Data Science Certification Module 3

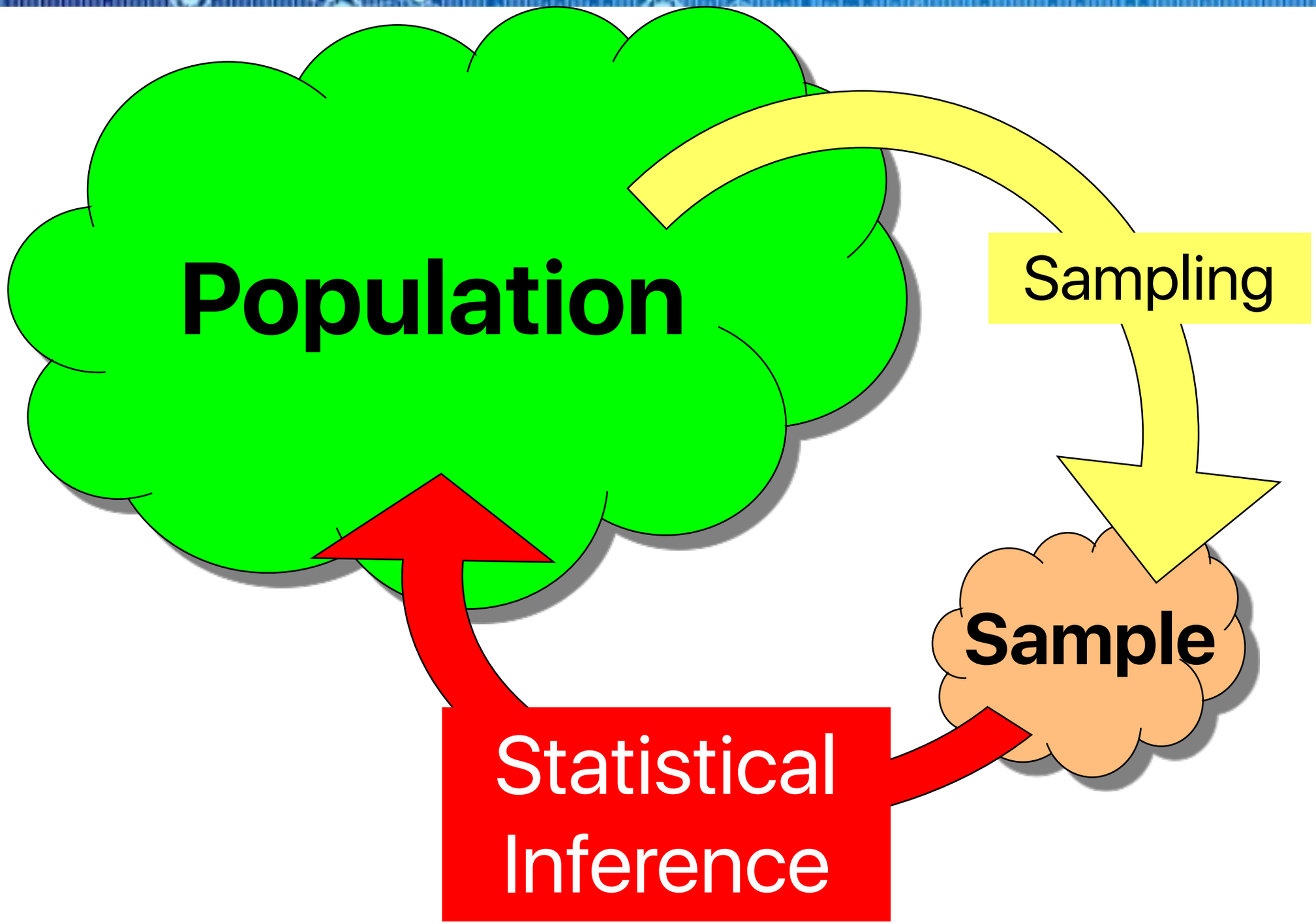
Data Science Certification

Sample versus Population

- A **population** includes all individuals or objects of interest.

- A **sample** is all the cases that we have collected data on (a subset of the population).

- **Statistical inference** is the process of using data from a sample to gain information about the population.


Population
Sampling
Sample
Statistical Inference

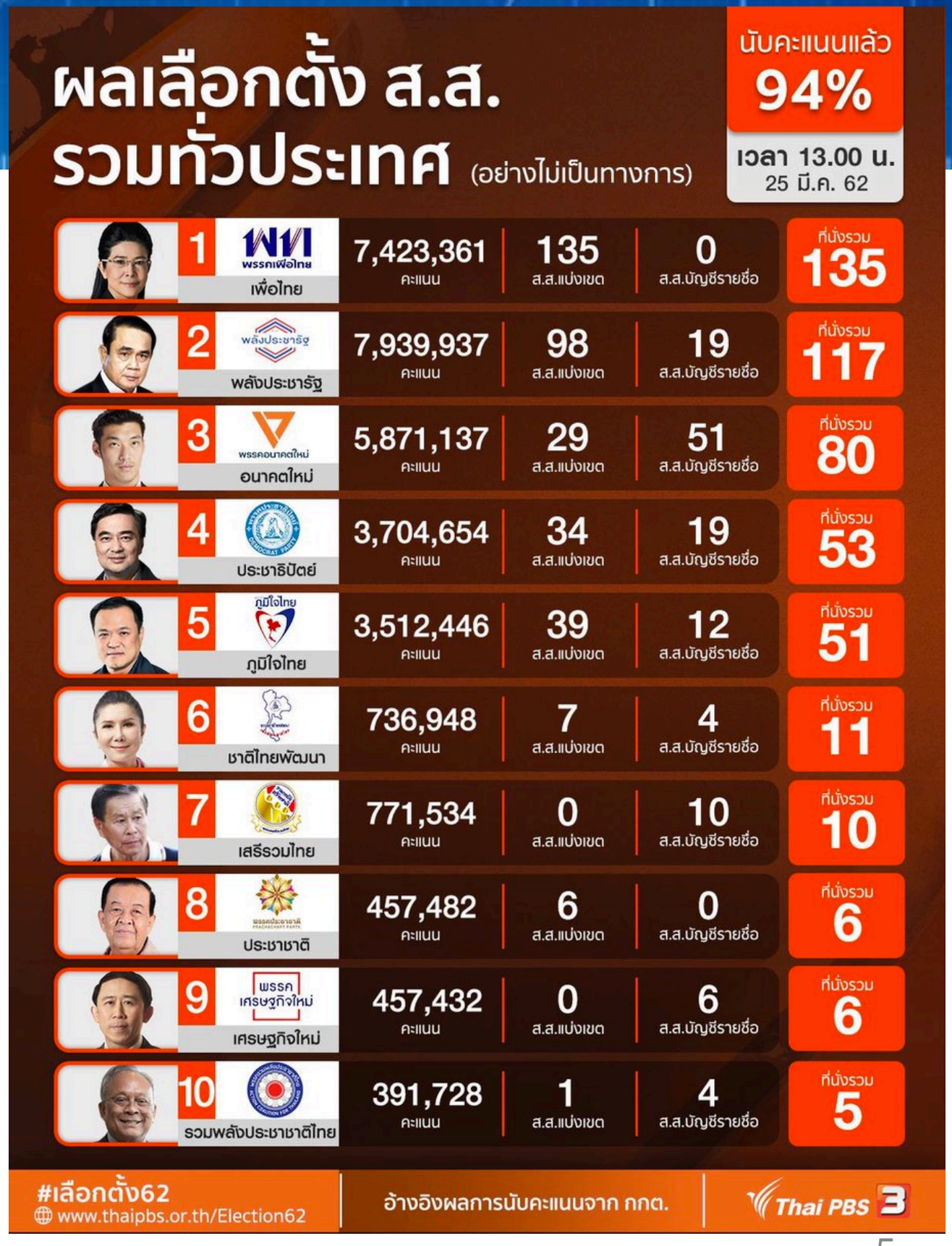
Most Important to You

- Suppose researchers studying student in Bangkok use the question like what the students find important

- What is the sample?
- What is the population?

- Can the sample data be generalized to make inferences about the population? Why or why not?


The di·a·logue by TBIS
thedialoguebytbs
รังสิตโพล วันที่ 5 มี.ค.2562
กรุงเทพโพล วันที่ 4-6 มี.ค. 2562
นิด้าโพล วันที่ 4-7 ก.พ. 2562
ผลเฉลี่ยทั้ง 3 โพล
พรรคเพื่อไทย 15.2% 21.7% 36.5% 24.5%
พรรคพลังประชารัฐ 20.4% 19.0% 22.6% 20.7%
พรรคประชาธิปัตย์ 17.2% 15.5% 15.2% 16.0%
พรรคอนาคตใหม่ 7.6% 12.0% 8.2% 9.2%
พรรคเสรีรวมไทย 6.6% 3.9% 5.0% 5.2%
พรรคภูมิใจไทย 8.8% 1.7% 0.9% 3.8%
พรรคชาติไทยพัฒนา 4.2% 1.0% 1.4% 2.2%
อื่นๆ 20% 25.2% 10.2% 18.4%
คิดเหมือนกันหรือเปล่า?
สรุป 3 ผลโพล แต่ละพรรคได้เท่าไหร่?
What do the polls say?
General Election Poll 2019 Comparison
ที่มา bangkokpoll nidapoll springnews


ผลเลือกตั้ง ส.ส. รวมทั่วประเทศ (อย่างไม่เป็นทางการ)
นับคะแนนแล้ว 94%
เวลา 13.00 น. 25 มี.ค. 62
1 เพื่อไทย 7,423,361 คะแนน 135 ส.ส.แบ่งเขต 0 ส.ส.บัญชีรายชื่อ ที่นั่งรวม 135
2 พลังประชารัฐ 7,939,937 คะแนน 98 ส.ส.แบ่งเขต 19 ส.ส.บัญชีรายชื่อ ที่นั่งรวม 117
3 อนาคตใหม่ 5,871,137 คะแนน 29 ส.ส.แบ่งเขต 51 ส.ส.บัญชีรายชื่อ ที่นั่งรวม 80
4 ประชาธิปัตย์ 3,704,654 คะแนน 34 ส.ส.แบ่งเขต 19 ส.ส.บัญชีรายชื่อ ที่นั่งรวม 53
5 ภูมิใจไทย 3,512,446 คะแนน 39 ส.ส.แบ่งเขต 12 ส.ส.บัญชีรายชื่อ ที่นั่งรวม 51
6 ชาติไทยพัฒนา 736,948 คะแนน 7 ส.ส.แบ่งเขต 4 ส.ส.บัญชีรายชื่อ ที่นั่งรวม 11
7 เสรีรวมไทย 771,534 คะแนน 0 ส.ส.แบ่งเขต 10 ส.ส.บัญชีรายชื่อ ที่นั่งรวม 10
8 ประชาชาติ 457,482 คะแนน 6 ส.ส.แบ่งเขต 0 ส.ส.บัญชีรายชื่อ ที่นั่งรวม 6
9 เศรษฐกิจใหม่ 457,432 คะแนน 0 ส.ส.แบ่งเขต 6 ส.ส.บัญชีรายชื่อ ที่นั่งรวม 6
10 รวมพลังประชาชาติไทย 391,728 คะแนน 1 ส.ส.แบ่งเขต 4 ส.ส.บัญชีรายชื่อ ที่นั่งรวม 5
#เลือกตั้ง62
www.thaipbs.or.th/Election62
อ้างอิงผลการนับคะแนนจาก กกต.
Thai PBS

Dewey Defeats Truman?

Dewey Defeats Truman?

The paper was published before the conclusion of the 1948 presidential election, and was based on the results of a large telephone poll which showed Dewey sweeping Truman

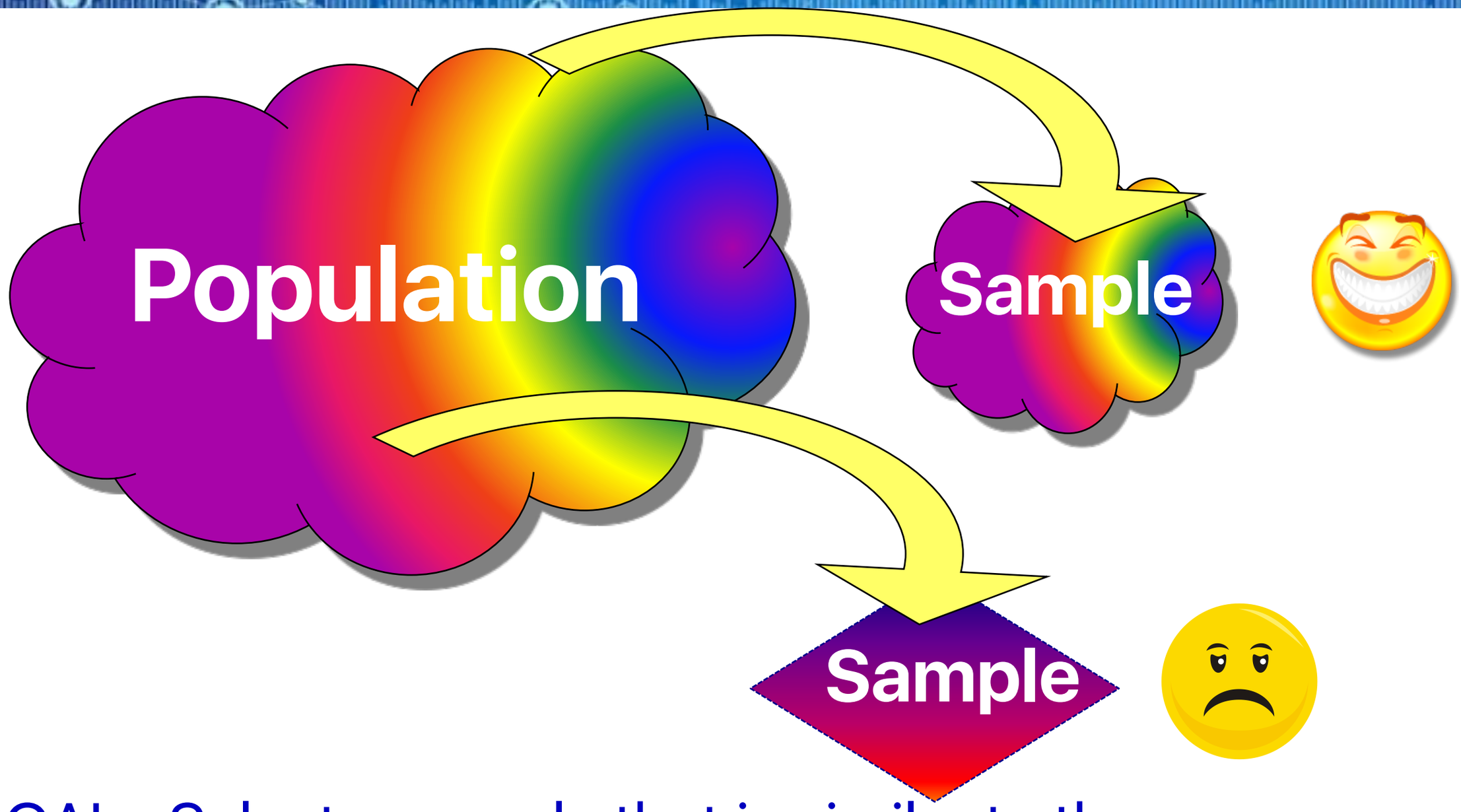
However, Harry S. Truman won the election

What went wrong?

Sampling Bias

- **Sampling bias** occurs when the method of selecting a sample causes the sample to differ from the population in some relevant way.
- If sampling bias exists, we cannot trust generalizations from the sample to the population

Sampling



GOAL: Select a sample that is similar to the population, only smaller

Random Sampling

How can we make sure to avoid sampling bias?

Take a RANDOM sample!

Imagine putting the names of all the units of the population into a hat, and drawing out names at random to be in the sample

More often, we use technology

Random Sampling

Before the 2008 election, the Gallup Poll took a **random sample** of 2,847 Americans. 52% of those sampled supported Obama

In the actual election, 53% voted for Obama

Random sampling is a very powerful tool!!!

Random vs Non-Random Sampling

- Random samples have averages that are centered around the correct number
- Non-random samples may suffer from sampling bias, and averages may not be centered around the correct number
- Only random samples can truly be trusted when making generalizations to the population!

Simple Random Sample

In a **simple random sample**, each unit of the population has the same chance of being selected, regardless of the other units chosen for the sample

Realities of Sampling

- While a random sample is ideal, often it isn't feasible. A list of the entire population may not be available, or it may be impossible or too difficult to contact all members of the population.
- Sometimes, your population of interest has to be altered to something more feasible to sample from. Generalization of results are limited to the population that was actually sampled from.
- In practice, think hard about potential sources of sampling bias, and try your best to avoid them

Non-Random Samples

Suppose you want to estimate the average number of hours that Chulalongkorn University's students spend studying each week. Which of the following is the best method of sampling?

(a) Go to the Chula library and ask all the students there how much they study

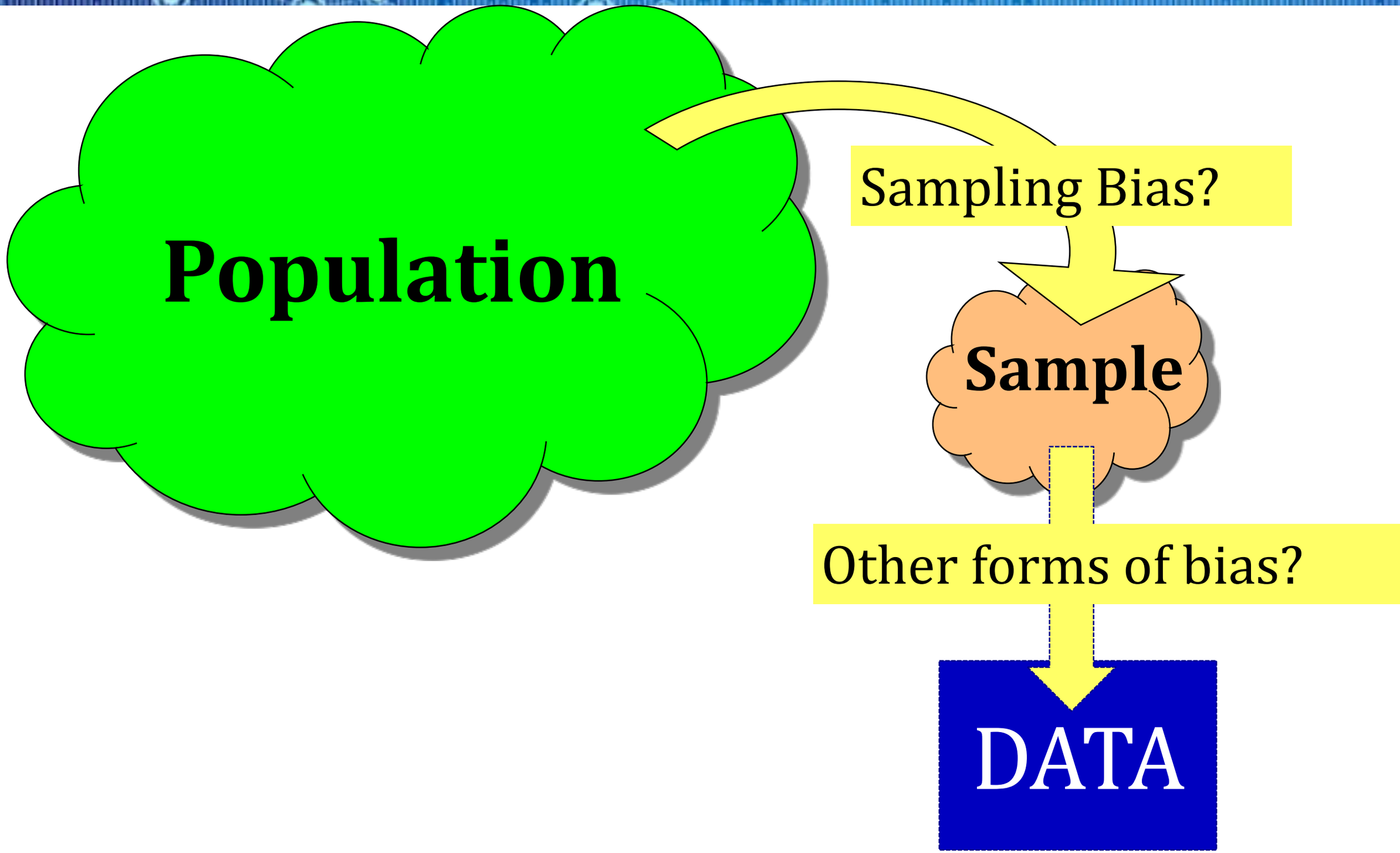
(b) Email all Chula students asking how much they study, and use all the data you get

(c) Stand outside the Samyan Midtown and ask everyone going in how much they study

Alcohol, Marijuana, and Driving

- The Federal Office of Road Safety in Australia conducted a study on the effects of alcohol and marijuana on performance
- Volunteers who responded to advertisements for the study on rock radio stations were given a random combination of the two drugs, then their performance was observed
  - What is the sample? What is the population?
  - Is there sampling bias?
  - Will the results be informative and/or do you think the study is worth conducting?

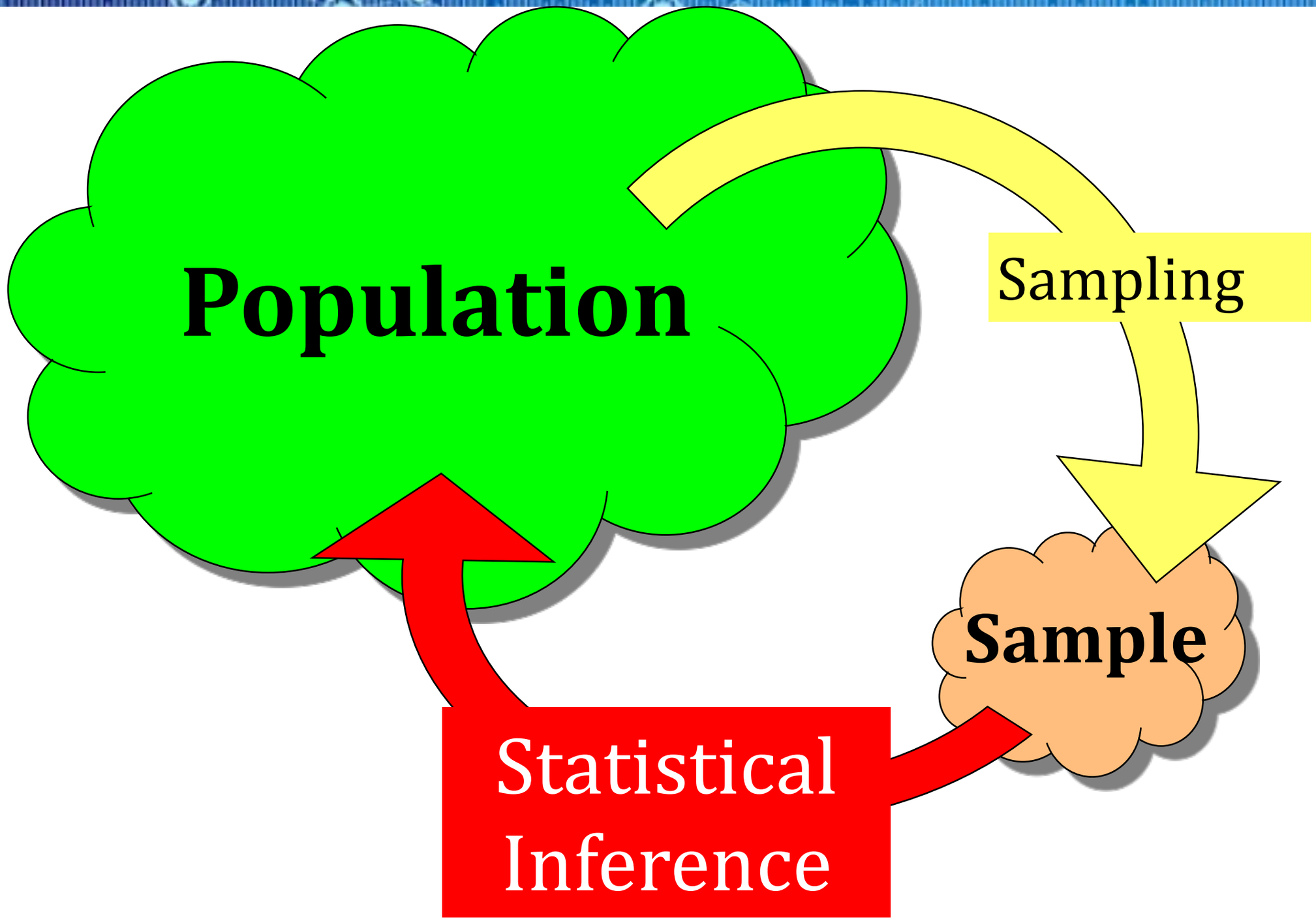
Data Collection and Bias



Other Forms of Bias

- Even with a random sample, data can still be biased, especially when collected on humans
- Other forms of bias to watch out for in data collection:
  - Question wording
  - Context
  - Inaccurate responses
  - Many other possibilities – examine the specifics of each study!

Confidence Interval : Sampling Distribution


Population
Sampling
Sample
Statistical Inference

Statistical Inference

Statistical inference is the process of drawing conclusions about the entire population based on information in a sample.

Statistic and Parameter

A ***parameter*** is a number that describes some aspect of a population.

A ***statistic*** is a number that is computed from data in a sample.

- We usually have a sample statistic and want to use it to make inferences about the population parameter

Population

PARAMETERS

Sampling

Sample

STATISTICS

Statistical Inference

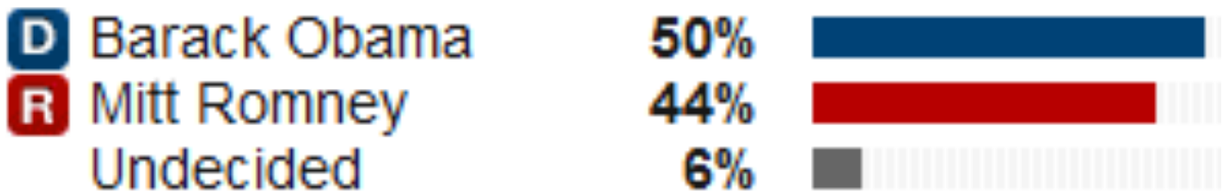
Parameter versus Statistic

	Parameter	Statistic
Mean	μ mu	$\bar{x}$ x-bar
Proportion	p	$\hat{p}$ p-hat
Std. Dev.	σ sigma	s
Correlation	ρ	r
Slope	β rho	b

Election Polling

Before the 2012 presidential election, 1000 registered voters were asked who they plan to vote for in the 2012 presidential election

What proportion of voters planned to vote for Obama?



$\hat{p} = 0.50$ $p = ???$

http://www.politico.com/p/2012-election/polls/president

Point Estimate

We use the statistic from a sample as a ***point estimate*** for a population parameter.

- Point estimates will not match population parameters exactly, but they are our best guess, given the data

Election Polls

Actually, several polls were conducted over this time frame (9/7/12 – 9/9/12):

National '12 President General Election

Washington Post-ABC News
09/07/2012-09/09/2012
710 likely voters

D	Barack Obama	49%
R	Mitt Romney	48%
	No opinion	3%

National '12 President General Election

Public Policy Polling/SIEU/Daily Kos
09/07/2012-09/09/2012
1000 registered voters

D	Barack Obama	50%
R	Mitt Romney	44%
	Undecided	6%

National '12 President General Election

CNN/ORC International
09/07/2012-09/09/2012
709 likely voters

D	Barack Obama	52%
R	Mitt Romney	46%
	Neither	2%
	No opinion	1%

http://www.politico.com/p/2012-election/polls/president

IMPORTANT POINTS

- Sample statistics **vary** from sample to sample. (they will not match the parameter exactly)

- **KEY QUESTION**: For a given sample statistic, what are plausible values for the population parameter? How much uncertainty surrounds the sample statistic?

- **KEY ANSWER**: It depends on how much the statistic varies from sample to sample!

Many Samples

- To see how statistics vary from sample to sample, let's take many samples and compute many statistics!

Reese's Pieces

- What proportion of Reese's pieces are orange?
- Take a random sample of 10 Reese's pieces
- What is your sample proportion? ⇒ dotplot
- Give a range of plausible values for the population proportion
- You just made your first sampling distribution!

Sampling Distribution

A ***sampling distribution*** is the distribution of sample statistics computed for different samples of the same size from the same population.

- A sampling distribution shows us how the sample statistic varies from sample to sample

Center and Shape

Center: If samples are randomly selected, the sampling distribution will be centered around the population parameter.

Shape: For most of the statistics we consider, if the sample size is large enough the sampling distribution will be symmetric and bell-shaped.

Sampling Caution

- If you take **random samples**, the sampling distribution will be centered around the true population parameter

- If sampling bias exists (if you do not take random samples), your sampling distribution may give you bad information about the true parameter

- "The. Polls. Have. Stopped. Making. Any. Sense."

Sampling Distribution

- We've learned about center and shape, but remember what we really care about is ***variability*** of the sampling distribution
- Remember our key question and answer: to assess uncertainty of a statistic, we need to know how much the statistic varies from sample to sample!
- The variability of the sample statistic is so important that it gets it's own name…

Standard Error

The **standard error** of a statistic, SE, is the standard deviation of the sample statistic

- The standard error measures how much the statistic varies from sample to sample
- The standard error can be calculated as the standard deviation of the sampling distribution

Standard Error

The more the statistic varies from sample to sample, the

a) higher

b) lower

the standard error.

The standard error measures how much the statistic varies from sample to sample.

Sample Size Matters!

As the sample size increases, the variability of the sample statistics tends to decrease and the sample statistics tend to be closer to the true value of the population parameter

- For larger sample sizes, you get less variability in the statistics, so less uncertainty in your estimates

Sample Size

Suppose we were to take samples of size 10 and samples of size 100 from the same population, and compute the sample means. Which sample means would have the *higher* standard error?

a) The sample means using n = 10
b) The sample means using n = 100

Smaller sample sizes give more variability, so a higher standard error

Interval Estimate

An ***interval estimate*** gives a range of plausible values for a population parameter.

Margin of Error

One common form for an interval estimate is

statistic ± margin of error

where the ***margin of error*** reflects the precision of the sample statistic as a point estimate for the parameter.

Margin of Error

The higher the standard deviation of the sampling distribution, the

a) higher
b) lower

the margin of error.

The higher the variability in the statistic, the higher the uncertainty in the statistic.

Confidence Interval

A ***confidence interval*** for a parameter is an interval computed from sample data by a method that will capture the parameter for a specified proportion of all samples

- The success rate (proportion of all samples whose intervals contain the parameter) is known as the **confidence level**
- A 95% confidence interval will contain the true parameter for 95% of all samples

Sampling Distribution

If you had access to the sampling distribution, how would you find the margin of error to ensure that intervals of the form

statistic ± margin of error

would capture the parameter for 95% of all samples?

95% Confidence Interval

If the sampling distribution is relatively symmetric and bell-shaped, a 95% confidence interval can be estimated using

statistic ± 2 × SE

Interpreting a Confidence Interval

95% of all samples yield intervals that contain the true parameter, so we say we are "95% sure" or "95% confident" that one interval contains the truth.

"We are 95% confident that the true proportion of all Americans that considered the economy a 'top priority' in January 2012 is between 0.84 and 0.88"

Carbon in Forest Biomass

Scientists hoping to curb deforestation estimate that the carbon stored in tropical forests in Latin America, sub-Saharan Africa, and southeast Asia has a total biomass of 247 gigatons.

To arrive at this estimate, they first estimate the mean amount of carbon per square kilometer. Based on a sample of size n = 4079 inventory plots, the sample mean is tons with a standard

Saatchi, S.S. et. al. "Benchmark Map of Forest Carbon Stocks in Tropical Regions Across Three Continents," *Proceedings of the National Academy of Sciences,* 5/31/11.

Carbon in Forest Biomass

95% CI: 11,600 ± 2·1000 = (9,600, 13,600)

We are 95% confident that the average amount of carbon stored in each square kilometer of tropical forest is between 9,600 and 13,600 tons.

Reese's Pieces

Each of you will create a 95% confidence interval based off your sample. If you all sampled randomly, and all create your CI correctly, what percentage of your intervals do you expect to include the true p?

a) 95%

b) 5%

c) All of them

d) None of them

Confidence Intervals

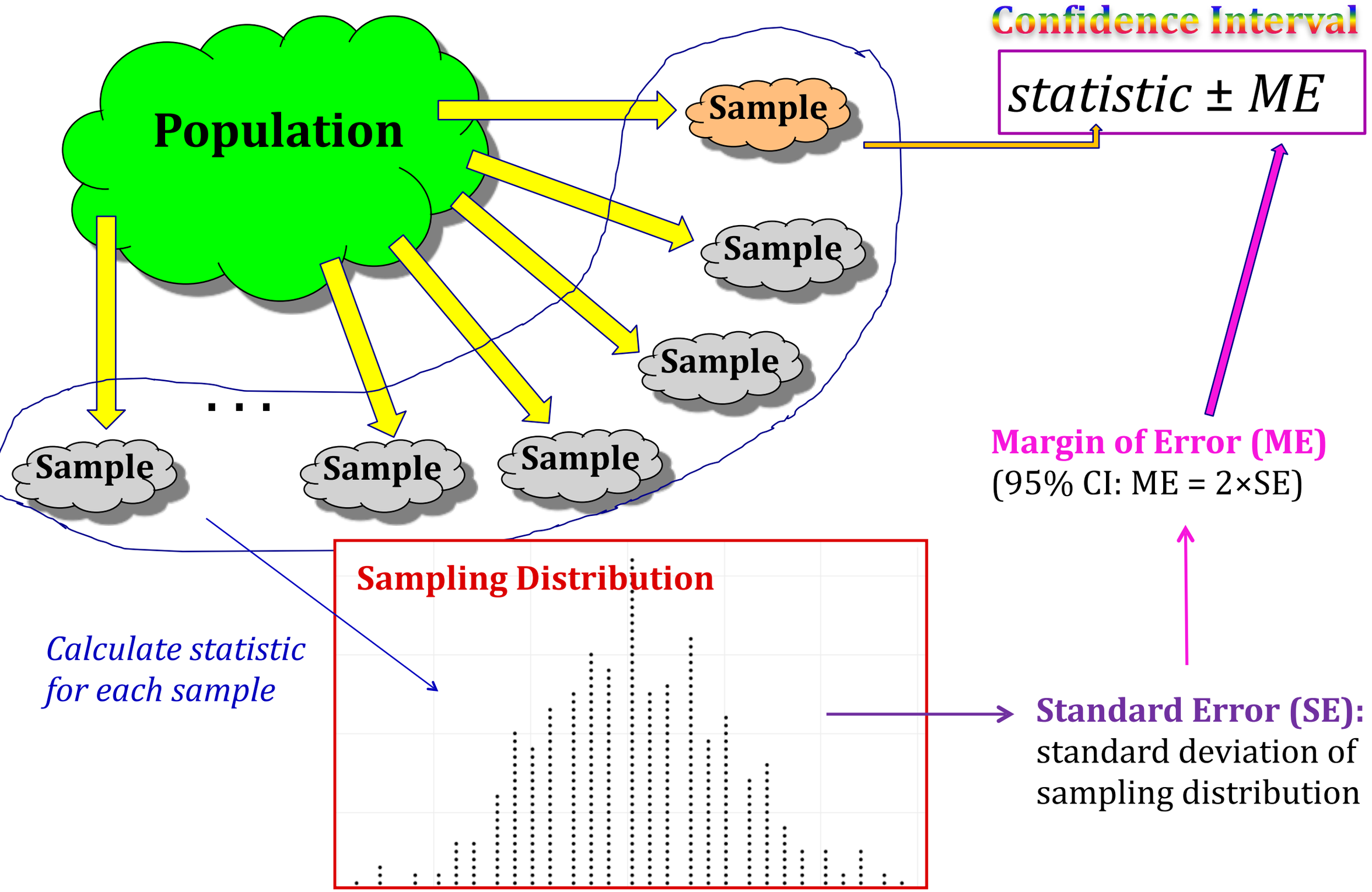
If context were added, which of the following would be an appropriate interpretation for a 95% confidence interval:

a) "we are 95% sure the interval contains the parameter"
b) "there is a 95% chance the interval contains the parameter"
c) Both (a) and (b)
d) Neither (a) or (b)

95% of all samples yield intervals that contain the true parameter, so we say we are "95% sure" or "95% confident" that one interval contains the truth.

We can't make probabilistic statements such as (b) because the interval either contains the truth or it doesn't, and also the 95% pertains to all intervals that could be generated, not just the one you've created.

Confidence Intervals



Summary

- To create a plausible range of values for a parameter:
  - Take many random samples from the population, and compute the sample statistic for each sample
  - Compute the standard error as the standard deviation of all these statistics
  - Use statistic ± 2×SE

Hypothesis Testing

Fundamental of Hypothesis Testing

There two types of **statistical inferences**, **Estimation** and **Hypothesis Testing**

Hypothesis Testing: A hypothesis is a **claim** (assumption) about one or more **population** parameters.

- Average price of a six-pack in the U.S. is $\mu = \$4.90$
- The population mean monthly cell phone bill of this city is: $\mu = \$42$
- The average number of TV sets in U.S. Homes is equal to three; $\mu = 3$

It Is always about a population parameter, not about a sample statistic

Sample evidence is used to assess the probability that the claim about the population parameter is true

A. It starts with Null Hypothesis, H_0

$H_0: \quad =$

1. We begin with the assumption that H_0 is true and any difference between the sample statistic and true population parameter is due to chance and not a real (systematic) difference.
2. Similar to the notion of “innocent until proven guilty”
3. That is, “innocence” is a null hypothesis.

Null Hypo, Continued

4. Refers to the status quo
5. Always contains “=” , “≤” or “≥” sign
6. May or may not be rejected

B. Next we state the Alternative Hypothesis, H_1

1. Is the opposite of the null hypothesis
   1. e.g., The average number of TV sets in U.S. homes is not equal to 3 (H_1: $\mu \neq 3$)
2. Challenges the status quo
3. Never contains the “=” , “≤” or “≥” sign
4. May or may not be proven
5. Is generally the hypothesis that the researcher is trying to prove. Evidence is always examined with respect to **H_1,** never with respect to **H_0.**
6. We never “accept” **H_0,** we either “reject” or “not reject” it

Summary:

- In the process of hypothesis testing, the null hypothesis initially is assumed to be true
- Data are gathered and examined to determine whether the evidence is strong enough with respect to the alternative hypothesis to reject the assumption.
- In another words, the burden is placed on the researcher to show, using sample information, that the null hypothesis is false.
- If the sample information is sufficient enough in favor of the alternative hypothesis, then the null hypothesis is rejected. This is the same as saying if the persecutor has enough evidence of guilt, the “innocence is rejected.
- Of course, erroneous conclusions are possible, type I and type II errors.

Reason for Rejecting H_0

Illustration: Let say, we **assume** that average age in the US is 50 years (H0=50). If in fact this is the true (unknown) population mean, it is unlikely that we get a sample mean of 20. So, if we have a sample that produces an average of 20, them we **reject** that the null hypothesis that average age is 50. (note that we are rejecting our assumption or claim). (would we get 20 if the true population mean was 50? NO. That is why we reject 50)

How Is the Test done?

We use the distribution of a Test Statistic, such as Z or t as the criteria.

A. Rejection Region Method:

Divide the distribution into rejection and non-rejection regions

Defines the unlikely values of the sample statistic if the null hypothesis is true, the critical value(s)

Defines rejection region of the sampling distribution

Rejection region(s) is designated by α , (level of significance)

Typical values are .01, .05, or .10

α is selected by the researcher at the beginning

α provides the critical value(s) of the test

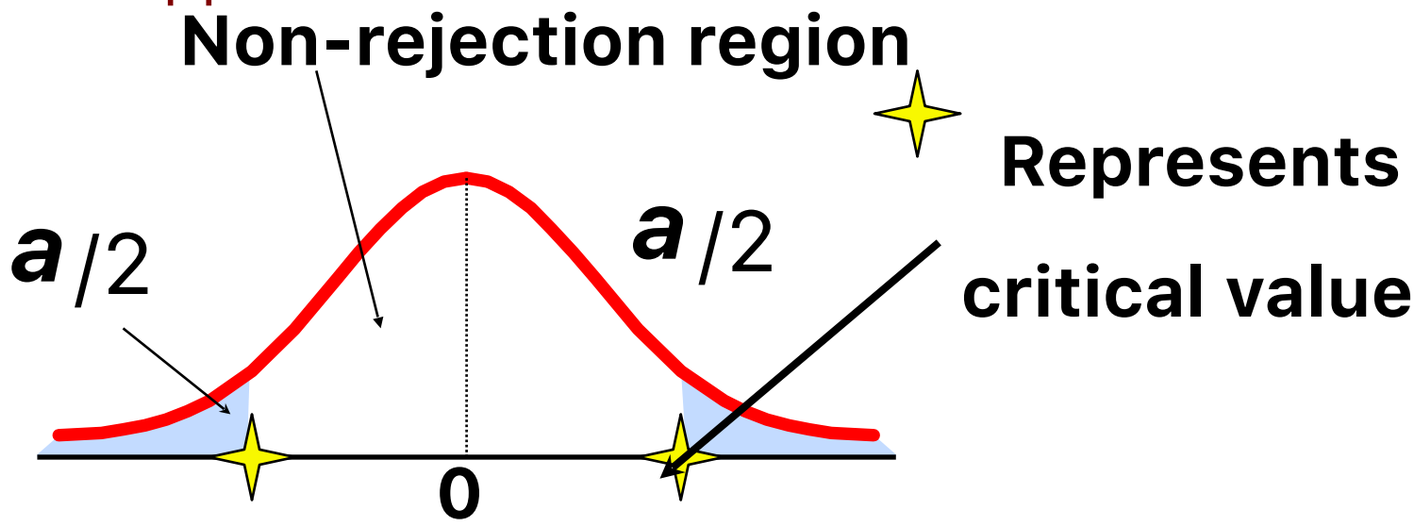
Rejection Region or Critical Value Approach:

Level of significance = α

H_0: μ = 12

H_1: μ ≠ 12

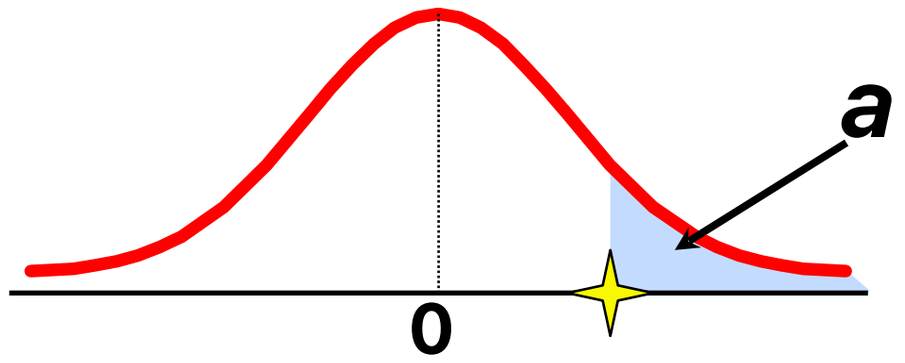
Two-tail test



H_0: μ ≤ 12

H_1: μ > 12

Upper-tail test

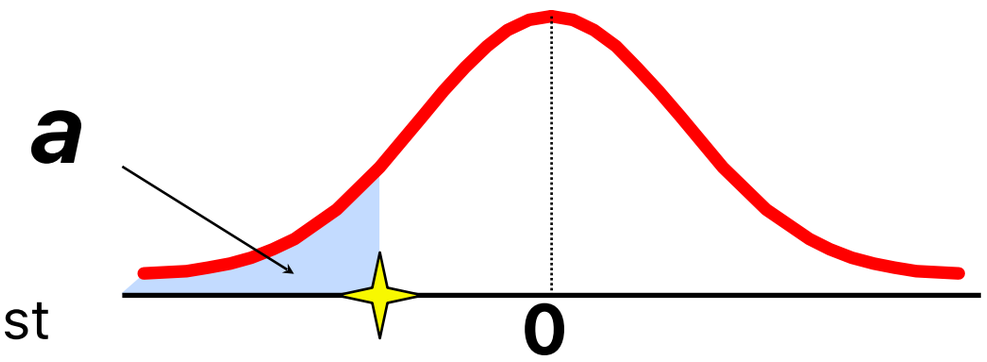


Rejection region is shaded

H_0: μ ≥ 12

H_1: μ < 12

Lower-tail test



P-Value Approach –

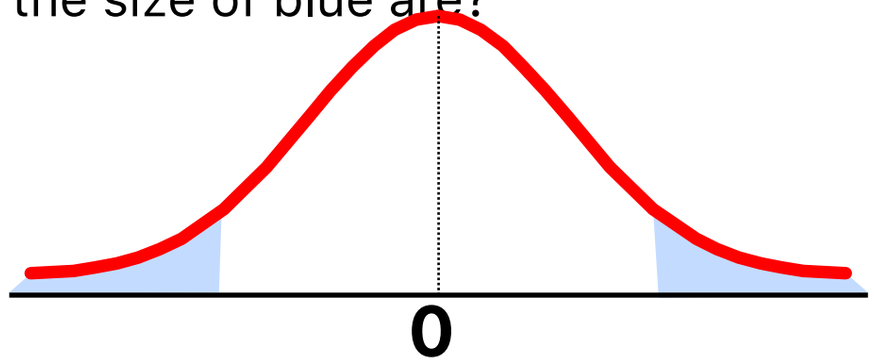
P-value=Max. Probability of (Type I Error), calculated from the sample.

Given the sample information what is the size of blue are?

H_0: μ = 12

H_1: μ ≠ 12

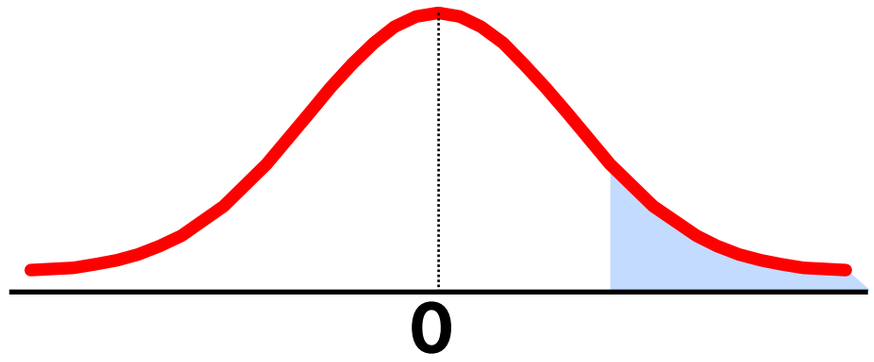
Two-tail test



H_0: μ ≤ 12

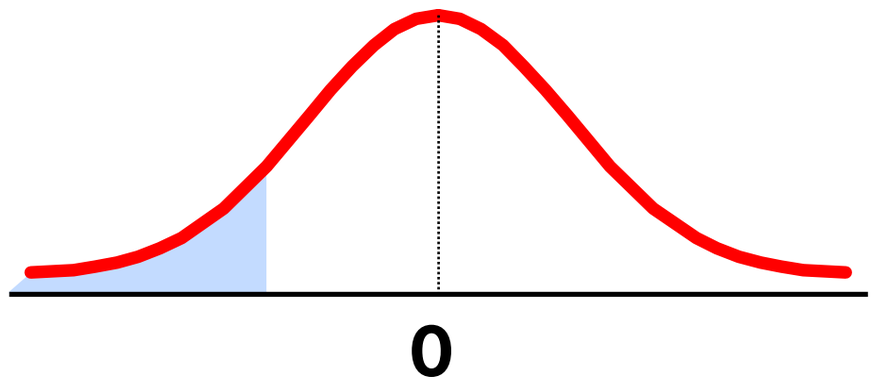
H_1: μ > 12

Upper-tail test



H_0: μ ≥ 12

H_1: μ < 12



Type I and II Errors:

The size of α , the rejection region, affects the risk of making different types of incorrect decisions.

Type I Error

Rejecting a true null hypothesis when it should NOT be rejected

Considered a serious type of error

The probability of Type I Error is α

It is also called level of significance of the test

Type II Error

Fail to reject a false null hypothesis that should have been rejected

The probability of Type II Error is β

Decision	Actual Situation			
	Hypothesis Testing		Legal System	
	H_0 True	H_0 False	Innocence	Not innocence
Do Not Reject H_0	**No Error** $(1-\alpha)$	**Type II Error** (β)	No Error (not guilty, found not guilty) $(1-\alpha)$	Type II Error (guilty, found not guilty) (β)
Reject H_0	**Type I Error** (α)	**No Error** $(1-\beta)$	Type I Error (Not guilty, found guilty) (α)	No Error (guilty, found guilty) $(1-\beta)$

Type I and Type II errors cannot happen at the same time

1. Type I error can only occur if H_0 is true

2. Type II error can only occur if H_0 is false

3. There is a tradeoff between type I and II errors. If the probability of type I error (α) increased, then the probability of type II error (β) declines.

4. When the difference between the hypothesized parameter and the actual true value is small, the probability of type two error (the non-rejection region) is larger.

5. Increasing the sample size, n, for a given level of α, reduces β

B. P-Value approach to Hypothesis Testing:

1. The rejection region approach allows you to examine evidence but restrict you to not more than a certain probability (say α = 5%) of rejecting a true H_0 by mistake.

2. The P-value approach allows you to use the information from the sample and then calculate the maximum probability of rejecting a true H_0 by mistake.

3. Another way of looking at P-value is the probability of observing a sample information of "A=11.5" when the true population parameter is "12=B". The P-value is the maximum probability of such mistake taking

4. That is to say that P-value is the smallest value of α for which H_0 can be rejected based on the sample information

5. Convert Sample Statistic (e.g., sample mean) to Test Statistic (e.g., Z statistic)

6. Obtain the p-value from a table or computer

7. Compare the p-value with α

If p-value $<$ α , reject H_0

If p-value $\geq$ α , do not reject H_0

Test of Hypothesis for the Mean

σ known

The test statistic is:

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

σ Unknown

The test statistic is:

$$t_{n-1} = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}$$

Steps to Hypothesis Testing

1. State the H_0 and H_1 clearly
2. Identify the test statistic (two-tail, one-tail, and Z or t distribution
3. Depending on the type of risk you are willing to take, specify the level of significance,
4. Find the decision rule, critical values, and rejection regions. If –CV<actual value (sample statistic) <+CV, then do not reject the H_0

Make statistical decision

Do not Reject H_0

Conclude H_0 may be true

Reject H_0

Conclude H_1 is “true”

(There is sufficient evidence of H1)

Make management/ business/administrative decision

When do we use a two-tail test?

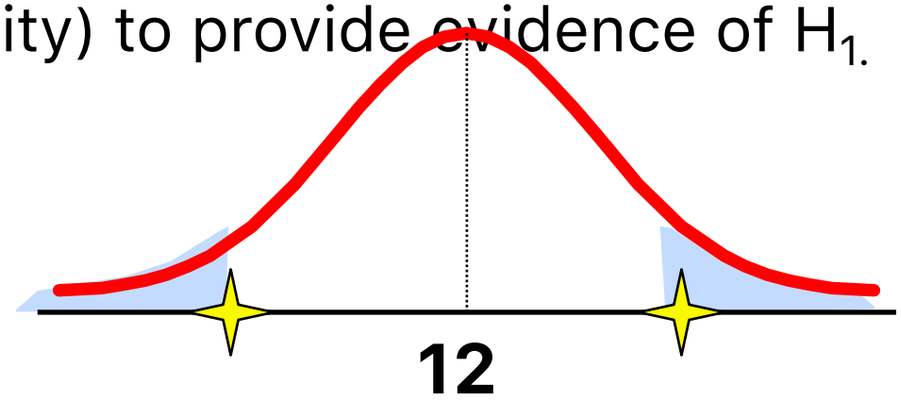
The answer depends on the question you are trying to answer.

A two-tail is used when the researcher has no idea which direction the study will go, interested in both direction. (example: testing a new technique, a new product, a new theory and we don't know the direction)

A new machine is producing 12 fluid once can of soft drink. The quality control manager is concern with cans containing too much or too little. Then, the test is a two-tailed test. That is the two rejection regions in tails is most likely (higher probability) to provide evidence of H_1.

$$H_0 : \mu = 12\ oz$$

$$H_1 : \mu \neq 12\ oz$$

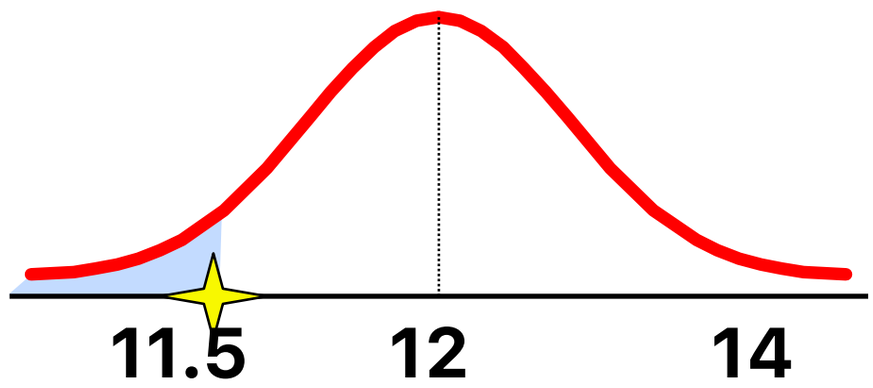


One-tail test is used when the researcher is interested in the direction.

Example: The soft-drink company puts a label on cans claiming they contain 12 oz. A consumer advocate desires to test this statement. She would assume that each can contains at least 12 oz and tries to find evidence to the contrary. That is, she examines the evidence for less than 12 0z.

What tail of the distribution is the most logical (higher probability) to find that evidence? The only way to reject the claim is to get evidence of less than 12 oz, left tail.

$$H_0 : \mu \geq 12\ oz$$
$$H_1 : \mu < 12\ oz$$



Review of Hypo. Testing

What is HT?

Probability of making erroneous conclusions

- Type I – only when Null Hypo is true
- Type II – only when Null Hypo is false

Two Approaches

- The Rejection or Critical Value Approach
- The P-value Approach (we calculate the observed level of significance)

Test Statistics

- Z- distribution if Population Std. Dev. is Know
- t-distribution if the Population Std. Dev. is unknown

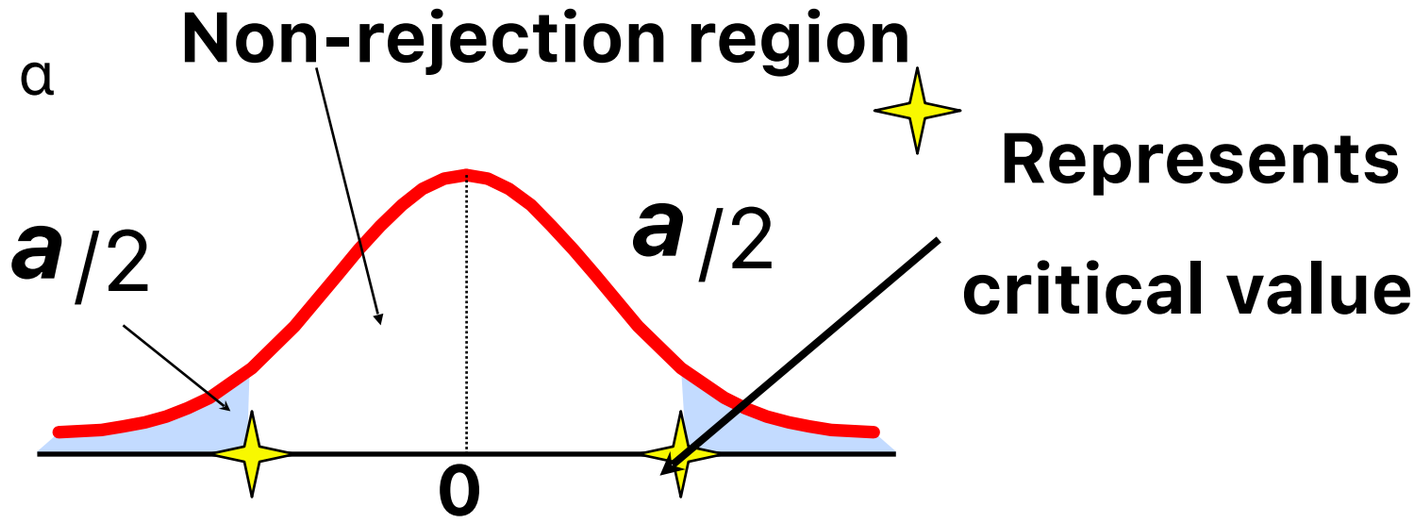
Rejection Region or Critical Value Approach:

The given level of significance = α

H_0: μ = 12

H_1: μ ≠ 12

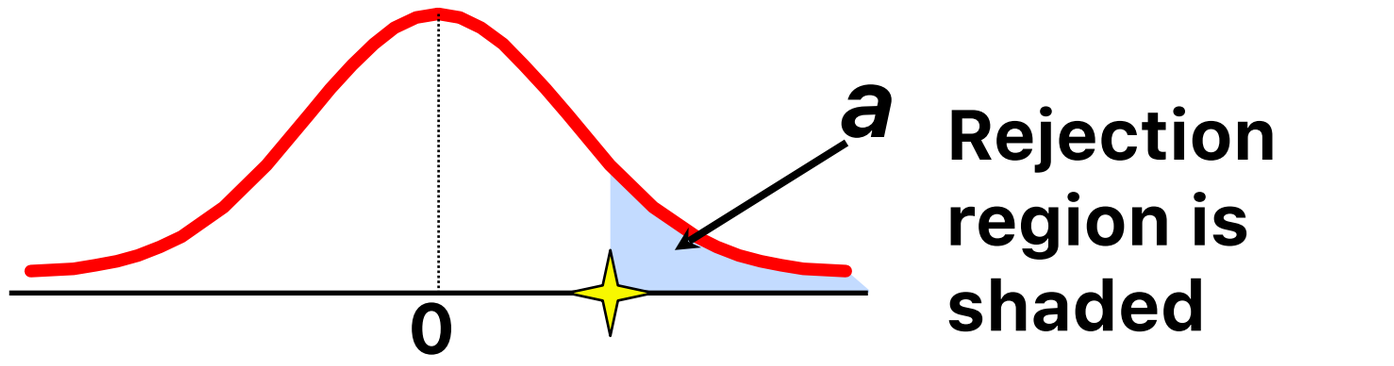
Two-tail test



H_0: μ ≤ 12

H_1: μ > 12

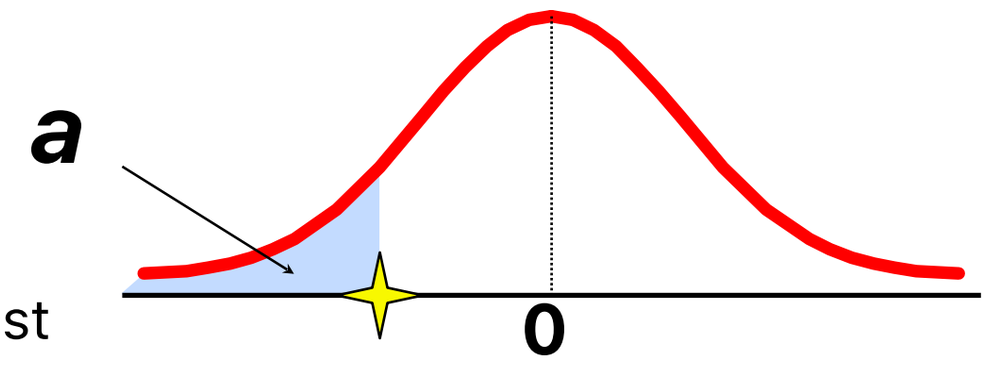
Upper-tail test



H_0: μ ≥ 12

H_1: μ < 12

Lower-tail test



P-Value Approach –

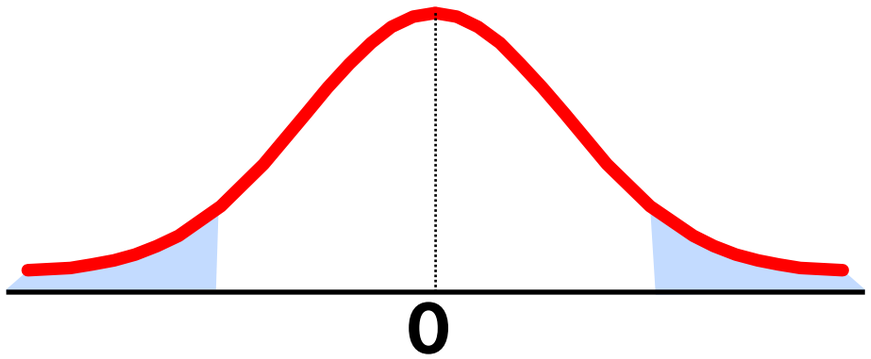
P-value=Max. Probability of (Type I Error), calculated from the sample.

Given the sample information what is the size of the blue areas? (The observed level of significance)

$H_0: \mu = 12$

$H_1: \mu \neq 12$

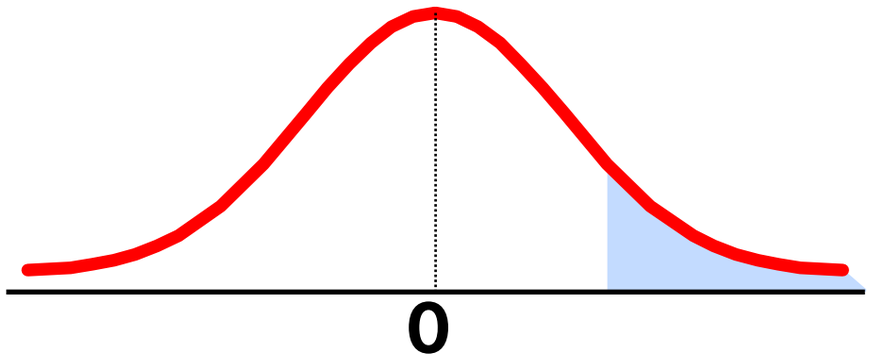
Two-tail test



$H_0: \mu \leq 12$

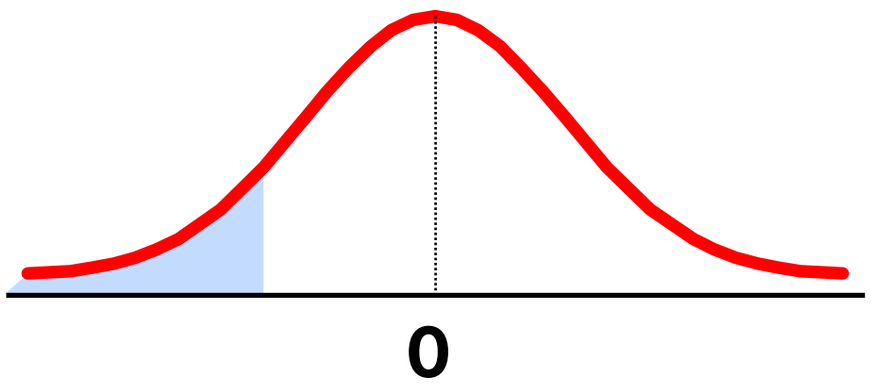
$H_1: \mu > 12$

Upper-tail test



$H_0: \mu \geq 12$

$H_1: \mu < 12$



Example 1:

Let's assume a sample of 25 cans produced a sample mean of 11.5 0z and the population std dev=1 0z.

Question 1:

At a 5% level of significance (that is allowing for a maximum of 5% prob. of rejecting a true null hypo), is there evidence that the population mean is different from 12 oz?

Null Hypo is:?
Alternative Hypo is?

Can both approaches be used to answer this question?

A: Rejection region approach: calculate the actual test statistics and compare it with the critical values

B: P-value approach: calculate the actual probability of type I error given the sample information. Then compare it with 1%, 5%, or 10% level of significance.

Interpretation of Critical Value/Rejection Region Approach:

Interpretation of P-value Approach:

Question 2:

At a 5% level of significance (that is allowing for a maximum of 5% prob. of rejecting a true null hypo), is the evidence that the population mean is less than 12 oz?

Null Hypo is:?

Alternative Hypo is?

Can both approaches be used to answer this question?

Interpretation of Critical Value Approach:

Interpretation of P-value Approach:

Question 3:

If in fact the pop. mean is 12 oz, what is the probability of obtaining a sample mean of 11.5 or less oz (sample size 25)? Null

Null Hypo is:?

Alternative Hypo is?

Question 4:

If in fact the pop. mean is 12 oz, and the sample mean is 11.5 (or less), what is the probability of erroneously rejecting the null hypo that the pop. mean is 12 oz?

Null Hypo is:?

Alternative Hypo is?

Can both approaches be used to answer these question?

Connection to Confidence Intervals

While the confidence interval estimation and hypothesis testing serve different purposes, they are based on same concept and conclusions reached by two methods are consistent for a two-tail test.

In CI method we estimate an interval for the population mean with a degree of confidence. If the estimated interval contains the hypothesized value under the hypothesis testing, then this is equivalent of not rejecting the null hypothesis. For example: for the beer sample with mean 5.20, the confidence interval is:

$P(4.61 \leq \mu \leq 5.78)=95\%$

Since this interval contains the Hypothesized mean ($4.90), we do not (did not) reject the null hypothesis at $\alpha = .05$

Did not reject and within the interval, thus consistent results.

t-test using R

Testing differences in mean between two groups

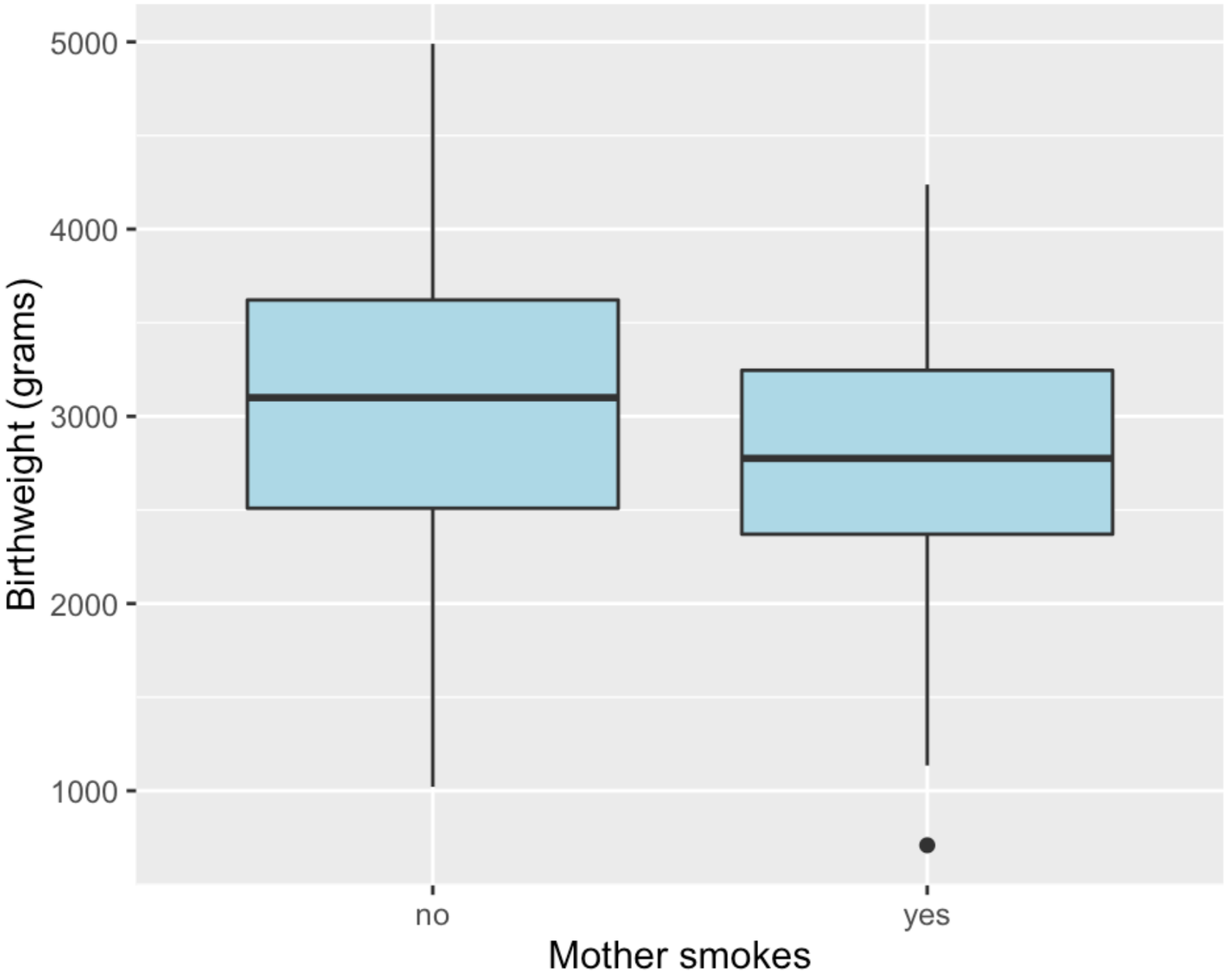
Let's begin by loading the packages we'll need to get started

```
library(MASS)
library(plyr)
library(ggplot2)
```

```
# Rename the columns to have more descriptive names
colnames(birthwt) <- c("birthwt.below.2500", "mother.age", "mother.weight",
    "race", "mother.smokes", "previous.prem.labor", "hypertension", "uterine.irr",
    "physician.visits", "birthwt.grams")

# Transform variables to factors with descriptive levels
birthwt <- transform(birthwt,
            race = as.factor(mapvalues(race, c(1, 2, 3),
                             c("white","black", "other"))),
            mother.smokes = as.factor(mapvalues(mother.smokes,
                             c(0,1), c("no", "yes"))),
            hypertension = as.factor(mapvalues(hypertension,
                             c(0,1), c("no", "yes"))),
            uterine.irr = as.factor(mapvalues(uterine.irr,
                             c(0,1), c("no", "yes")))
            )
```

To start, it always helps to plot things

```
# Create boxplot showing how birthwt.grams varies between
# smoking status
qplot(x = mother.smokes, y = birthwt.grams,
      geom = "boxplot", data = birthwt,
      xlab = "Mother smokes",
      ylab = "Birthweight (grams)",
      fill = I("lightblue"))
```



This plot suggests that smoking is associated with lower birth weight.

Let's compute a summary table

```
aggregate(birthwt.grams ~ mother.smokes, data = birthwt,
          FUN = function(x) {c(mean = mean(x), sd = sd(x))})
```

```
##   mother.smokes birthwt.grams.mean birthwt.grams.sd
## 1            no          3055.6957         752.6566
## 2           yes          2771.9189         659.6349
```

The standard deviation is good to have, but to assess statistical significance we really want to have the standard error (which the standard deviation adjusted by the group size).

```
aggregate(birthwt.grams ~ mother.smokes, data = birthwt,
          FUN = function(x) {c(mean = mean(x),
                               se = sd(x) / sqrt(length(x)))})
```

```
##   mother.smokes birthwt.grams.mean birthwt.grams.se
## 1            no         3055.69565         70.18559
## 2           yes         2771.91892         76.68100
```

t-test via t.test()

This difference is looking quite significant. To run a two-sample t-test, we can simple use the `t.test()` function.

```
birthwt.t.test <- t.test(birthwt.grams ~ mother.smokes, data = birthwt)
birthwt.t.test
```

```
##
##  Welch Two Sample t-test
##
## data:  birthwt.grams by mother.smokes
## t = 2.7299, df = 170.1, p-value = 0.007003
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   78.57486 488.97860
## sample estimates:
##  mean in group no mean in group yes
##          3055.696           2771.919
```

p-value

```
birthwt.t.test$p.value   # p-value
```

```
## [1] 0.007002548
```

```
birthwt.t.test$estimate  # group means
```

```
##  mean in group no mean in group yes
##          3055.696           2771.919
```

```
birthwt.t.test$conf.int  # confidence interval for difference
```

```
## [1]  78.57486 488.97860
## attr(,"conf.level")
## [1] 0.95
```

Workshop 3.1 : t-test

Testing means between two groups

(a) Using the Cars93 data and the t.test() function, run a t-test to see if average MPG.highway is different between US and non-US vehicles.

Try doing this both using the formula style input and the x, y style input.

(b) What is the confidence interval for the difference?

Analysis of Variance

- **Analysis of Variance (ANOVA)** compares the variability between groups to the variability within groups

Total Variability = **Variability Between Groups** + **Variability Within Groups**

Analysis of Variance

If the groups are actually different, then

a) the variability between groups should be higher than the variability within groups

b) the variability within groups should be higher than the variability between groups

F-Statistic

- The **F-statistic** is a ratio of the average variability between groups to the average variability within groups

- $$F = \frac{MSG}{MSE} = \frac{\text{average between group variability}}{\text{average within group variability}}$$

F-statistic

If there really is a difference between the groups, we would expect the F-statistic to be

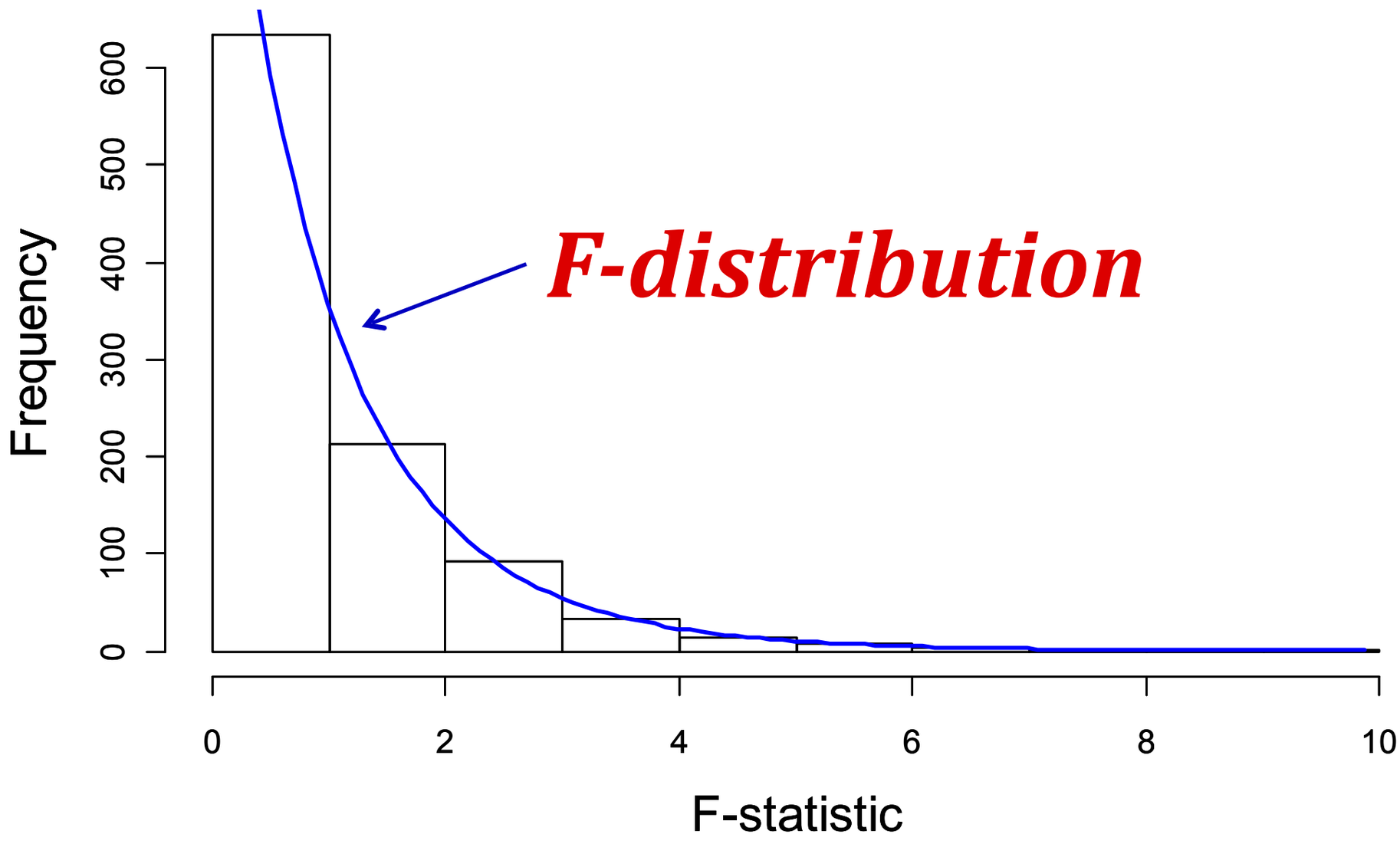
a) Higher than we would observe by random chance

b) Lower than we would observe by random chance

We have a test statistic. What else do we need to perform the hypothesis test?

A distribution of the test statistic assuming H_0 is true

F-distribution



Prepare data

```
library(MASS)
library(plyr)
library(ggplot2)
library(knitr)
```

```
# Rename the columns to have more descriptive names
colnames(birthwt) <- c("birthwt.below.2500", "mother.age", "mother.weight",
    "race", "mother.smokes", "previous.prem.labor", "hypertension", "uterine.irr",
    "physician.visits", "birthwt.grams")
```

```
# Transform variables to factors with descriptive levels
birthwt <- transform(birthwt,
            race = as.factor(mapvalues(race, c(1, 2, 3),
                             c("white","black", "other"))),
            mother.smokes = as.factor(mapvalues(mother.smokes,
                             c(0,1), c("no", "yes"))),
            hypertension = as.factor(mapvalues(hypertension,
                             c(0,1), c("no", "yes"))),
            uterine.irr = as.factor(mapvalues(uterine.irr,
                             c(0,1), c("no", "yes")))
            )
```

One-way ANOVA example

Question: Is there a significant association between race and birthweight?

Here's a table showing the mean and standard error of birthweight by race.

```
aggregate(birthwt.grams ~ race, data = birthwt, FUN = mean)
```

```
##    race birthwt.grams
## 1 black      2719.692
## 2 other      2805.284
## 3 white      3102.719
```

Terminology: a k-way ANOVA is used to assess whether the mean of an outcome variable is constant across all combinations of k factors. The most common examples are 1-way ANOVA (looking at a single factor), and 2-way ANOVA (looking at two factors).

We'll use the `aov()` function. For convenience, `aov()` allows you to specify a formula.

```
summary(aov(birthwt.grams ~ race, data = birthwt))
```

```
##              Df   Sum Sq Mean Sq F value  Pr(>F)
## race          2  5015725 2507863   4.913 0.00834 **
## Residuals   186 94953931  510505
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Workshop 3.2 : ANOVA

Let's form our favourite birthwt data set.

```
# Rename the columns to have more descriptive names
colnames(birthwt) <- c("birthwt.below.2500", "mother.age", "mother.weight",
    "race", "mother.smokes", "previous.prem.labor", "hypertension", "uterine.irr",
    "physician.visits", "birthwt.grams")

# Transform variables to factors with descriptive levels
birthwt <- transform(birthwt,
            race = as.factor(mapvalues(race, c(1, 2, 3),
                             c("white","black", "other"))),
            mother.smokes = as.factor(mapvalues(mother.smokes,
                             c(0,1), c("no", "yes"))),
            hypertension = as.factor(mapvalues(hypertension,
                             c(0,1), c("no", "yes"))),
            uterine.irr = as.factor(mapvalues(uterine.irr,
                             c(0,1), c("no", "yes")))
            )
```

(a) Create a new factor that categorizes the number of physician visits into three levels: 0, 1, 2, 3 or more.

Hint: One way of doing this is with mapvalues, by mapping all instances of 3, 4,... etc, to “3 or more”.

(b) Run an ANOVA to determine whether the average birth weight varies across number of physician visits.

Regression

DATA SCIENCE CERTIFICATION

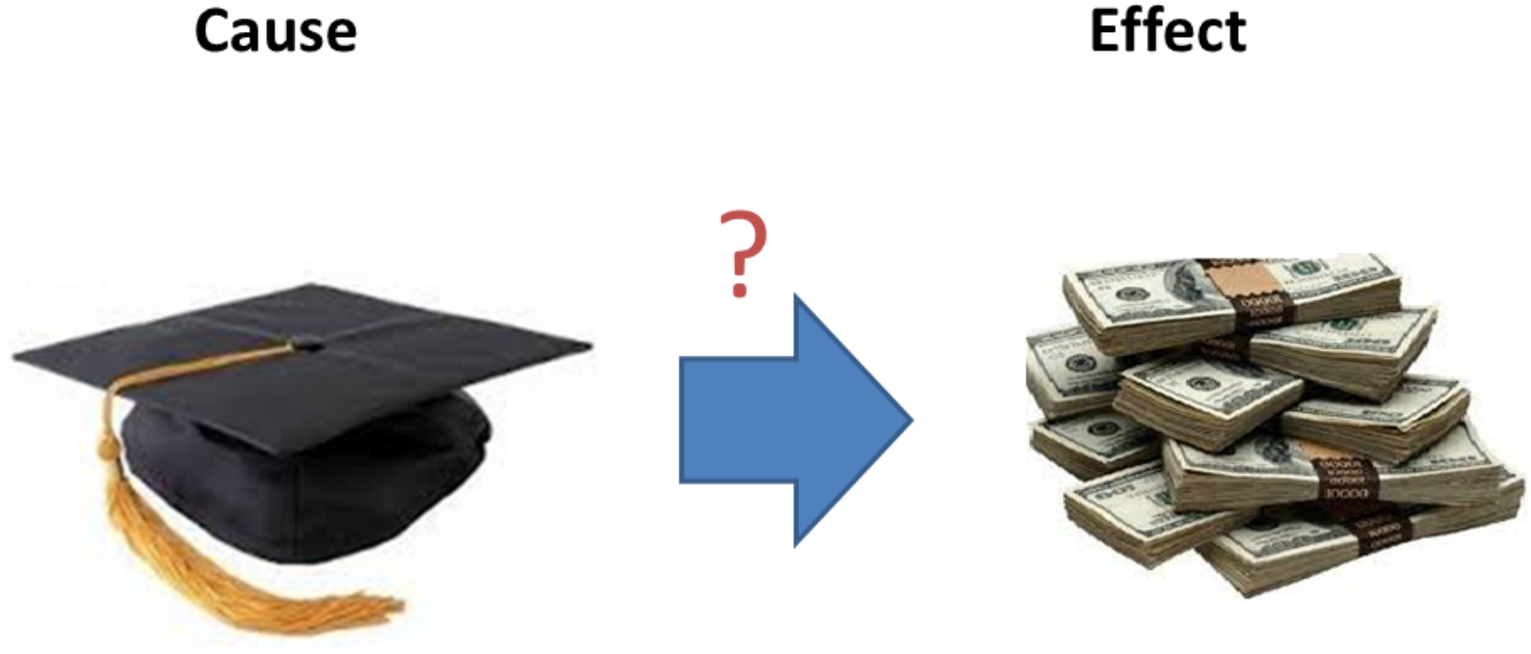
What’s the difference between Causality and Correlation?

Causation and Correlation are loosely used words in analytics. People tend to use these words interchangeably without knowing the fundamental logic behind them. Apparently, people get trapped in the phonetics of these words and end up using them at incorrect places.

Let’s understand the difference between Causation and Correlation using a few examples below. Analyze the following scenarios and tell us whether there is a causal relation between the two events (X and Y). Answers are provided below.

Example 1 : X – Tier of B-school college a student gets offer for => Y – Salary after the graduation

Hypothesis – Students going to premium B-schools get higher salaries on an average. Are these B-school a cause of getting better jobs?



Example 2 : X – Smoking Cigarettes => Y – Level of Mental Stress

Hypothesis – People who smoke are found to have higher level of stress.

Is smoking the reason of stress?

Cause **Effect**

?

Example 3 : X – Having Kids => Y – Maturity level

Hypothesis – People get more matured after having kids? Is having kids a cause of attaining higher maturity levels?

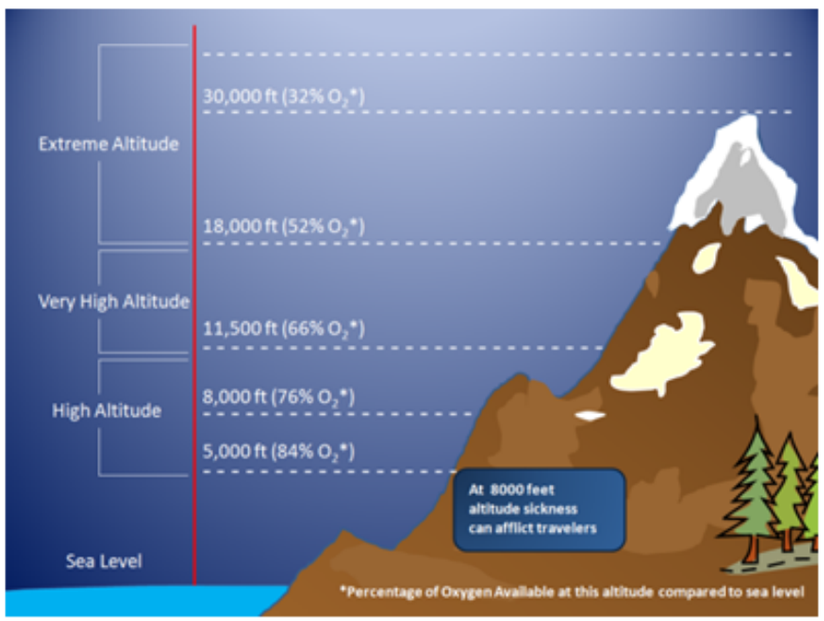
Cause **Effect**

?

Example 4 : X – Altitude => Y – Temperature

Hypothesis – We witness lower temperature at high altitudes. Which means, the higher you go, the colder it would become. Is higher altitude a cause of lower temperature?

Cause **Effect**



?

- **Example 1** : Causal relation does not exist. For instance, only ambitious and intelligent people are selected from elite B-schools who further get much higher salary than the average. Hence, even if these students did not study in Tier 1 B-School, he/she still might get more than the average salaries. Hence, we have alternate reasoning issue in this case.
- **Example 2** : Causal relation does not exist. We can reject hypothesis based on inverse causality. For instance, higher mental stress can actually influence a person to smoke.
- **Example 3** : Causal relation does not exist. Once again, we can reject hypothesis based on inverse causality. For instance, only mature people are likely to be prepared to have kids. We can also apply alternate reasoning with underlying cause as the age. Higher age leads to both, having kids and higher maturity levels.
- **Example 4** : Causal relation does exist. We definitely know that inverse causality is not possible. Also alternate reasoning or mutual independence can be rejected.

What are the keypoints in establishing causation?

Here are the key point (X = > Y) pairs used in establishing causation :

1. **Alternate Reasoning :** If there is an alternate reason (Z) which indeed can influence both X and Y (Z => X & Z => Y are true) , we can reject the hypothesis of X => Y.

2. **Inverse Causality :** If instead of X influencing Y, we have Y influencing X , we can reject X => Y hypothesis based on inverse causality.

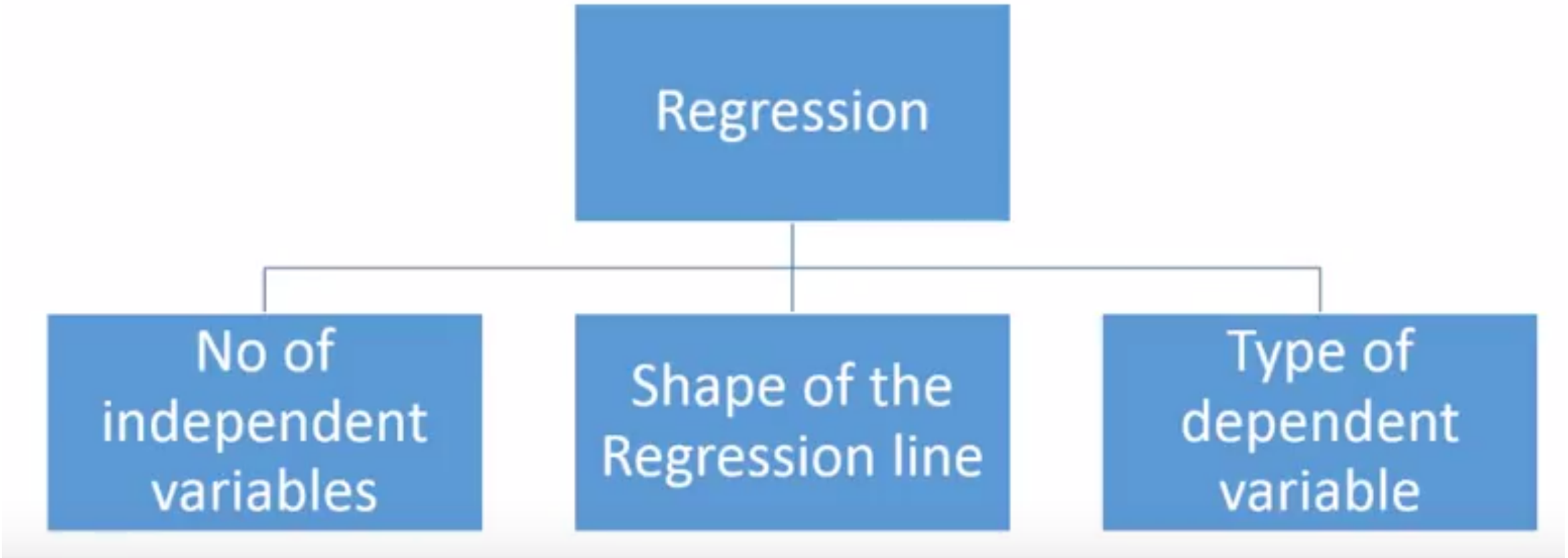
3. **Mutual independence :** Sometimes X and Y might just be correlated and nothing else. In such cases we reject hypothesis based on mutual independence.

How can we conclusively derive cause-effect relationship?

1. Randomized Experimental data : An experiment is often defined as random assignment of observational units to different conditions, and conditions differ by the treatment of observational units. Treatment is a generic term, which translates most easily in medical applications (e.g. patients are treated differently under different conditions), but it applies to other areas as well.

2. Observational data : If we do not have the luxury to do a randomized experiment, we are forced to work on existing data sources. These events have already happened without any control. Hence, the selection is not random.

Regression Analysis



Type of Regression

1. Linear Regression
2. Logistic Regression
3. Polynomial Regression
4. Stepwise Regression
5. Ridge Regression
6. Lasso Regression
7. ElasticNet Regression

Introduction to Regression Analysis

Regression analysis is used to:

1. Predict values of a dependent variable,Y, based on its relationship with values of at least one independent variable, X.

2. Explain the impact of changes in an independent variable on the dependent variable by estimating the **numerical value** of the relationship

Dependent variable: the variable we wish to explain

Independent variable: the variable used to explain the dependent variable

Simple Linear Regression Model

Only **one** independent variable (thus, simple), X

Relationship between X and Y is described by a linear function

Changes in Y are assumed to be caused by changes in X, that is,

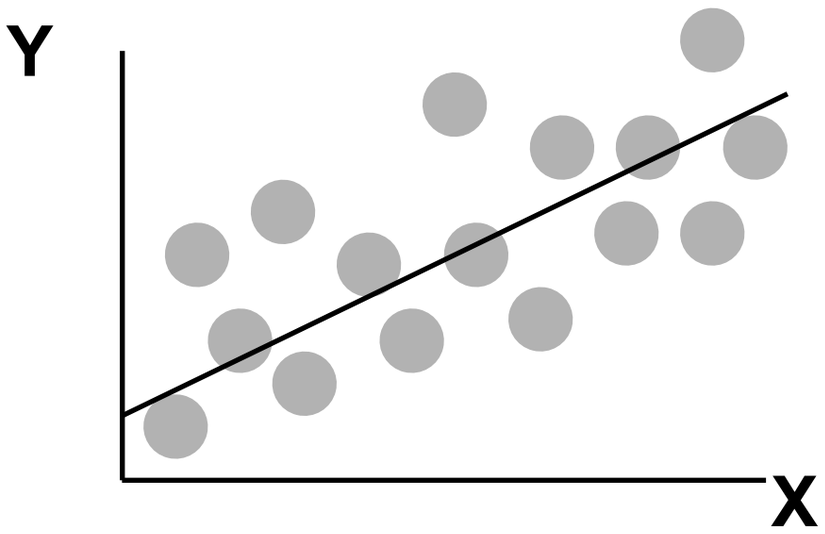
Change In **X** **Causes** Change in Y

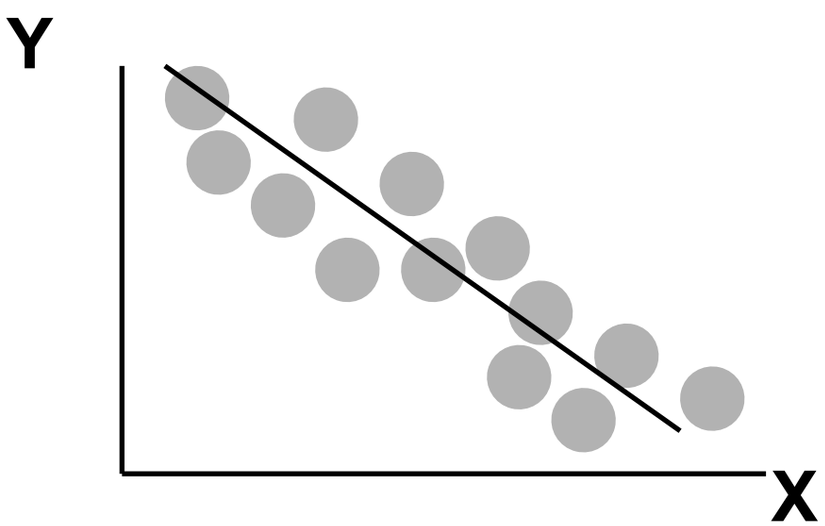
Important points before we start a regression analysis:

- The most important thing in deciding whether or not there is a relationship between X and Y is to have a systematic model that is based on **logical reasons**.
- Investigate the nature of the relationship between X and Y (use scatter diagram, covariance, correlation of coefficient)
- Remember that regression is not an exact or deterministic mathematical equation. It is a **behavioral relationship** that is subject to randomness.
- Remember that X is not the only thing that explains the behavior of Y. There are other factor that you may not have information about.
- All you are trying to do is to have an estimate of the relationship using the **best linear fit** possible

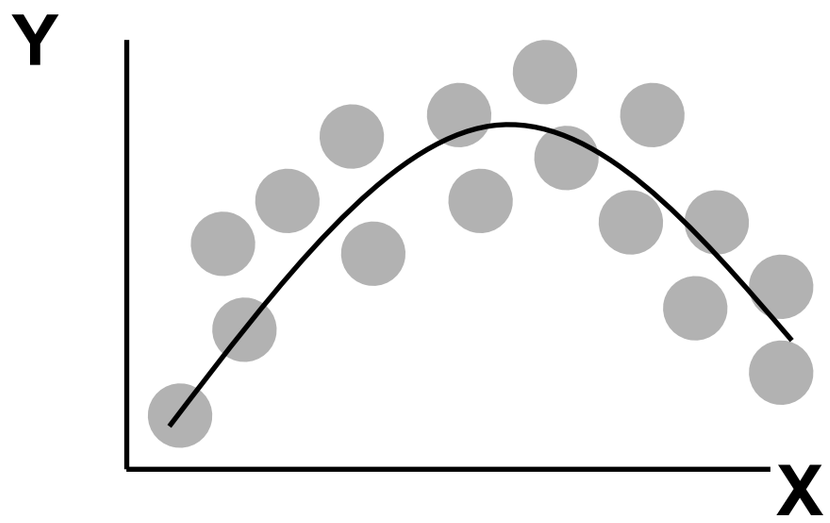
Types of Relationships

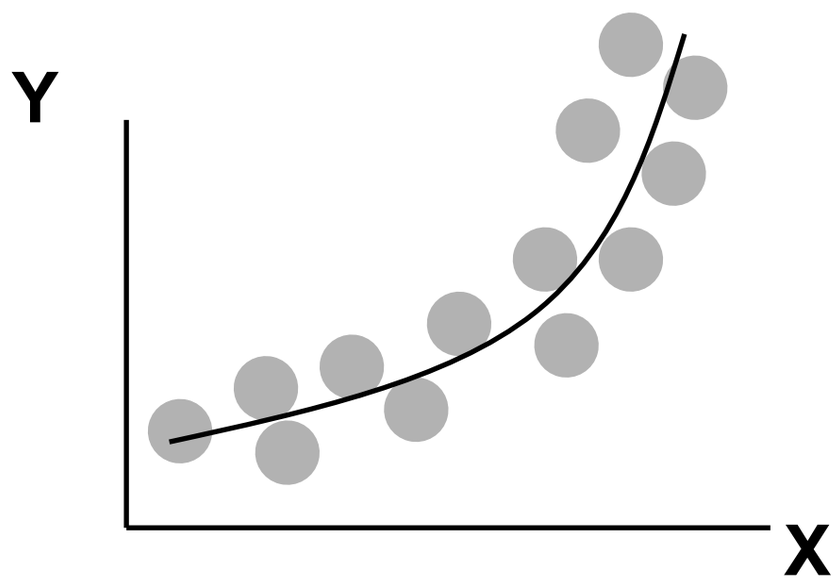
Linear relationships





Curvilinear relationships

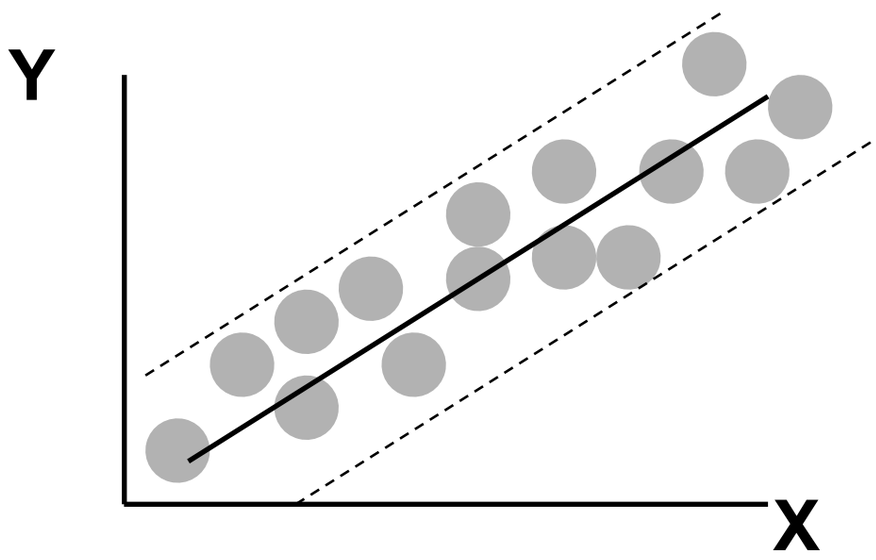


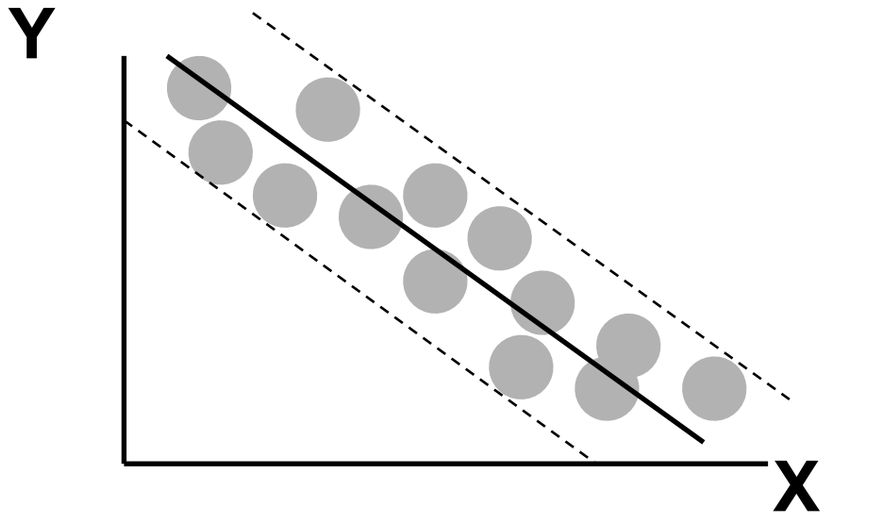


Types of Relationships

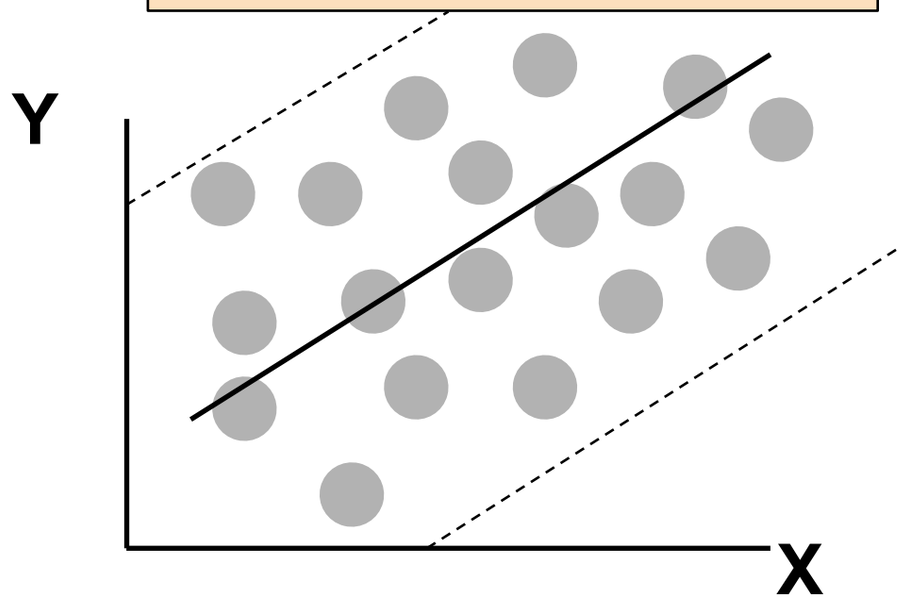
(continued)

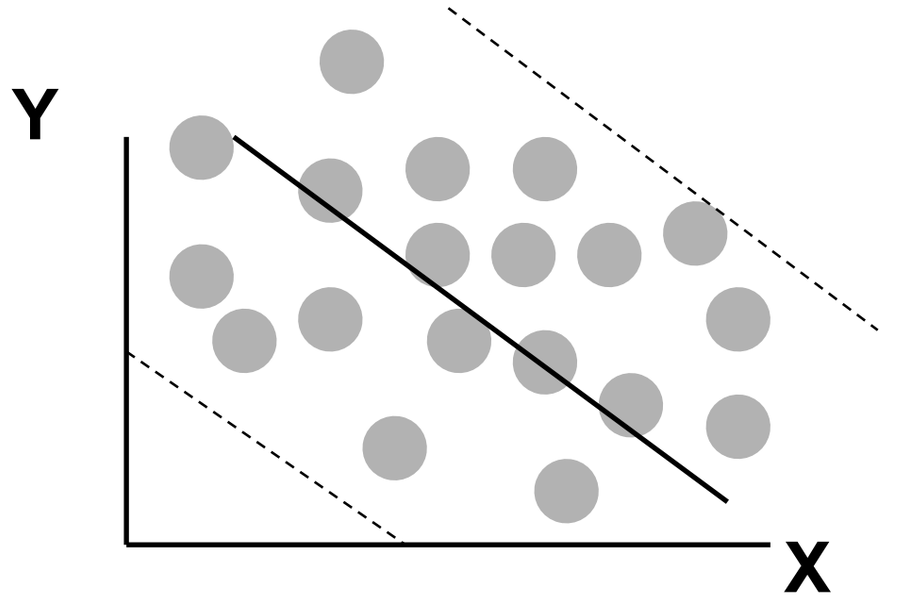
Strong relationships





Weak relationships

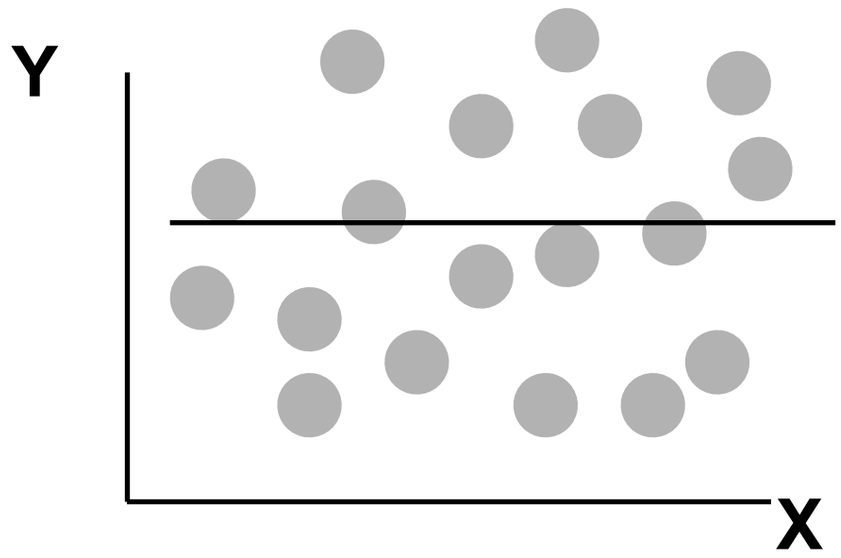


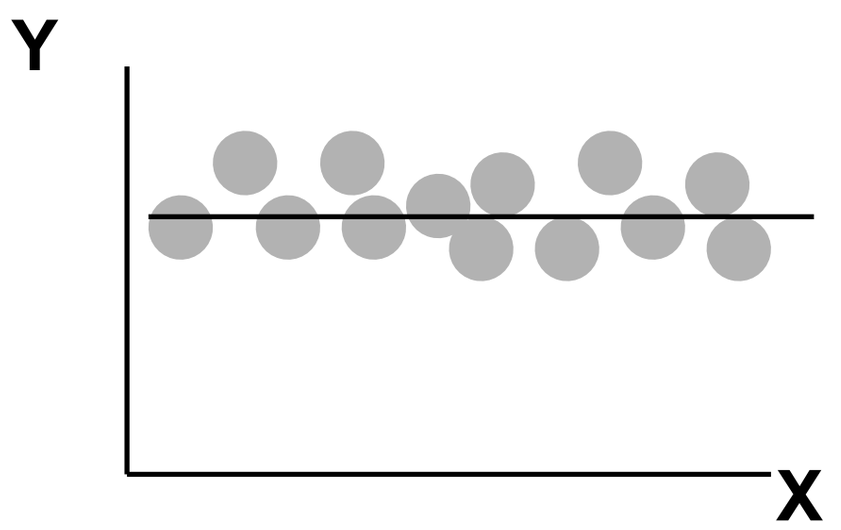


Types of Relationships

(continued)

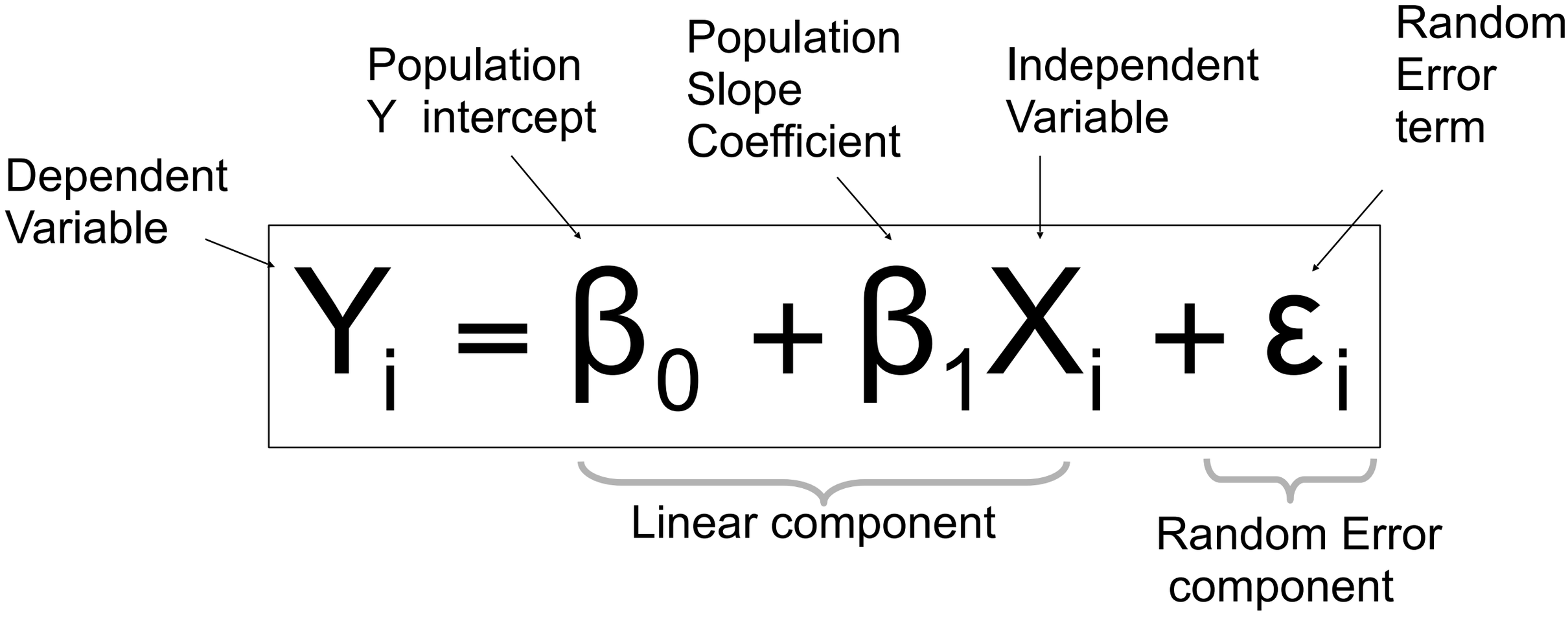
No relationship





Simple Linear Regression Conceptual Model

The population regression model: This is a conceptual model, a hypothesis, or a postulation



- The model to be estimated from sample data is:

$$Y_i = b_0 + b_1 X_i + e_i$$

Residual (random error from the sample)

- The actual estimated from the sample

Estimated (or predicted) Y value for observation i

Estimate of the regression intercept

Estimate of the regression slope

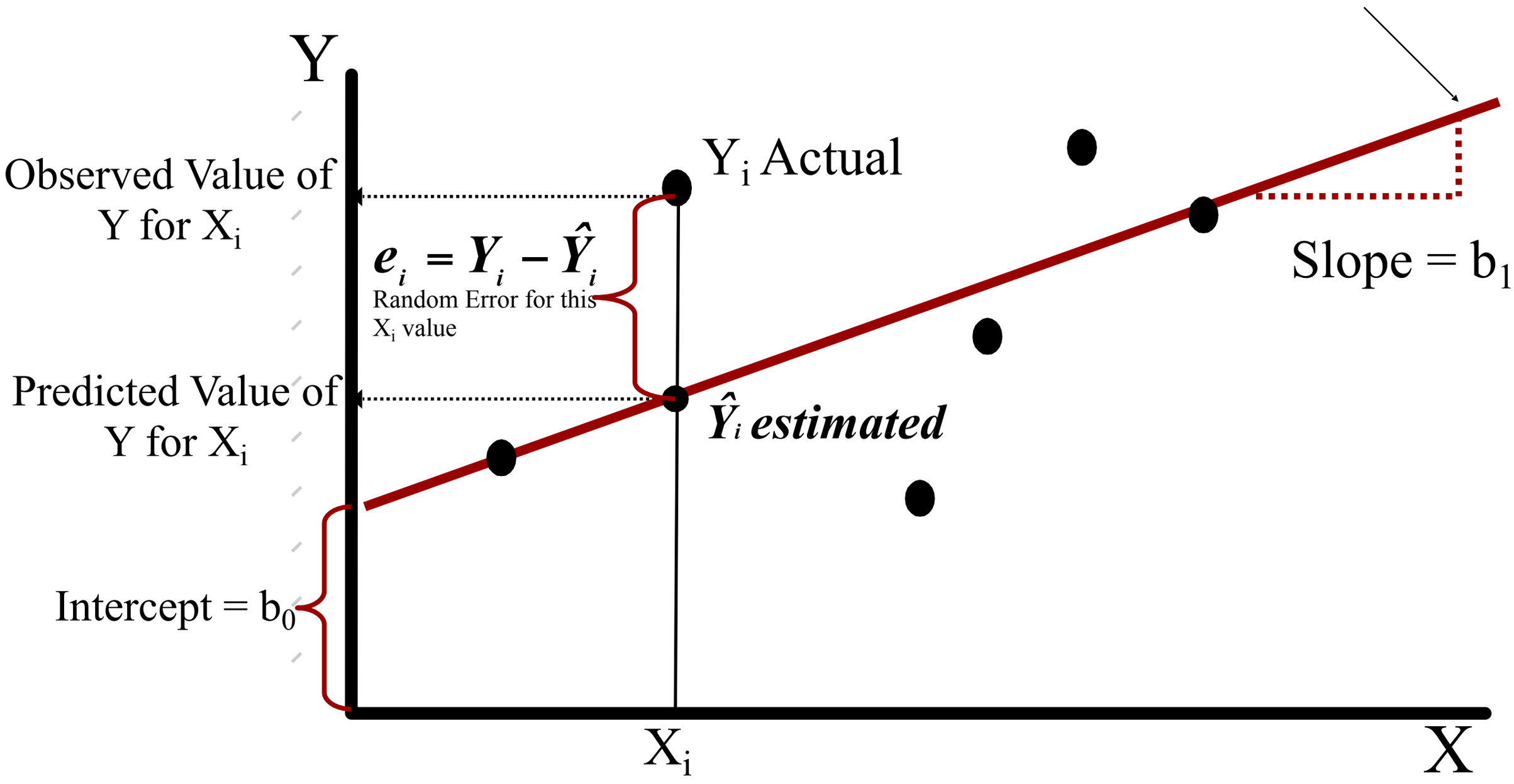
Value of X for observation i

$$\hat{Y}_i = b_0 + b_1 X_i$$

– Where $e_i = Y_i - \hat{Y}_i$

Simple Linear Regression Model

$$\hat{Y}_i = b_0 + b_1 X_i$$



Interpretation of the slope and the intercept

$$\beta_0 = E(Y \mid X = 0)\ ; \quad \beta_1 = \Delta\ \mathrm{E(Y|X)}/\Delta\ \mathrm{(X)};$$

b_0 is the estimated average value of Y when the value of X is b_0 zero

b_1 is the estimated change in the average value of Y as a result of a one-unit change in X

Units of measurement of X and Y are very important for the correct interpretation of the slop and the intercept

Example: $\widehat{App\ Val} = 165.03 + 6.93\ (Lot\ size)$

Predict the app. Value of a house with 10,000 s.f. lot size

$$\widehat{App\ Val} = 165.03 + 6.93\ (10) = \$234,330$$

How Good is this prediction?

How Good is the Model's prediction Power?

Total variation is made up of two parts:

$$SST = SSR + SSE$$

Total Sum of Squares	Regression Sum of Squares	Error Sum of Squares

$$SST = \sum (Y_i - \bar{Y})^2 \quad SSR = \sum (\hat{Y}_i - \bar{Y})^2 \quad SSE = \sum (Y_i - \hat{Y}_i)^2$$

where:

$\bar{Y}$ = Average value of the dependent variable

Y_i = Observed values of the dependent variable

$\hat{Y}_i$ = Predicted value of Y for the given X_i value

SST = total sum of squares

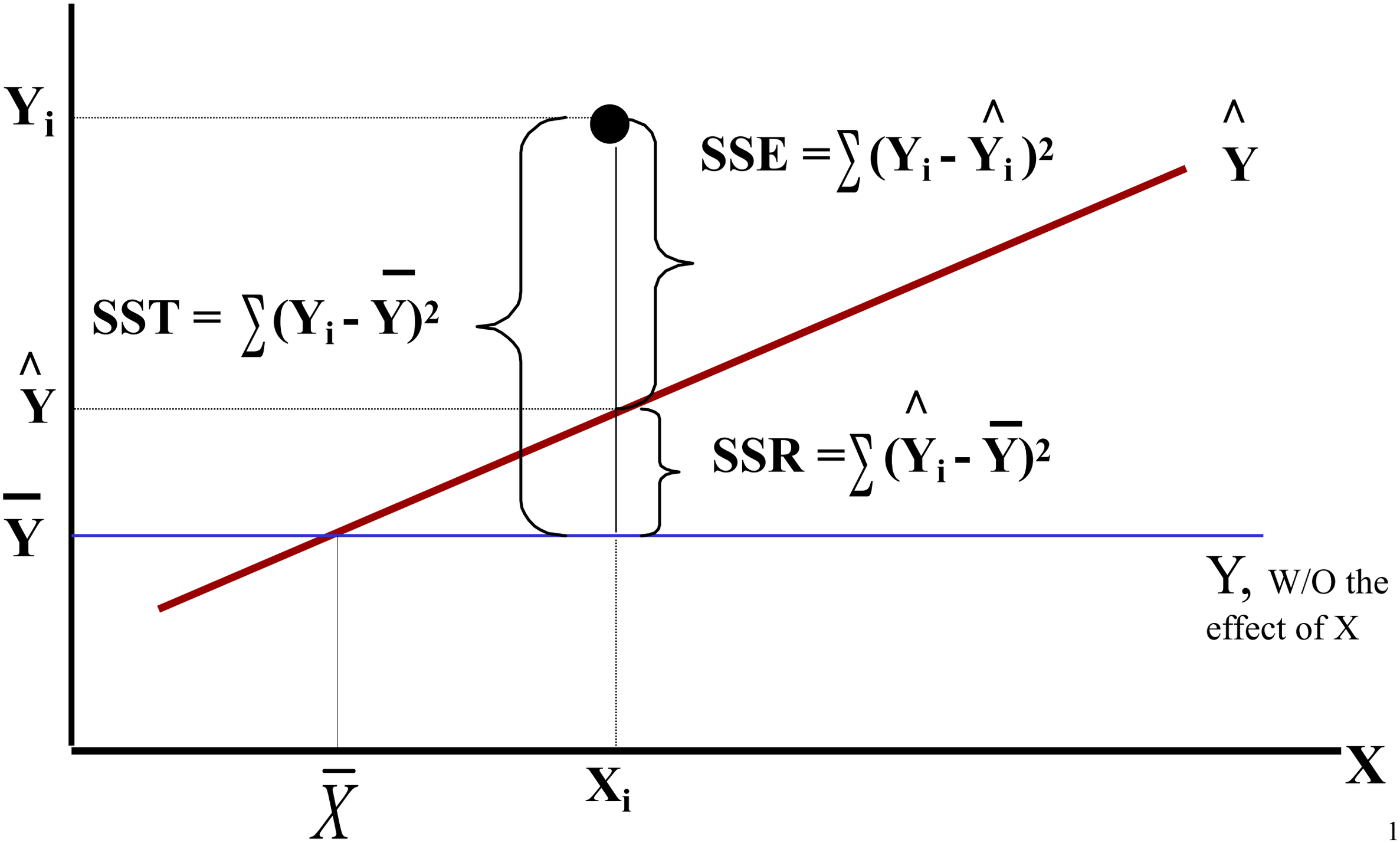
Measures total variation of the Y_i values around their mean

SSR = regression sum of squares (Explained)

Explained portion of total variation attributed to Y's relationship with X

SSE = error sum of squares (Unexplained)

Variation of Y values attributable to other factors than its relationship with X


Y_i
$\hat{Y}$
$\bar{Y}$
$\bar{X}$
X_i
X
$SST = \sum (Y_i - \bar{Y})^2$
$SSE = \sum (Y_i - \hat{Y}_i)^2$
$SSR = \sum (\hat{Y}_i - \bar{Y})^2$
$\hat{Y}$
Y, W/O the effect of X

How Good is the Model's prediction Power?

The coefficient of determination is the portion of the total variation in the dependent variable,Y, that is explained by variation in the independent variable, X

The coefficient of determination is also called r-squared and is denoted as r^2

$$r^2 = \frac{SSR}{SST} = \frac{regression\ sum\ of\ squares}{total\ sum\ of\ squares}$$

$$0 \leq r^2 \leq 1$$

Standard Error of Estimate

The standard deviation of the variation of observations around the regression line is estimated by

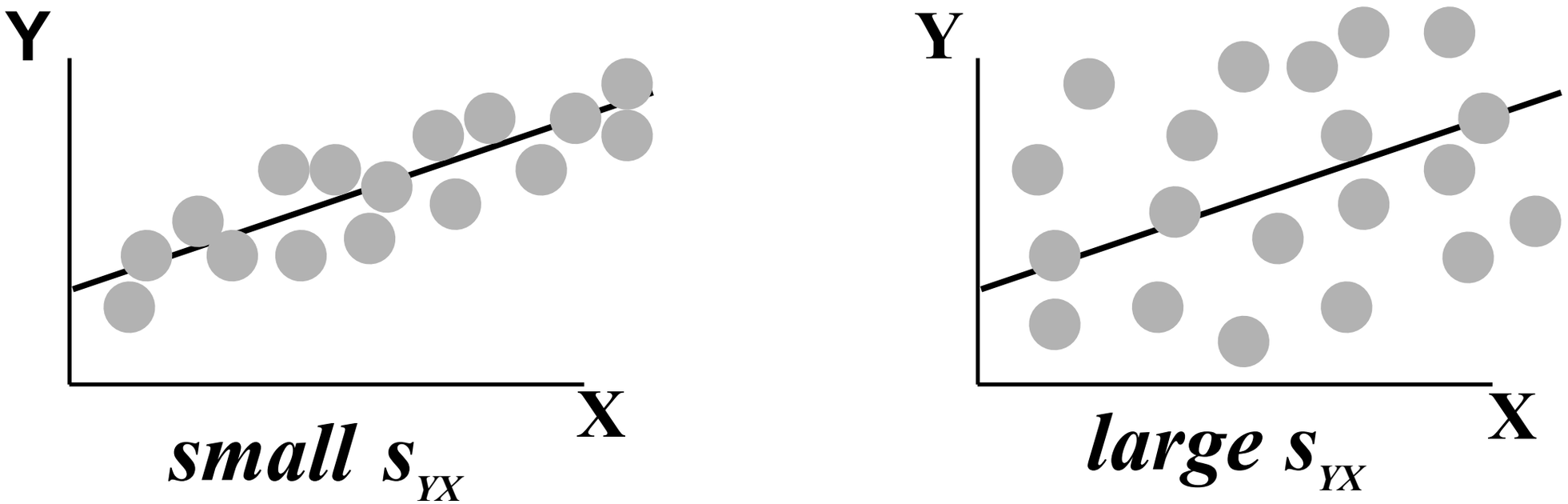
Where SSE = error sum of squares; n = sample size

$$S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}} = \sqrt{MSE}$$

The concept is the same as the standard deviation (average difference) around the mean of a univariate

Comparing Standard Errors

S_{YX} is a measure of the variation of observed Y values from the regression line



The magnitude of S_{YX} should always be judged relative to the size of the Y values in the sample data

i.e., S_{YX} = \$36.34K is moderately small relative to house prices in the \$200 - \$300K range (average 215K)

Assumptions of Regression

Normality of Error

Error values (ε) are normally distributed for any given value of X

Homoscedasticity

The probability distribution of the errors has constant variance

Independence of Errors

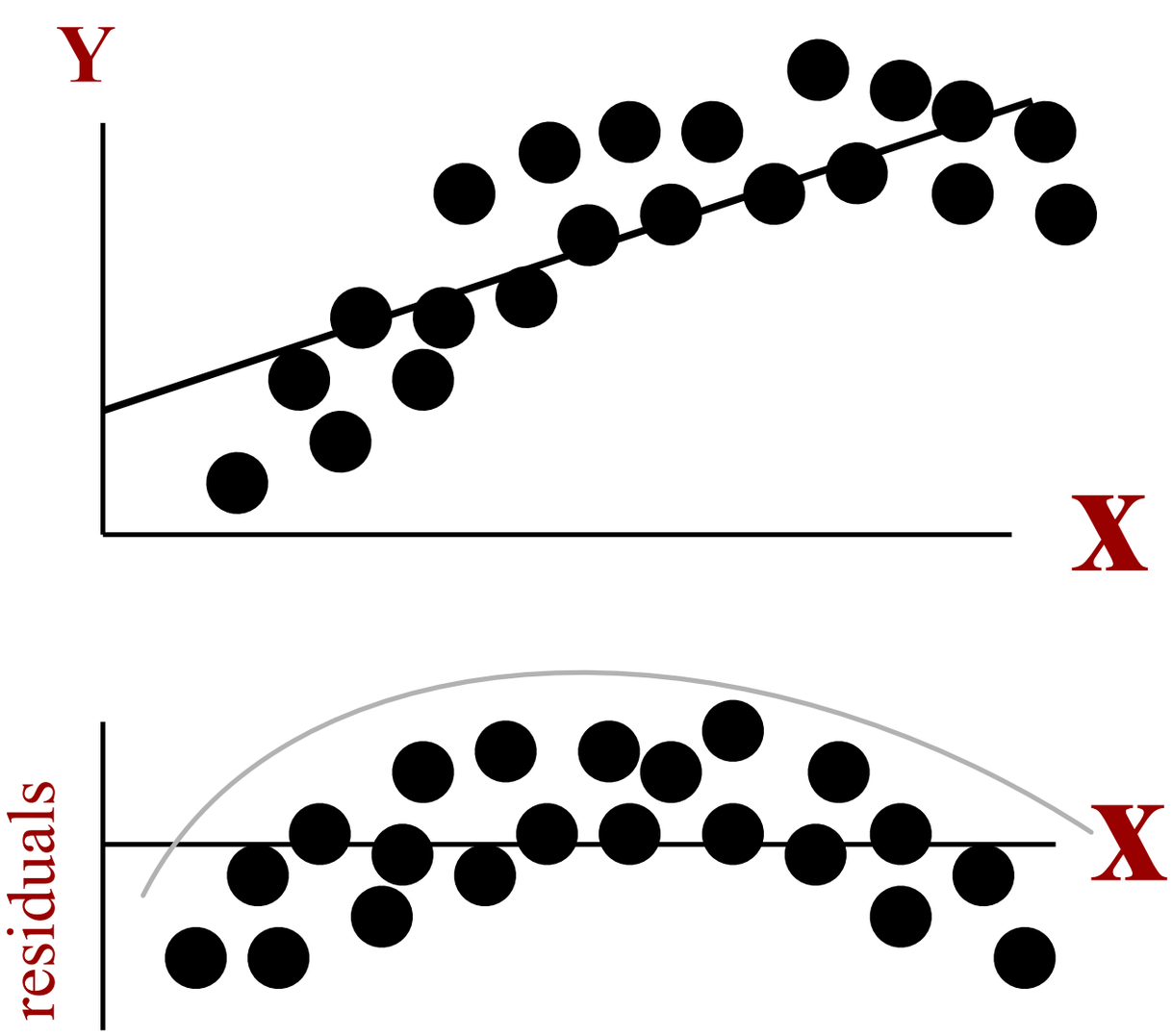
Error values are statistically independent

How to investigate the appropriateness of the fitted model

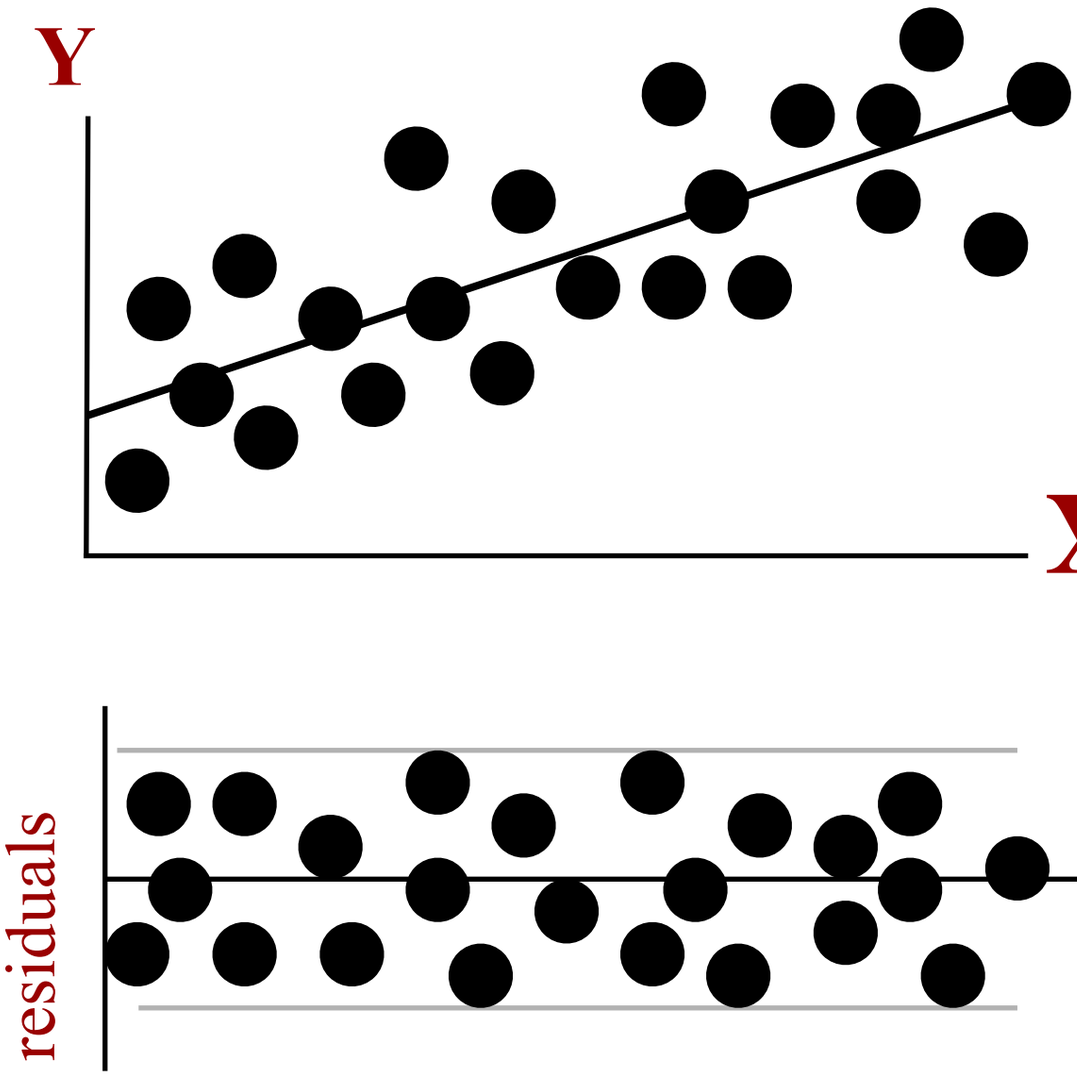
- The residual for observation i, e_i, is the difference between its observed and predicted value;
- Check the assumptions of regression by examining the residuals $$\boldsymbol{e}_i = \boldsymbol{Y}_i - \hat{\boldsymbol{Y}}_i$$
  - Examine for linearity assumption
  - Examine for constant variance for all levels of X (homoscedasticity)
  - Evaluate normal distribution assumption
  - Evaluate independence assumption
- Graphical Analysis of Residuals

  Can plot residuals vs. X

Residual Analysis for Linearity

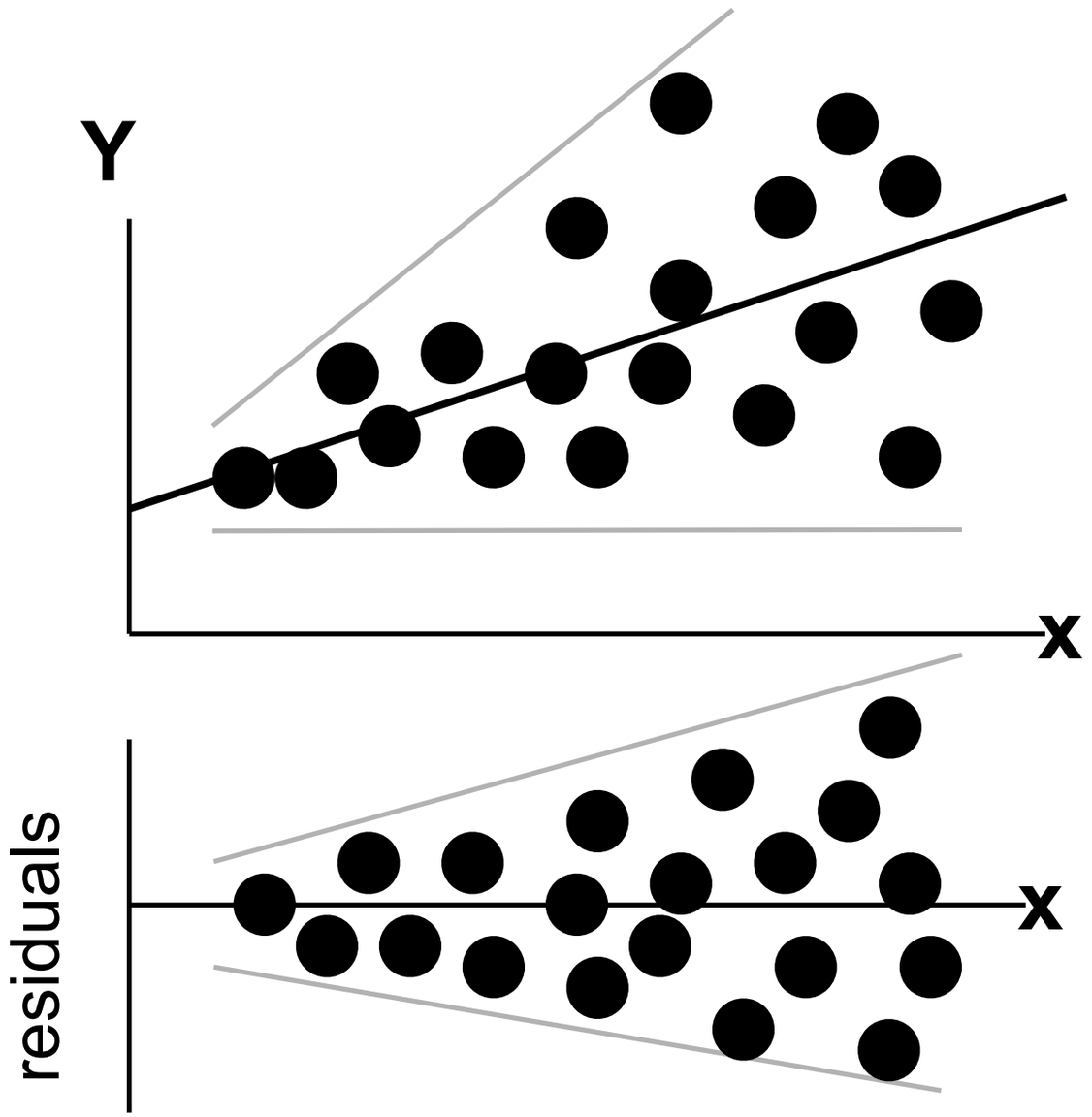


Not Linear

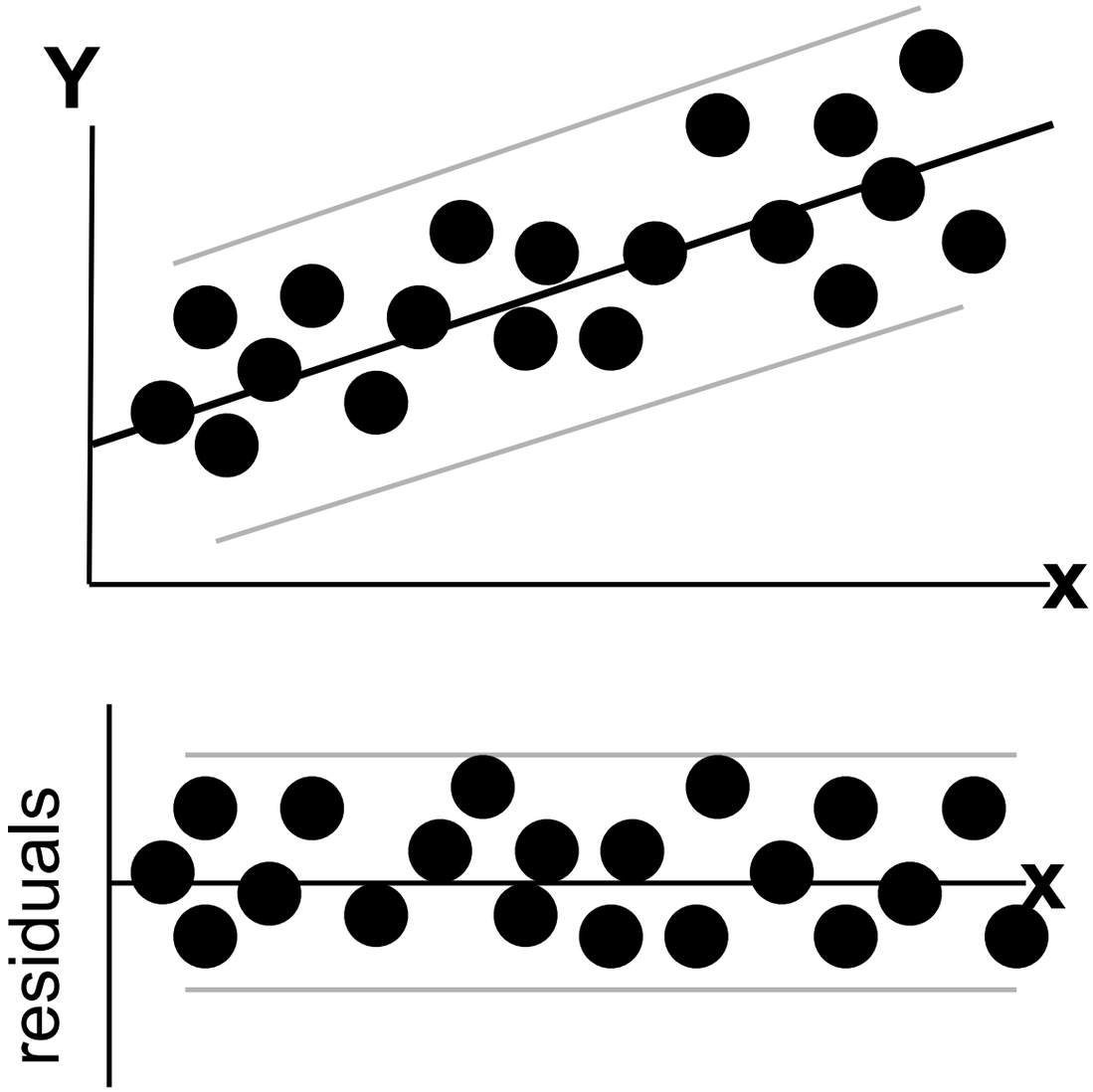


Linear

Residual Analysis for Homoscedasticity



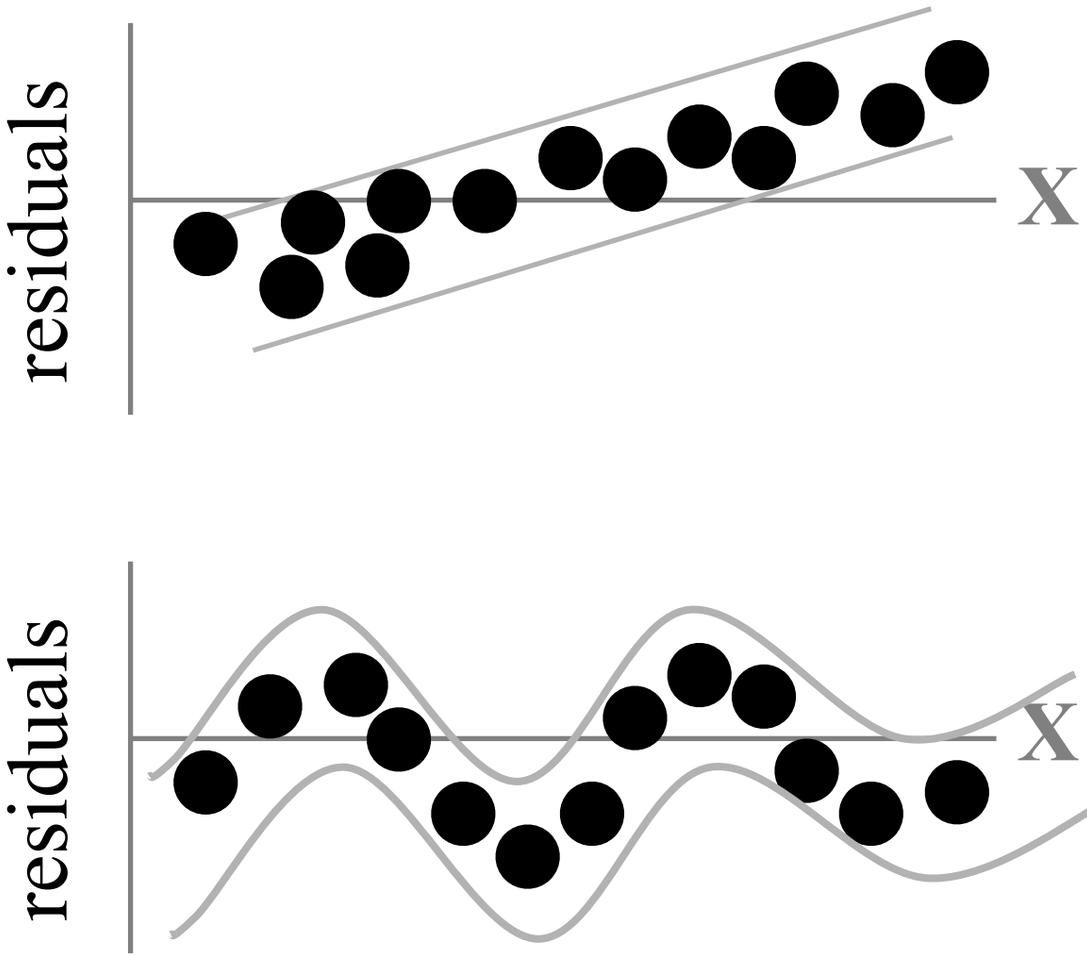
Non-constant variance



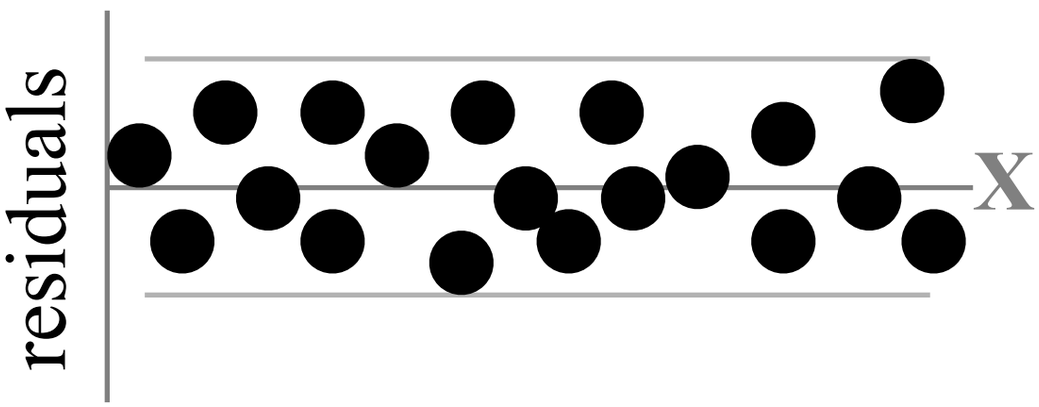
Constant variance

Residual Analysis for Independence

Not Independent



Independent



Regression

Linear Regression

Let's first import the data set and see what we're working with.

```
# Import data set
crime <- read.table("crime_simple.txt", sep = "\t", header = TRUE)
```

The variable names that this data set comes with are very confusing, and even misleading.

R: Crime rate: # of offenses reported to police per million population

Age: The number of males of age 14-24 per 1000 population

S: Indicator variable for Southern states (0 = No, 1 = Yes)

Ed: Mean # of years of schooling x 10 for persons of age 25 or older

Ex0: 1960 per capita expenditure on police by state and local government

Ex1: 1959 per capita expenditure on police by state and local government

LF: Labor force participation rate per 1000 civilian urban males age 14-24

M: The number of males per 1000 females

N: State population size in hundred thousands

NW: The number of non-whites per 1000 population

U1: Unemployment rate of urban males per 1000 of age 14-24

U2: Unemployment rate of urban males per 1000 of age 35-39

W: Median value of transferable goods and assets or family income in tens of $

X: The number of families per 1000 earning below 1/2 the median income

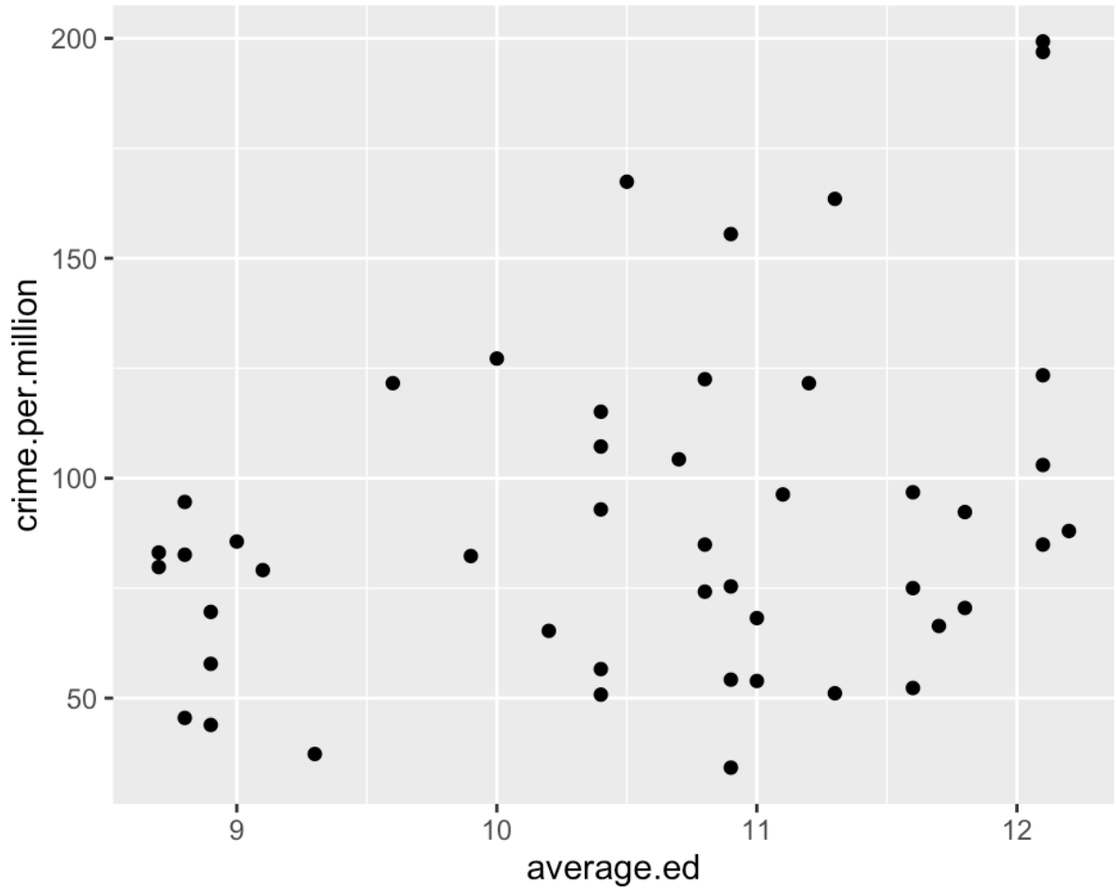
We really need to give these variables better names

```
# Assign more meaningful variable names
colnames(crime) <- c("crime.per.million", "young.males", "is.south", "average.ed",
                     "exp.per.cap.1960", "exp.per.cap.1959", "labour.part",
                     "male.per.fem", "population", "nonwhite",
                     "unemp.youth", "unemp.adult", "median.assets", "num.low.salary")

# Convert is.south to a factor
# Divide average.ed by 10 so that the variable is actually average education
# Convert median assets to 1000's of dollars instead of 10's
crime <- transform(crime, is.south = as.factor(is.south),
                          average.ed = average.ed / 10,
                          median.assets = median.assets / 100)

# print summary of the data
summary(crime)
```

```
##  crime.per.million  young.males    is.south   average.ed
##  Min.   : 34.20    Min.   :119.0   0:31     Min.   : 8.70
##  1st Qu.: 65.85    1st Qu.:130.0   1:16     1st Qu.: 9.75
##  Median : 83.10    Median :136.0            Median :10.80
##  Mean   : 90.51    Mean   :138.6            Mean   :10.56
##  3rd Qu.:105.75    3rd Qu.:146.0            3rd Qu.:11.45
##  Max.   :199.30    Max.   :177.0            Max.   :12.20
##  exp.per.cap.1960 exp.per.cap.1959  labour.part     male.per.fem
##  Min.   : 45.0    Min.   : 41.00   Min.   :480.0   Min.   : 934.0
##  1st Qu.: 62.5    1st Qu.: 58.50   1st Qu.:530.5   1st Qu.: 964.5
##  Median : 78.0    Median : 73.00   Median :560.0   Median : 977.0
##  Mean   : 85.0    Mean   : 80.23   Mean   :561.2   Mean   : 983.0
##  3rd Qu.:104.5    3rd Qu.: 97.00   3rd Qu.:593.0   3rd Qu.: 992.0
##  Max.   :166.0    Max.   :157.00   Max.   :641.0   Max.   :1071.0
##    population        nonwhite      unemp.youth      unemp.adult
##  Min.   :  3.00   Min.   :  2.0   Min.   : 70.00   Min.   :20.00
##  1st Qu.: 10.00   1st Qu.: 24.0   1st Qu.: 80.50   1st Qu.:27.50
##  Median : 25.00   Median : 76.0   Median : 92.00   Median :34.00
##  Mean   : 36.62   Mean   :101.1   Mean   : 95.47   Mean   :33.98
##  3rd Qu.: 41.50   3rd Qu.:132.5   3rd Qu.:104.00   3rd Qu.:38.50
##  Max.   :168.00   Max.   :423.0   Max.   :142.00   Max.   :58.00
##  median.assets   num.low.salary
##  Min.   :2.880   Min.   :126.0
##  1st Qu.:4.595   1st Qu.:165.5
##  Median :5.370   Median :176.0
##  Mean   :5.254   Mean   :194.0
##  3rd Qu.:5.915   3rd Qu.:227.5
##  Max.   :6.890   Max.   :276.0
```

First step: some plotting and summary statistics

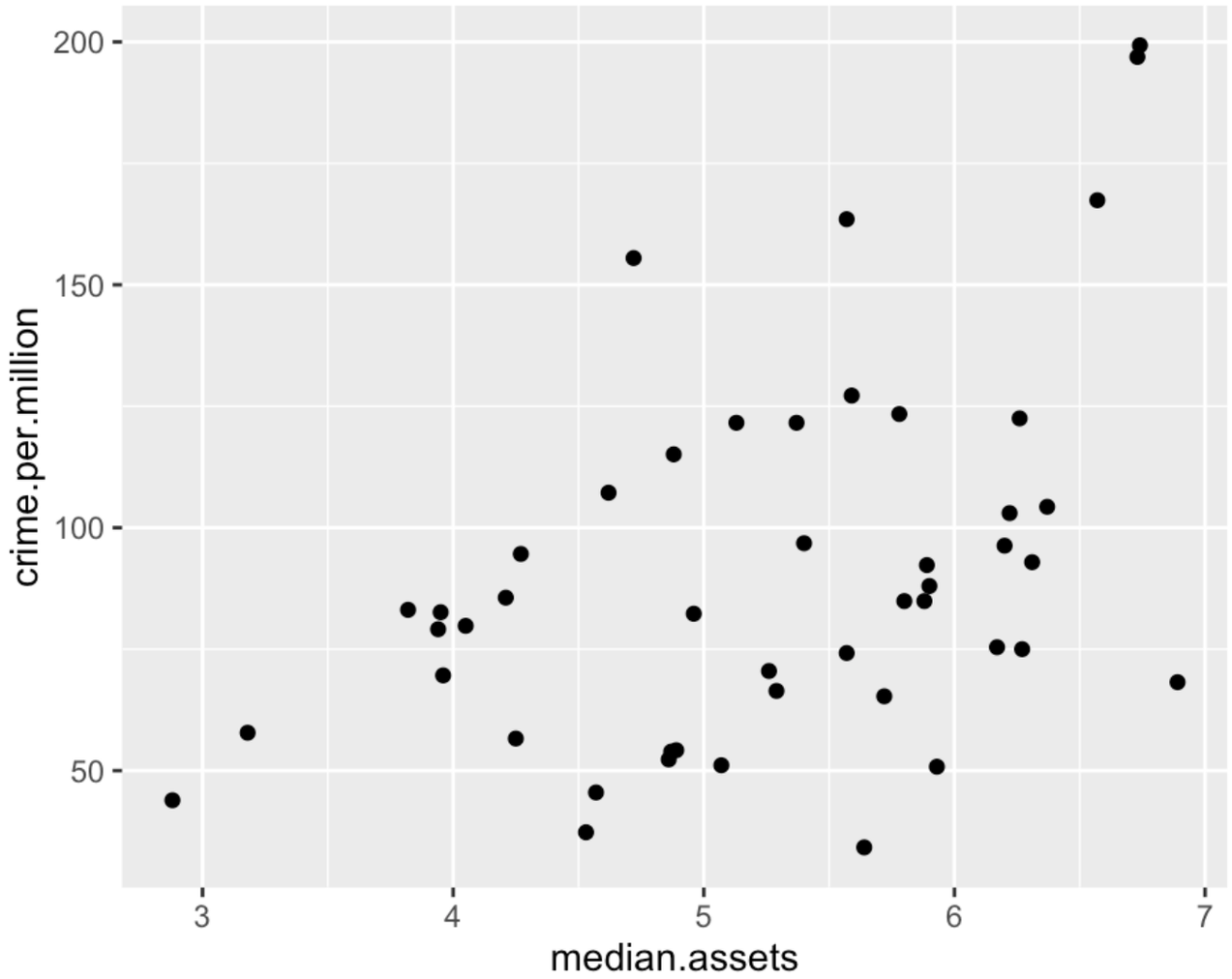
```
# Scatter plot of outcome (crime.per.million) against average.ed
qplot(average.ed, crime.per.million, data = crime)
```




```
# correlation between education and crime
with(crime, cor(average.ed, crime.per.million))
```

```
## [1] 0.3228349
```

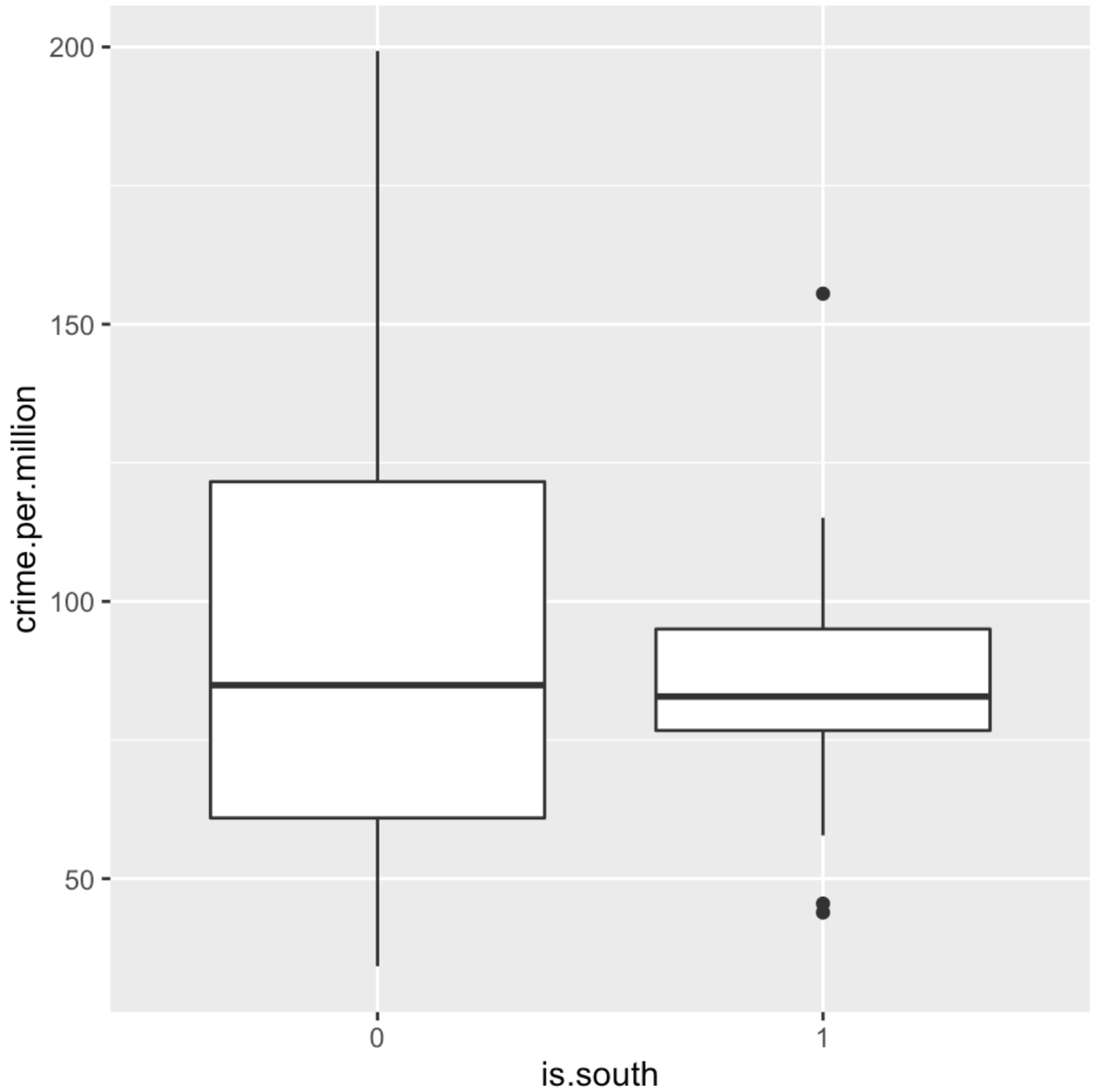
```
# Scatter plot of outcome (crime.per.million) against median.assets
qplot(median.assets, crime.per.million, data = crime)
```




```
# correlation between education and crime
with(crime, cor(median.assets, crime.per.million))
```

```
## [1] 0.4413199
```

```
# Boxplots showing crime rate broken down by southern vs non-southern state
qplot(is.south, crime.per.million, geom = "boxplot", data = crime)
```



Constructing a regression model

```
crime.lm <- lm(crime.per.million ~ ., data = crime)
# Summary of the linear regression model
crime.lm
```

```
##
## Call:
## lm(formula = crime.per.million ~ ., data = crime)
##
## Coefficients:
##      (Intercept)       young.males         is.south1        average.ed
##       -6.918e+02         1.040e+00        -8.308e+00         1.802e+01
## exp.per.cap.1960  exp.per.cap.1959       labour.part      male.per.fem
##        1.608e+00        -6.673e-01        -4.103e-02         1.648e-01
##       population          nonwhite       unemp.youth       unemp.adult
##       -4.128e-02         7.175e-03        -6.017e-01         1.792e+00
##    median.assets    num.low.salary
##        1.374e+01         7.929e-01
```

```
summary(crime.lm)
```

```
##
## Call:
## lm(formula = crime.per.million ~ ., data = crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.884 -11.923  -1.135  13.495  50.560
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -6.918e+02  1.559e+02  -4.438 9.56e-05 ***
## young.males      1.040e+00  4.227e-01   2.460  0.01931 *
## is.south1       -8.308e+00  1.491e+01  -0.557  0.58117
## average.ed       1.802e+01  6.497e+00   2.773  0.00906 **
## exp.per.cap.1960  1.608e+00  1.059e+00   1.519  0.13836
## exp.per.cap.1959 -6.673e-01  1.149e+00  -0.581  0.56529
## labour.part      -4.103e-02  1.535e-01  -0.267  0.79087
## male.per.fem      1.648e-01  2.099e-01   0.785  0.43806
## population       -4.128e-02  1.295e-01  -0.319  0.75196
## nonwhite          7.175e-03  6.387e-02   0.112  0.91124
## unemp.youth      -6.017e-01  4.372e-01  -1.376  0.17798
## unemp.adult       1.792e+00  8.561e-01   2.093  0.04407 *
## median.assets     1.374e+01  1.058e+01   1.298  0.20332
## num.low.salary    7.929e-01  2.351e-01   3.373  0.00191 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.94 on 33 degrees of freedom
## Multiple R-squared:  0.7692, Adjusted R-squared:  0.6783
## F-statistic: 8.462 on 13 and 33 DF,  p-value: 3.686e-07
```

```
options(scipen=4)  # Set scipen = 0 to get back to default
```

```
summary(crime.lm)
```

```
##
## Call:
## lm(formula = crime.per.million ~ ., data = crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.884 -11.923  -1.135  13.495  50.560
##
## Coefficients:
##                    Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)      -691.837588  155.887918  -4.438 0.0000956 ***
## young.males         1.039810    0.422708   2.460   0.01931 *
## is.south1          -8.308313   14.911588  -0.557   0.58117
## average.ed         18.016011    6.496504   2.773   0.00906 **
## exp.per.cap.1960    1.607818    1.058667   1.519   0.13836
## exp.per.cap.1959   -0.667258    1.148773  -0.581   0.56529
## labour.part        -0.041031    0.153477  -0.267   0.79087
## male.per.fem        0.164795    0.209932   0.785   0.43806
## population         -0.041277    0.129516  -0.319   0.75196
## nonwhite            0.007175    0.063867   0.112   0.91124
## unemp.youth        -0.601675    0.437154  -1.376   0.17798
## unemp.adult         1.792263    0.856111   2.093   0.04407 *
## median.assets      13.735847   10.583028   1.298   0.20332
## num.low.salary      0.792933    0.235085   3.373   0.00191 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.94 on 33 degrees of freedom
## Multiple R-squared:  0.7692, Adjusted R-squared:  0.6783
## F-statistic: 8.462 on 13 and 33 DF,  p-value: 0.0000003686
```

Exploring the lm object

What kind of output do we get when we run a linear model (`lm`) in R?

```
# List all attributes of the linear model
attributes(crime.lm)
```

```
## $names
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "contrasts"     "xlevels"       "call"          "terms"
## [13] "model"
##
## $class
## [1] "lm"
```

```
# coefficients
crime.lm$coef
```

```
##      (Intercept)      young.males        is.south1       average.ed
##   -691.837587905      1.039809653     -8.308312889     18.016010601
## exp.per.cap.1960 exp.per.cap.1959      labour.part     male.per.fem
##      1.607818377     -0.667258285     -0.041031047      0.164794968
##       population         nonwhite      unemp.youth      unemp.adult
##     -0.041276887      0.007174688     -0.601675298      1.792262901
##    median.assets   num.low.salary
##     13.735847285      0.792932786
```

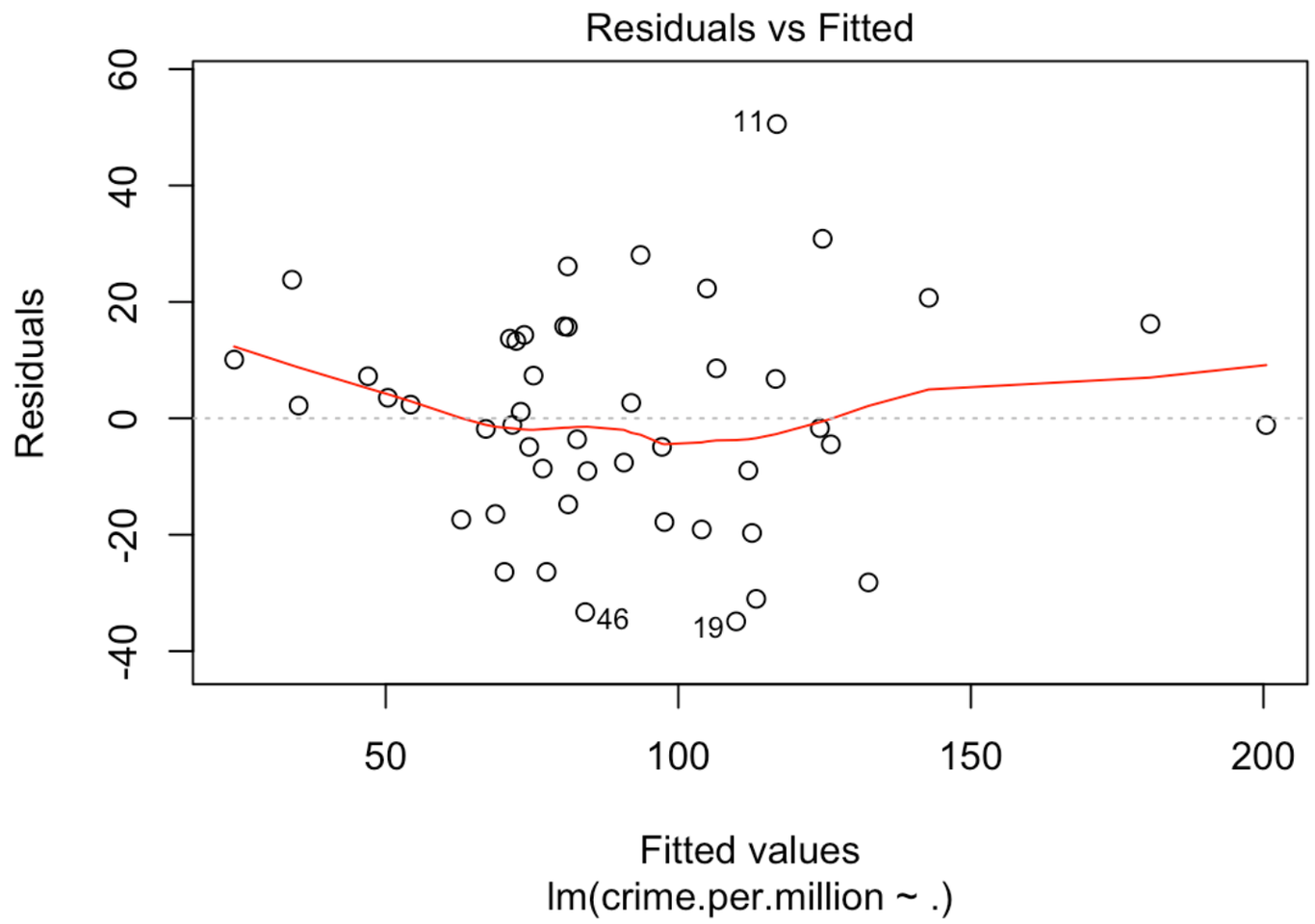
None of the attributes seem to give you p-values. Here's what you can do to get a table that allows you to extract p-values.

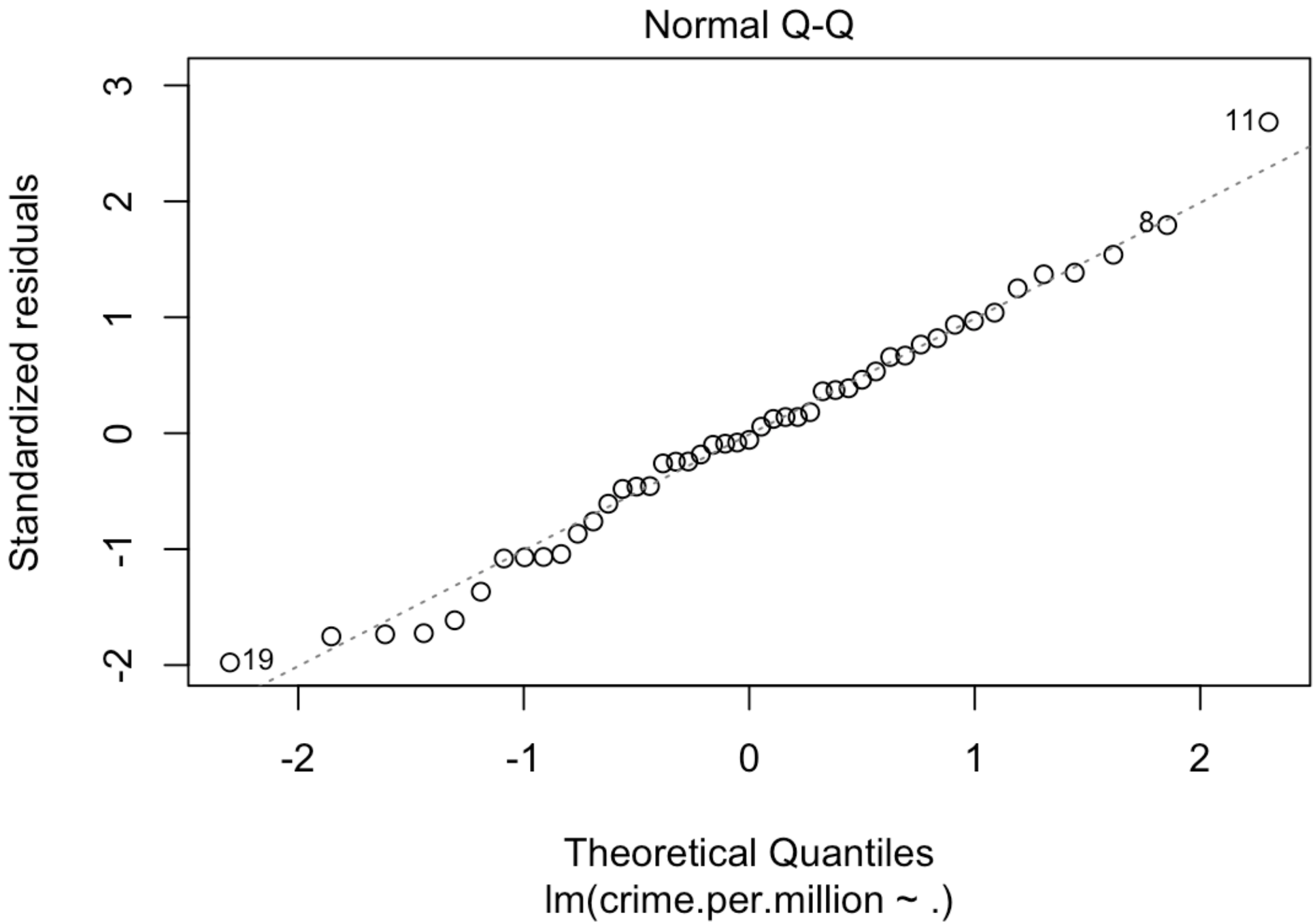
```
# Pull coefficients element from summary(lm) object
round(summary(crime.lm)$coef, 3)
```

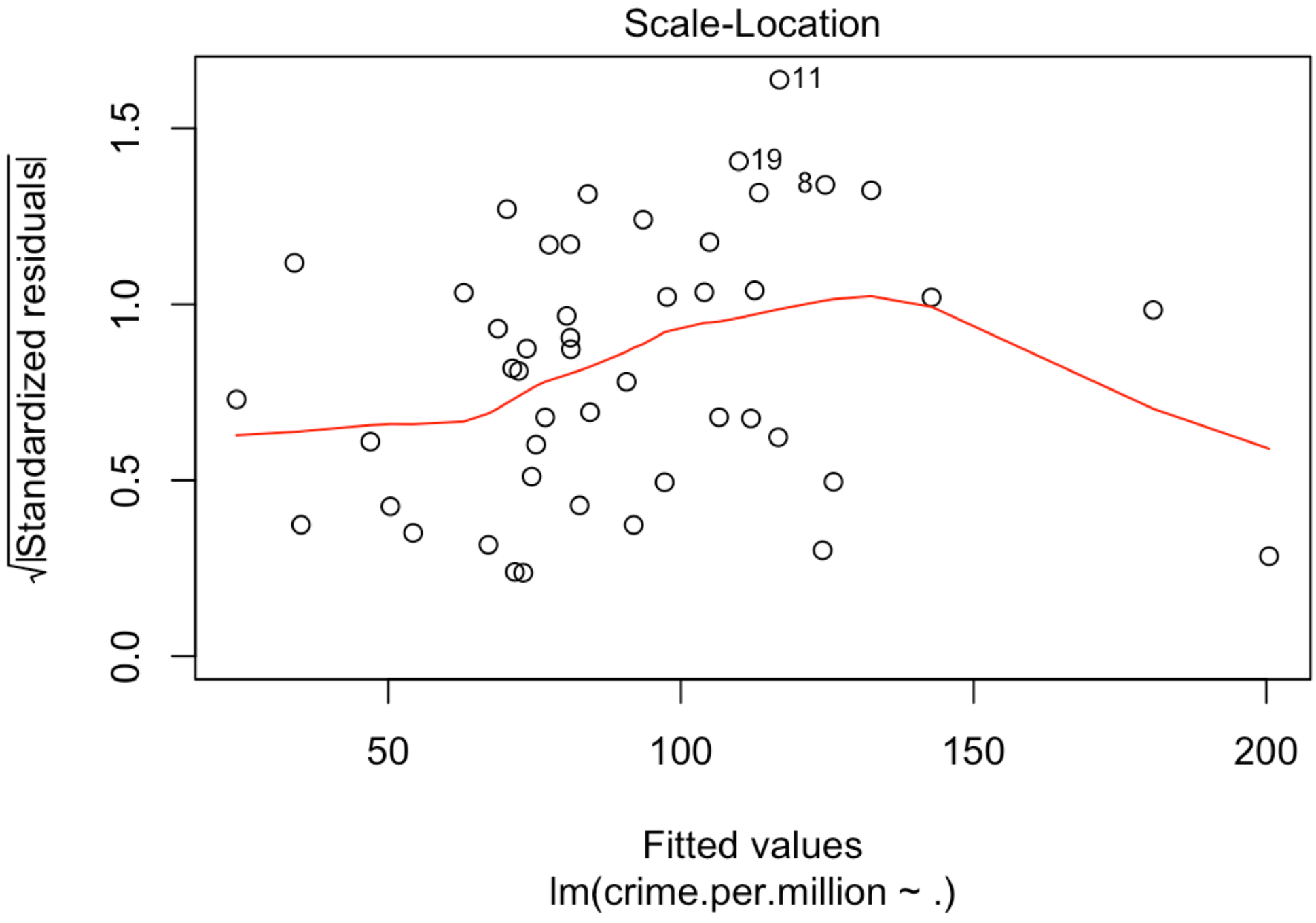
```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -691.838    155.888  -4.438    0.000
## young.males         1.040      0.423   2.460    0.019
## is.south1          -8.308     14.912  -0.557    0.581
## average.ed         18.016      6.497   2.773    0.009
## exp.per.cap.1960    1.608      1.059   1.519    0.138
## exp.per.cap.1959   -0.667      1.149  -0.581    0.565
## labour.part        -0.041      0.153  -0.267    0.791
## male.per.fem        0.165      0.210   0.785    0.438
## population         -0.041      0.130  -0.319    0.752
## nonwhite            0.007      0.064   0.112    0.911
## unemp.youth        -0.602      0.437  -1.376    0.178
## unemp.adult         1.792      0.856   2.093    0.044
## median.assets      13.736     10.583   1.298    0.203
## num.low.salary      0.793      0.235   3.373    0.002
```

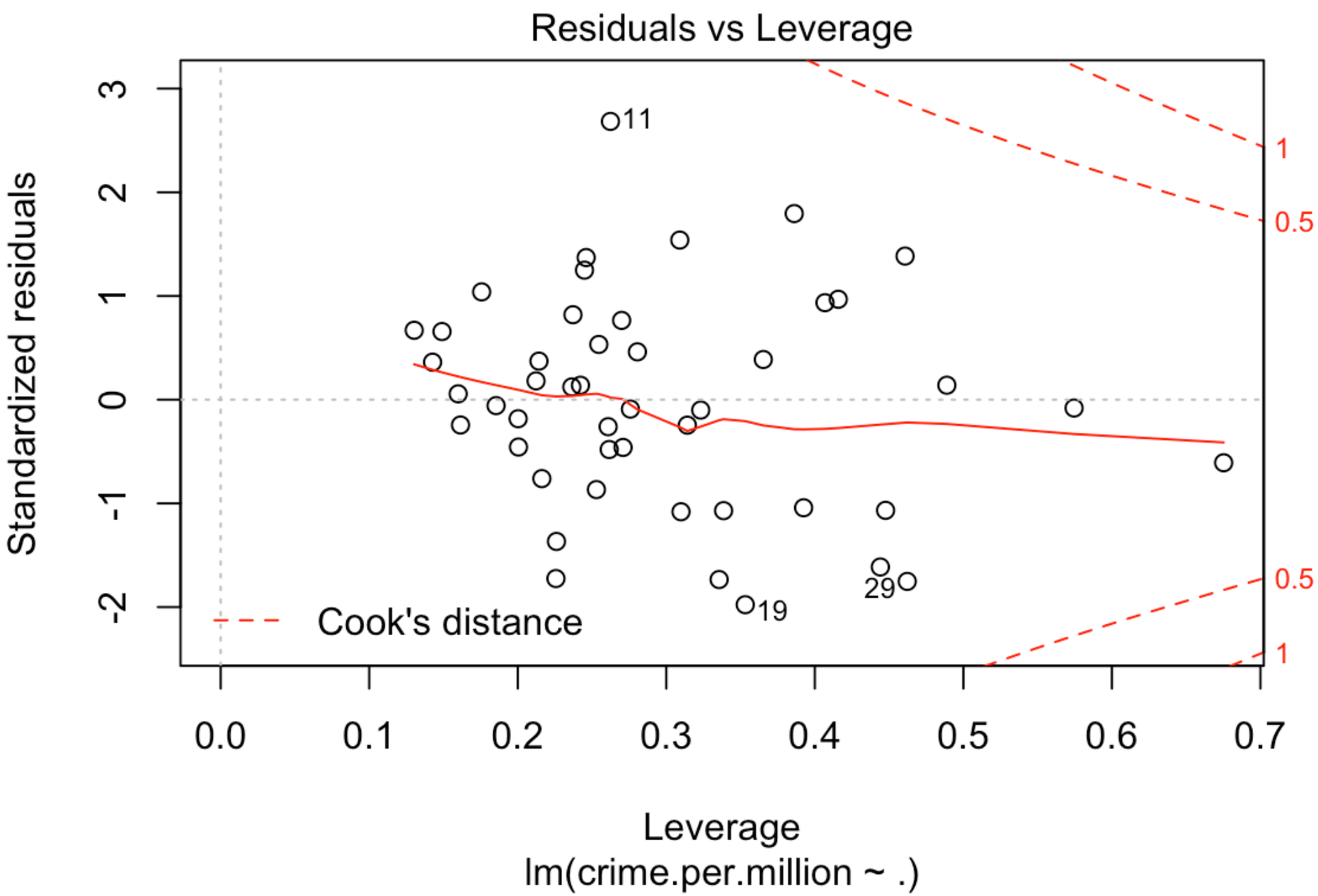
Plotting the lm object

```
plot(crime.lm)
```




Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
11
8
19
-2
-1
0
1
2
Theoretical Quantiles
lm(crime.per.million ~ .)


Scale-Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
11
19
8
50
100
150
200
Fitted values
lm(crime.per.million ~ .)


Residuals vs Leverage
Standardized residuals
Leverage
lm(crime.per.million ~ .)
Cook's distance
11
19
29
1
0.5
0.5
1
3
2
1
0
-1
-2
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

Residuals vs Fitted Plot

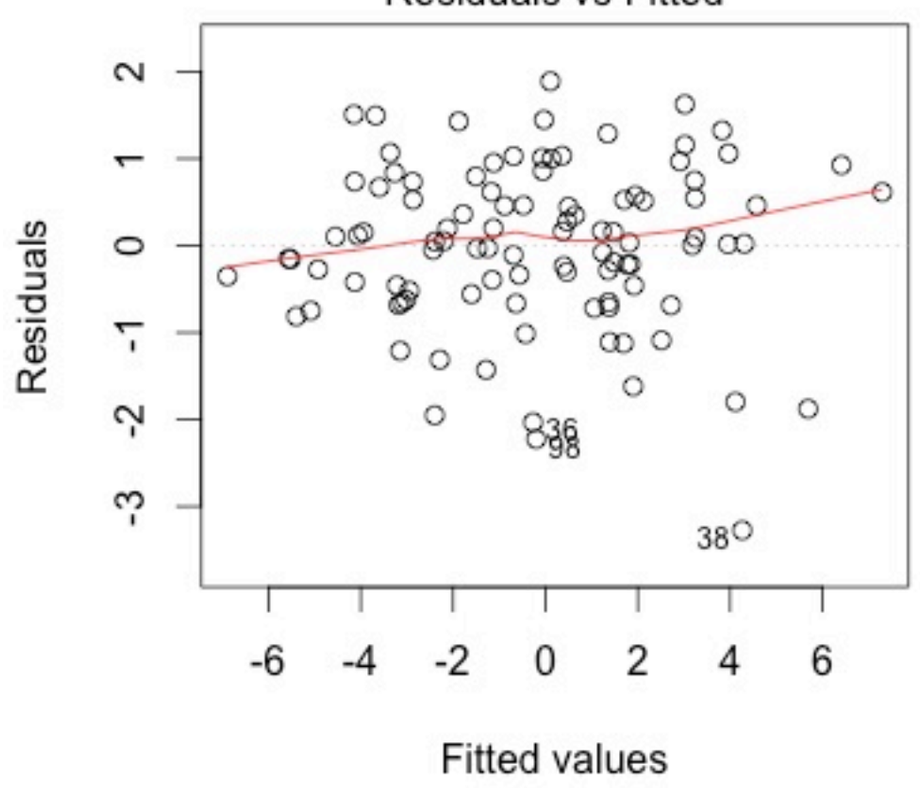
Residuals vs. Fitted When a linear model is appropriate, we expect

1 the residuals will have constant variance when plotted against fitted values; and

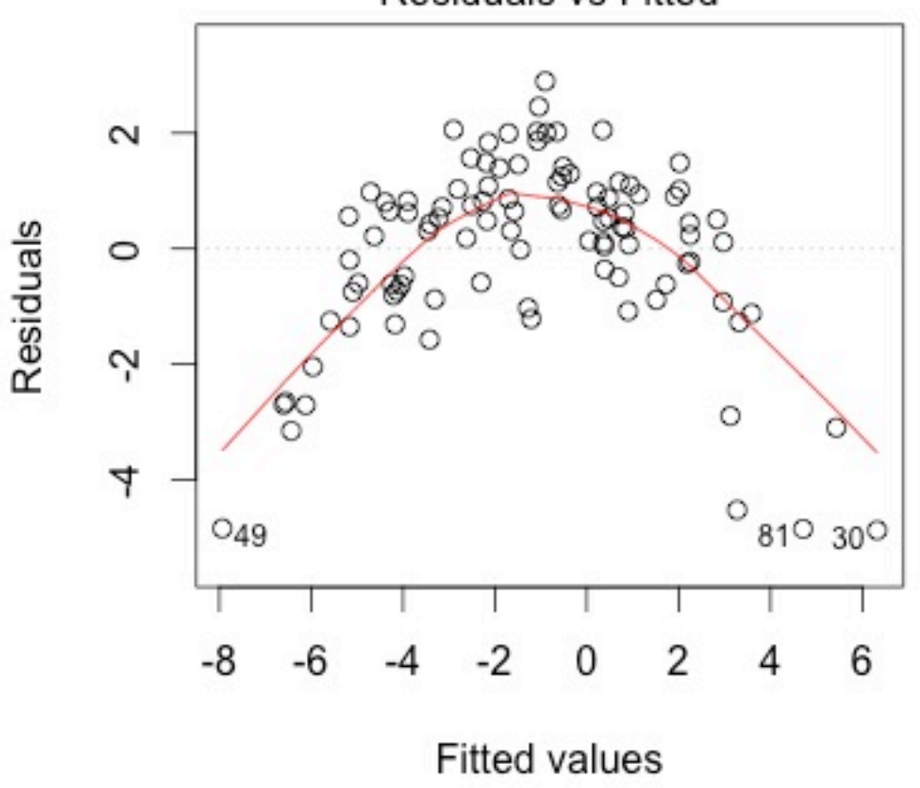
2 the residuals and fitted values will be uncorrelated.

If there are clear trends in the residual plot, or the plot looks like a funnel, these are clear indicators that the given linear model is inappropriate.

Case 1


Residuals vs Fitted
Residuals
Fitted values
-6
-4
-2
0
2
4
6
-3
-2
-1
0
1
2
36
98
38

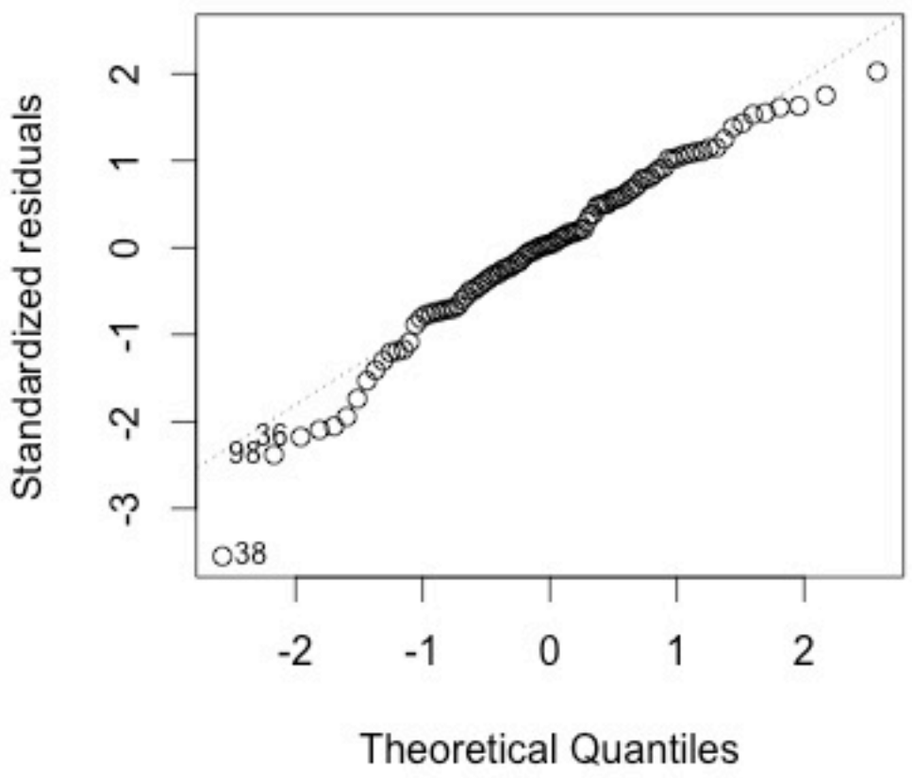
Case 2


Residuals vs Fitted
Residuals
Fitted values
-8
-6
-4
-2
0
2
4
6
-4
-2
0
2
49
81
30

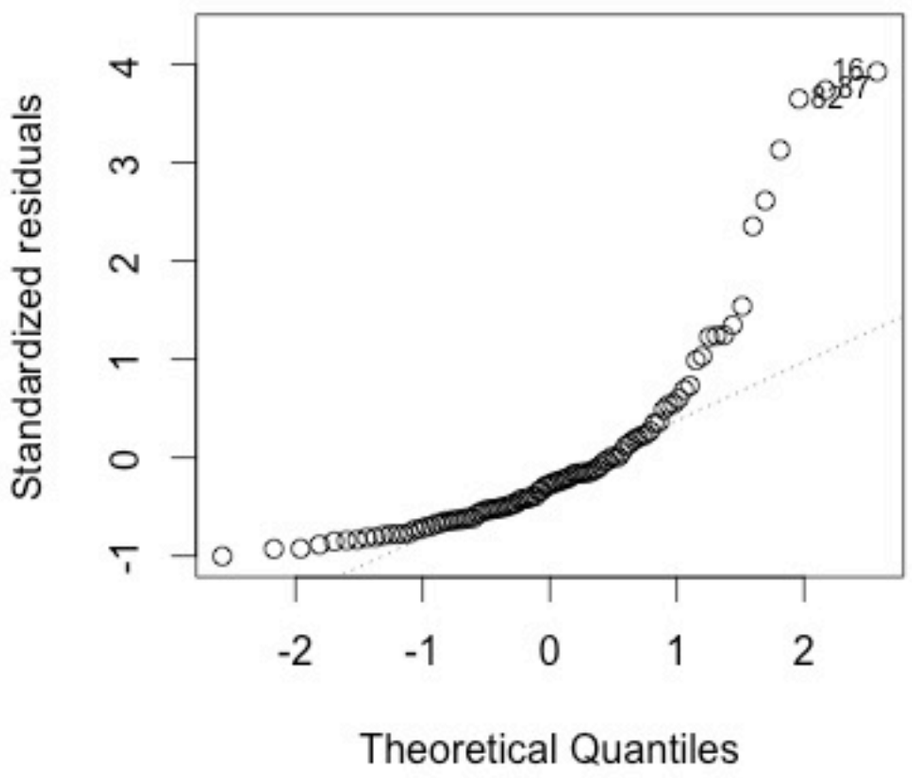
Normal QQ Plot

Normal QQ plot You can use a linear model for prediction even if the underlying normality assumptions don’t hold. However, in order for the p-values to be believable, the residuals from the regression must look approximately normally distributed.

Case 1


Normal Q-Q
Standardized residuals
Theoretical Quantiles
-3
-2
-1
0
1
2
38
98
36

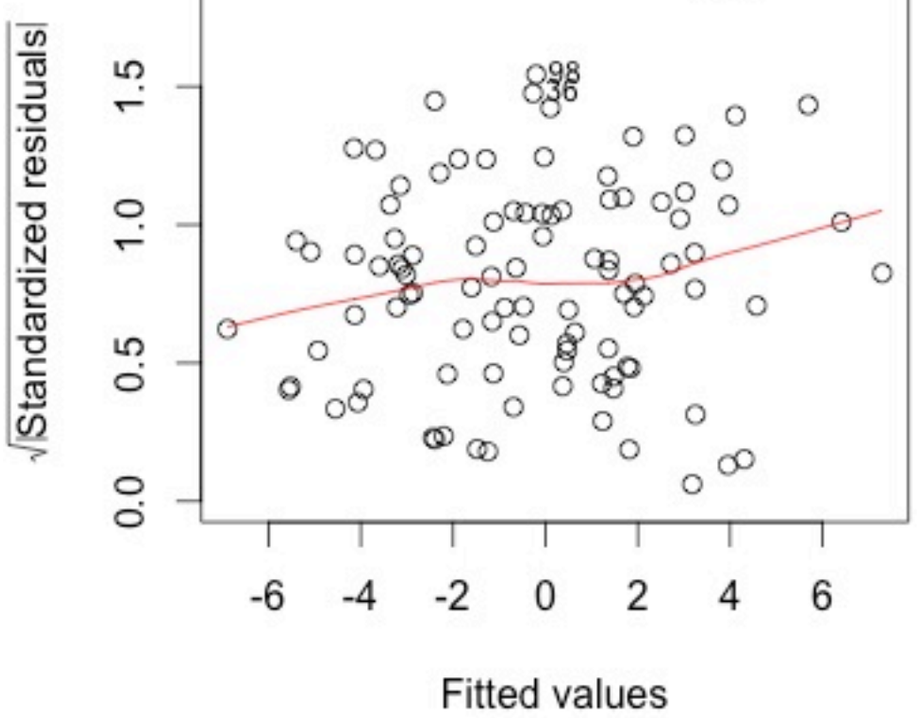
Case 2


Normal Q-Q
Standardized residuals
Theoretical Quantiles
-1
0
1
2
3
4
16
82
87

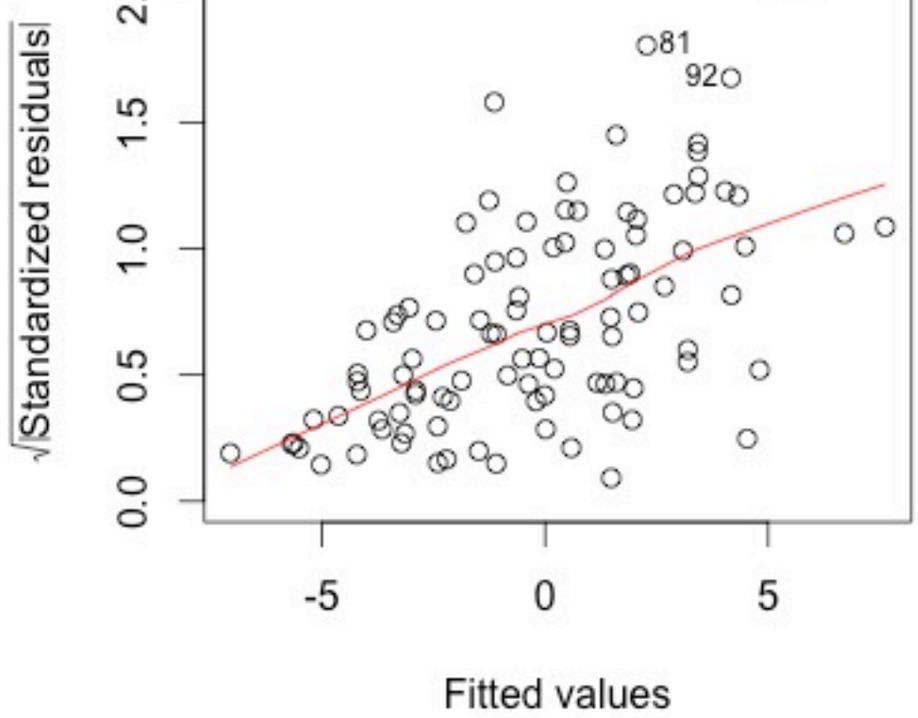
Scale-Location Plot

Scale-location plot This is another version of the residuals vs fitted plot. There should be no discernible trends in this plot.

Case 1

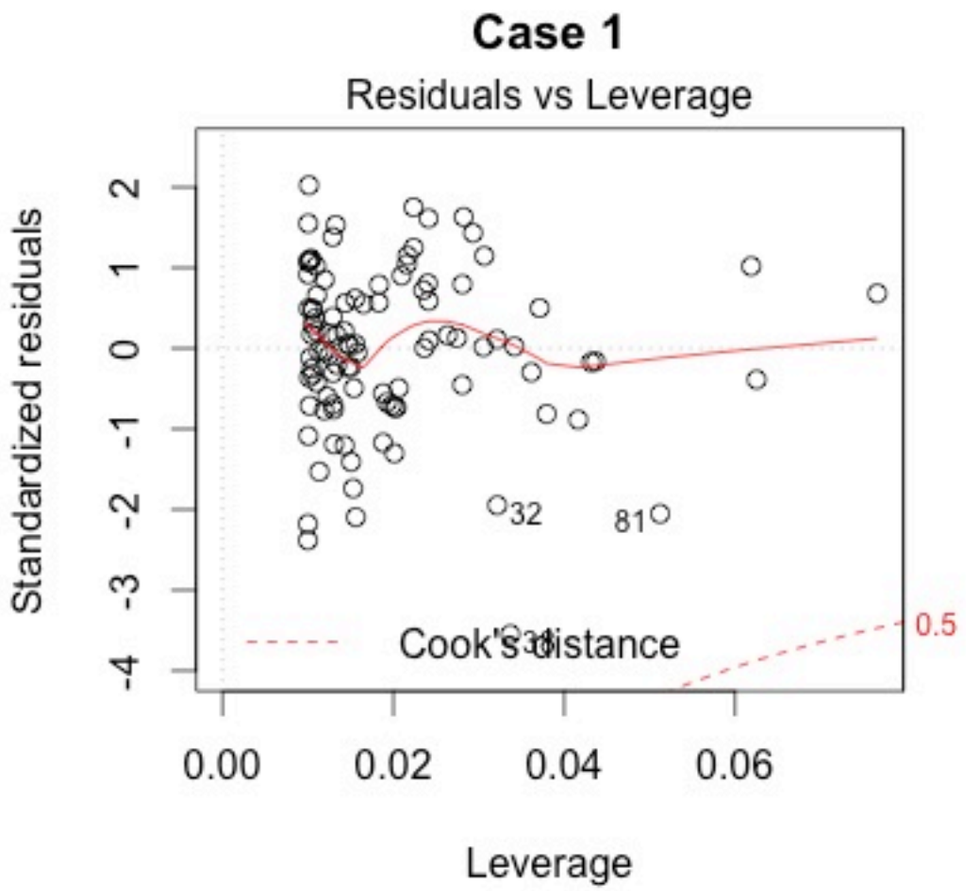

Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
-6
-4
-2
0
2
4
6
Fitted values
38
98
36

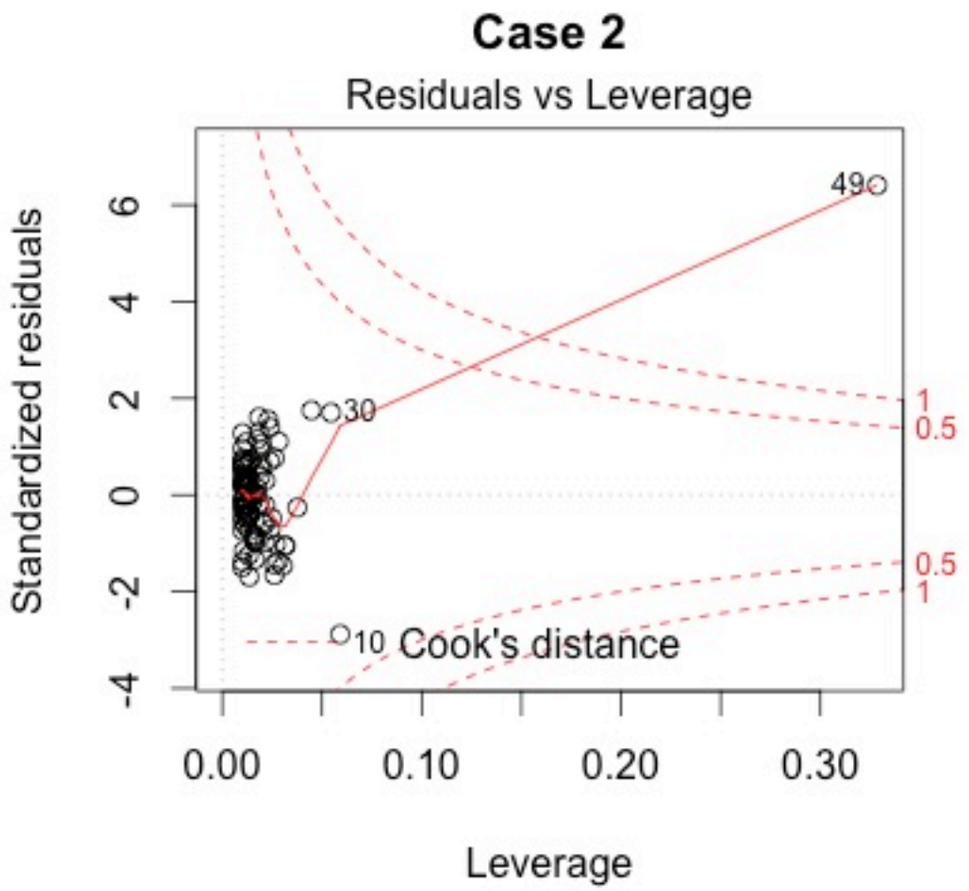
Case 2


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
-5
0
5
Fitted values
98
81
92

Residuals vs Leverage

Residuals vs Leverage. Leverage is a measure of how much an observation influenced the model fit. It’s a one-number summary of how different the model fit would be if the given observation was excluded, compared to the model fit where the observation is included. Points with high residual (poorly described by the model) and high leverage (high influence on model fit) are outliers. They’re skewing the model fit away from the rest of the data, and don’t really seem to fit with the rest of the data.


Case 1
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
0.5
32
81
0.00
0.02
0.04
0.06
-4
-3
-2
-1
0
1
2


Case 2
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
49
30
10
1
0.5
0.5
1
0.00
0.10
0.20
0.30
-4
-2
0
2
4
6

Workshop 3.4: Regression

We'll begin by loading some packages.

```
library(MASS)
library(plyr)
```

Interaction terms in regression

```
# Building up the familiar birthwt data...

# Rename the columns to have more descriptive names
colnames(birthwt) <- c("birthwt.below.2500", "mother.age", "mother.weight",
    "race", "mother.smokes", "previous.prem.labor", "hypertension", "uterine.irr",
    "physician.visits", "birthwt.grams")

# Transform variables to factors with descriptive levels
birthwt <- transform(birthwt,
            race = as.factor(mapvalues(race, c(1, 2, 3),
                                c("white","black", "other"))),
            mother.smokes = as.factor(mapvalues(mother.smokes,
                                c(0,1), c("no", "yes"))),
            hypertension = as.factor(mapvalues(hypertension,
                                c(0,1), c("no", "yes"))),
            uterine.irr = as.factor(mapvalues(uterine.irr,
                                c(0,1), c("no", "yes")))
            )
```

Workshop 3.4 : Regression

(a) Run a linear regression to better understand how birthweight varies with the mother’s age and smoking status (do not include interaction terms).

(b) What is the coefficient of mother.age in your regression? How do you interpret this coefficient?

(c) How many coefficients are estimated for the mother’s smoking status variable? How do you interpret these coefficients?

Issues in Linear Regression

- Multicollinearity
- Non-linear relationship between the outcome and the predictors
- Confounding variables

Issues in Linear Regression

- **Multicollinearity**
  - **Multicollinearity**, where collinearity exists between three or more variables even if no pair of variables has a particularly high correlation. This means that there is redundancy between predictor variables.
  - In the presence of multicollinearity, the solution of the regression model becomes unstable.
  - multicollinearity can assessed by computing a score called the variance inflation factor (or VIF), which measures how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

- The smallest possible value of VIF is one (absence of multicollinearity). As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity.
- When faced to multicollinearity, the concerned variables should be removed, since the presence of multicollinearity implies that the information that this variable provides about the response is redundant in the presence of the other variables

Model performance metrics

In regression model, the most commonly known evaluation metrics include:

1. **R-squared (R2)**, which is the proportion of variation in the outcome that is explained by the predictor variables. In multiple regression models, R2 corresponds to the squared correlation between the observed outcome values and the predicted values by the model. The Higher the R-squared, the better the model.

Model performance metrics

2. **Root Mean Squared Error (RMSE)**, which measures the average error performed by the model in predicting the outcome for an observation. Mathematically, the RMSE is the square root of the mean squared error (MSE), which is the average squared difference between the observed actual outome values and the values predicted by the model. So, MSE = mean((observeds - predicteds)^2) and RMSE = sqrt(MSE). **The lower the RMSE, the better the model.**

Model performance metrics

3. **Residual Standard Error (RSE),** also known as the model sigma, is a variant of the RMSE adjusted for the number of predictors in the model. The lower the RSE, the better the model. In practice, the difference between RMSE and RSE is very small, particularly for large multivariate data.
4. **Mean Absolute Error (MAE),** like the RMSE, the MAE measures the prediction error. Mathematically, it is the average absolute difference between observed and predicted outcomes, MAE = mean(abs(observeds - predicteds)). MAE is less sensitive to outliers compared to RMSE.

5. **AIC stands for (Akaike's Information Criteria)**, a metric developed by the Japanese Statistician, Hirotugu Akaike, 1970. The basic idea of AIC is to penalize the inclusion of additional variables to a model. It adds a penalty that increases the error when including additional terms. The lower the AIC, the better the model.
6. **BIC (or Bayesian information criteria)** is a variant of AIC with a stronger penalty for including additional variables to the model.

Confounding Variables

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.671e+01  8.506e-01  78.427  < 2e-16 ***
Electricity 1.899e-04  1.553e-04   1.223    0.224
GDP         2.455e-04  4.844e-05   5.068 1.52e-06 ***
```

Once GDP is accounted for, electricity use is no longer a significant predictor of life expectancy.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.414e+01  1.272e+00  42.573  < 2e-16 ***
Cell        1.354e-01  1.419e-02   9.539  < 2e-16 ***
GDP         1.884e-04  3.465e-05   5.439 1.95e-07 ***
```

Even after accounting for GDP, cell phone subscriptions per capita is still a significant predictor of life expectancy.

Confounding Variables (review)

- Multiple regression is one potential way to account for confounding variables

- This is most commonly used in practice across a wide variety of fields, but is quite sensitive to the conditions for the linear model (particularly linearity)

- You can only account for confounding variables that you have data on, so it is still very hard to make true causal conclusions without a randomized experiment

Workshop 3.5 - Stepwise Regression

From birthstone dataset on MASS, try stepwise regression to find regression model for Birth Weight as dependent variable and all other variables as independent variables. Try “forward”, “backward” and “both direction” to see all model.

Which is the best model, explain? What is the final regression model?

Polynomial Regression

Data Science Certification

See Jupyter notebook

Workshop 3.6 Polynomial Regression

use library(ISLR) to get access to Wage dataset

- Using non-linear regression model to find model that can explain “wage” (dependent variable) using “age” (independent variable)

- What is the best approach for this dataset?

Probability

Data Science Certification

Event

- An **event** is something that either happens or doesn't happen, or something that either is true or is not true

- Examples:
  - A randomly selected card is a Heart
  - The response variable Y > 90
  - A randomly selected person is male
  - It rains today

Probability

- The **probability** of event A, P(A), is the probability that A will happen

- Probability is always between 0 and 1

- Probability always refers to an event

- P(A) = 1 means A will definitely happen
- P(A) = 0 means A will definitely not happen

Probability Examples

- Y = number of siblings. P(Y = 1) = 0.481 (based on survey data)

- P(Gender = male) = 0.506

- P(it rains today) = 0.3 (www.weather.com)

Sexual Orientation

- What are the sexual orientation demographics of US adults?

- We need data!

- Data collected in 2009 on a random sample of American adults (National Survey of Sexual Health and Behavior)

Sexual Orientation

	Male	Female	Total
Heterosexual	2325	2348	4673
Homosexual	105	23	128
Bisexual	66	92	158
Other	25	58	83
Total	2521	2521	5042

Herbenick D, Reece M, Schick V, Sanders SA, Dodge B, and Fortenberry JD (2010). *Sexual behavior in the United States: Results from a national probability sample of men and women ages 14–94*. Journal of Sexual Medicine;7(suppl 5):255–265.

Sexual Orientation

	Male	Female	Total
Heterosexual	2325	2348	4673
Homosexual	105	23	128
Bisexual	66	92	158
Other	25	58	83
Total	2521	2521	5042

What is the probability that an American adult is homosexual?

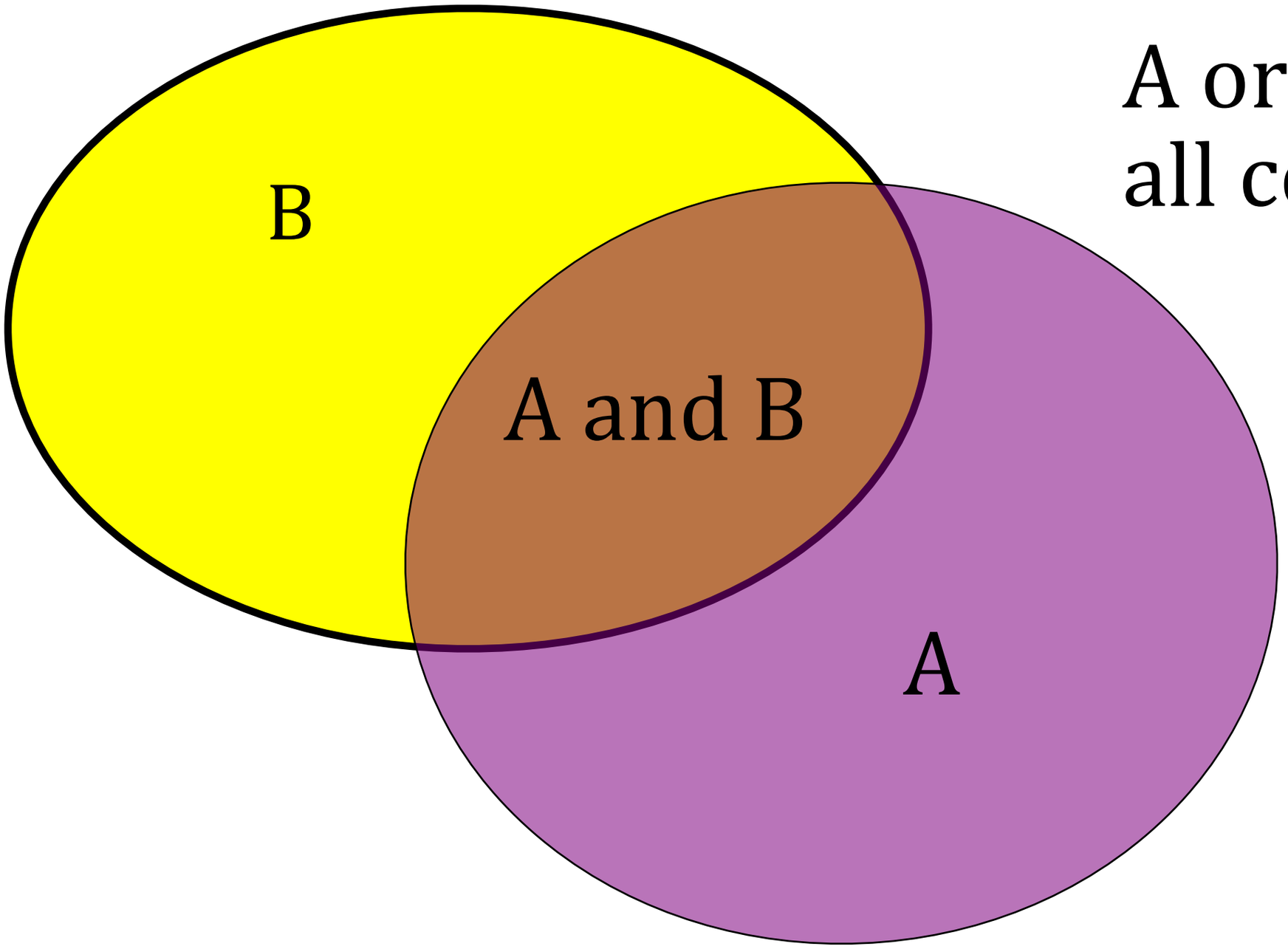
a) 128/5042 = 0.025
b) 128/4673 = 0.027
c) 105/2521 = 0.04
d) I got a different answer

Two Events

- **P(A and B)** is the probability that both events A and B will happen

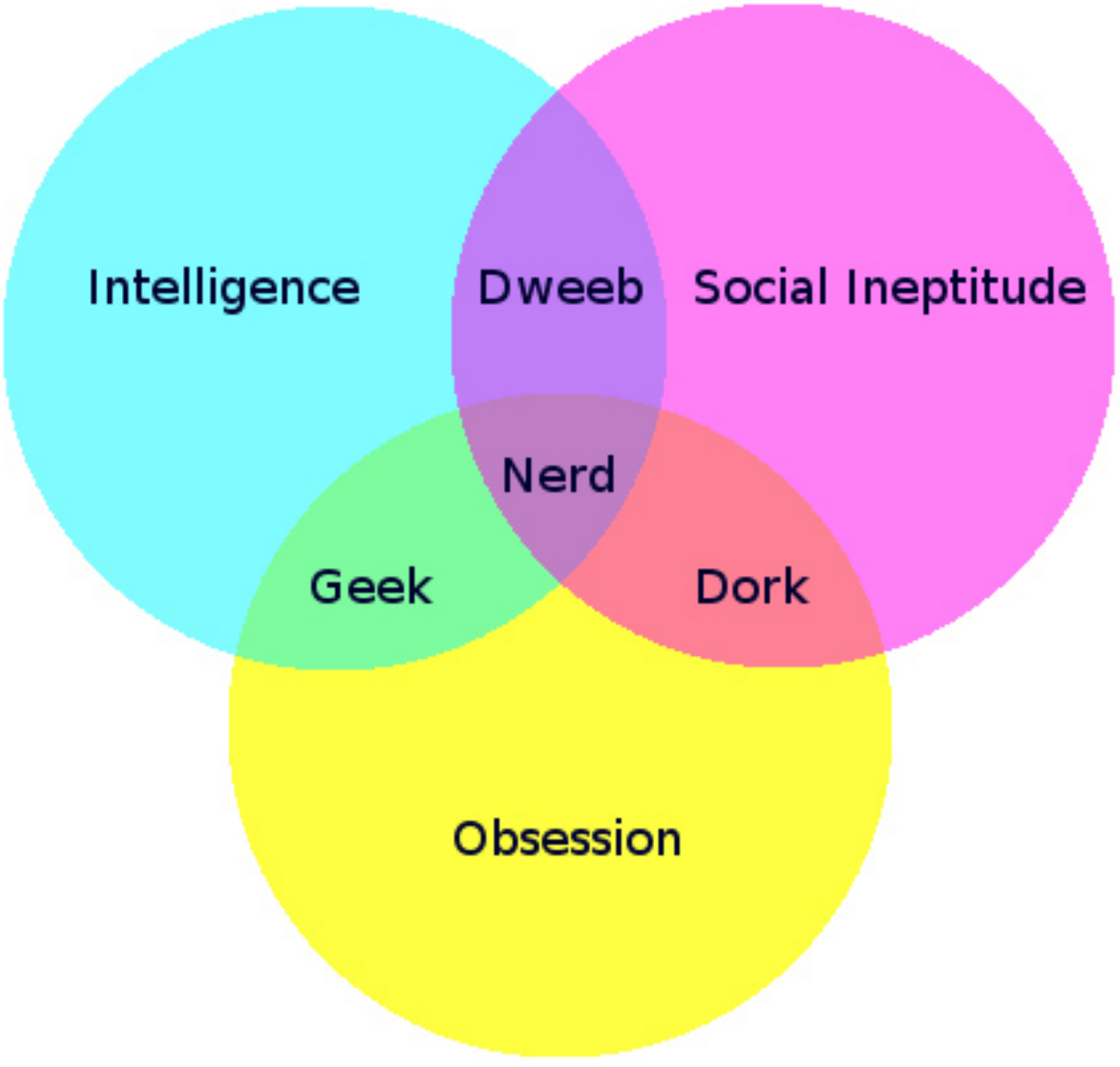
- **P(A or B)** is the probability that either event A or event B will happen

Two Events



A or B:
all color

Venn Diagram



Sexual Orientation

	Male	Female	Total
Heterosexual	2325	2348	4673
Homosexual	105	23	128
Bisexual	66	92	158
Other	25	58	83
Total	2521	2521	5042

What is the probability that an American adult is male and homosexual?

a) 105/128 = 0.82
b) 105/2521 = 0.04
c) 105/5042 = 0.021
d) I got a different answer

Sexual Orientation

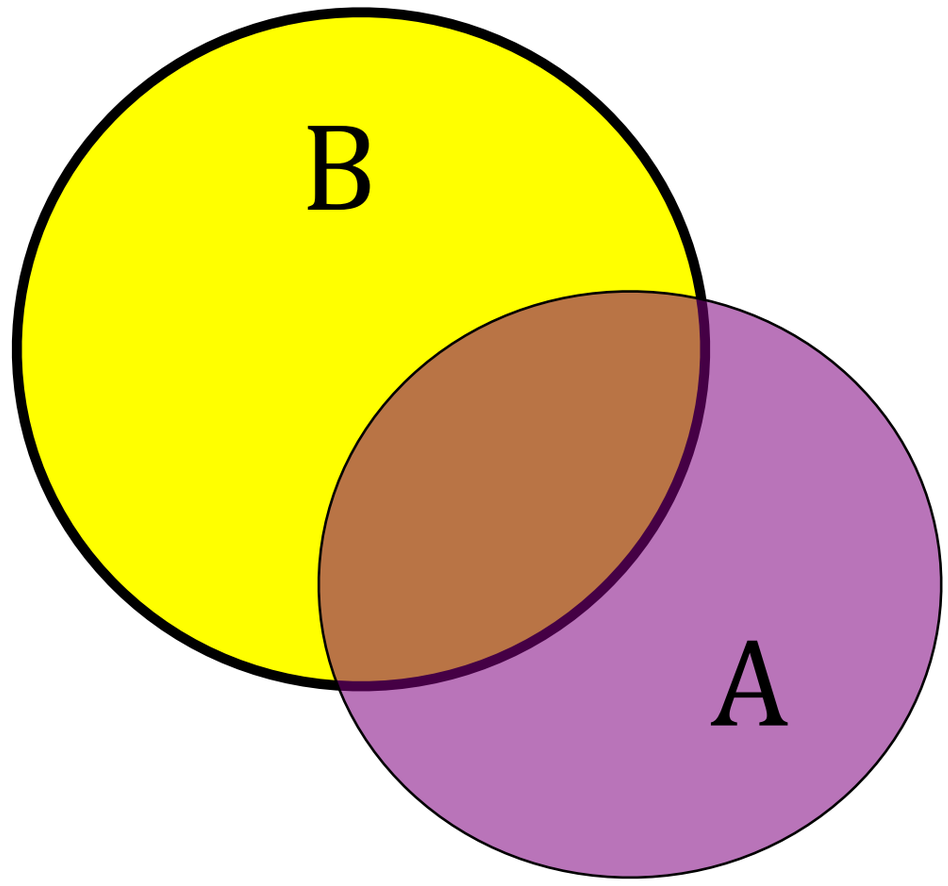
	Male	Female	Total
Heterosexual	2325	2348	4673
Homosexual	105	23	128
Bisexual	66	92	158
Other	25	58	83
Total	2521	2521	5042

What is the probability that an American adult is female or bisexual?

a) 2679/5042 = 0.531
b) 2587/5042 = 0.513
c) 92/2521 = 0.036
d) I got a different answer

P(A or B)

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$



Sexual Orientation

	Male	Female	Total
Heterosexual	2325	2348	4673
Homosexual	105	23	128
Bisexual	66	92	158
Other	25	58	83
Total	2521	2521	5042

What is the probability that an American adult is not heterosexual?

a) 369/5042 = 0.073
b) 2587/5042 = 0.513
c) 92/2521 = 0.036
d) I got a different answer

P(not A)

$$P(\text{not } A) = 1 - P(A)$$

Caffeine

Based on last year's survey data, 52% of students drink caffeine in the morning, 48% of students drink caffeine in the afternoon, and 37% drink caffeine in the morning and the afternoon. What percent of students do not drink caffeine in the morning or the afternoon?

a) 63%
b) 37%
c) 100%
d) 50%

P(not(morning or afternoon)
= 1 – P(morning or afternoon)
= 1 – [P(morning)+P(afternoon)
– P(morning and afternoon)]
= 1 – [0.52 + 0.48 – 0.37]
= 1 – 0.63
= 0.37

Conditional Probability

- **P(A if B)** is the probability of A, if we know B has happened
- This is read in multiple ways:
  - “probability of A if B”
  - “probability of A given B”
  - “probability of A conditional on B”
- You may also see this written as P(A | B)

Sexual Orientation

	Male	Female	Total
Heterosexual	2325	2348	4673
Homosexual	105	23	128
Bisexual	66	92	158
Other	25	58	83
Total	2521	2521	5042

What is the probability that an American adult male is homosexual?

a) 105/128 = 0.82
b) 105/2521 = 0.04
c) 105/5042 = 0.021
d) I got a different answer

P(homosexual if male)

Sexual Orientation

	Male	Female	Total
Heterosexual	2325	2348	4673
Homosexual	105	23	128
Bisexual	66	92	158
Other	25	58	83
Total	2521	2521	5042

What is the probability that an American adult homosexual is male?

a) 105/128 = 0.82
b) 105/2521 = 0.04
c) 105/5042 = 0.021
d) I got a different answer

P(male if homosexual)

Conditional Probability

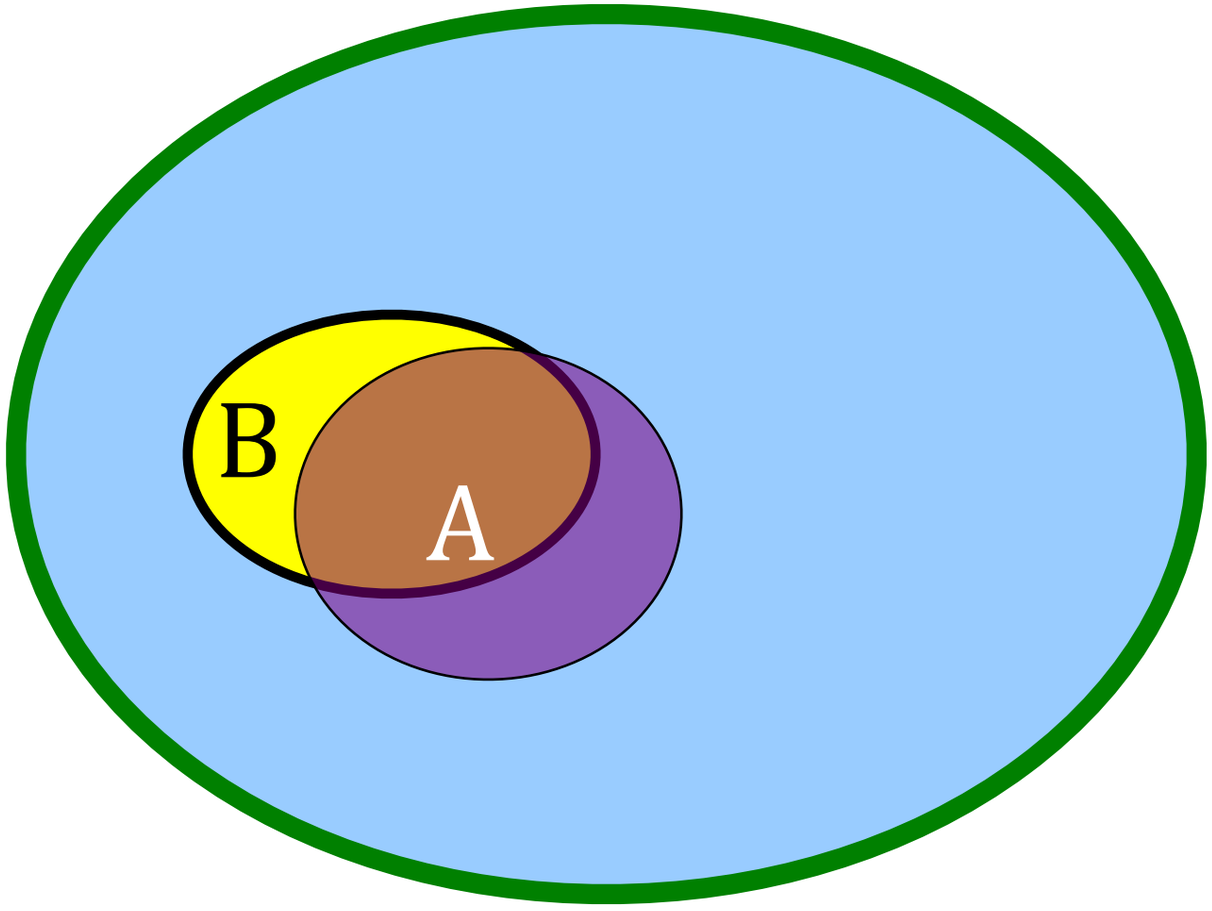
$$P(A \text{ if } B) \neq P(B \text{ if } A)$$

P(homosexual if male) = 0.04

P(male if homosexual) = 0.82

Conditional Probability

$$P(A \text{ if } B) = \frac{P(A \text{ and } B)}{P(B)}$$



Caffeine

Based on last year's survey data, 52% of students drink caffeine in the morning, 48% of students drink caffeine in the afternoon, and 37% drink caffeine in the morning and the afternoon. What percent of students who drink caffeine in the morning also drink caffeine in the afternoon?

a) 77%
b) 37%
c) 71%

P(afternoon if morning)
= P(afternoon and morning)/P(morning)
= 0.37/0.52
= 0.71

Helpful Tip

If the table problems are easier for your than the sentence problems, try to first convert what you know into a table.

52% of students drink caffeine in the morning, 48% of students drink caffeine in the afternoon, and 37% drink caffeine in the morning and the afternoon

	Caffeine Afternoon	No Caffeine Afternoon	Total
Caffeine Morning	37	15	52
No Caffeine Morning	11	37	48
Total	48	52	100

P(afternoon if morning) = 37/52 = 0.71

P(A and B)

$$P(A \text{ if } B) = \frac{P(A \text{ and } B)}{P(B)}$$

$$P(A \text{ and } B) = P(A \text{ if } B)P(B)$$

Chula Rank and Experience

60% of students rank their university experience as "Excellent," and Chula was the first choice school for 59% of those who ranked their Chula experience as excellent. What percentage of stat students had Chula as a first choice and rank their experience here as excellent?

a) 60%
b) 59%
c) 35%
d) 41%

P(first choice and excellent)

= P(first choice if excellent)P(excellent)

= 0.59 × 0.60

= 0.354

Summary

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

$$P(A \text{ and } B) = P(A \text{ if } B)P(B)$$

$$P(\text{not } A) = 1 - P(A)$$

$$P(A \text{ if } B) = \frac{P(A \text{ and } B)}{P(B)}$$

$$P(A \text{ if } B) \neq P(B \text{ if } A)$$

Disjoint Events

- Events A and B are **disjoint** or **mutually exclusive** if only one of the two events can happen

- Think of two events that are disjoint, and two events that are not disjoint.

Disjoint Events

If A and B are disjoint, then

a) P(A or B) = P(A) + P(B)

b) P(A and B) = P(A)P(B)

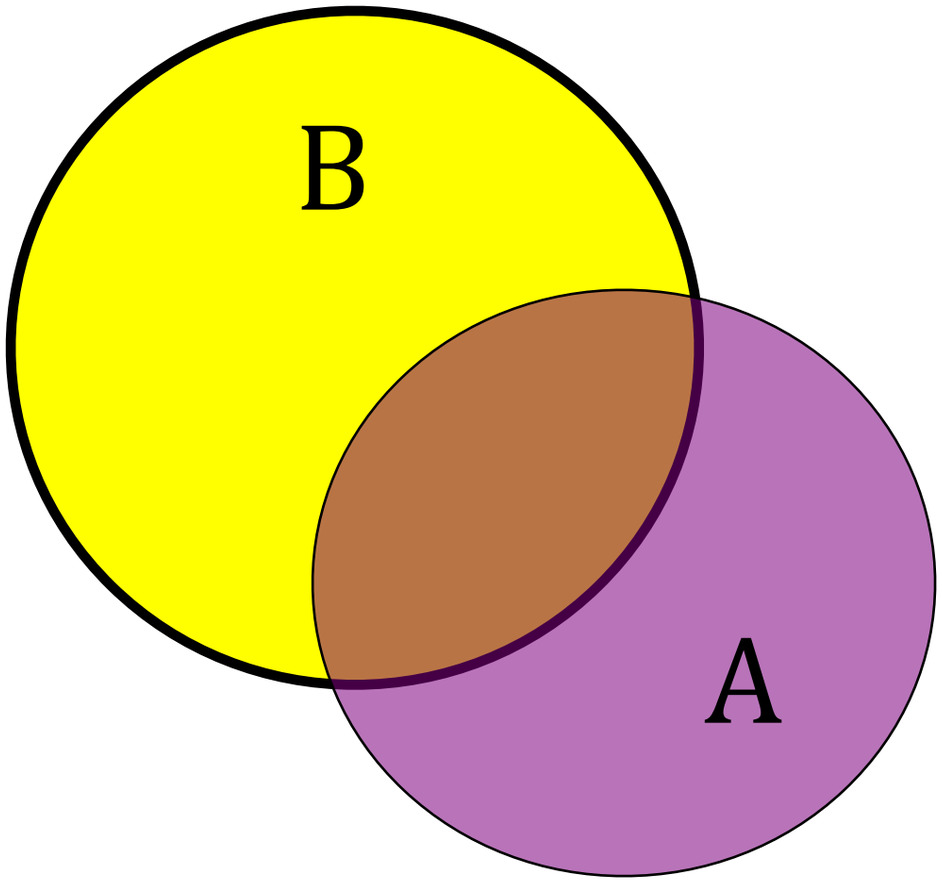
P(A or B) = P(A) + P(B) – P(A and B)

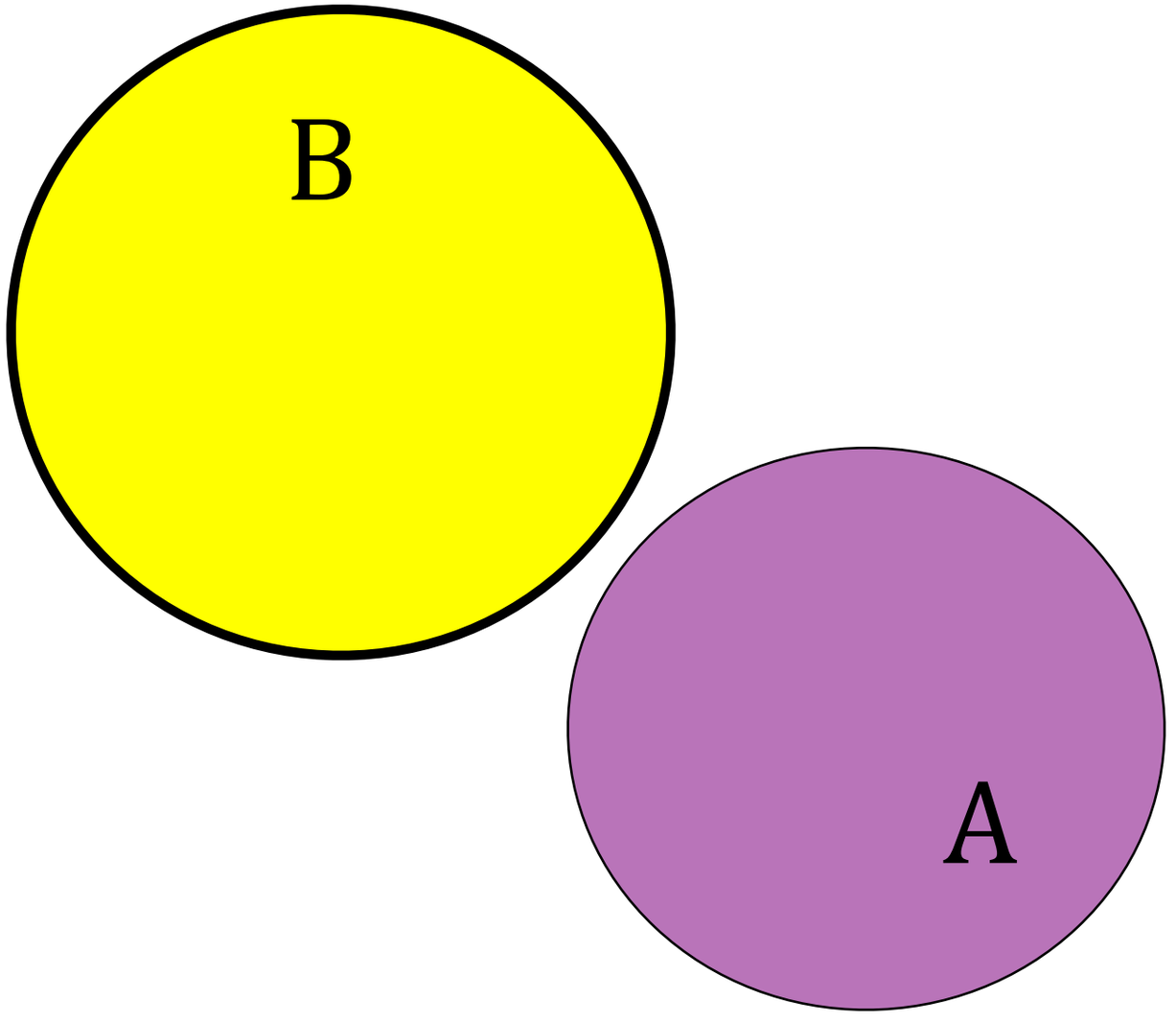
If A and B are disjoint, then both cannot happen, so P(A and B) = 0.

P(A or B)

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

SPECIAL CASE:
If A and B are disjoint:
$P(A \text{ or } B) = P(A) + P(B)$





Independence

- Events A and B are **independent** if P(A if B) = P(A).
- Intuitively, knowing that event B happened does not change the probability that event A happened.
- Think of two events that are independent, and two events that are not independent.

Independent Events

If A and B are independent, then

a) P(A or B) = P(A) + P(B)

b) P(A and B) = P(A)P(B)

P(A and B) = P(A if B)P(B)

If A and B are independent, then P(A if B) = P(A), so P(A and B) = P(A)P(B)

P(A and B)

$$P(A \text{ and } B) = P(A \text{ if } B)P(B)$$

If A and B are independent,

then $P(A \text{ if } B) = P(A)$, so

SPECIAL CASE:

If A and B are independent,

$$P(A \text{ and } B) = P(A)P(B)$$

Disjoint and Independent

Assuming that P(A) > 0 and P(B) > 0, then disjoint events are

a) Independent
b) Not independent
c) Need more information to determine whether the events are also independent

If A and B are disjoint, then A cannot happen if B has happened, so P(A if B) = 0.

If P(A) > 0, then P(A if B) ≠ P(A) so A and B are not independent.

Probabilities and Distributions with R

Data Science Certification

Generate random samples from a normal Distribution

```
set.seed(124)
```

```
norm <- rnorm(100)
```

```
norm[1:10]
```

```
##  [1] -1.3851  0.0383 -0.7630  0.2123  1.4255  0.7445  0.7002 -0.2294
##  [9]  0.1971  1.2072
```

```
mean(norm)
```

```
## [1] 0.00962
```

```
sd(norm)
```

```
## [1] 0.884
```

```r
set.seed(124)
norm <- rnorm(100, 2, 5)
norm[1:10]
```

```
##  [1] -4.925  2.192 -1.815  3.062  9.128  5.722  5.501  0.853  2.985  8.036
```

```r
mean(norm)
```

```
## [1] 2.05
```

```r
sd(norm)
```

```
## [1] 4.42
```

Generating random samples from other distributions

- other common

  distributions: runif, rpois, rmvnorm, rnbinom, rbinom, rbeta, rchisq, rexp, r

  gamma, rlogis, rstab, rt, rgeom, rhyper, rwilcox, rweibull
- For example, the **rpois** function is the random number generator for the

  Poisson distribution and it has only the parameter argument **lambda.**

  The **rbinom** function is the random number generator for the binomial

  distribution and it takes two arguments: **size** and **prob**

```r
# Generating a random sample from a Poisson distribution with lambda=3
set.seed(124)
pois <- rpois(100, lambda = 3)
pois[1:10]
```

```
##  [1] 1 2 3 2 2 2 3 3 6 2
```

```r
mean(pois)
```

```
## [1] 2.83
```

```r
var(pois)
```

```
## [1] 2.34
```

```r
# Generating a random sample from a Binomial distribution with size=20 and
# prob=.2
set.seed(124)
binom <- rbinom(100, 20, 0.2)
binom[1:10]
```

```
##  [1] 2 3 4 3 3 3 4 4 7 3
```

```r
mean(binom)
```

```
## [1] 3.85
```

```r
sd(binom)
```

```
## [1] 1.6
```

Prob package

```
> tosscoin(1)

  toss1
1     H
2     T
```

```
> rolldie(1)

  X1
1  1
2  2
3  3
4  4
5  5
6  6
```

```
> tosscoin(3)

  toss1 toss2 toss3
1     H     H     H
2     T     H     H
3     H     T     H
4     T     T     H
5     H     H     T
6     T     H     T
7     H     T     T
8     T     T     T
```

Sampling from Urns

Ordered, With Replacement

```
> urnsamples(1:3, size = 2, replace = TRUE, ordered = TRUE)

  X1 X2
1  1  1
2  2  1
3  3  1
4  1  2
5  2  2
6  3  2
7  1  3
8  2  3
9  3  3
```

Ordered, without replacement

```
> urnsamples(1:3, size = 2, replace = FALSE, ordered = TRUE)

  X1 X2
1  1  2
2  2  1
3  1  3
4  3  1
5  2  3
6  3  2
```

Unordered, Without Replacement

```
> urnsamples(1:3, size = 2, replace = FALSE, ordered = FALSE)

  X1 X2
1  1  2
2  1  3
3  2  3
```

Defining a Probability Space

The Equally Likely Model

```
> outcomes <- rolldie(1)

  X1
1  1
2  2
3  3
4  4
5  5
6  6

> p <- rep(1/6, times = 6)

[1] 0.1666667 0.1666667 0.1666667 0.1666667 0.1666667 0.1666667
```

The probspace() function is designed to save us some time in many of the most common situations

```
> probspace(outcomes, probs = p)
  X1     probs
1  1 0.1666667
2  2 0.1666667
3  3 0.1666667
4  4 0.1666667
5  5 0.1666667
6  6 0.1666667
```

Functions for Finding Subsets

```
> x <- 1:10
> y <- 3:7
> y %in% x
```

```
[1] TRUE TRUE TRUE TRUE TRUE
```

```
> S <- rolldie(4)
> subset(S, isin(S, c(2,2,6), ordered = TRUE))

     X1 X2 X3 X4
188   2  2  6  1
404   2  2  6  2
620   2  2  6  3
836   2  2  6  4
1052  2  2  6  5
1088  2  2  1  6
1118  2  1  2  6
1123  1  2  2  6
1124  2  2  2  6
```

Set Union, Intersection, and Difference

Name	Denoted	Defined by elements	Code
Union	$A \cup B$	in A or B or both	`union(A,B)`
Intersection	$A \cap B$	in both A and B	`intersect(A,B)`
Difference	$A \backslash B$	in A but not in B	`setdiff(A,B)`

```
> S <- cards()
> A <- subset(S, suit == "Heart")
> B <- subset(S, rank %in% 7:9)
```

```
> union(A,B)

   rank     suit
6     7     Club
7     8     Club
8     9     Club
19    7  Diamond
20    8  Diamond
21    9  Diamond
27    2    Heart
28    3    Heart
29    4    Heart
30    5    Heart
31    6    Heart
32    7    Heart
33    8    Heart
34    9    Heart
35   10    Heart
36    J    Heart
37    Q    Heart
```

```
> intersect(A,B)

   rank  suit
32    7 Heart
33    8 Heart
34    9 Heart
```

```
> setdiff(A,B)

   rank  suit
27    2 Heart
28    3 Heart
29    4 Heart
30    5 Heart
31    6 Heart
35   10 Heart
36    J Heart
37    Q Heart
38    K Heart
39    A Heart
```

Calculating Probabilities

```
> S <- cards(makespace = TRUE)
> A <- subset(S, suit == "Heart")
> B <- subset(S, rank %in% 7:9)
```

Now it is easy to calculate

```
> Prob(A)
```

```
[1] 0.25
```

Note that we can get the same answer with

```
> Prob(S, suit == "Heart")
```

```
[1] 0.25
```

Conditional Probability

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \quad \text{if } \mathbb{P}(B) > 0.$$

```
> Prob(S, suit=="Heart", given = rank %in% 7:9)

[1] 0.25
```

```
> Prob(B, given = A)

[1] 0.2307692
```

Example

Consider an urn with 10 balls inside, 7 of which are red and 3 of which are green. Select 3 balls successively from the urn. Let A = {1st ball is red}, B= {2nd ball is red} and C={3rd ball is red}. Then

IP (all 3 balls are red) = IP(A∩B∩C) = 7/10 x 6/9 x 5/8 ≈0.2917

```
> library(prob)
> L <- rep(c("red", "green"), times = c(7, 3))
> M <- urnsamples(L, size = 3, replace = FALSE, ordered = TRUE)
> N <- probspace(M)
```

Example

Toss ten coins. What is the probability of observing at least one Head?

```
> S <- tosscoin(10, makespace = TRUE)
> A <- subset(S, isrep(S, vals = "T", nrep = 10))
> 1 - prob(A)
[1] 0.9990234
```

Workshop 3.6

Consider two urns, the first with 5 red balls and 3 green balls, and the second with 2 red balls and 6 green balls. Your friend randomly selects one ball from the first urn and transfers it to the second urn, without disclosing the color of the ball. You select one ball from the second urn. What is the probability that the selected ball is red?

Use R code to answer it.

Bayesian Inference

A 40-year old woman participates in routine screening and has a positive mammography. What's the probability she has cancer?

a) 0-10%
b) 10-25%
c) 25-50%
d) 50-75%
e) 75-100%

Breast Cancer Screening

1% of women at age 40 who participate in routine screening have breast cancer.

80% of women with breast cancer get positive mammographies.

9.6% of women without breast cancer get positive mammographies.

A 40-year old woman participates in routine screening and has a positive mammography. What's the probability she has cancer?

A 40-year old woman participates in routine screening and has a positive mammography. What's the probability she has cancer?

a) 0-10%

b) 10-25%

c) 25-50%

d) 50-75%

e) 75-100%

Breast Cancer Screening

A 40-year old woman participates in routine screening and has a positive mammography. What's the probability she has cancer?

What is this asking for?

a) P(cancer if positive mammography)
b) P(positive mammography if cancer)
c) P(positive mammography if no cancer)
d) P(positive mammography)
e) P(cancer)

Bayes Rule

- We know P(positive mammography if cancer)… how do we get to P(cancer if positive mammography)?
- How do we go from P(A if B) to P(B if A)?

$$P(A \text{ if } B) = \frac{P(B \text{ if } A)P(A)}{P(B)}$$

$$P(A \text{ if } B) = \frac{P(A \text{ and } B)}{P(B)}$$

$$\Rightarrow P(A \text{ and } B) = P(A \text{ if } B)P(B)$$

$$P(A \text{ and } B) = P(B \text{ if } A)P(A)$$

$$\Rightarrow P(A \text{ if } B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(B \text{ if } A)P(A)}{P(B)}$$

$$\Rightarrow P(A \text{ if } B) = \frac{P(B \text{ if } A)P(A)}{P(B)}$$ <- **Bayes Rule**

Rev. Thomas Bayes

1702 - 1761

Breast Cancer Screening

$$P(\text{cancer if positive}) = \frac{P(\text{positive if cancer})P(\text{cancer})}{P(\text{positive})}$$

- 1% of women at age 40 who participate in routine screening have breast cancer.
- 80% of women with breast cancer get positive mammographies.
- 9.6% of women without breast cancer get positive mammographies.

P(positive)

How do we figure out P(positive)?

We know:

80% of women with breast cancer get positive mammographies.

9.6% of women without breast cancer get positive mammographies.

We need to average these two numbers, weighted by the proportion of people with breast cancer:

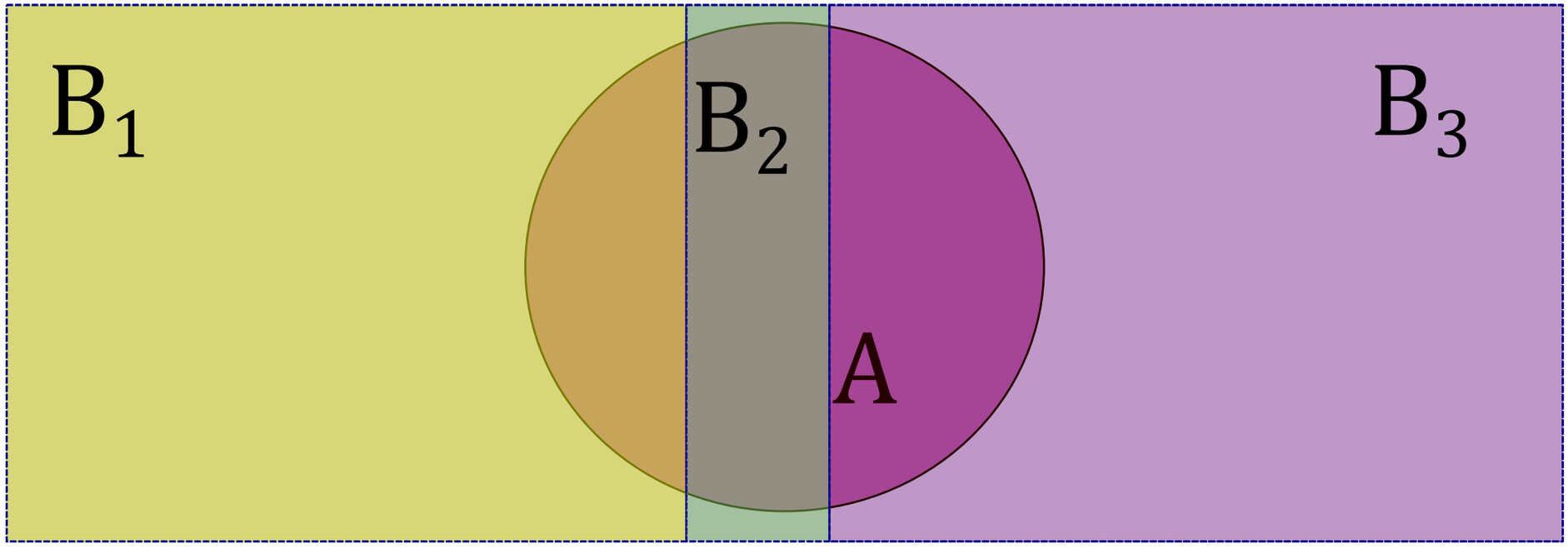
$$P(\text{positive if cancer})P(\text{cancer}) + P(\text{positive if no cancer})P(\text{no cancer})$$

$$0.8 \times P(\text{cancer}) + 0.096 \times P(\text{no cancer})$$

Law of Total Probability

- If events B_1 through B_k are *disjoint* and together make up all possibilities, then

$$P(A) = P(A \text{ and } B_1) + P(A \text{ and } B_2) + \ldots + P(A \text{ and } B_k)$$

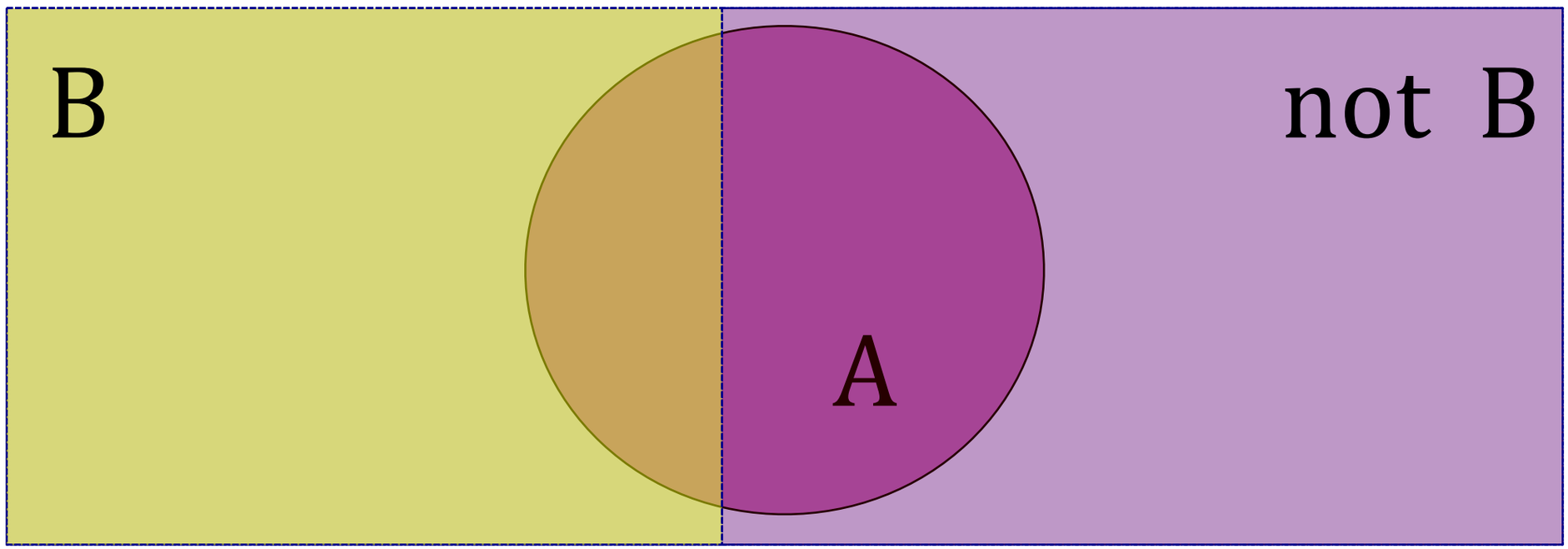


$$P(A) = P(A \text{ if } B_1)P(B_1) + P(A \text{ if } B_2)P(B_2) + \ldots + P(A \text{ if } B_k)P(B_k)$$

Law of Total Probability

- Special case: B and not B

$$P(A) = P(A \text{ and } B) + P(A \text{ and not B})$$



$$P(A) = P(A \text{ if } B)\,P(B) + P(A \text{ if not } B)\,P(\text{not } B)$$

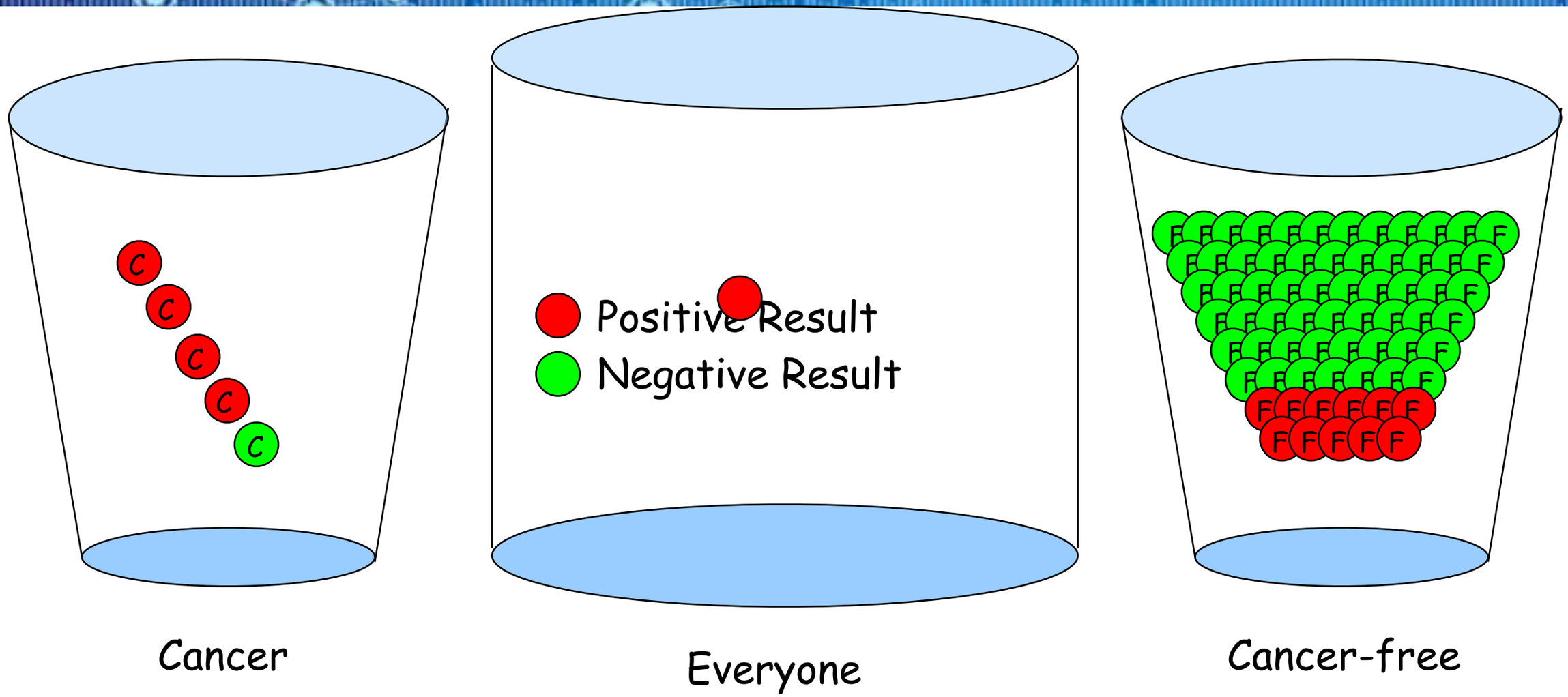
P(positive)

Use the law of total probability to find P(positive).

Breast Cancer Screening

$$P(\text{cancer if positive}) = \frac{P(\text{positive if cancer})P(\text{cancer})}{P(\text{positive})}$$

- 1% of women at age 40 who participate in routine screening have breast cancer.
- 80% of women with breast cancer get positive mammographies.
- 9.6% of women without breast cancer get positive mammographies.



- We randomly pick a ball from the Everyone bin.
- If the ball is red/positive, is it more likely to be from the Cancer or Cancer-free bin?

100,000 women in the population

1%

99%

1000 have cancer

99,000 cancer-free

80%

20%

9.6%

90.4%

800 test positive

200 test negative

9,504 test positive

89,496 test negative

Thus, 800/(800+9,504) = 7.8% of positive results have cancer

Hypotheses

H_0 : no cancer

H_a : cancer

Data: positive mammography

p-value = P(statistic as extreme as observed if H_0 true)

= P(positive mammography if no cancer)

= 0.096

The probability of getting a positive mammography just by random chance, if the woman does not have cancer, is 0.096.

Hypotheses

H_0 : no cancer

H_a : cancer

Data: positive mammography

You don’t really want the p-value, you want the probability that the woman has cancer!

You want P(H_0 true if data), not P(data if H_0 true)

Hypotheses

H_0 : no cancer

H_a : cancer

Data: positive mammography

Using Bayes Rule:

P(H_a true if data) = P(cancer if data) = 0.078

P(H_0 true if data) = P(no cancer I data) = 0.922

This tells a very different story than a p-value of 0.096!

Frequentist Inference

- ***Frequentist Inference*** considers what would happen if the data collection process (sampling or experiment) was repeated many times

- Probability is considered to be the proportion of times an event would happen if repeated many times

- In frequentist inference, we condition on some unknown truth, and find the probability of our data given this unknown truth

Frequentist Inference

- Everything we have done so far is based on frequentist inference

- A confidence interval is created to capture the truth for a specified proportion of all samples

- A p-value is the proportion of times you would get results as extreme as those observed, if the null hypothesis were true

Bayesian Inference

- ***Bayesian inference*** does not think about repeated sampling or repeating the experiment, but only what you can tell from your single observed data set

- Probability is considered to be the subjective degree of belief in some statement

- In Bayesian inference we condition on the data, and find the probability of some unknown parameter, given the data

Fixed and Random

- In frequentist inference, the parameter is considered fixed and the sample statistic is random

- In Bayesian inference, the statistic is considered fixed, and the parameter is considered random

Bayesian Inference

Frequentist: P(data if truth)

Bayesian: P(truth if data)

- How are they connected?

$$P(truth \text{ if } data) = \frac{P(data \text{ if } truth)P(truth)}{P(data)}$$

Bayesian Inference

POSTERIOR Probability

PRIOR Probability

$$P(truth \text{ if } data) = \frac{P(data \text{ if } truth)P(truth)}{P(data)}$$

- Prior probability: probability of a statement being true, before looking at the data

- Posterior probability: probability of the statement being true, after updating the prior probability based on the data

Breast Cancer

- Before getting the positive result from her mammography, the prior probability that the woman has breast cancer is 1%

- Given data (the positive mammography), update this probability using Bayes rule:

$$\frac{P(data \text{ if } truth)P(truth)}{P(data)} = \frac{0.8 \times 0.01}{0.103} = 0.078$$

- The posterior probability of her having breast cancer is 0.078.

Paternity

• A woman is pregnant. However, she slept with two different guys (call them Al and Bob) close to the time of conception, and does not know who the father is.

• What is the prior probability that Al is the father?

• The baby is born with blue eyes. Al has brown eyes and Bob has blue eyes. Update based on this information to find the posterior probability that Al is the father.

Eye Color

- In reality eye color comes from several genes, and there are several possibilities but let's simplify here:
  - Brown is dominant, blue is recessive
  - One gene comes from each parent
  - BB, bB, Bb would all result in brown eyes
  - Only bb results in blue eyes
- To make it a bit easier: You know that Al's mother and the mother of the child both have blue eyes.

Paternity

What is the probability that Al is the father?

a) 1/2
b) 1/3
c) 1/4
d) 1/5
e) No idea

Bayesian Inference

- Why isn't everyone a Bayesian?

???

$$P(truth \text{ if } data) = \frac{P(data \text{ if } truth)P(truth)}{P(data)}$$

- Need some "prior belief" for the probability of the truth

- Also, until recently, it was hard to be a Bayesian (needed complicated math.) Now, we can let computers do the work for us!

Inference

Both kinds of inference have the same goal, and it is a goal fundamental to statistics:

to use information from the data to gain information about the unknown truth

Bayesian Theory using R

Example

- Suppose we have a test for the flu that is positive 90% of the time when tested on a flu patient (P(test + | flu) = 0.9), and is negative 95% of the time when tested on a healthy person (P(test - | no flu) = 0.95).
- We also know that the flu is affecting about 1% of the population (P(flu)=0.01).
- You go to the doctor and test positive.
- What is the chance that you truly have the flu?

```r
flu <- sample(c('No','Yes'), size=100000, replace=TRUE, prob=c(0.99,0.01))
test <- rep(NA, 100000) # create a dummy variable first
test[flu=='No'] <- sample(c('Neg','Pos'), size=sum(flu=='No'), replace=TRUE, prob=c(0.95,0.05))
test[flu=='Yes'] <- sample(c('Neg','Pos'), size=sum(flu=='Yes'), replace=TRUE, prob=c(0.1, 0.9))

mean(flu[test=='Pos']=='Yes')
```

```
## [1] 0.154525
```

Workshop 3.7

Suppose we go to test skin cancer at one of the hospital. Doctor take our skin sample. We have knowledge that if skin sample test is positive, there is chance at true positive p(+|cancer) at 99% and false positive p(+|~cancer) 5%.

Before we go to hospital, we google and find out that every 10,000 Thai, there is change to have this skin cancer at 100 person.

Use Bayes Theorem to find p(cancer|+)