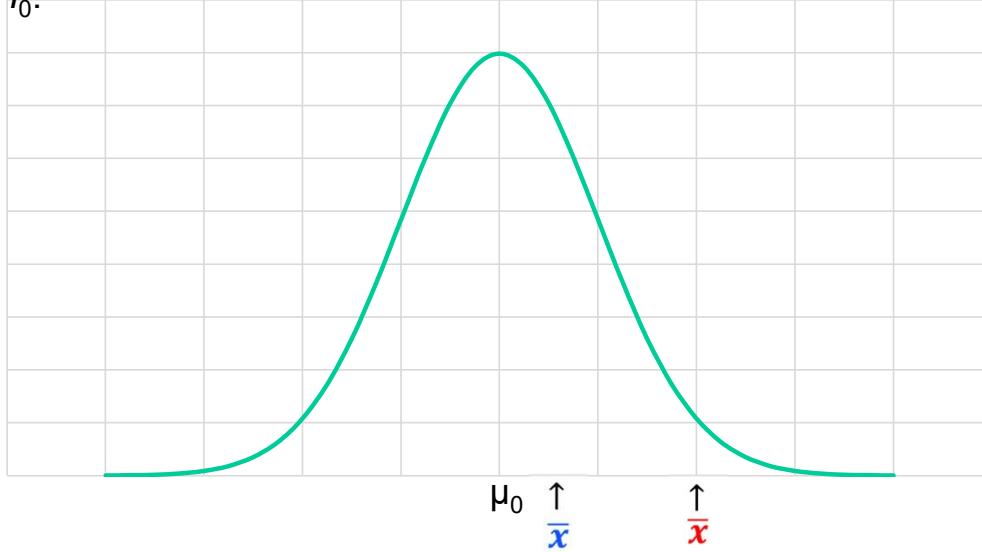
- This is the hypothesis that the researcher wishes to **support**. If we are conducting a study and want to use a hypothesis test to **support** our claim, the claim must be worded so that it becomes the alternative.
 - Null hypothesis and alternative hypothesis are complementary each other. But we can always use **equality** only in a null hypothesis. See slides 12 – 15.
- ❖ A **hypothesis test** is a procedure for testing H_0 versus the alternative H_A to decide whether H_0 should be rejected in favor of H_A .

Rare Event Rule for Inferential Statistics

If, under a **given assumption** (H_0 is true in the hypothesis testing problems), the probability of a particular observed event is extremely small, we conclude that the assumption is probably not correct.

Rare Event Rule for Inferential Statistics

Example: Test $H_0: \mu = \mu_0$ ($H_0: \mu \leq \mu_0$) versus $H_1: \mu > \mu_0$. The following curve is the sampling distribution of the sample mean under H_0 .



Components of a Hypothesis Test

1. The null hypothesis, H_0 , and the alternative hypothesis, H_A .

Pharmaceuticals:

$$H_0: \mu = \mu_0 \quad (\mu \leq \mu_0)$$

$$H_A: \mu > \mu_0$$

2. The test statistic:

A single statistic (quantity) calculated from the sample which will allow us to reject or not reject H_0 . It is constructed by **converting the point estimate** (such as \bar{x} and \hat{p}) to a score (such as t and z).

Components of a Hypothesis Test

3. Use appropriate method to make a decision as to reject or not to reject H_0 .

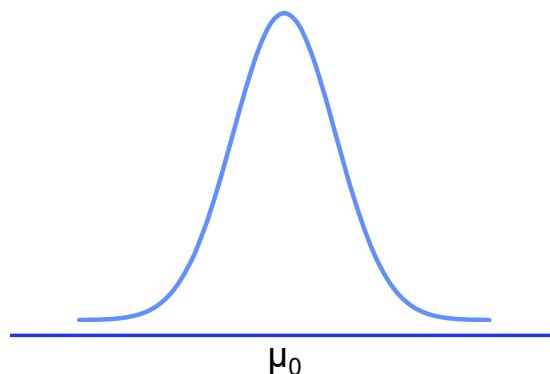
- i) **Critical value(s) method:** determine the Rejection Region
- ii) **P-value** method (technology required)

4. Conclusion:

A statement that uses simple nontechnical wording that addresses the original claim.

Caution: Null Hypothesis - Equal sign only

- The null hypothesis H_0 is **contradictory** to H_A . For example, if the statement we want to support $H_A: \mu > \mu_0$ is false, then $H_0: \mu \leq \mu_0$, which corresponds to a **family** of sampling distributions, is true.
- We **always** (in this textbook) write the null hypothesis as
 - H_0 : parameter = hypothesized value
 - We **always** use a single fixed value in the construction of a test statistics to have a single sampling distribution.
 - If $H_0: \mu = \mu_0$ (versus $H_A: \mu > \mu_0$) is rejected, $H_0: \mu \leq \mu_0$ will be rejected as well.



Concept Question

Carter Motor Company claims that its new sedan, the Libra, will average better than 23 miles per gallon in the city. Use μ , the true average mileage of the Libra.

- A. $H_0: \mu < 23$ $H_a: \mu \geq 23$
- B. $H_0: \mu = 23$ $H_a: \mu < 23$
- C. $H_0: \mu > 23$ $H_a: \mu \leq 23$
- D. $H_0: \mu = 23$ $H_a: \mu > 23$

One and two-sided alternatives

- The null hypothesis is **always** a single value of the parameter (**always** using the “=” symbol). For example, with respect to population mean,

$$H_0 : \mu = \mu_0$$

- The alternative hypothesis (the symbolic form must use one of the symbols: $<$, $>$, \neq) can be two sided:

$$H_A : \mu \neq \mu_0$$

- Or one-sided:

$$H_A : \mu > \mu_0$$

$$H_A : \mu < \mu_0$$

Caution: Null Hypothesis - Equal sign only

Signs in H_0 and H_1 and Tails of a Test

	Two-sided Test	Left-sided Test	Right-sided Test
Sign in the null hypothesis H_0	=	= or \geq	= or \leq
Sign in the alternative hypothesis H_1	\neq	<	>

- The null hypothesis always has an *equal to* (=), or a *greater than or equal to* (\geq), or a *less than or equal to* (\leq) sign, and the alternative hypothesis always has a *not equal to* (\neq) or a *less than* ($<$) or a *greater than* ($>$) sign.

Accept Versus Fail to Reject

- Some texts use “accept the null hypothesis.”
- We **never** say “accept the null hypothesis.”
- We are not proving the null hypothesis when the null hypothesis not rejected. Fail to reject says more correctly that the available evidence is not strong enough to warrant rejection of the null hypothesis.

Conclusion

- Never conclude a hypothesis test with a statement of “reject the null hypothesis” or “fail to reject the null hypothesis.”
- Always make sense of the conclusion with a statement that uses **simple nontechnical wording** that addresses the original claim.
 - ✓ If the null hypothesis is rejected, there is sufficient evidence to support the alternative.
 - ✓ If the null hypothesis is NOT rejected, there is insufficient evidence to support the alternative.

Concept Question

Which of the following do NOT correctly describe an alternative hypothesis about a population mean μ ?

- A. $H_1: \mu > 1.62$
- B. $H_1: \mu = 1.62$
- C. $H_1: \mu < 1.62$
- D. $H_1: \mu \neq 1.62$
- E. $H_1: \mu \geq 1.62$

Two Types of Errors

There are two types of errors which can occur in a statistical hypothesis test.

Actual Fact Our Decision	H ₀ true	H ₀ false (H _A true)
H ₀ false (Reject H ₀)	Type I Error	Correct
H ₀ true (Fail to reject H ₀)	Correct	Type II Error

Define:

$$\alpha = P(\text{Type I error}) = P(\text{reject } H_0 \text{ when } H_0 \text{ is true})$$

$$\beta = P(\text{Type II error}) = P(\text{fail to reject } H_0 \text{ when } H_0 \text{ is false})$$

Two Types of Errors

H_0 : The defendant is innocent

		Defendant Innocent	Defendant Guilty
Reject Presumption of Innocence (Guilty Verdict)	Defendant Innocent	Type I Error	Correct
	Defendant Guilty	Correct	Type II Error
Fail to Reject Presumption of Innocence (Not Guilty Verdict)	Correct	Type II Error	

Two Types of Errors

Significance level α :

α is the maximum probability of committing type I error (making the mistake of rejecting H_0 when it is true) that one can tolerate.

- ❖ Usually, the significance level is $\alpha = 0.01$, $\alpha = 0.05$ or $\alpha = 0.1$ though the level in any particular problem will depend on the seriousness of a type I error.

Two Types of Errors

We want to keep the probabilities of error as small as possible.

- The probability of type I error, α , is called the **significance level** of a hypothesis test. It is controlled by the experimenter.
- The value of β can be calculated if we fail to reject a **false null hypothesis**.

Controlling Type I and Type II Errors

- For any fixed sample size n , a decrease in α will cause an increase in β . Conversely, an increase in α will cause a decrease in β .
- For any fixed α , an increase in the sample size n will cause a decrease in β .
- To decrease both α and β , increase the sample size.

DIY Example (Type I and II errors)

Pharmaceuticals:

$H_0: \mu = \mu_0$ ($\mu \leq \mu_0$) mean potency μ does not exceed the minimum allowable potency μ_0

$H_A: \mu > \mu_0$ mean potency μ exceeds the minimum allowable potency μ_0

Type I error = Reject H_0 | $H_0 =$

Type II error = Fail to reject H_0 | $H_A =$

Methods of testing a hypothesis

- After a **test statistic** is determined, we can then test using one of the three methods to decide if H_0 can be rejected:
 - 1) **critical value(s)** of the test statistic
 - 2) using a *p*-value (technology required, to be discussed).
 - 3) Confidence interval method (reading only)
- ❖ We'll first introduce the **critical value approach** to hypothesis testing.

Methods of testing a hypothesis

How do we determine the rejection region using critical value method?

Rejection region : Assuming that H_0 is true, the rejection region is the set of all values of the test statistic in which the null hypothesis H_0 should be rejected.

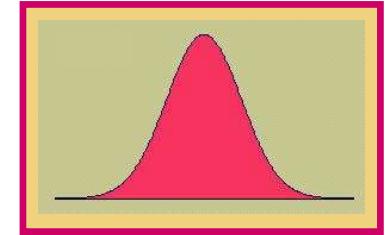
- Depends on the significance level, α .

Key Concepts

1. Components of a Hypothesis Test.
2. Type I and Type II error.
3. Two-sided and one-sided test.

Section 2. A Hypothesis Test for the Mean –Critical Value Method

Test of a Population Mean, μ : σ unknown



- Take a random sample of size n from a population with mean μ and unknown standard deviation σ .
- We assume that either or both
 1. The population is approximately Normally distributed
 2. sample size large ($n \geq 30$)
- Consider one of the three alternatives.

$H_0: \mu = \mu_0$ versus one of $H_A: \mu > \mu_0$

$H_A: \mu < \mu_0$

$H_A: \mu \neq \mu_0$

One-Sample t -Test for the Mean

- $H_0: \mu = \mu_0$
- Test statistic

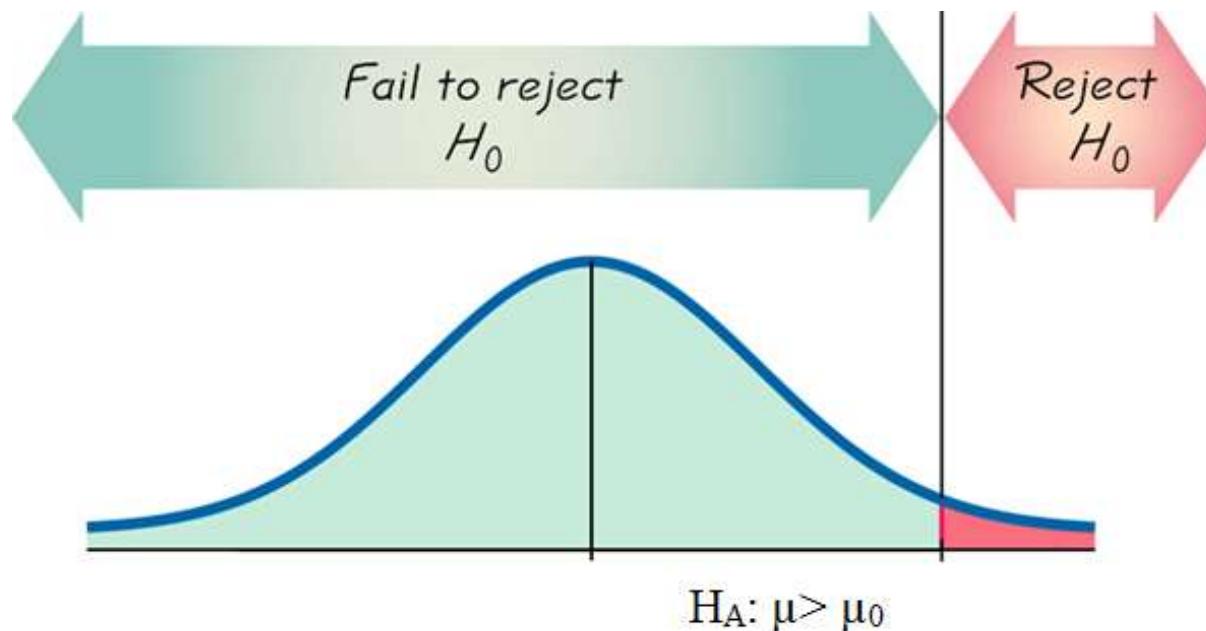
$$t_0 = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}.$$

where Standard Error of \bar{x} : $SE(\bar{x}) = s / \sqrt{n}$

- When the conditions are met and H_0 is true, the test statistic t_0 follows the **Student's t** Model with $df = n-1$.

Rejection Region

- For example, consider $H_A: \mu > \mu_0$
- If H_0 is true, the value of \bar{x} should be close to μ_0 , and t_0 will be close to 0. If H_0 is false, \bar{x} will be much larger than μ_0 , and t_0 will be much larger than 0, indicating that we should reject H_0 .



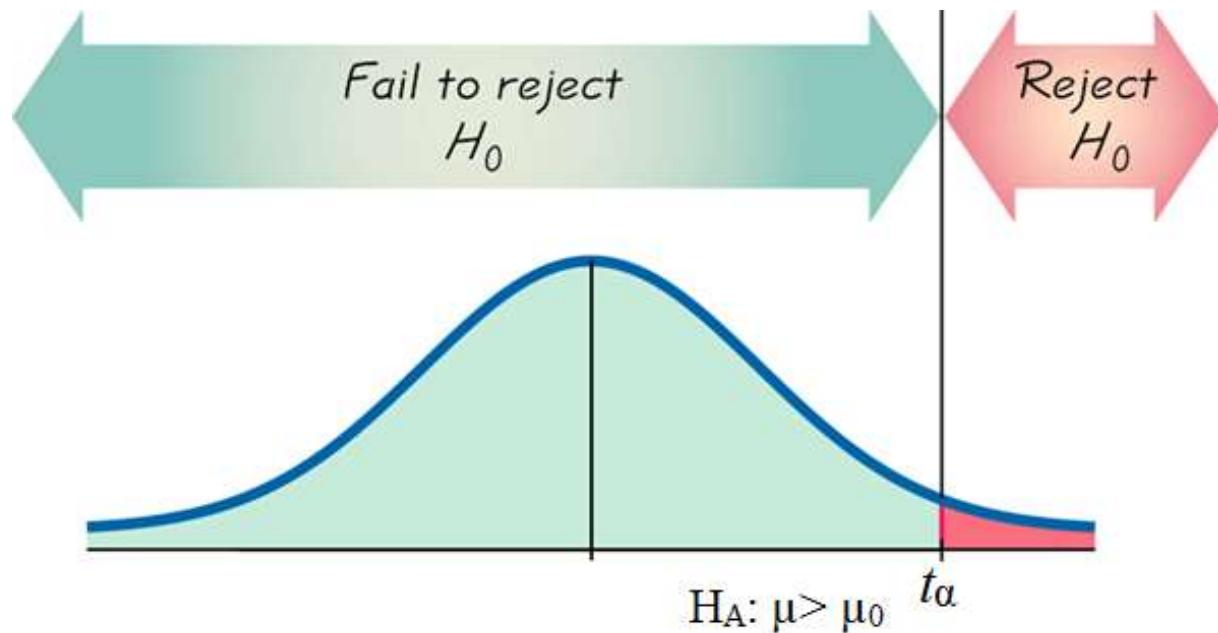
Rejection Region

- The value of the test statistic that separates the rejection region (rejecting H_0) and non-rejection region (not rejecting H_0) is called the **critical value**.
- **Recall:** t_α will denote the t axis for which α of the area under the t curve lies to the right of t_α .

Once the significance level α is fixed, the Rejection Region is set such that the probability of making a type I error is α .

Rejection Region

$H_A: \mu > \mu_0$

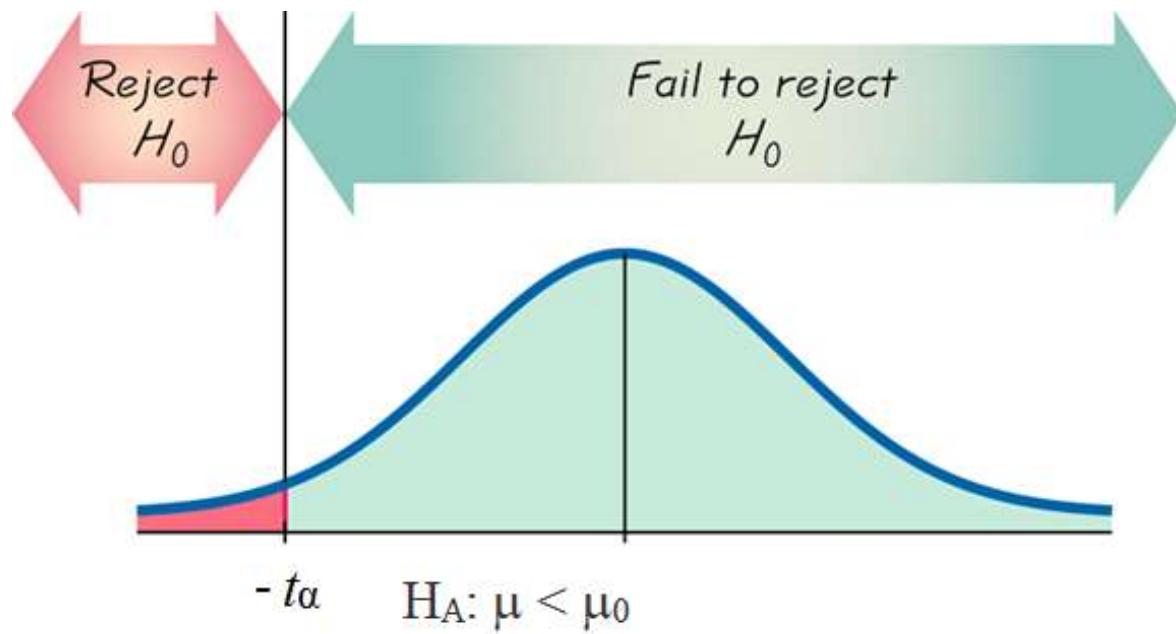


Rejection Region of a right-sided test:

Reject H_0 only if $t_0 \geq t_\alpha$.

Rejection Region

$$H_A: \mu < \mu_0$$

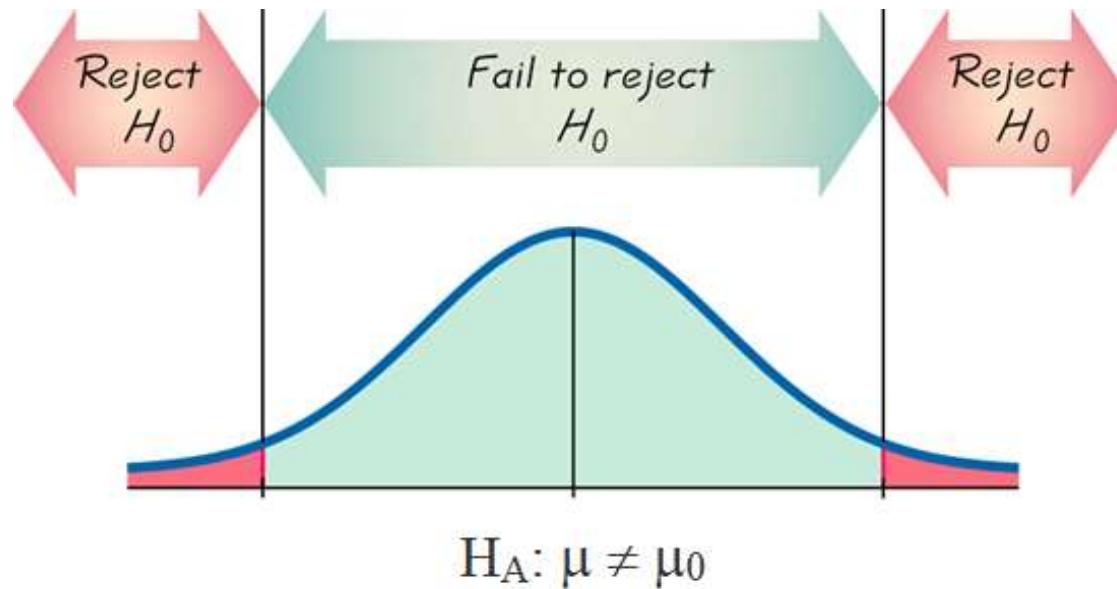


Rejection Region of a left-sided test:

Reject H_0 only if $t_0 \leq -t_\alpha$.

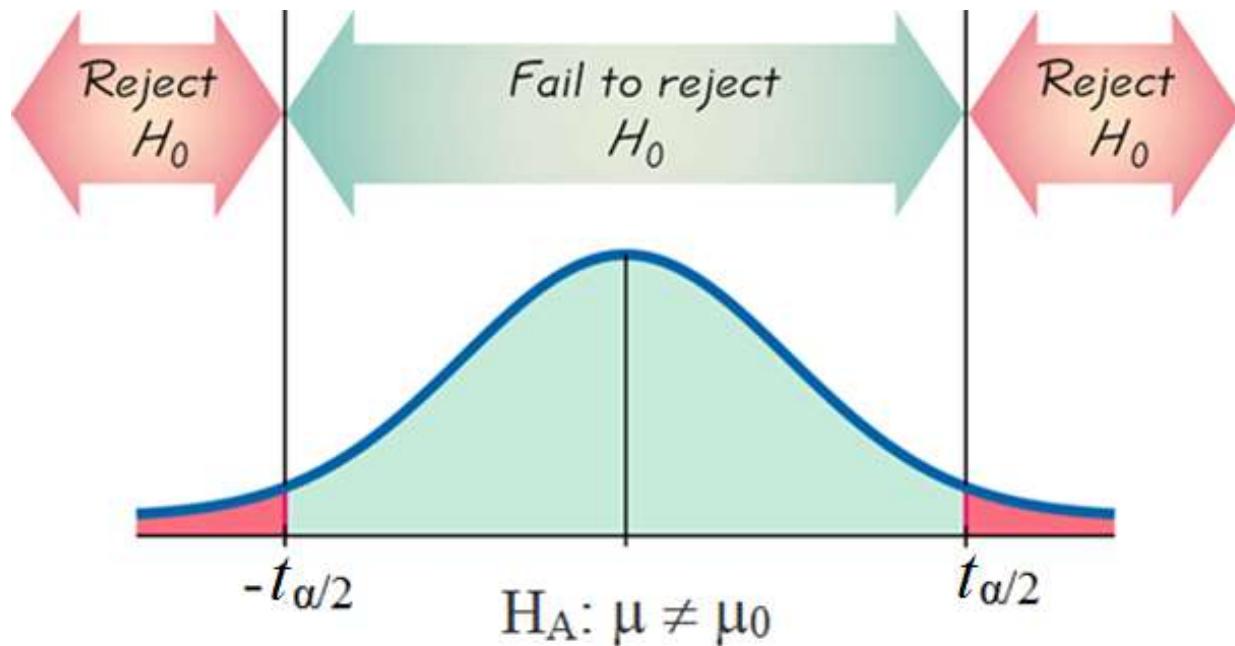
Rejection Region

- Now consider $H_A: \mu \neq \mu_0$
- If H_0 is true, the value of \bar{x} should be close to μ_0 , and t_0 will be close to 0. If H_0 is false, \bar{x} will be much larger or smaller than μ_0 , and t_0 will be much larger or smaller than 0, indicating that we should reject H_0 .



Rejection Region

$$H_A: \mu \neq \mu_0$$



Rejection Region of a two-sided test:

Reject H_0 only if $t_0 \leq -t_{\alpha/2}$ or $t_0 \geq t_{\alpha/2}$.

Summary of critical value method of t-test with significance level α

To test $H_0 : \mu = \mu_0$ versus H_A : one or two sided at level α ,

(1) $H_A : \mu > \mu_0$, then the rejection region is

$$\{t : t \geq t_\alpha\};$$

(2) $H_A : \mu < \mu_0$, then the rejection region is

$$\{t : t \leq -t_\alpha\};$$

(3) $H_A : \mu \neq \mu_0$, then the rejection region is

$$\{t : t \leq -t_{\alpha/2} \text{ or } t \geq t_{\alpha/2}\}.$$

That H_0 is rejected or not depends on if the observed test statistic t_0 is in the rejection region.

Calculation Question

Assume that the data has a normal distribution and the number of observations is 15. Find the critical t value used to test a null hypothesis: $\alpha=0.08$ for a two-sided test.

Example

Listed below are the measured radiation emissions (in W/kg) corresponding to a sample of cell phones. We assume the sample is a simple random sample.

Use a **0.05** level of significance to test the claim that cell phones have a mean radiation level that is less than 1.00 W/kg.

0.38	0.55	1.54	1.55	0.50	0.60	0.92	0.96	1.00	0.86	1.46
------	------	------	------	------	------	------	------	------	------	------

The summary statistics are:

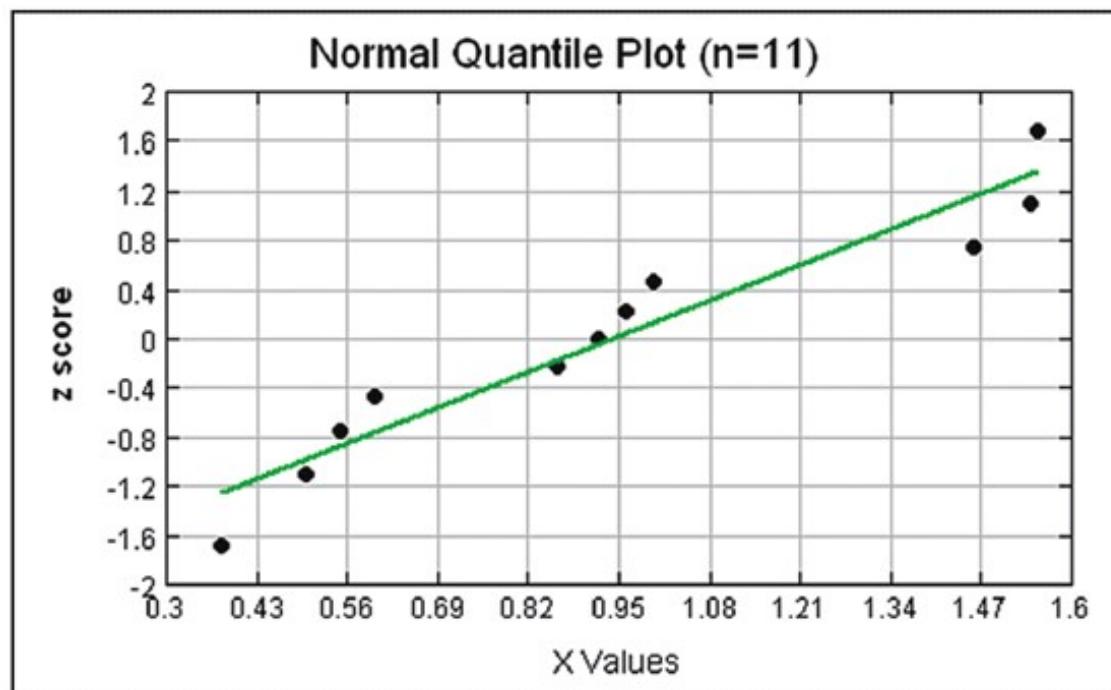
$$\bar{x} = 0.938 \text{ and } s = 0.423$$

Requirement Check:

The sample size is $n = 11$, which is not greater than 30, so we must check a **Normal quantile plot** for Normality.

Example - Continued

The points are reasonably close to a straight line and there is no other pattern, so we conclude the data appear to be from a Normally distributed population.



Example - Continued

Step 1: The claim that cell phones have a mean radiation level less than 1.00 W/kg is expressed as $\mu < 1.00$ W/kg. This is the claim we want to support.

The hypotheses are written as:

$$H_0 : \mu = 1.00 \text{ W/kg}$$

$$H_1 : \mu < 1.00 \text{ W/kg}$$

Example - Continued

Step 2: The stated level of significance is $\alpha = 0.05$. Because the claim is about a population mean μ , the statistic most relevant to this test is the sample mean: \bar{x} . We calculate the test statistic

$$t = \frac{\bar{x} - 1.00}{s / \sqrt{n}} = \frac{0.938 - 1.00}{0.423 / \sqrt{11}} = -0.486$$

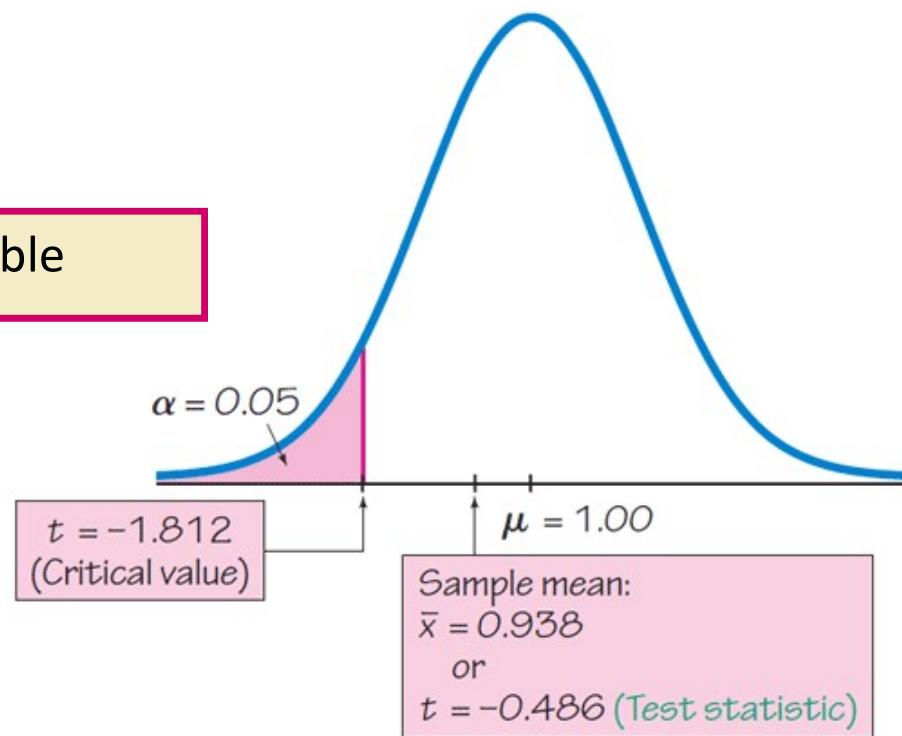
Example - Continued

Then we find the critical value:

Step 3: Because the test statistic of $t = -0.486$ does not fall in the rejection region bounded by the critical value of $t = -1.812$, fail to reject the null hypothesis.

$$qt(0.05, 10) = -1.812 \text{ or t-critical value table}$$

$$R.R. = \{t: t \leq -1.812\}$$



Example - Continued

Step 4: Because we fail to reject the null hypothesis, we conclude that there is not sufficient evidence to support the claim that cell phones have a mean radiation level that is less than 1.00 W/kg.

A Simulation Study: Understanding Probability of Type I error (significance level) α

Simulation Scheme:

- Test
 $H_0: \mu = \mu_0 = 0$ versus $H_A: \mu > 0$
at the significance level $\alpha=0.05$
- Consider iter=100 random samples of size $n=15$ from a standard Normal population (That is, H_0 is true).
- count=the number of times that H_0 is rejected
- Empirical significance level =count/iter
- Repeat the above using more samples such as iter=10000.

A Simulation Study: Understanding Probability of Type I error (significance level) α

<https://www.statcrunch.com/applets/type3&htmean>

Or use the R code in the next slide.

A Simulation Study: Understanding Probability of Type I error (significance level) α

R code

```
mu0=0;
n = 15;
alpha = 0.05;
Counter = 0;
iter = 10000; #number of samples
for (i in 1:iter)
{
  x = rnorm(n, mean=0, sd=1); #Data generation
  talpha=qt(p=alpha, df=n-1, lower.tail = FALSE); #critical value
  ttest=mean(x)/(sd(x)/sqrt(n)); #test statistics
  if(ttest>=talpha) Counter = Counter + 1;
  #Raising counter if we reject H0
  #test = t.test(x, alternative = "greater", mu=mu0); # t-test
  #if(test$p.value <=alpha) Counter = Counter + 1;
}
Counter;
Empalpha = Counter / iter; #Calculating empirical alpha
Empalpha;
```

Key Concepts

t-Test of a Population Mean, μ : σ unknown

✓ **Critical value method**

Understanding Probability of Type I error

(significance level) α

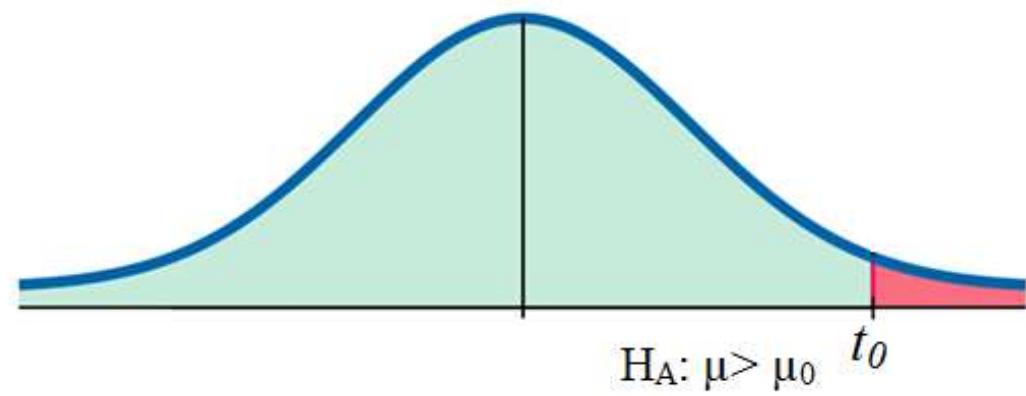
Section 3

A Hypothesis Test for the Mean – p-value method

p-value method: Likely or Unlikely?

$$H_A : \mu > \mu_0$$

$$t_0 = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}.$$



Rare Event Rule for Inferential Statistics

If, under a **given assumption** (H_0 is true in the hypothesis testing problems), the probability of a particular observed event is extremely small, we conclude that the assumption is probably not correct.

p-value method: Likely or Unlikely?

p-value : The **probability** of observing, just by *chance*, a test statistic as **extreme** as or more extreme than the one observed in the direction of H_A .

- **p-value** measures the strength of the evidence against H_0 .
- **Remark.** Corresponding to an observed value of a test statistic, the **p-value** is the *lowest level of significance* at which H_0 could have been rejected. So **p-value** is also called **observed significance**.

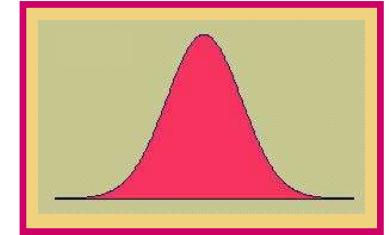
p-value method: Likely or Unlikely?

The p-Value and Surprise

The p-value is the probability of seeing sampling data (or even more unlikely data) given the null hypothesis is true.

- Tells us **how surprised** we would be to get these data **given H_0 is true**.
- **P-value small:** H_0 is not true.
- **P-value not small enough:** Not a surprise. Data consistent with the model. Do not reject H_0 .

Test of a Population Mean, μ : σ unknown



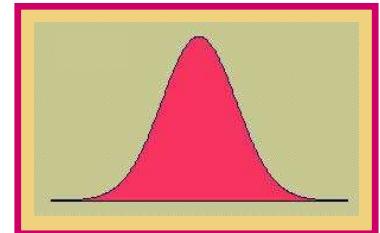
- Take a random sample of size n from a population with mean μ and unknown standard deviation σ .
- We assume that either or both
 1. The population is approximately Normally distributed
 2. sample size large ($n \geq 30$)
- Consider one of the three alternatives.

$H_0: \mu = \mu_0$ versus one of $H_A: \mu > \mu_0$

$H_A: \mu < \mu_0$

$H_A: \mu \neq \mu_0$

One-Sample t -Test for the Mean



- $H_0: \mu = \mu_0$
- Test statistic

$$t_0 = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}.$$

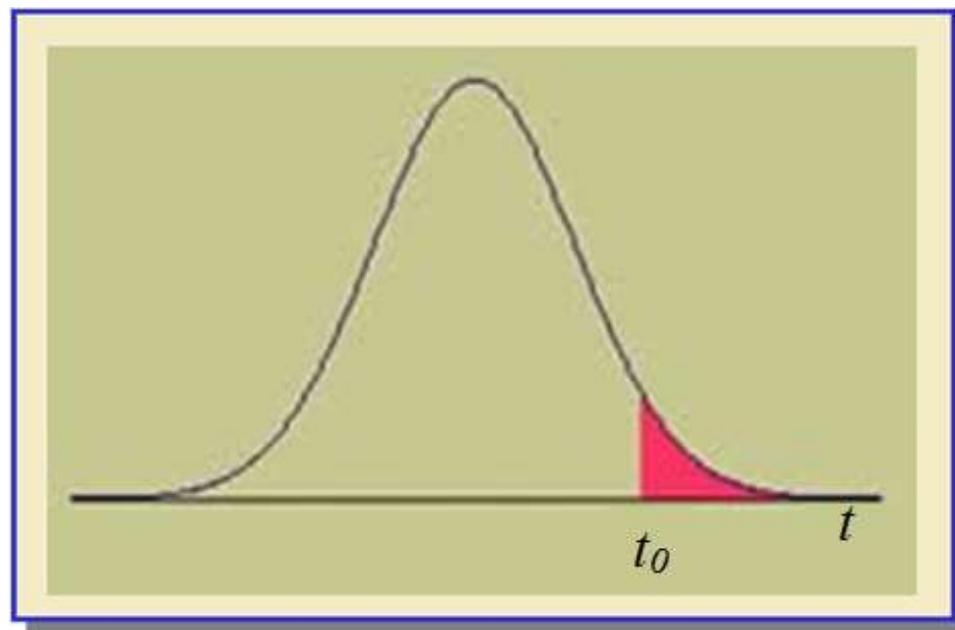
where Standard Error of \bar{x} : $SE(\bar{x}) = s / \sqrt{n}$

- When the conditions are met and H_0 is true, the test statistic t_0 follows the **Student's t** model with $df = n-1$.

Calculation of p -value – one-tailed t-test

If the alternative is one-sided $H_A: \mu > \mu_0$, then

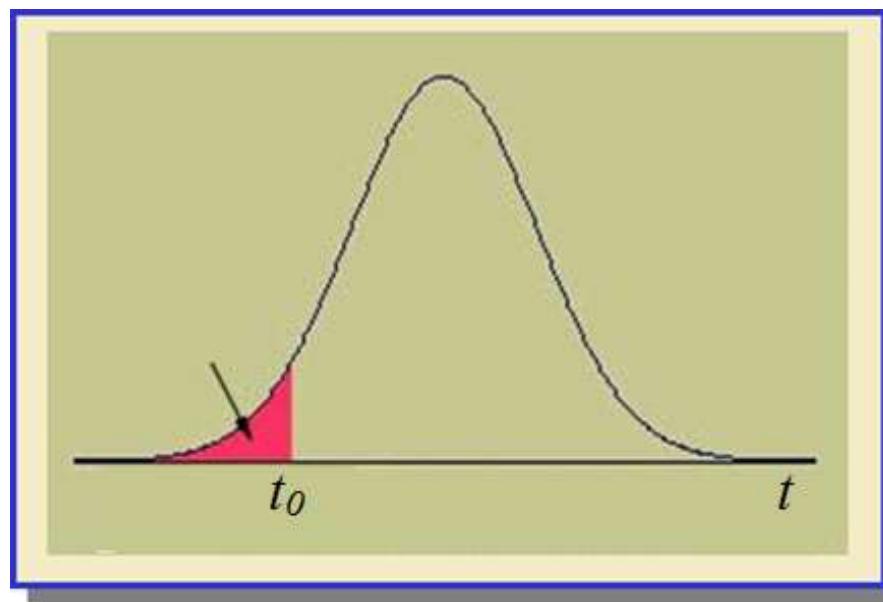
$$p\text{-value} = p_+ = P(t \geq t_0), \quad \text{where } t_0 = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$



Calculation of p -value – one-tailed t-test

Now if the alternative is one-sided $H_A : \mu < \mu_0$, then

$$p\text{-value} = p_- = P(t \leq t_0), \quad \text{where } t_0 = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

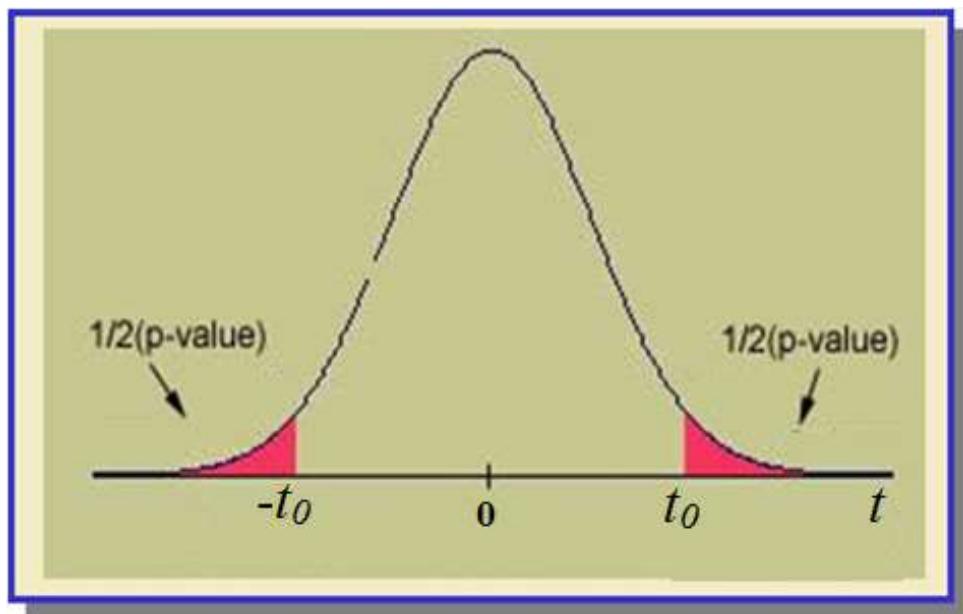


Calculation of p -value – two-tailed test

Now if the alternative is two-sided $H_A : \mu \neq \mu_0$, then

$$p\text{-value} = 2P(t \geq |t_0|), \text{ where } t_0 = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} \quad \text{or}$$

$$p\text{-value} = 2\min(p_+, p_-) = 2 \min\{P(t \geq t_0), P(t \leq t_0)\}$$



suppose $t_0 > 0$

That is, the p -value of the two-tailed test is **twice** of the **smaller** tail area.

P-value method: When is H_0 rejected?

- Suppose we have a *pre-assigned* significance level α ,

- If $p\text{-value} \leq \alpha$, reject H_0 . We report that the results are statistically significant at level α
- If $p\text{-value} > \alpha$, fail to reject H_0 . We report that the results are not significant at level α .

P-value method: When is H_0 rejected?

- If the *p*-value is very **small**, H_0 can be rejected.
- **Statistical significance:**
 - If the *p*-value is less than or equal to 0.01, reject H_0 . The results are **highly significant**.
 - If the *p*-value is between 0.01 and 0.05, reject H_0 . The results are **statistically significant**.
 - If the *p*-value is between 0.05 and 0.10, do not reject H_0 . But, the results are tending towards **significance**.
 - If the *p*-value is greater than 0.10, do not reject H_0 . The results are **not statistically significant**.

When the p-Value is Not Small

Wrong

- Accept H_0 since we did not prove H_0 .

Right

- Fail to reject H_0
- There is insufficient evidence to reject H_0 .
- H_0 may or may not be true.

Summary: calculation of *p*-values of t-test

Let the calculated observed test statistic be

$$t_0 = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

- (1) $H_A : \mu > \mu_0$, then $p\text{-value} = p_+ = P(t \geq t_0);$
- (2) $H_A : \mu < \mu_0$, then $p\text{-value} = p_- = P(t \leq t_0);$
- (3) $H_A : \mu \neq \mu_0$, then $p\text{-value} = 2P(|t| \geq |t_0|) = 2\min(p_+, p_-)$
 $= 2 \min\{P(t \geq t_0), P(t \leq t_0)\}$

Example 1

Listed below are the measured radiation emissions (in W/kg) corresponding to a sample of cell phones. We assume the sample is a simple random sample.

Use a 0.05 level of significance to test the claim that cell phones have a mean radiation level that is less than 1.00 W/kg. Use the **p-value method**.

0.38	0.55	1.54	1.55	0.50	0.60	0.92	0.96	1.00	0.86	1.46
------	------	------	------	------	------	------	------	------	------	------

The summary statistics are:

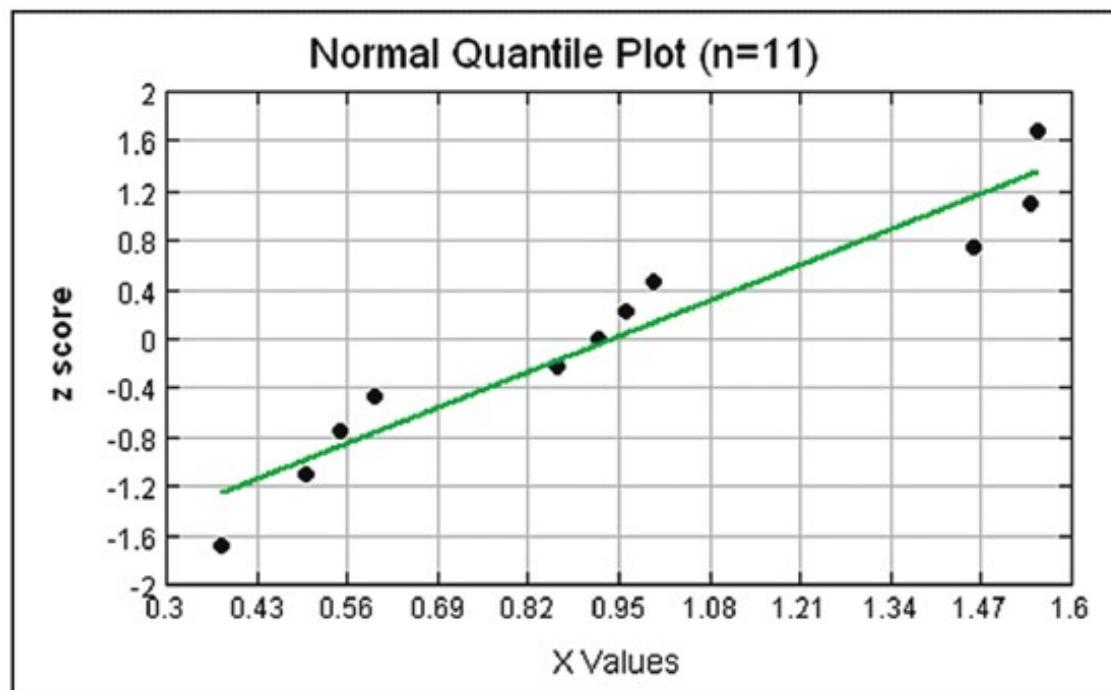
$$\bar{x} = 0.938 \text{ and } s = 0.423$$

Requirement Check:

The sample size is $n = 11$, which is not greater than 30, so we must check a **Normal quantile plot** for Normality.

Example 1- Continued

The points are reasonably close to a straight line and there is no other pattern, so we conclude the data appear to be from a Normally distributed population.



Example 1- Continued

Step 1: The claim that cell phones have a mean radiation level less than 1.00 W/kg is expressed as $\mu < 1.00$ W/kg. This is the claim we want to support.

The hypotheses are written as:

$$H_0 : \mu = 1.00 \text{ W/kg}$$

$$H_1 : \mu < 1.00 \text{ W/kg}$$

Example 1- Continued

Step 2: The stated level of significance is $\alpha = 0.05$. Because the claim is about a population mean μ , the statistic most relevant to this test is the sample mean: \bar{x} . We calculate the test statistic

$$t = \frac{\bar{x} - 1.00}{s / \sqrt{n}} = \frac{0.938 - 1.00}{0.423 / \sqrt{11}} = -0.486$$

Example 1- Continued

Then we find the p-value using the **R** (with $df=n-1=10$)

Step 3: Because the p-value of the test = $0.319 > 0.05$.
We fail to reject the null hypothesis.

See handout for Step 2 and 3 using R!

Step 4: Because we fail to reject the null hypothesis, we **conclude** that there is not sufficient evidence to support the claim that cell phones have a mean radiation level that is less than 1.00 W/kg.

R code:

```
data=c(0.38,0.55,1.54,1.55,0.5,0.6,0.92,0.96,1,0.86,1.46)
t.test(data, mu=1, alternative = "less")
```

One Sample t-test

data: data

t = -0.48485, df = 10, p-value = 0.3191

alternative hypothesis: true mean is less than 1

95 percent confidence interval:

-Inf 1.169269

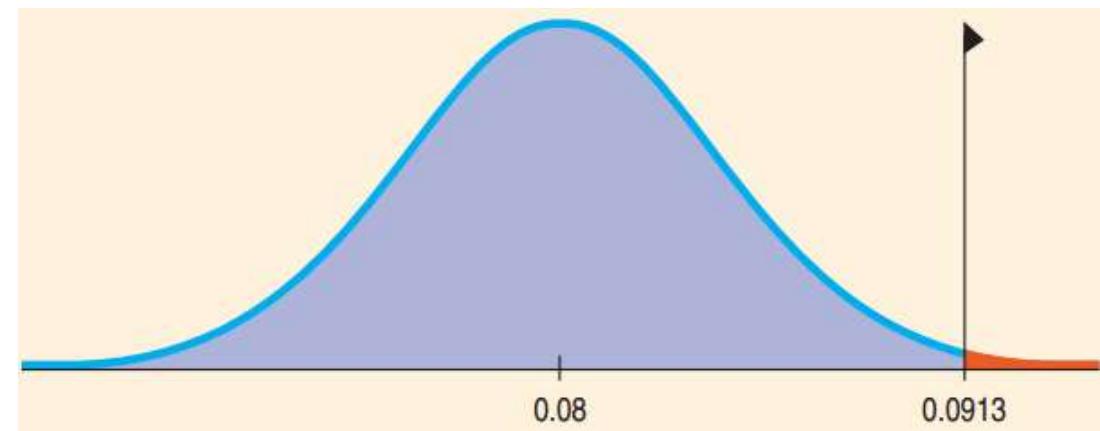
sample estimates:

mean of x

0.9381818

Example 2 - Are the Salmon Unsafe?

- EPA recommended mirex screening is 0.08 ppm.
 - Are farmed salmon contaminated beyond the permitted EPA level?
 - Recap: Sampled $n=150$ salmon. Mean 0.0913 ppm, Standard Deviation 0.0495 ppm.
 - Test $H_0: \mu = 0.08$ versus $H_A: \mu > 0.08$



Example 2 - Are the Salmon Unsafe?

- One-Sample t -Test for the Mean

- $n = 150, df = 149, \bar{y} = 0.0913, s = 0.0495$

- Test statistic:

$$t = \frac{\bar{y} - 0.08}{s/\sqrt{n}} = \frac{0.0913 - 0.08}{0.0495/\sqrt{150}} \approx 2.796$$

- $P(t_{149} > 2.796) = 0.00293$ (right-tail probability)
 - Since the p-value is so low, reject H_0 and conclude that the population mean mirex level does exceed the EPA screening value.

Calculation Question

A simple random sample of 15-year old boys from one city is obtained and their weights (in pounds) are listed below. Test the claim that these sample weights come from a population with a mean greater than 147 lb.

146	140	160	151	134	189
157	144	175	127	164	

What is the p-value of the test?

```
x=c(146, 140, 160, 151, 134, 189, 157, 144, 175, 127, 164)  
t.test(x, mu=147, alternative = "greater")
```

Calculation Question

A simple random sample of 15-year old boys from one city is obtained and their weights (in pounds) are listed below. Test the claim that these sample weights come from a population with a mean equal to 147 lb.

146	140	160	151	134	189
157	144	175	127	164	

What is the p-value of the test?

Calculation Question

A simple random sample of 15-year old boys from one city is obtained and their weights (in pounds) are listed below. Test the claim that these sample weights come from a population with a mean greater than or equal to 147 lb.

146	140	160	151	134	189
157	144	175	127	164	

What is the p-value of the test?

Confidence Intervals for Hypothesis Tests (Optional)

- **Recall: Confidence Intervals**

- ✓ Start with data and find plausible values for the parameter.
- ✓ Only 2-sided symmetric CIs were discussed
 - ✓ We need 1-sided CI for one-sided alternatives
- **Example:** $H_0: \mu = \mu_0$ versus $H_A: \mu \neq \mu_0$ at level α

Confidence Intervals for Hypothesis Tests

Example: $H_0: \mu = \mu_0$ versus $H_A: \mu \neq \mu_0$ at level α

- Test $H_0: \mu = \mu_0$ versus one of the three alternatives at significance level α .
- Construct a **($1 - \alpha$)100% two-sided symmetric confidence interval for a two-sided test** or a **($1 - \alpha$)100% one-sided confidence interval for a one-sided test** (not discussed before).
- Then, if the value of the parameter under H_0 falls in this interval, we do not reject H_0 . Otherwise we reject H_0 in favor of H_A .

Confidence Intervals for Hypothesis Tests

Example 1 - Revisited

Listed below are the measured radiation emissions (in W/kg) corresponding to a sample of cell phones. We assume the sample is a simple random sample.

Use a **0.05 level of significance** to test the claim that cell phones have a mean radiation level that is less than 1.00 W/kg. Use the **confidence interval method**.

0.38	0.55	1.54	1.55	0.50	0.60	0.92	0.96	1.00	0.86	1.46
------	------	------	------	------	------	------	------	------	------	------

The summary statistics are:

$$n=11, \bar{x} = 0.938 \text{ and } s = 0.423$$

Confidence Intervals for Hypothesis Tests

Example 1- Revisited

Step 1:

$$H_0 : \mu = 1.00 \text{ W/kg}$$

$$H_1 : \mu < 1.00 \text{ W/kg}$$

Confidence Intervals for Hypothesis Tests

Example 1- Revisited

Step 2: Since H_1 is one-sided, we construct a $(1-\alpha)$
 $=1-0.05=95\%$ confidence interval ($df=10$) of μ :

$$\begin{aligned} (-\infty, \bar{x} + t_\alpha \frac{s}{\sqrt{n}}) &= (-\infty, \bar{x} + t_{0.05} \frac{s}{\sqrt{n}}) \\ &= (-\infty, 0.938 + 1.8125 \frac{0.4229}{\sqrt{11}}) \\ &= (-\infty, 1.169) \end{aligned}$$

Since 1 is included in the 95% confidence interval we do NOT reject the null hypothesis.

Confidence Intervals for Hypothesis Tests

Example 1- Revisited

Step 3: Because we fail to reject the null hypothesis, we conclude that there is not sufficient evidence to support the claim that cell phones have a mean radiation level that is less than 1.00 W/kg.

Key Concepts

T-Test of a Population Mean, μ : σ unknown

- **p-value method** (technology must be used)
- **Confidence Interval method**

Section 4

Hypothesis Testing for the Proportion

Concept Question

A researcher claims that 72% of voters favor gun control. Identify the null hypothesis H_0 and the alternative hypothesis H_1 .

- A. $H_0: p = 0.72, H_1: p > 0.72$
- B. $H_0: p = 0.72, H_1: p \neq 0.72$
- C. $H_0: p \neq 0.72, H_1: p = 0.72$
- D. $H_0: p = 0.72, H_1: p < 0.72$

Large Sample Test

Recall that the test statistics we have used are of the form

$$\frac{\text{statistic} - \text{hypothesized value}}{\text{standard error of statistic}}$$

Large Sample Test

To test $H_0 : p = p_0$ versus one or two sided alternative

$$(1) \ H_A : p > p_0$$

$$(2) \ H_A : p < p_0$$

$$(3) \ H_A : p \neq p_0$$

Requirements for Testing Claims About a Population Proportion p

- 1) The sample observations are a simple **random sample** (The conditions for a **binomial distribution** are satisfied).
- 2) **10% Condition:** $n/N < 10\%$.
- 3) The conditions $np_0 \geq 10$ and $n(1-p_0) \geq 10$ are both satisfied, so **the distribution of the sample proportions** under H_0 can be approximated by a normal distribution with

$$\mu = p_0 \quad \text{and} \quad \sigma = \sqrt{\frac{p_0(1-p_0)}{n}}.$$

Test Statistic

Test $H_0 : p = p_0$ versus H_a : one or two sided
using the test statistic

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

using p -value or critical value method
based on standard Normal distribution

Note that $se(\hat{p}) = \sqrt{p_0(1 - p_0)/n}$ under $H_0: p = p_0$

- **Critical Value method (in exams)**
- p -value method
- Confidence interval method

Caution

Don't confuse a **p -value** with a **proportion p** .

- p -value = probability of getting a test statistic at least as extreme as the one representing sample data = observed significance level
- p = population proportion

Summary of critical value method of z-test with significance level α

To test $H_0 : p = p_0$ versus H_a : one or two sided at level α ,

(1) $H_A : p > p_0$, then the rejection region is

$$\{z : z \geq z_\alpha\};$$

(2) $H_A : p < p_0$, then the rejection region is

$$\{z : z \leq -z_\alpha\};$$

(3) $H_A : p \neq p_0$, then the rejection region is

$$\{z : z \leq -z_{\alpha/2} \text{ or } z \geq z_{\alpha/2}\}.$$

That H_0 is rejected or not depends on if the observed test statistic z_0 is in the rejection region.

Calculation of *p*-values of z-test

Let the observed test statistic be

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

- (1) $H_A : p > p_0$, then $p-value = p_+ = P(z \geq z_0)$;
- (2) $H_A : p < p_0$, then $p-value = p_- = P(z \leq z_0)$;
- (3) $H_A : p \neq p_0$, then $p-value = 2P(|z| \geq |z_0|) = 2\min(p_+, p_-)$
 $= 2 \min\{P(z \geq z_0), P(z \leq z_0)\}$

A Simulation Study

<https://www.statcrunch.com/applets/type3&htprop>

Confidence Intervals for Hypothesis Testing (Optional)

To test $H_0 : p = p_0$ versus H_a : one or two sided at level α ,

- (1) For a two-sided test, we construct **1- α** level 2-sided symmetric C.I. of p

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \quad \text{for } H_A : p \neq p_0$$

- (2) For a one-sided test, we construct **1- α** level 1-sided C.I. of p

$$\left(0, \quad \hat{p} + z_\alpha \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \right) \quad \text{for } H_A : p < p_0$$

$$\left(\hat{p} - z_\alpha \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \quad 1 \right) \quad \text{for } H_A : p > p_0$$

Confidence Intervals for Hypothesis Testing (Optional)

Caution

The critical value method and the p -value method are equivalent and will yield the same result.

- It is possible that critical value method and the p -value method might yield a different conclusion than the confidence interval method since the confidence interval uses an estimated standard deviation based upon the **sample proportion**

$$z = \frac{\hat{p} - p}{\sqrt{\hat{p}(1 - \hat{p}) / n}}$$

- A good strategy is to use a confidence interval to **estimate** a population proportion, but use the p -value or critical value method for **testing a claim** about the proportion.

Example (2-sided test)

Based on information from the National Cyber Security Alliance, 93% of computer owners believe they have antivirus programs installed on their computers.

In a random sample of 400 scanned computers, it is found that 380 of them (or 95%) actually have antivirus software programs.

Use the sample data from the scanned computers to test the claim that 93% of computers have antivirus software. Use $\alpha = 0.05$.

Example - Continued

Requirement check:

1. The 400 computers are randomly selected.
2. There is a fixed number of independent trials with two categories (computer has an antivirus program or does not).
3. 10% rule is satisfied
4. The requirements $np \geq 10$ and $n(1-p) \geq 10$ are both satisfied with $n = 400$

$$np = (400)(0.93) = 372$$

$$nq = (400)(0.07) = 28$$

Example - Continued

Step 1. The original claim that 93% of computers have antivirus software can be expressed as $p = 0.93$. The opposite of the original claim is $p \neq 0.93$ since no direction is specified.

The hypotheses are written as:

$$H_0 : p = 0.93$$

$$H_1 : p \neq 0.93$$

Example - Continued

Step 2. We construct and calculate the test statistic

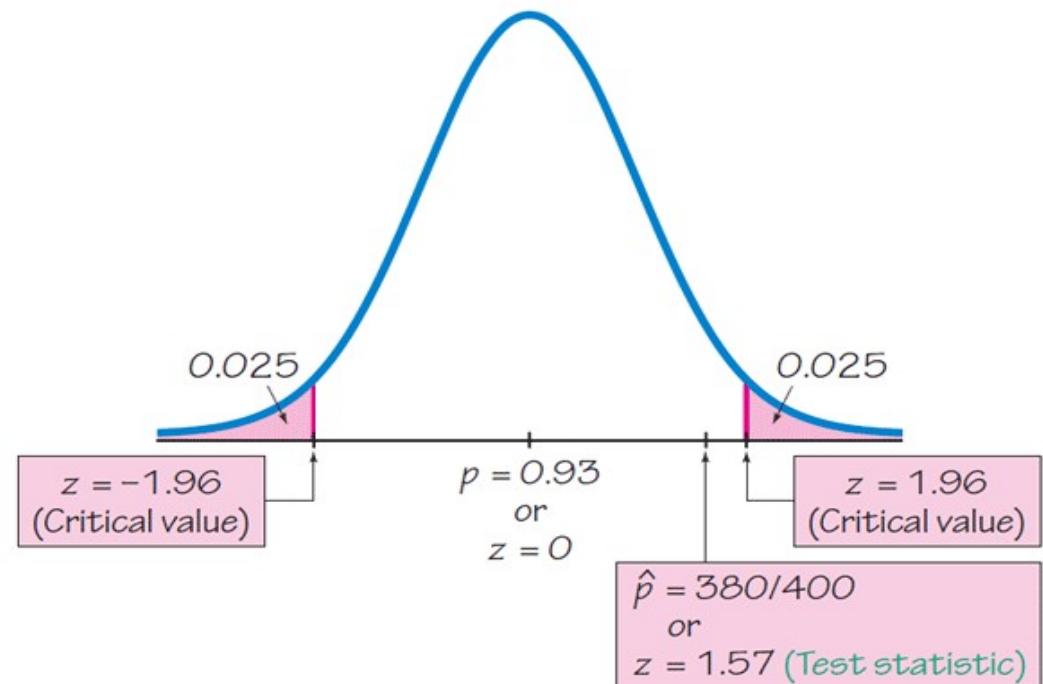
$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} = \frac{380/400 - 0.93}{\sqrt{0.93 \cdot 0.07 / 400}} \simeq 1.57$$

Example - Continued

Critical Value Method: Steps 1 and 2 are the same as for the p -value method.

Step 3. The test statistic is computed to be $z = 1.57$. We now find the critical values, with the critical region having an area of $\alpha = 0.05$, split equally in both tails.

$$RR = \{Z: Z \leq -1.96 \text{ or } Z \geq 1.96\}$$



Example - Continued

Critical Value Method:

Because the test statistic does not fall in the critical region, we fail to reject the null hypothesis.

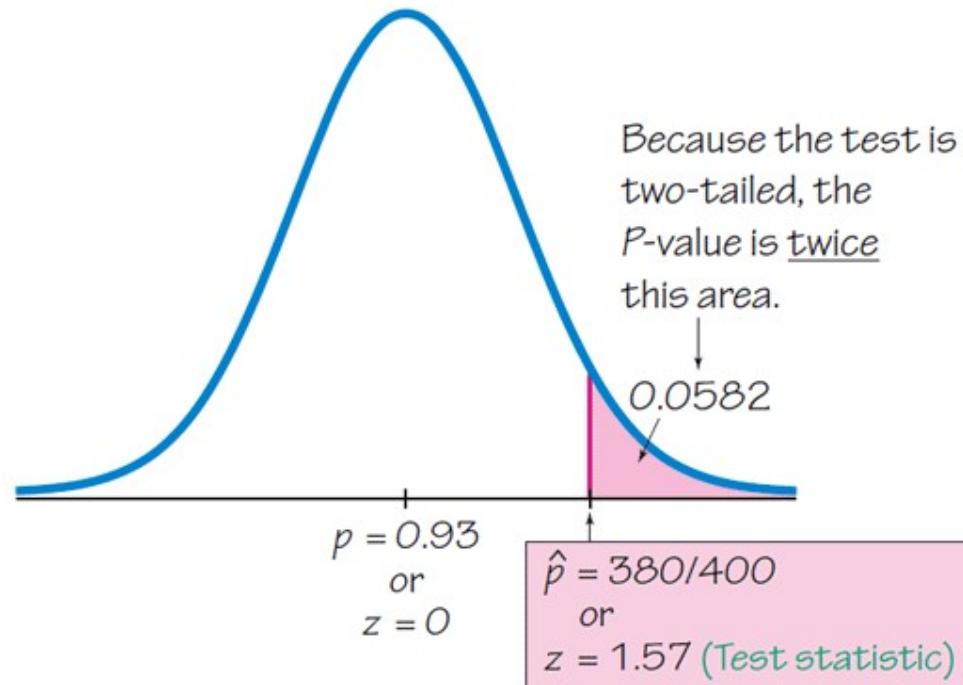
Step 4. We conclude that there is not sufficient sample evidence to warrant rejection of the claim that 93% of computers have antivirus programs.

Example - Continued

Step 3 (*p*-Value Method). Because the hypothesis test is two-tailed with a test statistic of $z = 1.57$, the *p*-value is twice the area to the right of $z = 1.57$.

The *p*-value is twice 0.0582, or 0.1164 (or use R).

➤ See **handout** for Step 2 and 3 using R



- Because the p -value of 0.1164 is greater than the significance level of $\alpha = 0.05$, we fail to reject the null hypothesis that 93% computers have antivirus software.

Data: $x=380$, $n=400$

R code:

```
prop.test(x=380, n=400, p=0.93, alternative = "two.sided", correct = FALSE)
```

```
> prop.test(x=380, n=400, p=0.93, alternative = "two.sided", correct = FALSE)
```

1-sample proportions test without continuity correction

```
data: 380 out of 400, null probability 0.93
X-squared = 2.4578, df = 1, p-value = 0.1169
alternative hypothesis: true p is not equal to 0.93
95 percent confidence interval:
 0.9240364 0.9674026
sample estimates:
 p
0.95
```

Example – Continued (Optional)

Confidence Interval Method:

The claim of $p = 0.93$ can be tested at the $\alpha = 0.05$ level of significance with a **95%** confidence interval since it is a two-sided test.

Using manual calculation, we get:

$$0.928642 < p < 0.971358$$

This interval **contains $p = 0.93$** , so we do not have sufficient evidence to warrant the rejection of the claim that 93% of computers have antivirus programs.

Calculation Question

A medical school claims that more than 28% of its students plan to go into general practice. It is found that among a random sample of 130 students, 32.3% of them plan to go into general practice. Find the p -value for a test of the school's claim.

- A. 0.137
- B. 0.274
- C. 0.308
- D. 0.863

```
prop.test(x=42, n=130, p=0.28,  
alternative = "greater", correct = FALSE)
```

Concept Question

A medical school claims that less than 28% of its students plan to go into general practice. It is found that among a random sample of 130 students, 32.3% of them plan to go into general practice. Find the p -value for a test of the school's claim based on the answer of the 1st question.

- A. 0.137
- B. 0.274
- C. 0.308
- D. 0.863

Concept Question

A medical school claims that 28% of its students plan to go into general practice. It is found that among a random sample of 130 students, 32.3% of them plan to go into general practice. Find the p -value for a test of the school's claim based on the answer of the 1st question..

- A. 0.137
- B. 0.274
- C. 0.308
- D. 0.863

Example 2 (left-sided test)

Cracking Rate < 20%?

- After a new engineering process the cracking rate of 400 casts **fell** to **17%**. Is this due to the new engineering or just random chance? Is this sufficient to show that the cracking rate is less than 20%? Test at significance level **$\alpha=0.05$** .



• Step 1.

- Null Hypothesis: Nothing has changed

- $\bullet \quad H_0: p = p_0 = 0.20$

- Alternative Hypothesis:

- $\bullet \quad H_A: p < p_0 = 0.20$

Example 2 continued

- Checking Conditions: $n = 400$, $p_0 = 0.20$

✓ $np_0 = (400)(0.20) = 80 \geq 10$

✓ $n(1-p_0) = (400)(0.80) = 320 \geq 10$

✓ Independence

✓ $n/N < 10\%$

Example 2 continued

Step 2. Calculate the test statistic ($q_0=1-p_0=0.8$)

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} = \frac{0.17 - 0.20}{\sqrt{0.2 \cdot 0.8 / 400}} = -1.5$$

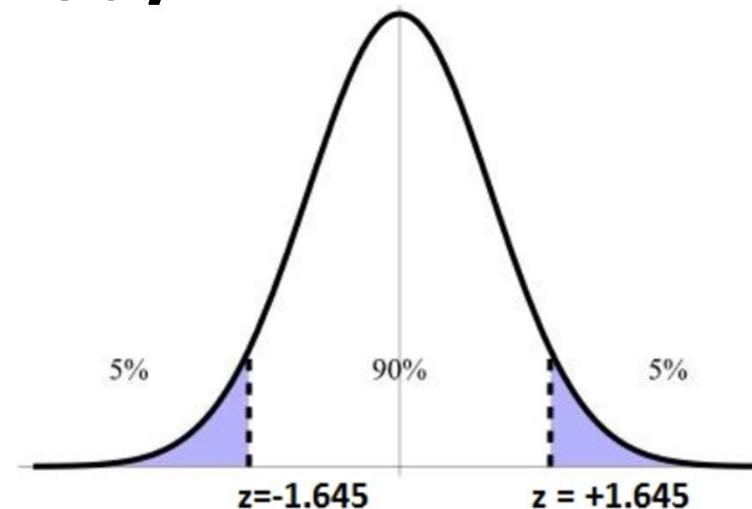
•**Note:** Use p_0 but not \hat{p} to find standard deviation.

Example 2 continued

- Step 3 (Critical-value method).

- The rejection region is

$$RR = \{Z: Z \leq -1.645\}$$



The test statistic $z = -1.5$ is not in RR . So we fail to reject H_0 .

Example 2 continued

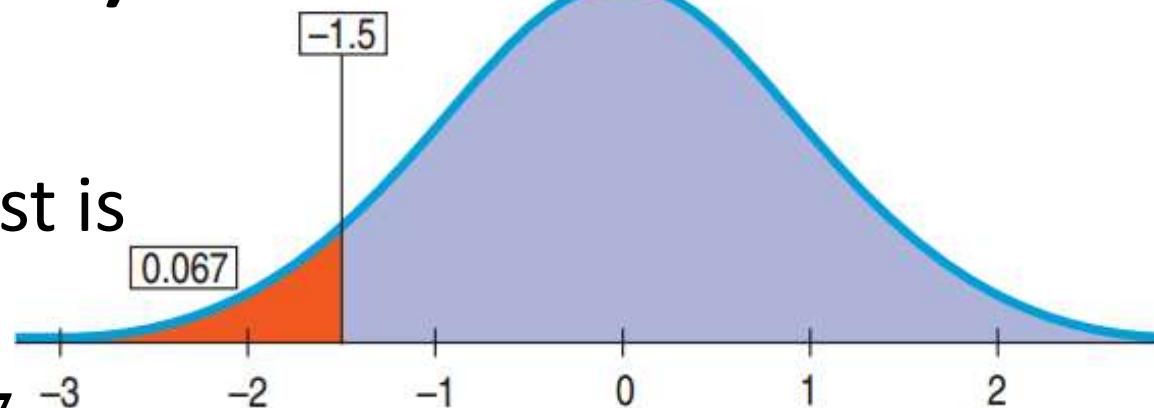
Step 4. We conclude that there is not sufficient evidence to indicate that the cracking rate is less than 20%.

Example 2 continued

- Step 3 (*p*-Value Method).

– The *p*-value of the test is

$$P(Z \leq -1.5) \approx 0.067$$



which is larger than $\alpha=0.05$. So we fail to reject H_0 .
(use technology to obtain $P(Z \leq -1.5)$)

step 2 and step 3 using R

$$n=400, \hat{p} = 0.17 \text{ and thus } x = n\hat{p} = 68$$

```
> prop.test(x=68, n=400, p=0.2, alternative = "less", correct = FALSE)
```

1-sample proportions test without continuity correction

```
data: 68 out of 400, null probability 0.2
X-squared = 2.25, df = 1, p-value = 0.06681
alternative hypothesis: true p is less than 0.2
95 percent confidence interval:
 0.0000000 0.2030859
sample estimates:
 p
0.17
```

The p-value of 0.06681 is larger than $\alpha=0.05$. So we fail to reject H_0 .

Example 3(right-sided test): Home Field Advantage?



- Is there a home field advantage in baseball?
 - The home team won 1277 of the 2429 (52.57%) games played in the season.
 - Is there evidence to suggest that the home team wins more than 50%? Test at level $\alpha=0.025$.

• Step 1.

Let p = proportion of home team wins and we are testing

- $H_0: p = p_0 = 0.50$ versus
- $H_A: p > 0.50$

Example 3 continued

- **Check the conditions**

- ✓ **Independence Assumption:** Questionable, the 2011 season may be representative of all games past and future.
- ✓ **10% Condition:** $n/N < 10\%$. The 2011 season is less than 10% of all games played past and future.
- ✓ **Success/Failure Condition:**
 - $np_0 = (2429)(0.5) \geq 10$
 - $n(1-p_0) = (2429)(0.5) \geq 10$

Example 3 continued

Step 2: Calculate the test statistic

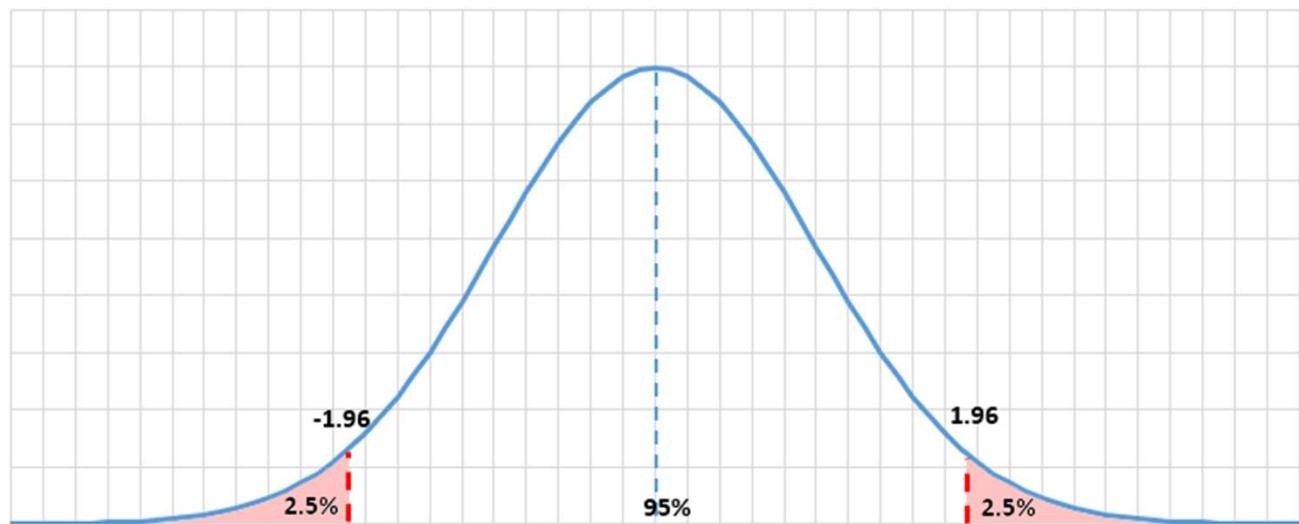
$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} = \frac{0.5257 - 0.5}{\sqrt{0.5 \cdot (1 - 0.5) / 2429}} \approx 2.53$$

Example 3 continued

- Step 3: Critical value method

The rejection region is

$$RR = \{Z: Z \geq 1.96\} \text{ because } \alpha = 0.025$$



The test statistic $z \approx 2.53$ is in the RR. So we reject H_0 .

Example 3 continued

Step 4. Conclusion:

—There is reasonable evidence that the true proportion of home team wins is greater than 50%.
And there appears to be a home field advantage.

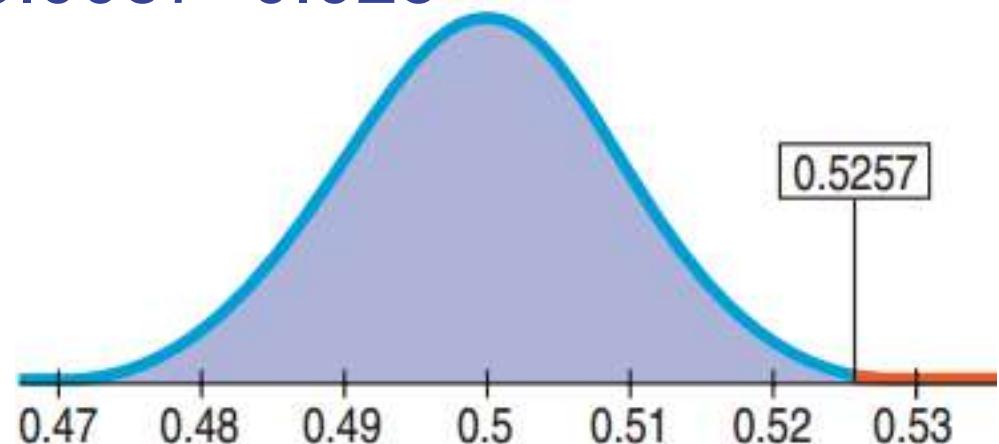
Example 3 continued

- Step 3: (p-value method)

The p-value of the test is

$$\text{p-value} \approx P(Z \geq 2.53) \approx 0.0057 < 0.025$$

So H_0 is rejected.



(Again, we use R for step 2 and step 3)

step 2 and step 3 using R:

```
> prop.test(x=1277, n=2429, p=0.5, alternative = "greater", correct = FALSE)

1-sample proportions test without continuity correction

data: 1277 out of 2429, null probability 0.5
X-squared = 6.4327, df = 1, p-value = 0.005602
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
0.5090463 1.0000000
sample estimates:
p
0.5257308
```

The p-value of the test 0.0056 is less than the significance level 0.025, so H_0 is rejected.

What Can Go Wrong?

- Don't forget to check the conditions.
 - Randomization, independence, and sample size
- Don't base your H_0 or H_A on what you see in the data.
 - Changing the null/alternative hypothesis after looking at the data is cheating.
- Don't accept the null hypothesis.
 - You can only say you don't have evidence to reject H_0 .
- If you fail to reject H_0 don't expect a larger sample would reject H_0 .
 - Check the confidence interval. If p_0 is not close to the lower limit or upper limit, then a larger sample will unlikely be worthwhile.

Key Concepts

Z-Test of a Population Proportion

- Critical value method
- p -value method
- Confidence interval method