INTRODUCTION TO DATA SCIENCE

This lecture is based on course by
M. Cetinkaya-Rundel, Duke University, Data Analysis and Statistical Inference
and book by
M. Cetinkaya-Rundel and J. Handrin, „Introduction to Modern Statistics”

https://www.openintro.org/book/stat/

08/01, 15/01, 22/01 2025

WFAiS UJ, Informatyka Stosowana
I stopień studiów

Modern statistics

- **Scientists seek to answer questions using rigorous methods and careful observations**
- **These observations – collected from the likes of field notes, surveys, and experiments – form the blackbone of statistical investigations and are called data.**
- **Statistics is the study of how best collect, analyze and draw conclusions from data.**

Introduction to data

Data basics

Observations, variables and data matrice

Table 1.3 displays six rows of a dataset for 50 randomly sampled loans offered through Lending Club, which is a peer-to-peer lending company. This dataset will be referred to as `loan50`.

Table 1.3: Six observations from the `loan50` dataset.

	loan_amount	interest_rate	term	grade	state	total_income	homeownership
1	22,000	10.90	60	B	NJ	59,000	rent
2	6,000	9.92	36	B	CA	60,000	rent
3	25,000	26.30	36	E	SC	75,000	mortgage
4	6,000	9.92	36	B	CA	75,000	rent
5	25,000	9.43	60	B	OH	254,000	mortgage
6	6,400	9.92	36	B	IN	67,000	mortgage

Each row in the table represents a single loan. The formal name for a row is a **case** or **observation** or **unit of observation**. The columns represent characteristics of each loan, where each column is referred to as a **variable**. For example, the first row represents a loan of \$22,000 with an interest rate of 10.90%, where the borrower is based in New Jersey (NJ) and has an income of \$59,000.

The data in Table 1.3 represent a **data frame**, which is a convenient and common way to organize data, especially if collecting data in a spreadsheet. A data frame where each row is a unique case (observational unit), each column is a variable, and each cell is a single value is commonly referred to as **tidy data** (Wickham 2014).

Data basics

In practice, it is especially important to ask clarifying questions to ensure important aspects of the data are understood. For instance, it is always important to be sure we know what each variable means and its units of measurement. Descriptions of the variables in the `loan50` dataset are given in Table 1.4.

Table 1.4: Variables and their descriptions for the `loan50` dataset.

Variable	Description
`loan_amount`	Amount of the loan received, in US dollars.
`interest_rate`	Interest rate on the loan, in an annual percentage.
`term`	The length of the loan, which is always set as a whole number of months.
`grade`	Loan grade, which takes a values A through G and represents the quality of the loan and its likelihood of being repaid.
`state`	US state where the borrower resides.
`total_income`	Borrower's total income, including any second income, in US dollars.
`homeownership`	Indicates whether the person owns, owns but has a mortgage, or rents.

Data basics

- **Type of variables**

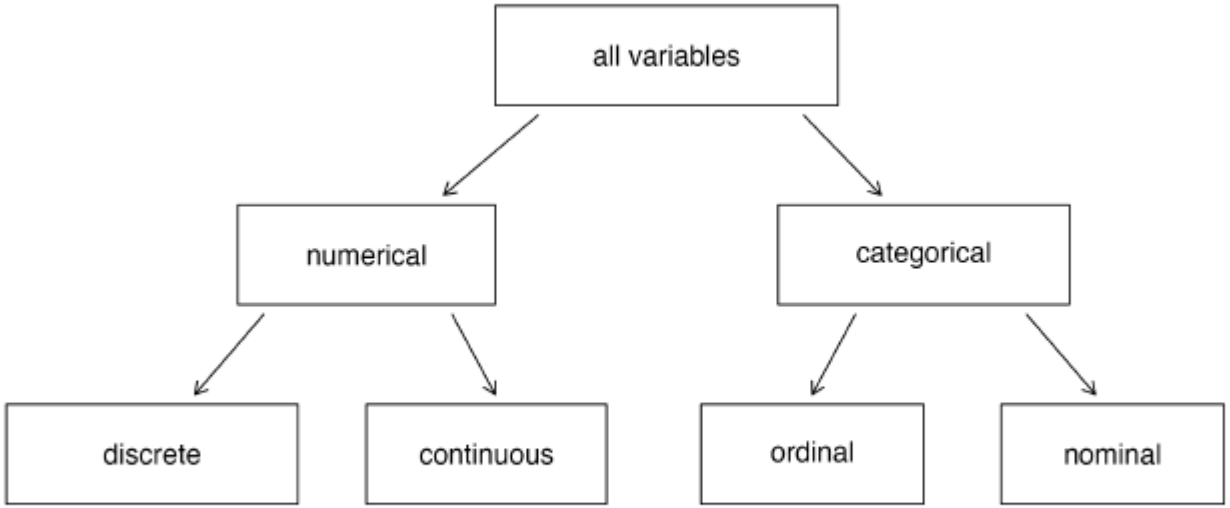


Figure 1.1: Breakdown of variables into their respective types.

Relationships between variables

Many analyses are motivated by a researcher looking for a relationship between two or more variables. A social scientist may like to answer some of the following questions:

> Does a higher-than-average increase in county population tend to correspond to counties with higher or lower median household incomes?
>
> If homeownership in one county is lower than the national average, will the percent of housing units that are in multi-unit structures in that county tend to be above or below the national average?
>
> How much can the median education level explain the median household income for counties in the US?

To answer these questions, data must be collected, such as the `county` dataset shown in Table 1.5. Examining **summary statistics** can provide numerical insights about the specifics of each of these questions. Alternatively, graphs can be used to visually explore the data, potentially providing more insight than a summary statistic.

Relationships between variables

GUIDED PRACTICE

We consider data for 3,142 counties in the United States, which includes the name of each county, the state where it resides, its population in 2017, the population change from 2010 to 2017, poverty rate, and nine additional characteristics. How might these data be organized in a data frame?[4]

Table 1.5: Six observations and six variables from the `county` dataset.

name	state	pop2017	pop_change	unemployment_rate	median_edu
Autauga County	Alabama	55,504	1.48	3.86	some_college
Baldwin County	Alabama	212,628	9.19	3.99	some_college
Barbour County	Alabama	25,270	-6.22	5.90	hs_diploma
Bibb County	Alabama	22,668	0.73	4.39	hs_diploma
Blount County	Alabama	58,013	0.68	4.02	hs_diploma
Bullock County	Alabama	10,309	-2.28	4.93	hs_diploma

Table 1.6: Variables and their descriptions for the `county` dataset.

Variable	Description
`name`	Name of county.
`state`	Name of state.
`pop2000`	Population in 2000.
`pop2010`	Population in 2010.
`pop2017`	Population in 2017.
`pop_change`	Population change from 2010 to 2017 (in percent).
`poverty`	Percent of population in poverty in 2017.
`homeownership`	Homeownership rate, 2006-2010.
`multi_unit`	Multi-unit rate: percent of housing units that are in multi-unit structures, 2006-2010.
`unemployment_rate`	Unemployment rate in 2017.
`metro`	Whether the county contains a metropolitan area, taking one of the values yes or no.
`median_edu`	Median education level (2013-2017), taking one of the values below_hs, hs_diploma, some_college, or bachelors.
`per_capita_income`	Per capita (per person) income (2013-2017).
`median_hh_income`	Median household income.
`smoking_ban`	Describes the type of county-level smoking ban in place in 2010, taking one of the values none, partial, or comprehensive.

Relationships between variables

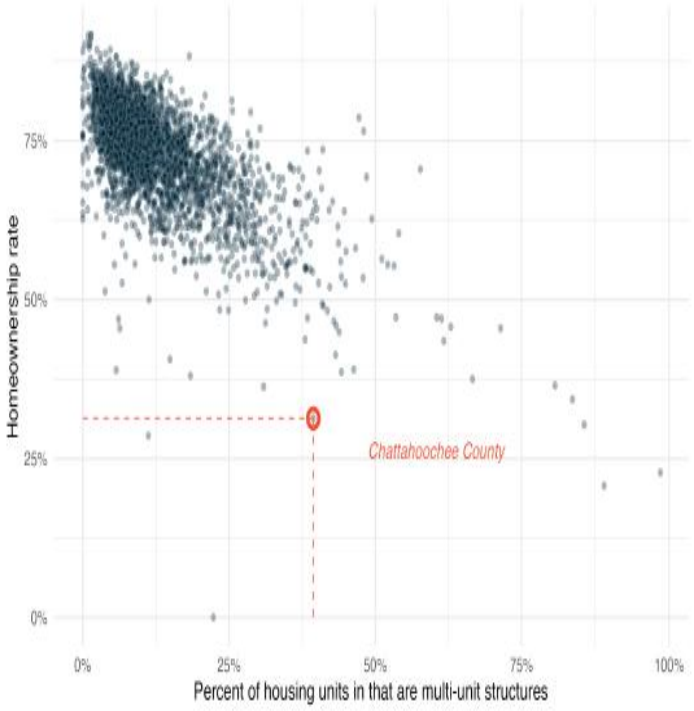


Figure 1.2: A scatterplot of homeownership versus the percent of housing units that are in multi-unit structures for US counties. The highlighted dot represents Chattahoochee County, Georgia, which has a multi-unit rate of 39.4% and a homeownership rate of 31.3%.

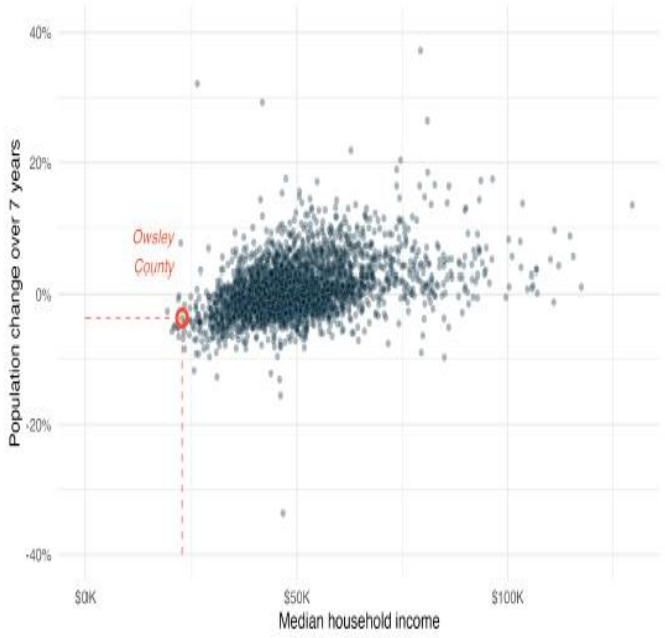


Figure 1.3: A scatterplot showing population change against median household income. Owsley County of Kentucky is highlighted, which lost 3.63% of its population from 2010 to 2017 and had median household income of $22,736.

Relationships between variables

Associated or independent, not both.

A pair of variables are either related in some way (associated) or not (independent). No pair of variables is both associated and independent.

Explanatory and response variables.

When we suspect one variable might causally affect another, we label the first variable the explanatory variable and the second the response variable. We also use the terms explanatory and response to describe variables where the response might be predicted using the explanatory even if there is no causal relationship.

explanatory variable $\rightarrow$ *might affect* $\rightarrow$ response variable

For many pairs of variables, there is no hypothesized relationship, and these labels would not be applied to either variable in such cases.

Observational studies and experiments

There are two primary types of data collection: experiments and observational studies.

When researchers want to evaluate the effect of particular traits, treatments, or conditions, they conduct an **experiment**.

To check if there really is a causal relationship between the explanatory and the response variable researchers identify a sample of individuals and split them into groups. The individuals in each group are *assigned* a treatment. When individuals are randomly assigned to a group, the experiment is called a **randomized experiment**. Random assignment organizes the participants in a study into groups that are roughly equal on all aspects, thus allowing us to control for any confounding variables that might affect the outcome

Researchers perform an **observational study** when they collect data in a way that does not directly interfere with how the data arise. For instance, researchers may collect information via surveys, review medical or company records, or follow a **cohort** of many similar individuals to form hypotheses about why certain diseases might develop. In each of these situations, researchers merely observe the data that arise. In general, observational studies can provide evidence of a naturally occurring association between variables, but they cannot by themselves show a causal connection as they do not offer a mechanism for controlling for confounding variables.

Observational studies and experiments

Association $\neq$ Causation.

In general, association does not imply causation. An advantage of a randomized experiment is that it is easier to establish causal relationships with such a study. The main reason for this is that observational studies do not control for confounding variables, and hence establishing causal relationships with observational studies requires advanced statistical methods

Study design

Sampling principles and strategies

The first step in conducting research is to identify topics or questions that are to be investigated. A clearly laid out research question is helpful in identifying what subjects or cases should be studied and what variables are important. It is also important to consider *how* data are collected so that the data are reliable and help achieve the research goals.

A proficient analyst will have a good sense of the types of data they are working with and how to visualize the data in order to gain a complete understanding of the variables. Equally important, however, is the data source.

Sampling principles and strategies

Population and sample

Consider the following three research questions:

1. What is the average mercury content in swordfish in the Atlantic Ocean?
2. Over the last five years, what is the average time to complete a degree for Duke undergrads?
3. Does a new drug reduce the number of deaths in patients with severe heart disease?

Each research question refers to a target **population**. In the first question, the target population is all swordfish in the Atlantic Ocean, and each fish represents a case. Oftentimes, it is not feasible to collect data for every case in a population. Collecting data for an entire population is called a **census**. A census is difficult because it is too expensive to collect data for the entire population, but it might also be because it is difficult or impossible to identify the entire population of interest! Instead, a sample is taken. A **sample** is the data you have. Ideally, a sample is a small fraction of the population. For instance, 60 swordfish (or some other number) in the population might be selected, and this sample data may be used to provide an estimate of the population average and to answer the research question.

Experiments

Studies where the researchers assign treatments to cases are called **experiments**. When this assignment includes randomization, e.g., using a coin flip to decide which treatment a patient receives, it is called a **randomized experiment**. Randomized experiments are fundamentally important when trying to show a causal connection between two variables.

Principle of experimental design

1. **Controlling.** Researchers assign treatments to cases, and they do their best to **control** any other differences in the groups[3].

2. **Randomization.** Researchers randomize patients into treatment groups to account for variables that cannot be controlled.

3. **Replication.** The more cases researchers observe, the more accurately they can estimate the effect of the explanatory variable on the response. In a single study, we **replicate** by collecting a sufficiently large sample.

4. **Blocking.** Researchers sometimes know or suspect that variables, other than the treatment, influence the response. Under these circumstances, they may first group individuals based on this variable into **blocks** and then randomize cases within each block to the treatment groups. This strategy is often referred to as **blocking**.

Experiments

- **Blocking**

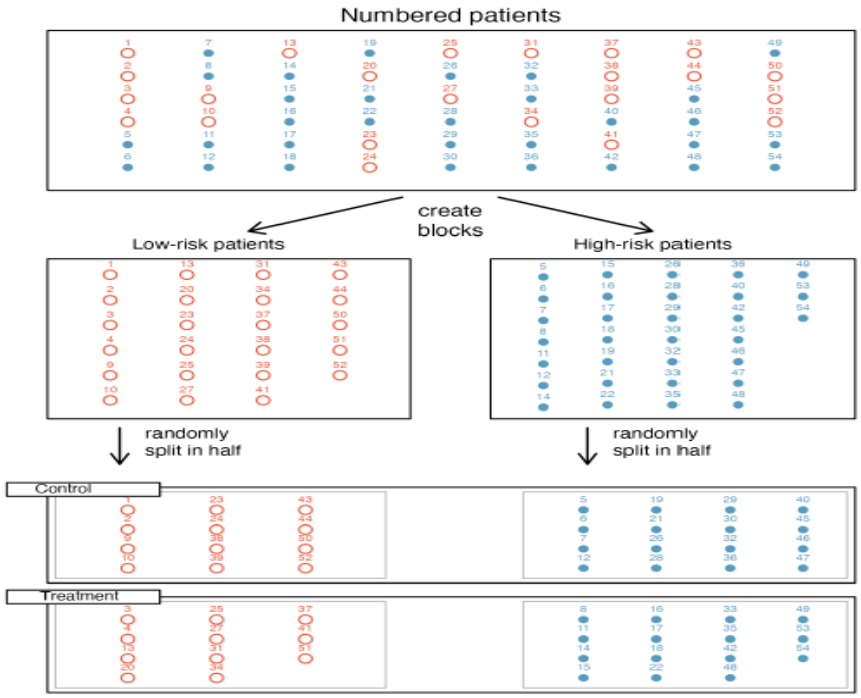


Figure 2.6: Blocking for patient risk. Patients are first divided into low-risk and high-risk blocks, then patients in each block are evenly randomized into the treatment groups. This strategy ensures equal representation of patients in each treatment group from both risk categories.

Experiments

Reducing bias in human experiments

Randomized experiments have long been considered to be the gold standard for data collection, but they do not ensure an unbiased perspective into the cause-and-effect relationship in all cases. Human studies are perfect examples where bias can unintentionally arise. Here we reconsider a study where a new drug was used to treat heart attack patients. In particular, researchers wanted to know if the drug reduced deaths in patients.

These researchers designed a randomized experiment because they wanted to draw causal conclusions about the drug's effect. Study volunteers[5] were randomly placed into two study groups. One group, the **treatment group**, received the drug. The other group, called the **control group**, did not receive any drug treatment.

Read more …. about placebo effect, blind study, double-blind setup

Observational studies

Studies where no treatment has been explicitly applied (or explicitly withheld) are called **observational studies**. For instance, studies on the loan data and county data described in Section 1.2 are would both be considered observational, as they rely on **observational data**.

Making causal conclusions based on experiments is often reasonable, since we can randomly assign the explanatory variable(s), i.e., the treatments. However, making the same causal conclusions based on observational data can be treacherous and is not recommended. Thus, observational studies are generally only sufficient to show associations or form hypotheses that can be later checked with experiments.

Suppose an observational study tracked sunscreen use and skin cancer, and it was found that the more sunscreen someone used, the more likely the person was to have skin cancer. Does this mean sunscreen *causes* skin cancer?

No! Some previous research tells us that using sunscreen actually reduces skin cancer risk, so maybe there is another variable that can explain this hypothetical association between sunscreen usage and skin cancer, as shown in Figure 2.7. One important piece of information that is absent is sun exposure. If someone is out in the sun all day, they are more likely to use sunscreen *and* more likely to get skin cancer. Exposure to the sun is unaccounted for in the simple observational investigation.

Observational studies

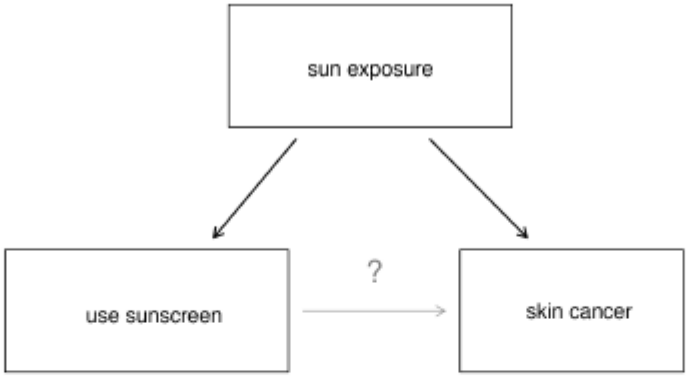
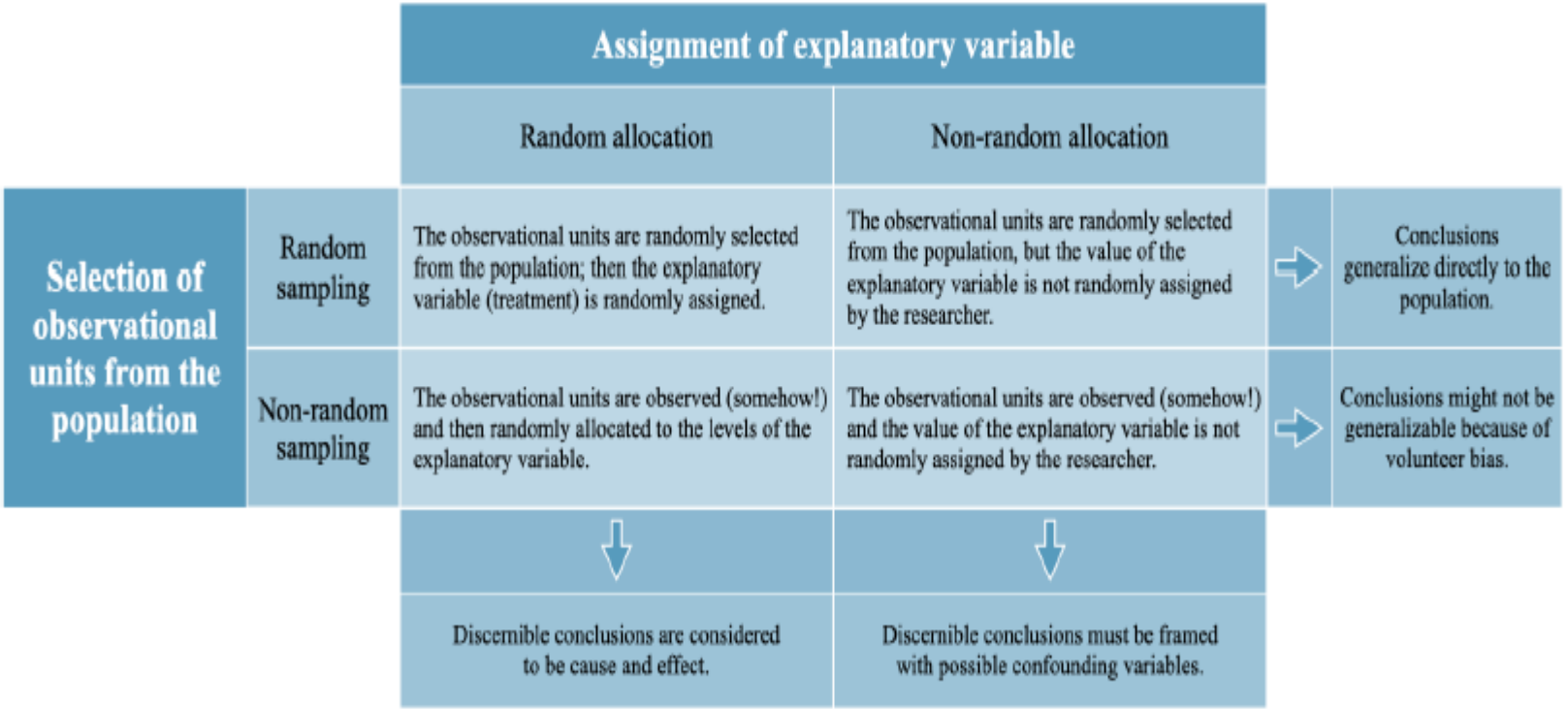


Figure 2.7: Sun exposure may be the root cause of both sunscreen use and skin cancer.

In this example, sun exposure is a confounding variable. The presence of confounding variables is what inhibits the ability for observational studies to make causal claims. While one method to justify making causal conclusions from observational studies is to exhaust the search for confounding variables, there is no guarantee that all confounding variables can be examined or measured.

Study design



		Assignment of explanatory variable		
		Random allocation	Non-random allocation	
Selection of observational units from the population	Random sampling	The observational units are randomly selected from the population; then the explanatory variable (treatment) is randomly assigned.	The observational units are randomly selected from the population, but the value of the explanatory variable is not randomly assigned by the researcher.	⇨ Conclusions generalize directly to the population.
	Non-random sampling	The observational units are observed (somehow!) and then randomly allocated to the levels of the explanatory variable.	The observational units are observed (somehow!) and the value of the explanatory variable is not randomly assigned by the researcher.	⇨ Conclusions might not be generalizable because of volunteer bias.
		⇩	⇩	
		Discernible conclusions are considered to be cause and effect.	Discernible conclusions must be framed with possible confounding variables.	

Figure 2.8: Analysis conclusions should be made carefully according to how the data were collected. Very few datasets come from the top left box because usually ethics require that random assignment of treatments can only be given to volunteers. Both representative (ideally random) sampling and experiments (random assignment of treatments) are important for how statistical conclusions can be made on populations.

Case study: Olympic 1500m

In this case study we introduce a dataset comparing Olympic and Paralympic gold medal finishers in the 1500m running competition (the Olympic "mile", if a bit shorter than a full mile).

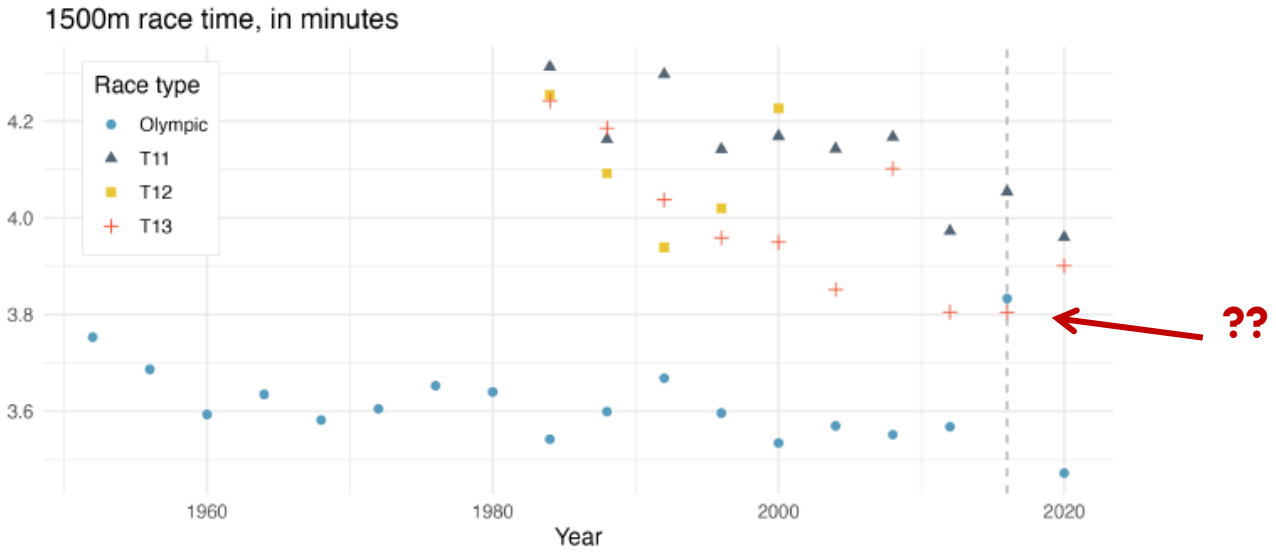


Figure 3.2: 1500m race time for Men's Olympic and Paralympic athletes. Dashed grey line represents the Rio Games in 2016.

The T11 athletes have almost complete visual impairment and are allowed to run with a guide-runner T12 and T13 athletes have some visual impairment

Case study: Olympic 1500m

Simson's paradox

Simpson's paradox is a description of three (or more) variables. The paradox happens when a third variable reverses the relationship between the first two variables.

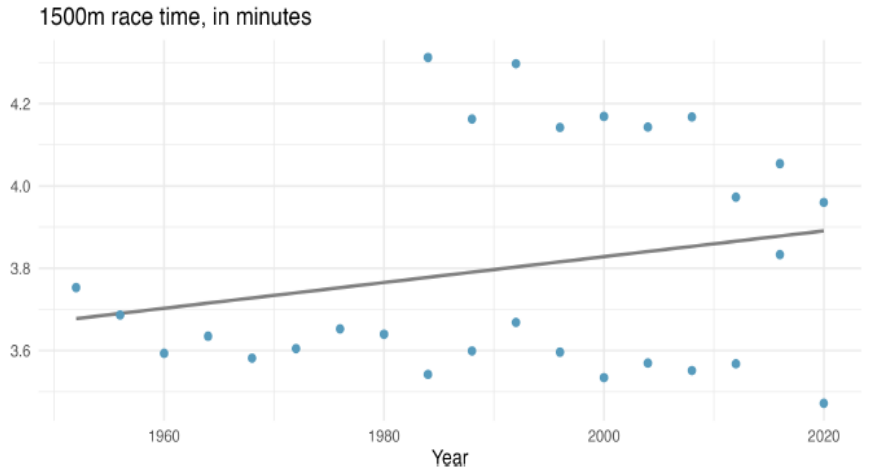


Figure 3.3: 1500m race time for Men's Olympic and Paralympic (T11) athletes. The line represents a line

Of course, both your eye and your intuition are likely telling you that it wouldn't make any sense to try to model all of the athletes together. Instead, a separate model should be run for each of the two types of Games: Olympic and Paralympic (T11).

Case study: Olympic 1500m

Simson's paradox

Simpson's paradox is a description of three (or more) variables. The paradox happens when a third variable reverses the relationship between the first two variables.

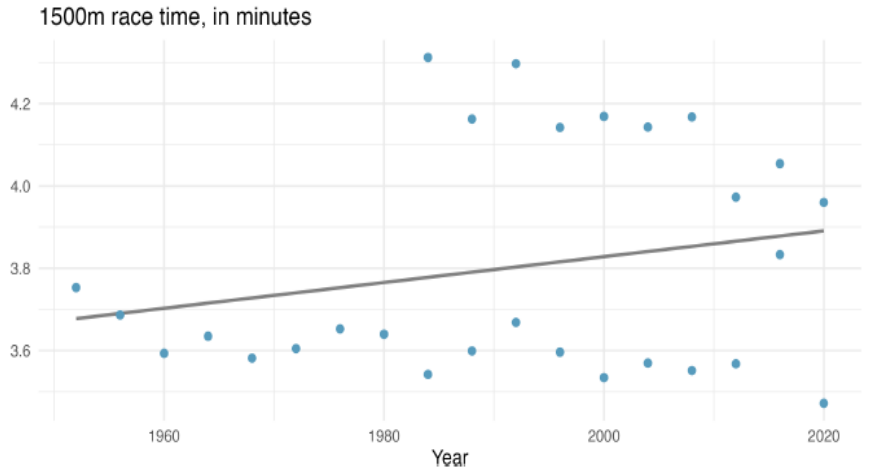


Figure 3.3: 1500m race time for Men's Olympic and Paralympic (T11) athletes. The line represents a line

Of course, both your eye and your intuition are likely telling you that it wouldn't make any sense to try to model all of the athletes together. Instead, a separate model should be run for each of the two types of Games: Olympic and Paralympic (T11).

Case study: Olympic 1500m

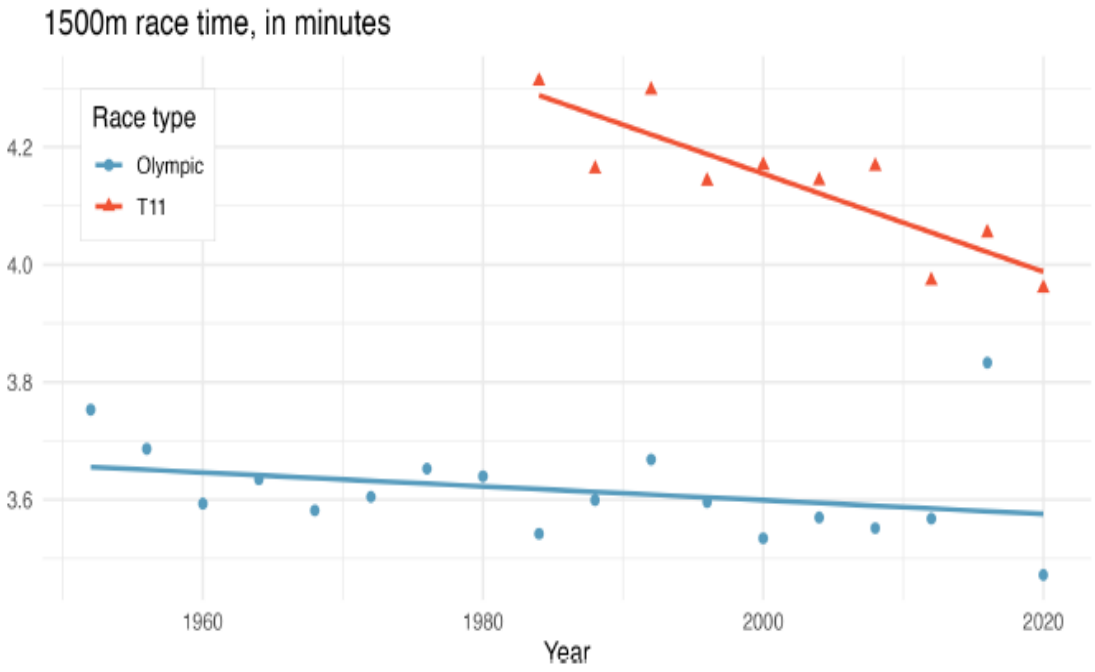


Figure 3.4: 1500m race time for Men's Olympic and Paralympic (T11) athletes. The best fit line is now fit separately to the Olympic and Paralympic athletes.

Case study: Olympic 1500m

Simpson's paradox.

Simpson's paradox happens when an association or relationship between two variables in one direction (e.g., positive) reverses (e.g., becomes negative) when a third variable is considered.

In the 1500m analysis, it would be most prudent to report the trends separately for the Olympic and the T11 athletes. However, in other situations, it might be better to aggregate the data and report the overall trend.

Exploratory data analysis

we skip it here, as covered it in the first lecture;

for interesting examples read chapters in the book

Regression modeling

- We skip most of it here, as covered during lectures in October; for more examples read chapters in the book
- Here included only what I found explained differently or not covered during lectures in October

Linear regression with single predictor

Linear regression is a very powerful statistical technique. Many people have some familiarity with regression models just from reading the news, where straight lines are overlaid on scatterplots. Linear models can be used for prediction or to describe the relationship between two numerical variables, assuming there is a linear relationship between them.

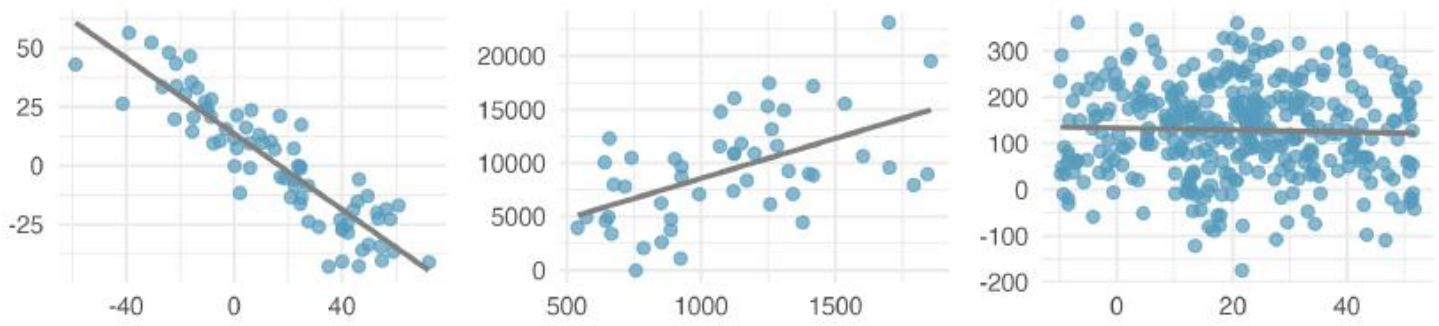


Figure 7.2: Three datasets where a linear model may be useful even though the data do not all fall exactly on the line.

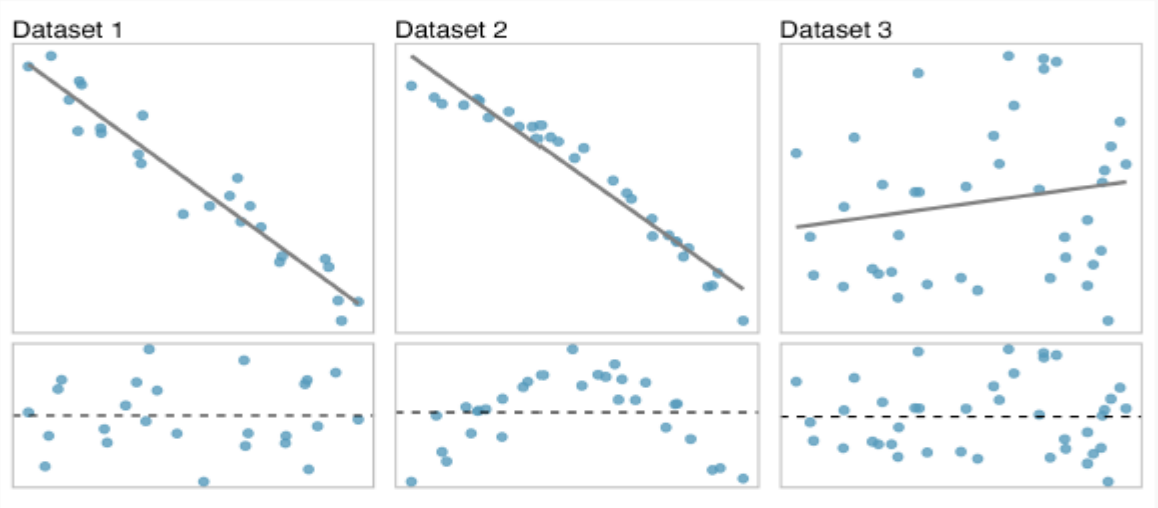
Residuals

Residuals are the leftover variation in the data after accounting for the model fit:

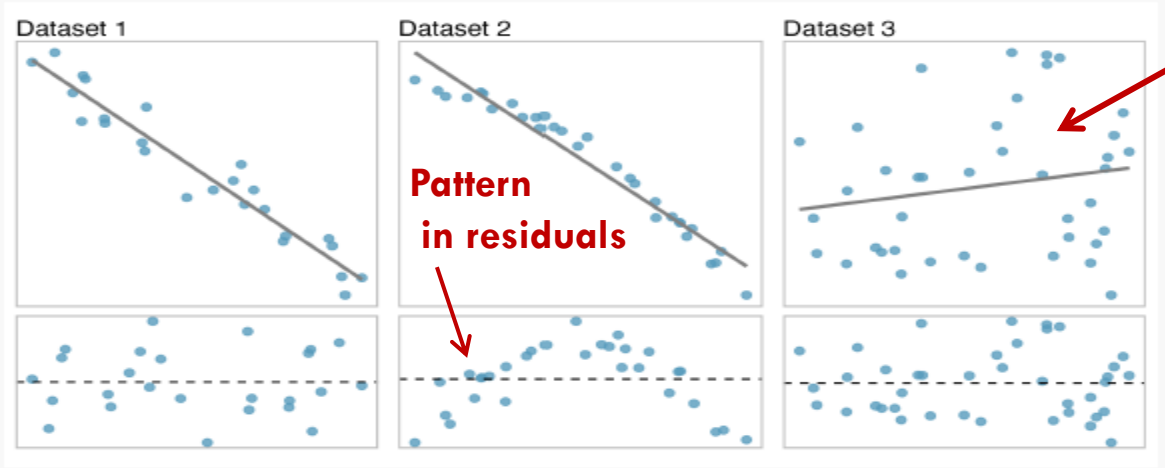
$$\text{Data} = \text{Fit} + \text{Residual}$$

EXAMPLE

One purpose of residual plots is to identify characteristics or patterns still apparent in data after fitting a model. The figure below shows three scatterplots with linear models in the first row and residual plots in the second row. Can you identify any patterns in the residuals?



Residuals



Dataset 1: the residuals show no obvious patterns. The residuals are scattered randomly around 0, represented by the dashed line.

Dataset 2: The second dataset shows a pattern in the residuals. There is some curvature in the scatterplot, which is more obvious in the residual plot. We should not use a straight line to model these data. Instead, a more advanced technique should be used to model the curved relationship, such as the variable transformations discussed in Section 5.7.

Dataset 3: The last plot shows very little upwards trend, and the residuals also show no obvious patterns. It is reasonable to try to fit a linear model to the data. However, it is unclear whether there is evidence that the slope parameter is different from zero. The point estimate of the slope parameter is not zero, but we might wonder if this could just be due to chance.

Describing linear relationship with correlations

Correlation: strength of a linear relationship.

Correlation which always takes values between -1 and 1, describes the strength and direction of the linear relationship between two variables. We denote the correlation by r.

The correlation value has no units and will not be affected by a linear change in the units (e.g., going from inches to centimeters).

We can compute the correlation using a formula

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} \frac{y_i - \bar{y}}{s_y}$$

where $\bar{x}$, $\bar{y}$, s_x, and s_y are the sample means and standard deviations for each variable.

Describing linear relationship with correlations

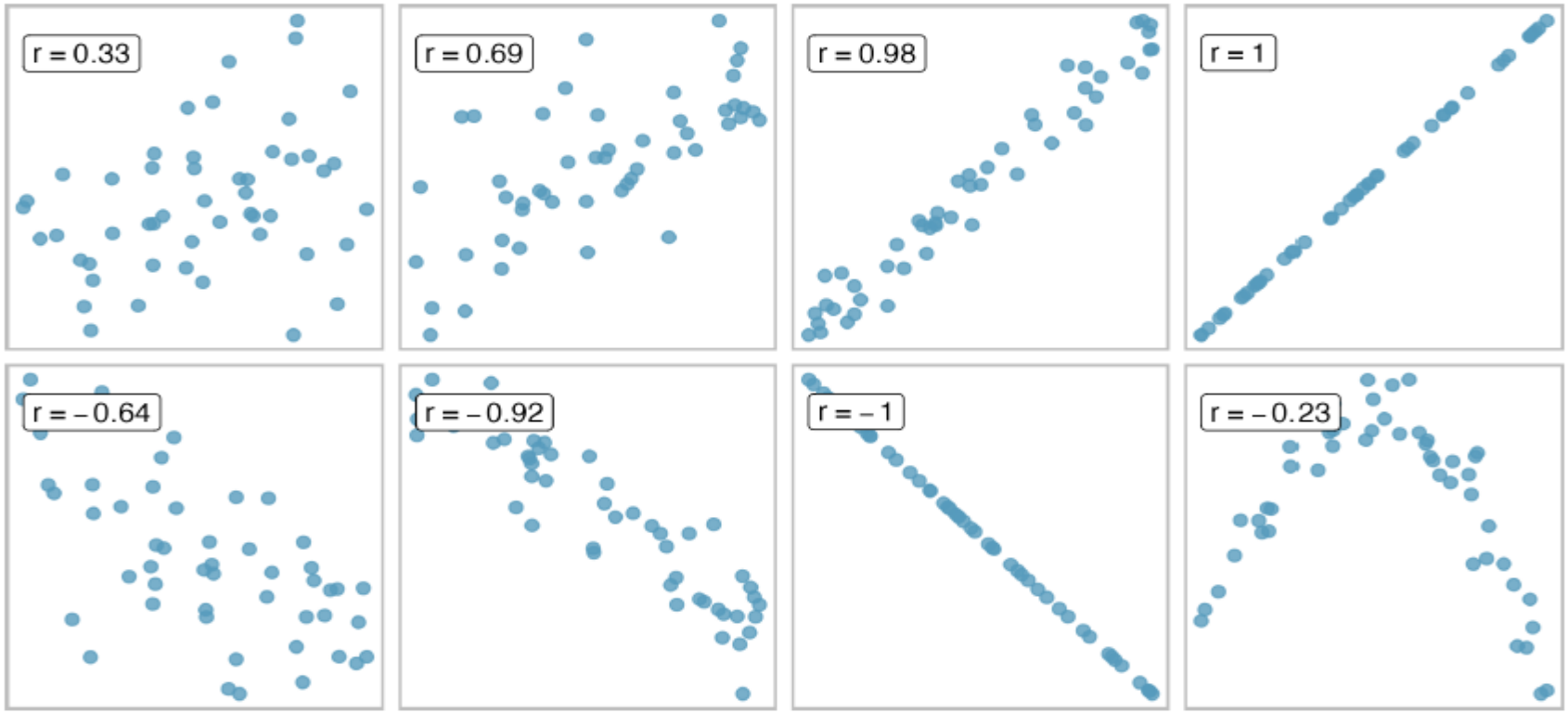


Figure 7.10: Sample scatterplots and their correlations. The first row shows variables with a positive relationship, represented by the trend up and to the right. The second row shows variables with a negative trend, where a large value in one variable is associated with a lower value in the other.

Describing linear relationship with correlations

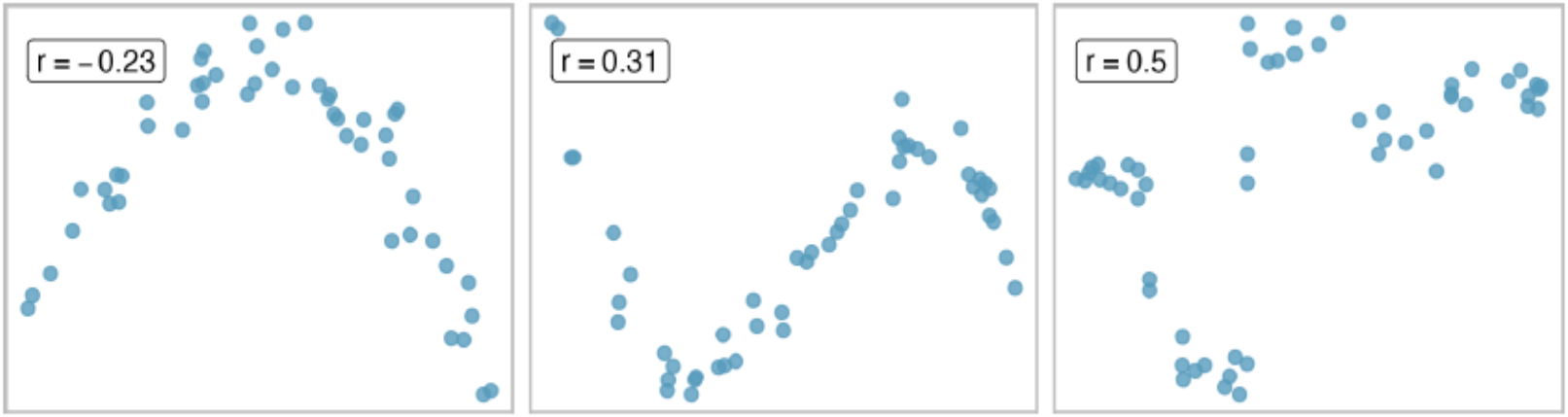


Figure 7.11: Sample scatterplots and their correlations. In each case, there is a strong relationship between the variables. However, because the relationship is not linear, the correlation is relatively weak.

Categorical predictors with two levels

Categorical variables are also useful in predicting outcomes. Here we consider a categorical predictor with two levels (recall that a *level* is the same as a *category*).

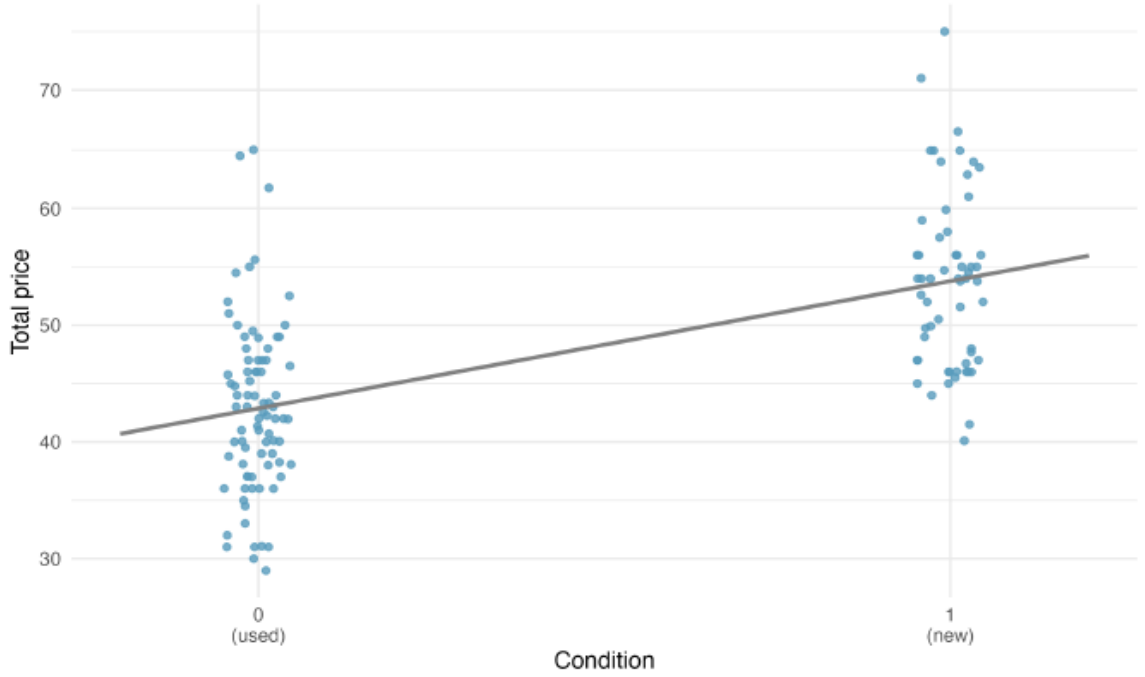


Figure 7.15: Total auction prices for the video game Mario Kart, divided into used ($x = 0$) and new ($x = 1$) condition games. The least squares regression line is also shown.

Categorical predictors with two levels

To incorporate the game condition variable into a regression equation, we must convert the categories into a numerical form. We will do so using an **indicator variable** called `condnew`, which takes value 1 when the game is new and 0 when the game is used. Using this indicator variable, the linear model may be written as

$$\widehat{\texttt{price}} = b_0 + b_1 \times \texttt{condnew}$$

The parameter estimates are given in Table 7.4.

Table 7.4: Least squares regression summary for the final auction price against the condition of the game.

term	estimate	std.error	statistic	p.value
(Intercept)	42.9	0.81	52.67	<0.0001
condnew	10.9	1.26	8.66	<0.0001

Using values from Table 7.4, the model equation can be summarized as

$$\widehat{\texttt{price}} = 42.87 + 10.9 \times \texttt{condnew}$$

Categorical predictors with two levels

Interpreting model estimates for categorical predictors.

The estimated intercept is the value of the outcome variable for the first category (i.e., the category corresponding to an indicator value of 0). The estimated slope is the average change in the outcome variable between the two categories.

The intercept is the estimated price when `condnew` has a value 0, i.e., when the game is in used condition. That is, the average selling price of a used version of the game is \$42.9. The slope indicates that, on average, new games sell for about \$10.9 more than used games.

Note that, fundamentally, the intercept and slope interpretations do not change when modeling categorical variables with two levels. However, when the predictor variable is binary, the coefficient estimates (b_0 and b_1) are directly interpretable with respect to the dataset at hand.

Coefficient of determination: R-squared

Sums of squares to measure variability in y.

We can measure the variability in the y values by how far they tend to fall from their mean, $\bar{y}$. We define this value as the total sum of squares, calculated using the formula below, where y_i represents each y value in the sample, and $\bar{y}$ represents the mean of the y values in the sample.

$$SST = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2.$$

Left-over variability in the y values if we know x can be measured by the sum of squared errors, or sum of squared residuals, calculated using the formula below, where $\hat{y}_i$ represents the predicted value of y_i based on the least squares regression.[8],

$$\begin{aligned} SSE &= (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2 \\ &= e_1^2 + e_2^2 + \cdots + e_n^2 \end{aligned}$$

The coefficient of determination can then be calculated as

$$R^2 = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

[8]The difference $SST - SSE$ is called the regression sum of squares, SSR, and can also be calculated as $SSR = (\hat{y}_1 - \bar{y})^2 + (\hat{y}_2 - \bar{y})^2 + \cdots + (\hat{y}_n - \bar{y})^2$. SSR represents the variation in y that was accounted for in our model.

Outliers

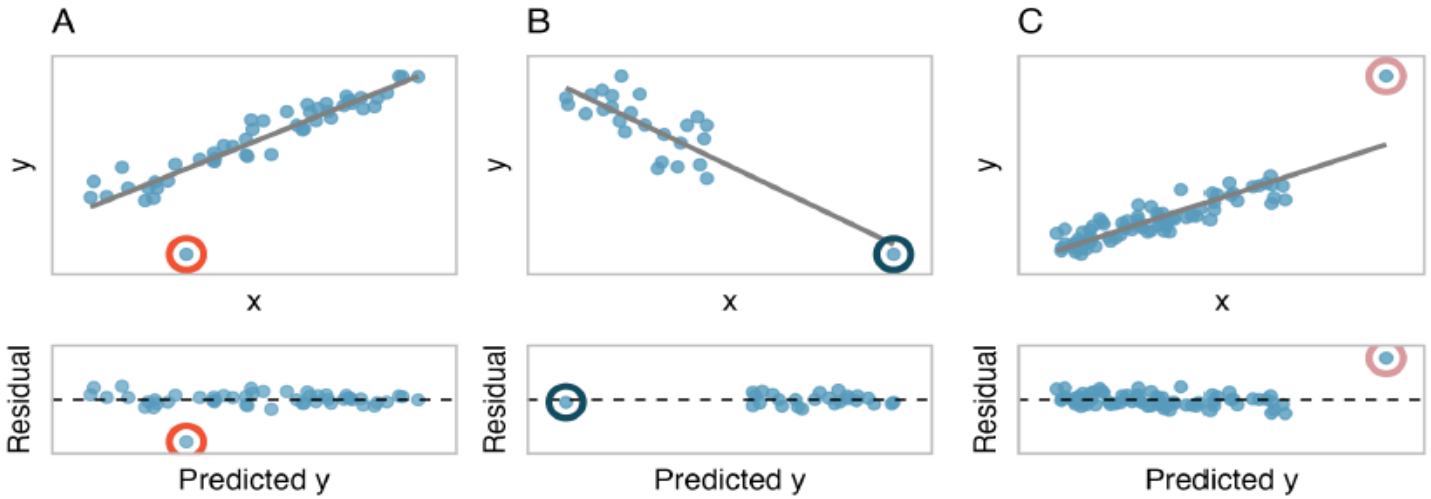
Types of outliers.

A point (or a group of points) that stands out from the rest of the data is called an outlier. Outliers that fall horizontally away from the center of the cloud of points are called leverage points. Outliers that influence on the slope of the line are called influential points.

Leverage.

Points that fall horizontally away from the center of the cloud tend to pull harder on the line, so we call them points with high leverage or leverage points.

Outliers

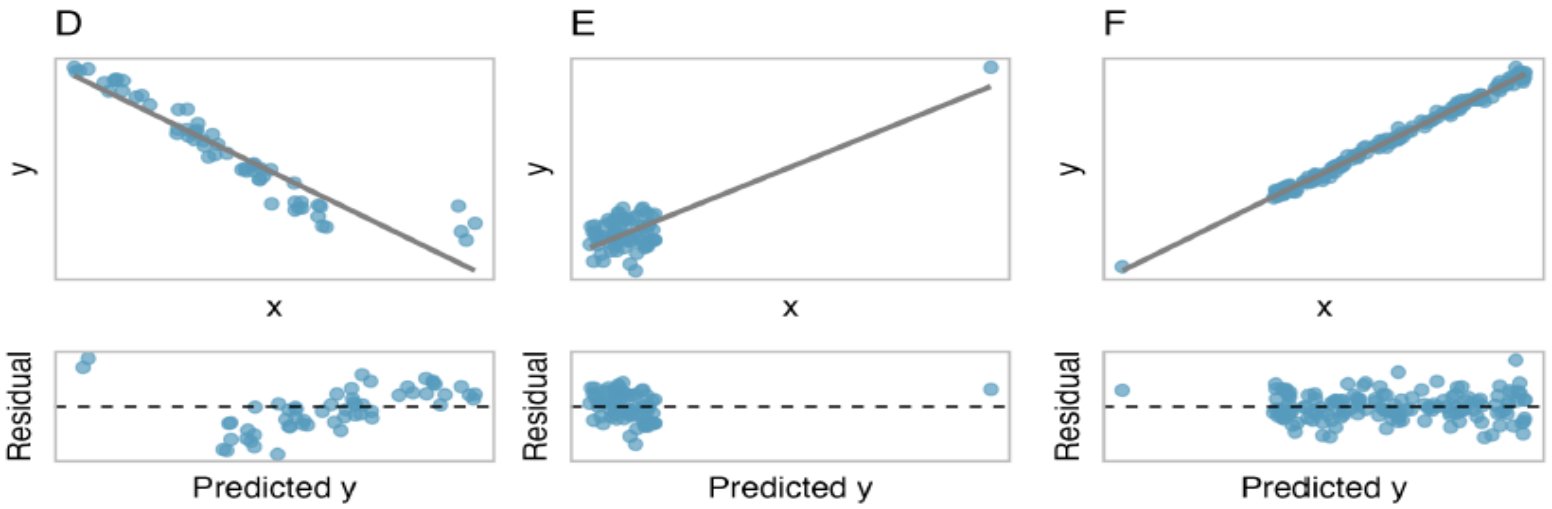


A: There is one outlier far from the other points (in the y direction and it is an outlier of the bivariate model), though it only appears to slightly influence the line.

B: There is one outlier on the right (in the x and y direction although it is not an outlier of the bivariate model), though it is quite close to the least squares line, which suggests it wasn't very influential.

C: There is one point far away from the cloud (in the x and y direction and an outlier of the bivariate model), and this outlier appears to pull the least squares line up on the right; examine how the line around the primary cloud does not appear to fit very well.

Outliers



D: There is a primary cloud and then a small secondary cloud of four outliers (with respect to both x and the bivariate model). The secondary cloud appears to be influencing the line somewhat strongly, making the least square line fit poorly almost everywhere. There might be an interesting explanation for the dual clouds, which is something that could be investigated.

E: There is no obvious trend in the main cloud of points and the outlier on the right (with respect to both x and y) appears to largely (and problematically) control the slope of the least squares line. The point creates a bivariate model when seemingly there is none.

F: There is one outlier far from the cloud (with respect to both x and y). However, it falls quite close to the least squares line and does not appear to be very influential (it is not outlying with respect to the bivariate model).

Outliers

A good practice for dealing with outlying observations is to produce two analyses: one with and one without the outlying observations. Presenting both analyses to a client and discussing the role of the outlying observations should lead you to a more holistic understanding of the appropriate model for the data.

Points that fall horizontally far from the line are points of high leverage; these points can strongly influence the slope of the least squares line. If one of these high leverage points does appear to actually invoke its influence on the slope of the line – as in Plots C, D, and E of Figure 7.16a and Figure 7.16b – then we call it an **influential point**. Usually we can say a point is influential if, had we fitted the line without it, the influential point would have been unusually far from the least squares line.

It is tempting to remove outliers. Don't do this without a very good reason. Models that ignore exceptional (and interesting) cases often perform poorly. For instance, if a financial firm ignored the largest market swings – the "outliers" – they would soon go bankrupt by making poorly thought-out investments.

Linear regression with multiple predictors

Multiple regression extends single predictor variable regression to the case that still has one response but many predictors (denoted x_1, x_2, x_3, ...). The method is motivated by scenarios where many variables may be simultaneously connected to an output.

Building on the ideas of one predictor variable in a linear regression model (from Chapter 7), a multiple linear regression model is now fit to two or more predictor variables. By considering how different explanatory variables interact, we can uncover complicated relationships between the predictor variables and the response variable. One challenge to working with multiple variables is that it is sometimes difficult to know which variables are most important to include in the model. Model building is an extensive topic, and we scratch the surface here by defining and utilizing the adjusted R^2 value.

Linear regression with multiple predictors

The dataset includes information on 10000 loans

Table 8.1: First six rows of the **loans** dataset.

interest_rate	verified_income	debt_to_income	credit_util	bankruptcy	term	credit_checks	issue_month
14.07	Verified	18.01	0.548	0	60	6	Mar-2018
12.61	Not Verified	5.04	0.150	1	36	1	Feb-2018
17.09	Source Verified	21.15	0.661	0	36	4	Feb-2018
6.72	Not Verified	10.16	0.197	0	36	0	Jan-2018
14.07	Verified	57.96	0.755	0	36	7	Mar-2018
6.72	Not Verified	6.46	0.093	0	36	6	Jan-2018

Variable	Description
`interest_rate`	Interest rate on the loan, in an annual percentage.
`verified_income`	Categorical variable describing whether the borrower's income source and amount have been verified, with levels Verified (source and amount verified), Source Verified (source only verified), and Not Verified.
`debt_to_income`	Debt-to-income ratio, which is the percentage of total debt of the borrower divided by their total income.
`credit_util`	Of all the credit available to the borrower, what fraction are they utilizing. For example, the credit utilization on a credit card would be the card's balance divided by the card's credit limit.
`bankruptcy`	An indicator variable for whether the borrower has a past bankruptcy in their record. This variable takes a value of '1' if the answer is *yes* and '0' if the answer is *no*.
`term`	The length of the loan, in months.
`issue_month`	The month and year the loan was issued, which for these loans is always during the first quarter of 2018.
`credit_checks`	Number of credit checks in the last 12 months. For example, when filing an application for a credit card, it is common for the company receiving the application to run a credit check.

Linear regression with multiple predictors

Indicator and categorical predictors

Let's start by fitting a linear regression model for interest rate with a single predictor indicating whether a person has a bankruptcy in their record:

$$\widehat{\texttt{interest\_rate}} = 12.34 + 0.74 \times \texttt{bankruptcy}$$

Table 8.4: Summary of a linear model for predicting **interest_rate** based on whether the borrower has a bankruptcy in their record. Degrees of freedom for this model is 9998.

term	estimate	std.error	statistic	p.value
(Intercept)	12.34	0.05	231.49	<0.0001
bankruptcy1	0.74	0.15	4.82	<0.0001

The variable takes one of two values: 1 when the borrower has a bankruptcy in their history and 0 otherwise. A slope of 0.74 means that the model predicts a 0.74% higher interest rate for those borrowers with a bankruptcy in their record.

Linear regression with multiple predictors

Indicator and categorical predictors

Suppose we had fit a model using a 3-level categorical variable, such as `verified_income`. The output from software is shown in Table 8.5. This regression output provides multiple rows for the variable. Each row represents the relative difference for each level of `verified_income`. However, we are missing one of the levels: `Not Verified`. The missing level is called the **reference level** and it represents the default level that other levels are measured against.

Table 8.5: Summary of a linear model for predicting `interest_rate` from the borrower's income source and amount verification. This predictor has three levels, which results in 2 rows in the regression output.

term	estimate	std.error	statistic	p.value
`(Intercept)`	11.10	0.08	137.2	<0.0001
`verified_incomeSource Verified`	1.42	0.11	12.8	<0.0001
`verified_incomeVerified`	3.25	0.13	25.1	<0.0001

Linear regression with multiple predictors

Table 8.5: Summary of a linear model for predicting `interest_rate` from the borrower's income source and amount verification. This predictor has three levels, which results in 2 rows in the regression output.

term	estimate	std.error	statistic	p.value
(Intercept)	11.10	0.08	137.2	<0.0001
verified_incomeSource Verified	1.42	0.11	12.8	<0.0001
verified_incomeVerified	3.25	0.13	25.1	<0.0001

EXAMPLE

How would we write an equation for this regression model?

The equation for the regression model may be written as a model with two predictors:

$$\widehat{\texttt{interest\_rate}} = 11.10 + 1.42 \times \texttt{verified\_income}_{\texttt{Source Verified}} + 3.25 \times \texttt{verified\_income}_{\texttt{Verified}}$$

We use the notation $\texttt{variable}_{\texttt{level}}$ to represent indicator variables for when the categorical variable takes a particular value. For example, $\texttt{verified\_income}_{\texttt{Source Verified}}$ would take a value of 1 if it was for a borrower that was source verified, and it would take a value of 0 otherwise. Likewise, $\texttt{verified\_income}_{\texttt{Verified}}$ would take a value of 1 if it was for a borrower that was verified, and 0 if it took any other value.

Linear regression with multiple predictors

When `verified_income` takes a value of `Not Verified`, then both indicator functions in the equation for the linear model are set to 0:

$$\widehat{\texttt{interest\_rate}} = 11.10 + 1.42 \times 0 + 3.25 \times 0 = 11.10$$

The average interest rate for these borrowers is 11.1%. Because the level does not have its own coefficient and it is the reference value, the indicators for the other levels for this variable all drop out.

When `verified_income` takes a value of `Source Verified`, then the corresponding variable takes a value of 1 while the other is 0:

$$\widehat{\texttt{interest\_rate}} = 11.10 + 1.42 \times 1 + 3.25 \times 0 = 12.52$$

The average interest rate for these borrowers is 12.52%.

Linear regression with multiple predictors

The higher interest rate for borrowers who have verified their income source or amount is surprising. Intuitively, we would think that a loan would look *less* risky if the borrower's income has been verified. However, note that the situation may be more complex, and there may be confounding variables that we didn't account for. For example, perhaps lenders require borrowers with poor credit to verify their income. That is, verifying income in our dataset might be a signal of some concerns about the borrower rather than a reassurance that the borrower will pay back the loan. For this reason, the borrower could be deemed higher risk, resulting in a higher interest rate.

[1]When `verified_income` takes a value of `Verified`, then the corresponding variable takes a value of 1 while the other is 0: $11.10 + 1.42 \times 0 + 3.25 \times 1 = 14.35$. The average interest rate for these borrowers is 14.35%.

[2]Each of the coefficients gives the incremental interest rate for the corresponding level relative to the `Not Verified` level, which is the reference level. For example, for a borrower whose income source and amount have been verified, the model predicts that they will have a 3.25% higher interest rate than a borrower who has not had their income source or amount verified.

[3]Relative to the `Not Verified` category, the `Verified` category has an interest rate of 3.25% higher, while the `Source Verified` category is only 1.42% higher. Thus, `Verified` borrowers will tend to get an interest rate about 3.25 higher than `Source Verified` borrowers.

Linear regression with multiple predictors

Predictors with several categories.

Software = R package

When fitting a regression model with a categorical variable that has k levels where $k > 2$, software will provide a coefficient for $k - 1$ of those levels. For the last level that does not receive a coefficient, this is the reference level, and the coefficients listed for the other levels are all considered relative to this reference level.

The higher interest rate for borrowers who have verified their income source or amount is surprising. Intuitively, we would think that a loan would look *less* risky if the borrower's income has been verified. However, note that the situation may be more complex, and there may be confounding variables that we didn't account for. For example, perhaps lenders require borrowers with poor credit to verify their income. That is, verifying income in our dataset might be a signal of some concerns about the borrower rather than a reassurance that the borrower will pay back the loan. For this reason, the borrower could be deemed higher risk, resulting in a higher interest rate.

Many predictors in the model

The world is complex, and it can be helpful to consider many factors at once in statistical modeling.

We want to construct a model that accounts not only for any past bankruptcy or whether the borrower had their income source or amount verified, but simultaneously accounts for all the variables in the `loans` dataset: `verified_income`, `debt_to_income`, `credit_util`, `bankruptcy`, `term`, `issue_month`, and `credit_checks`.

$$\begin{aligned}\widehat{\texttt{interest\_rate}} = b_0 &\\ &+ b_1 \times \texttt{verified\_income}_{\texttt{Source Verified}} + b_2 \times \texttt{verified\_income}_{\texttt{Verified}} \\ &+ b_3 \times \texttt{debt\_to\_income} + b_4 \times \texttt{credit\_util} \\ &+ b_5 \times \texttt{bankruptcy} + b_6 \times \texttt{term} \\ &+ b_7 \times \texttt{credit\_checks} + b_8 \times \texttt{issue\_month}_{\texttt{Jan-2018}} \\ &+ b_9 \times \texttt{issue\_month}_{\texttt{Mar-2018}}\end{aligned}$$

This equation represents a holistic approach for modeling all of the variables simultaneously. Notice that there are two coefficients for `verified_income` and two coefficients for `issue_month`, since both are 3-level categorical variables.

Many predictors in the model

The world is complex, and it can be helpful to consider many factors at once in statistical modeling.

The fitted model for the interest rate is given by:

$$\begin{aligned}\widehat{\texttt{interest\_rate}} = 1.89 \\ &+ 1.00 \times \texttt{verified\_income}_{\texttt{Source Verified}} + 2.56 \times \texttt{verified\_income}_{\texttt{Verified}} \\ &+ 0.02 \times \texttt{debt\_to\_income} + 4.90 \times \texttt{credit\_util} \\ &+ 0.39 \times \texttt{bankruptcy} + 0.15 \times \texttt{term} \\ &+ 0.23 \times \texttt{credit\_checks} + 0.05 \times \texttt{issue\_month}_{\texttt{Jan-2018}} \\ &- 0.04 \times \texttt{issue\_month}_{\texttt{Mar-2018}}\end{aligned}$$

If we count up the number of predictor coefficients, we get the *effective* number of predictors in the model; there are nine of those. Notice that the categorical predictor counts as two, once for each of the two levels shown in the model. In general, a categorical predictor with p different levels will be represented by $p-1$ terms in a multiple regression model. A total of seven variables were used as predictors to fit this model: `verified_income`, `debt_to_income`, `credit_util`, `bankruptcy`, `term`, `credit_checks`, `issue_month`.

Adjusted R-squared

Adjusted R-squared as a tool for model assessment.

The adjusted R-squared is computed as

$$R^2_{adj} = 1 - \frac{s^2_{\text{residuals}}/(n-k-1)}{s^2_{\text{outcome}}/(n-1)} = 1 - \frac{s^2_{\text{residuals}}}{s^2_{\text{outcome}}} \times \frac{n-1}{n-k-1}$$

where n is the number of observations used to fit the model and k is the number of predictor variables in the model. Remember that a categorical predictor with p levels will contribute $p-1$ to the number of variables in the model.

Stepwise selection using adjusted R^2 as the decision criteria is one of many commonly used model selection strategies. Stepwise selection can also be carried out with decision criteria other than adjusted R^2, such as p-values, AIC (Akaike information criterion) or BIC (Bayesian information criterion)

Alternatively, one could choose to include or exclude predictors from a model based on expert opinion or due to research focus. In fact, many statisticians discourage the use of stepwise regression *alone* for model selection and advocate, instead, for a more thoughtful approach that carefully considers the research focus and features of the data.

Logistic regression

In this chapter we introduce **logistic regression** as a tool for building models when there is a categorical response variable with two levels, e.g., yes and no. Logistic regression is a type of **generalized linear model (GLM)** for response variables where regular multiple regression does not work very well. GLMs can be thought of as a two-stage modeling approach. We first model the response variable using a probability distribution, such as the binomial or Poisson distribution. Second, we model the parameter of the distribution using a collection of predictors and a special form of multiple regression. Ultimately, the application of a GLM will feel very similar to multiple regression, even if some of the details are different.

Logistic regression: discrimination of hiring case

We will consider experiment data from a study that sought to understand the effect of race and sex on job application callback rates (Bertrand and Mullainathan 2003).

Table 9.1: List of all 36 unique names along with the commonly inferred race and sex associated with these names.

first_name	race	sex	first_name	race	sex	first_name	race	sex
Aisha	Black	female	Hakim	Black	male	Laurie	White	female
Allison	White	female	Jamal	Black	male	Leroy	Black	male
Anne	White	female	Jay	White	male	Matthew	White	male
Brad	White	male	Jermaine	Black	male	Meredith	White	female
Brendan	White	male	Jill	White	female	Neil	White	male
Brett	White	male	Kareem	Black	male	Rasheed	Black	male
Carrie	White	female	Keisha	Black	female	Sarah	White	female
Darnell	Black	male	Kenya	Black	female	Tamika	Black	female
Ebony	Black	female	Kristen	White	female	Tanisha	Black	female
Emily	White	female	Lakisha	Black	female	Todd	White	male
Geoffrey	White	male	Latonya	Black	female	Tremayne	Black	male
Greg	White	male	Latoya	Black	female	Tyrone	Black	male

Race and sex are protected classes in the United States, meaning they are not legally permitted factors for hiring or employment decisions.

The response variable of interest is whether there was a callback from the employer for the applicant

Logistic regression: discrimination of hiring case

Table 9.2: Descriptions of nine variables from the `resume` dataset. Many of the variables are indicator variables, meaning they take the value 1 if the specified characteristic is present and 0 otherwise.

variable	description
`received_callback`	Specifies whether the employer called the applicant following submission of the application for the job.
`job_city`	City where the job was located: Boston or Chicago.
`college_degree`	An indicator for whether the resume listed a college degree.
`years_experience`	Number of years of experience listed on the resume.
`honors`	Indicator for the resume listing some sort of honors, e.g. employee of the month.
`military`	Indicator for if the resume listed any military experience.
`has_email_address`	Indicator for if the resume listed an email address for the applicant.
`race`	Race of the applicant, implied by their first name listed on the resume.
`sex`	Sex of the applicant (limited to only man and woman), implied by the first name listed on the resume.

All of the attributes listed on each resume were randomly assigned, which means that no attributes that might be favorable or detrimental to employment would favor one demographic over another on these resumes. Importantly, due to the experimental nature of the study, we can infer causation between these variables and the callback rate, if substantial differences are found. Our analysis will allow us to compare the practical importance of each of the variables relative to each other.

Logistic regression

Logistic regression is a generalized linear model where the outcome is a two-level categorical variable. The outcome, Y_i, takes the value 1 (in our application, the outcome represents a callback for the resume) with probability p_i and the value 0 with probability $1 - p_i$. Because each observation has a slightly different context, e.g., different education level or a different number of years of experience, the probability p_i will differ for each observation. Ultimately, it is the **probability** of the outcome taking the value 1 (i.e., being a "success") that we model in relation to the predictor variables: we will examine which resume characteristics correspond to higher or lower callback rates.

Notation for a logistic regression model.

The outcome variable for a GLM is denoted by Y_i, where the index i is used to represent observation i. In the resume application, Y_i will be used to represent whether resume i received a callback ($Y_i = 1$) or not ($Y_i = 0$).

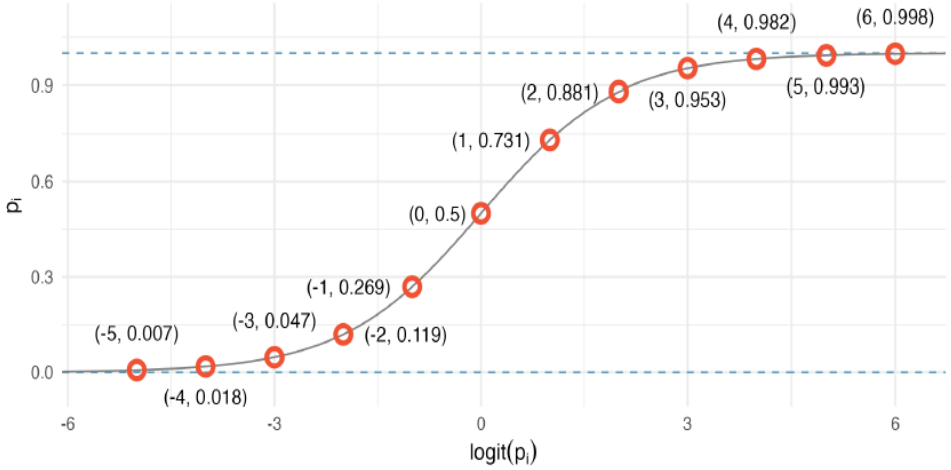
$$transformation(p_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i}$$

Logistic regression

We want to choose a **transformation** in the equation that makes practical and mathematical sense. A common transformation for p_i is the **logit transformation**, which may be written as

$$logit(p_i) = \log_e\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i}$$

The **logit transformation** is shown in Figure 9.1. Below, we rewrite the equation relating Y_i to its predictors using the logit transformation of p_i:



Modeling the probability of an event

We want to choose a **transformation** in the equation that makes practical and mathematical sense.

To convert from values on the logistic regression scale to the probability scale, we need to back transform and then solve for p_i:

$$\log_e\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}$$

$$\frac{p_i}{1-p_i} = e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}$$

$$p_i = (1-p_i)\, e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}$$

$$p_i = e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}} - p_i \times e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}$$

$$p_i + p_i\, e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}} = e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}$$

$$p_i(1 + e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}) = e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}$$

$$p_i = \frac{e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}}{1 + e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}}$$

As with most applied data problems, we substitute in the point estimates (the observed b_i) to calculate relevant probabilities.

Modeling the probability of an event

We start by fitting a model with a single predictor: **honors**. This variable indicates whether the applicant had any type of honors listed on their resume, such as employee of the month. A logistic regression model was fit using statistical software and the following model was found:

$$\log_e\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = -2.4998 + 0.8668 \times \textbf{honors}$$

a. If a resume is randomly selected from the study and it does not have any honors listed, what is the probability it resulted in a callback?

b. What would the probability be if the resume did list some honors?

---

a. If a randomly chosen resume from those sent out is considered, and it does not list honors, then **honors** takes the value of 0 and the right side of the model equation equals -2.4998. Solving for p_i: $\frac{e^{-2.4998}}{1+e^{-2.4998}} = 0.076$. Just as we labeled a fitted value of y_i with a "hat" in single-variable and multiple regression, we do the same for this probability: $\hat{p}_i = 0.076$.

b. If the resume had listed some honors, then the right side of the model equation is $-2.4998 + 0.8668 \times 1 = -1.6330$, which corresponds to a probability $\hat{p}_i = 0.163$. Notice that we could examine -2.4998 and -1.6330 in Figure 9.1 to estimate the probability before formally calculating the value.

Modeling the probability of an event

Table 9.6: Summary table for the logistic regression model for the resume callback example, where vari selection has been performed using AIC and `college_degree` has been dropped from the model.

term	estimate	std.error	statistic	p.value
(Intercept)	-2.72	0.16	-17.51	<0.0001
job_cityChicago	-0.44	0.11	-3.83	1e-04
years_experience	0.02	0.01	2.02	0.043
honors1	0.76	0.19	4.12	<0.0001
military1	-0.34	0.22	-1.60	0.1105
has_email_address1	0.22	0.11	1.97	0.0494
raceWhite	0.44	0.11	4.10	<0.0001
sexman	-0.20	0.14	-1.45	0.1473

The `race` variable had taken only two levels: `Black` and `White`. Based on the model results, what does the coefficient of the `race` variable say about callback decisions?

The coefficient shown corresponds to the level of `White`, and it is positive. The positive coefficient reflects a positive gain in callback rate for resumes where the candidate's first name implied they were White. The model results suggest that prospective employers favor resumes where the first name is typically interpreted to be White.

Modeling the probability of an event

EXAMPLE

Use the model summarized in Table 9.6 to estimate the probability of receiving a callback for a job in Chicago where the candidate lists 14 years experience, no honors, no military experience, includes an email address, and has a first name that implies they are a White male.

We can start by writing out the equation using the coefficients from the model:

$$log_e\left(\frac{\hat{p}}{1-\hat{p}}\right) = -2.7162 - 0.4364 \times \texttt{job\_city}_{\texttt{Chicago}} + 0.0206 \times \texttt{years\_experience} + 0.7634 \times \texttt{honors} - 0.3443 \times \texttt{military} + 0.2221 \times \texttt{email} + 0.4429 \times \texttt{race}_{\texttt{White}} - 0.1959 \times \texttt{sex}_{\texttt{man}}$$

Now we can add in the corresponding values of each variable for the individual of interest:

$$log_e\left(\frac{\hat{p}}{1-\hat{p}}\right) = -2.7162 - 0.4364 \times 1 + 0.0206 \times 14 + 0.7634 \times 0 - 0.3443 \times 0 + 0.2221 \times 1 + 0.4429 \times 1 - 0.1959 \times 1 = -2.3955$$

We can now back-solve for $\hat{p}$: the chance such an individual will receive a callback is about $\frac{e^{-2.3955}}{1+e^{-2.3955}} = 0.0835$.

Foundation for statistical inference

Statistical inference

- **The key concept in statistics is making conclusions about the population using information in a sample; the process is called statistical inference.**
- **By using computational methods as well as well develloped mathematical theory we can understand how one dataset differs from a different dataset –even if two dataset were collected under identical settings.**
- **Statistical inference is primarily concerned with qunatifying and understanding the uncertainty of parameter estimates. While the equations and details changes depending on the setting, the foundations for inference are the same through all the statistics.**

Idea of randomization test

- **Lets start with small case study:**
  - **gender discrimination**

- 48 male bank supervisors given the same personnel file, asked to judge whether the person should be promoted
- files were identical, except for gender of applicant
- random assignment
- 35 / 48 promoted
- are females are unfairly discriminated against?

Statistical inference: case study

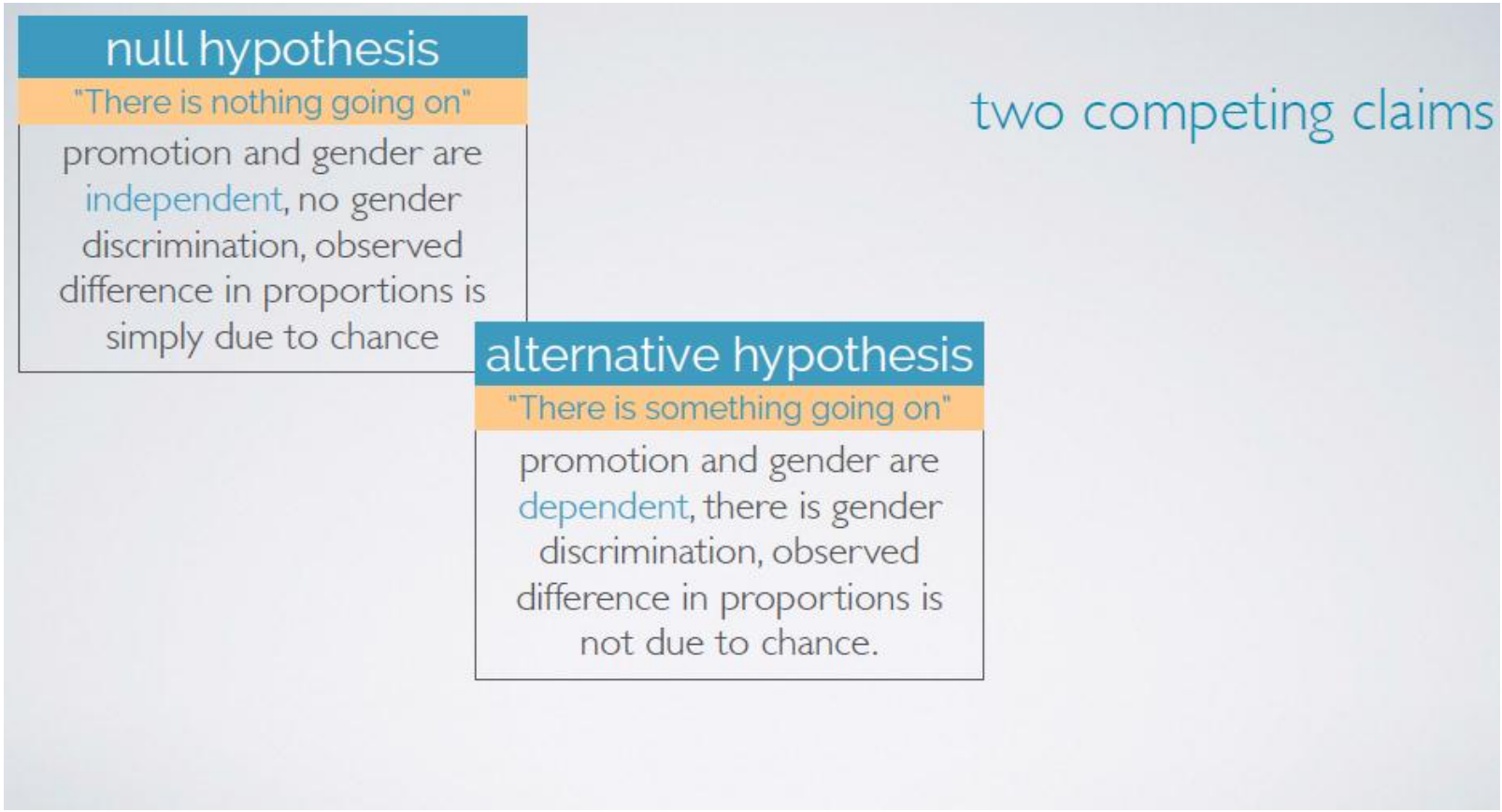
data

		promotion		
		promoted	not promoted	total
gender	male	21	3	24
	female	14	10	24
	total	35	13	48

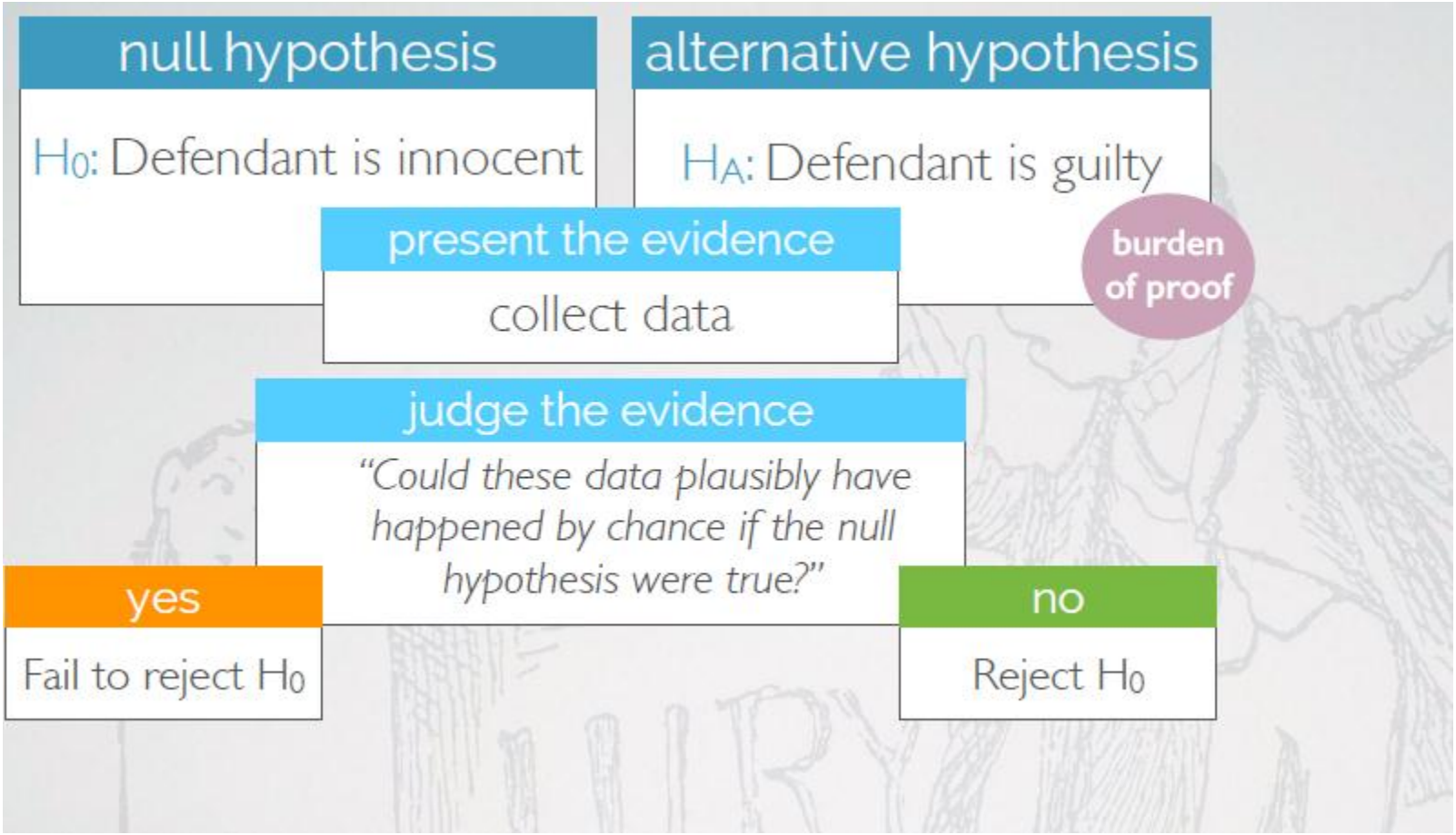
% of males promoted = 21/24 ≈ 88%

% of females promoted = 14/24 ≈ 58%

Statistical inference: case study



Statistical inference: case study



Statistical inference: case study

recap: hypothesis testing framework

- start with a null hypothesis (H_0) that represents the status quo
- set an alternative hypothesis (H_A) that represents the research question, i.e. what we're testing for
- conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation or theoretical methods
  - if the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, stick with the null hypothesis
  - if they do, then reject the null hypothesis in favor of the alternative

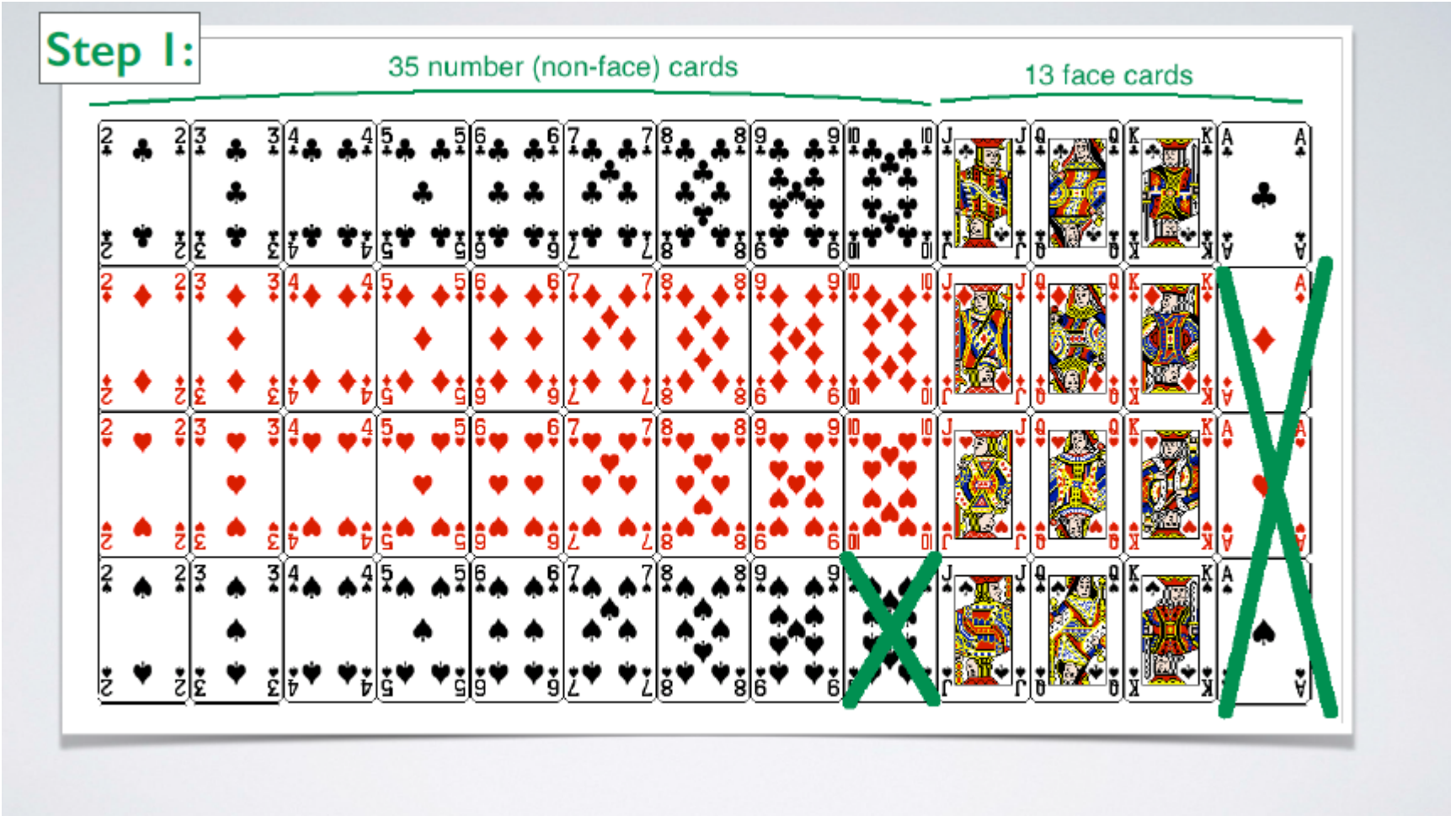
Statistical inference: case study

simulation scheme

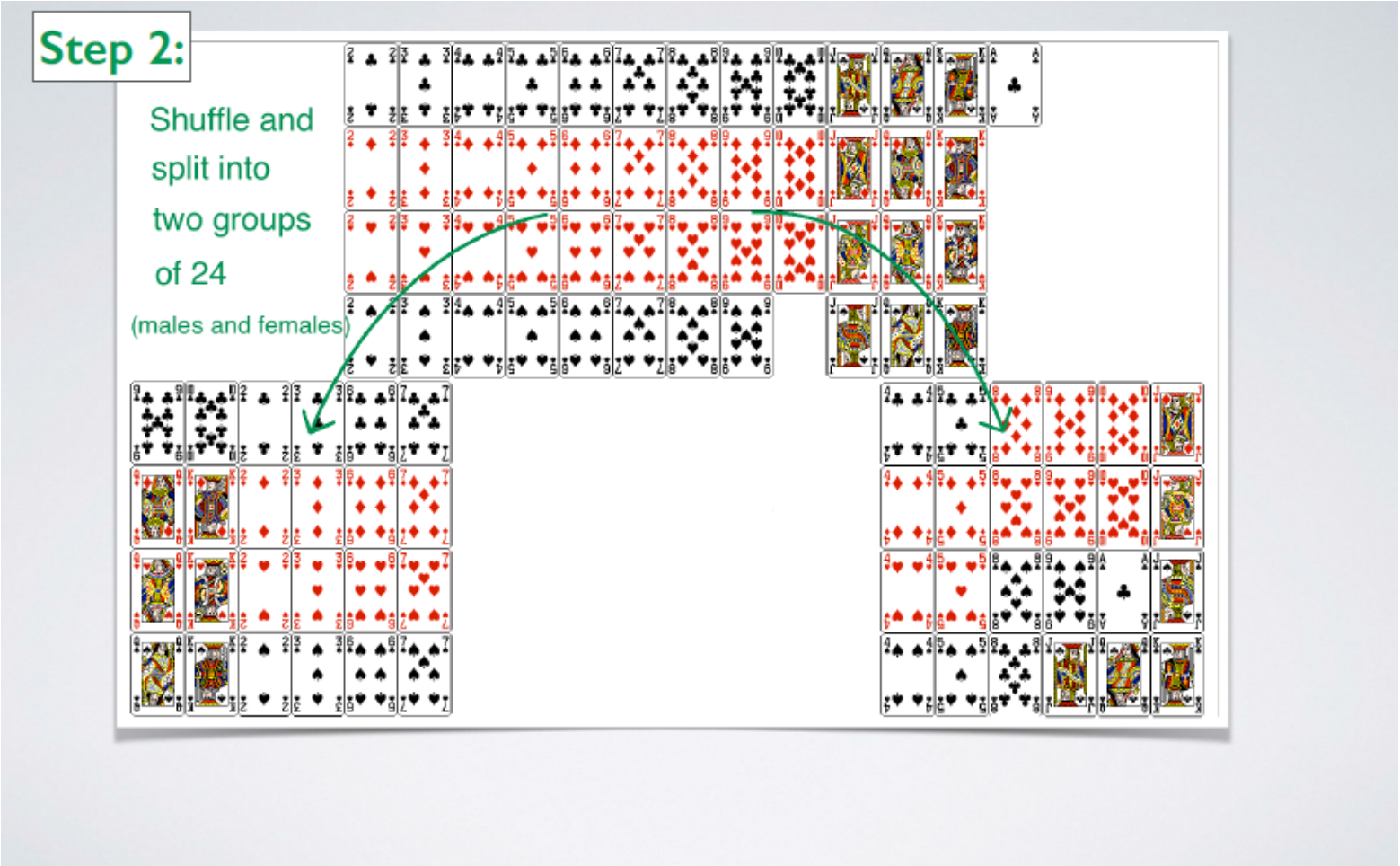
[use a deck of playing cards to simulate this experiment]

1. face card: not promoted, non-face card: promoted
   - set aside the jokers, consider aces as face cards
   - take out 3 aces → exactly 13 face cards left in the deck (face cards: A, K, Q, J)
   - take out a number card → 35 number (non-face) cards left in the deck (number cards: 2-10)
2. shuffle the cards, deal into two groups of size 24, representing males and females
3. count how many number cards are in each group (representing promoted files)
4. calculate the proportion of promoted files in each group, take the difference (male - female), and record this value
5. repeat steps 2 - 4 many times

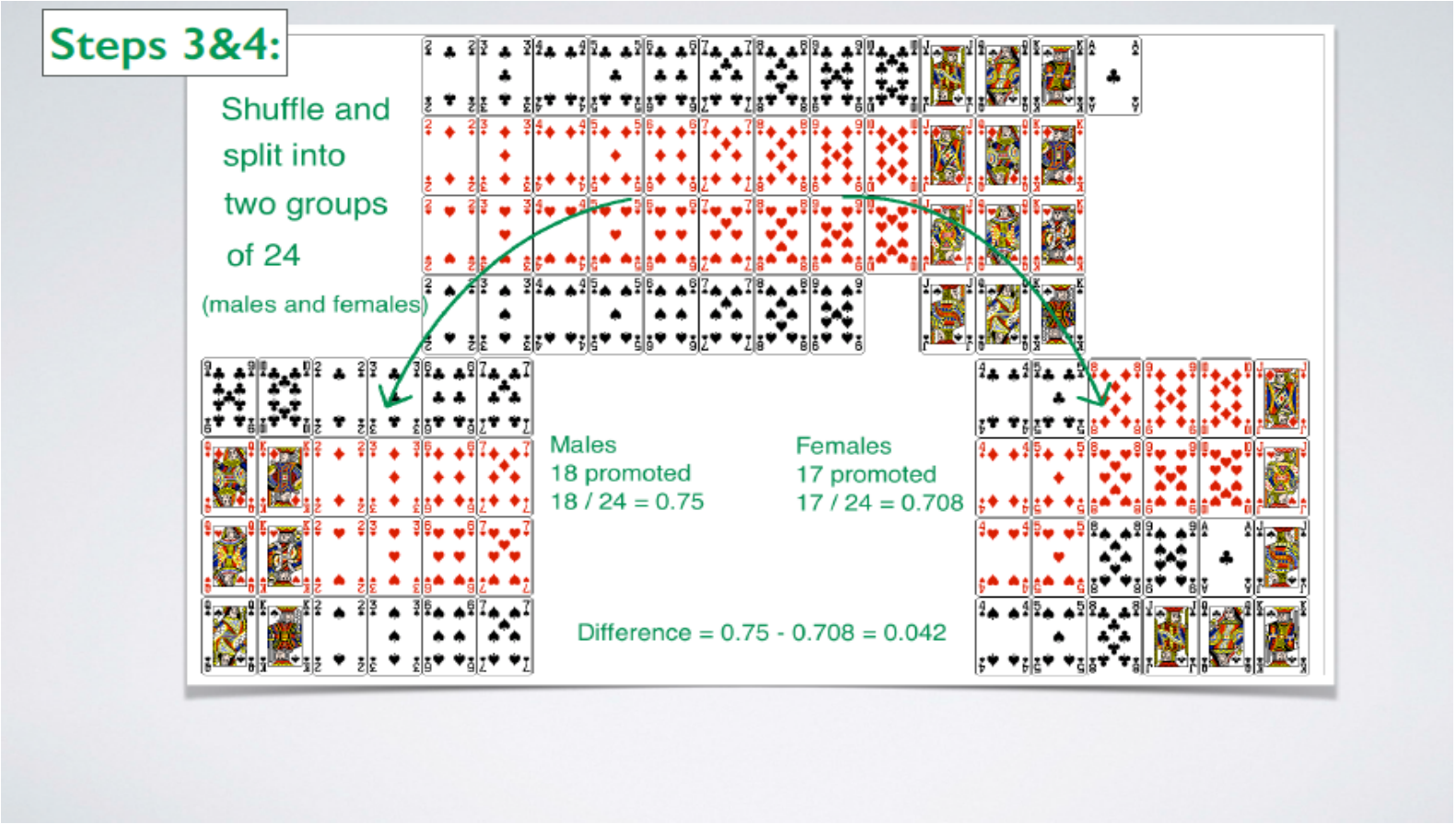
Statistical inference: case study

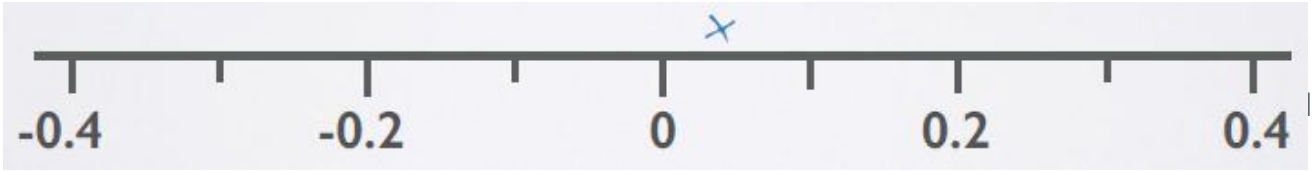


Statistical inference: case study

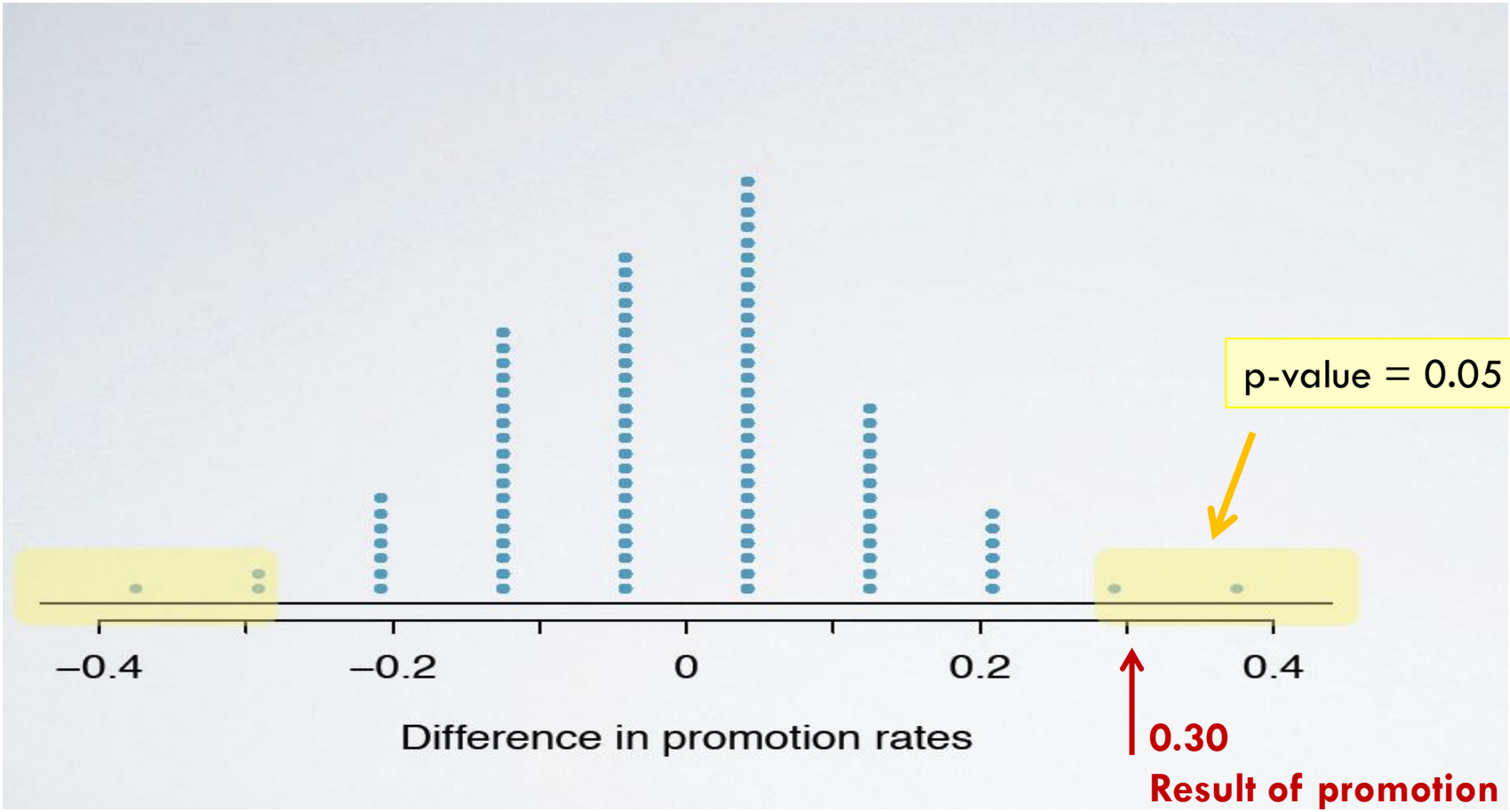


Statistical inference: case study





Statistical inference: case study



Statistical inference: case study

making a decision

- results from the simulations look like the data → the difference between the proportions of promoted files between males and females was due to chance (promotion and gender are independent)
- results from the simulations do not look like the data → the difference between the proportions of promoted files between males and females was not due to chance, but due to an actual effect of gender (promotion and gender are dependent)

Statistical inference: case study

summary

- set a null and an alternative hypothesis
- simulate the experiment assuming that the null hypothesis is true
- evaluated the probability of observing an outcome at least as extreme as the one observed in the original data

p-value

- and if this probability is low, reject the null hypothesis in favor of the alternative

Foundation for inference

Hypothesis testing with randomisation

Confidence intervals with bootstrapping

Inference with mathematical models

Hypothesis testing with randomisation

Hypotheses test

- **Hypotheses test, is a formal technique for evaluating two competing possibilities.**
  - **Null hypothesis represents either skeptical scenario or perspective with no difference.**
  - **Alternative hypotheses represents a new perspective such as possibility of relationship between variables or a treatement effect in an experiment.**
  - **The alternative hypotheses is usually the reason the scientists set out to do research in the first place**

Randomization testing procedure

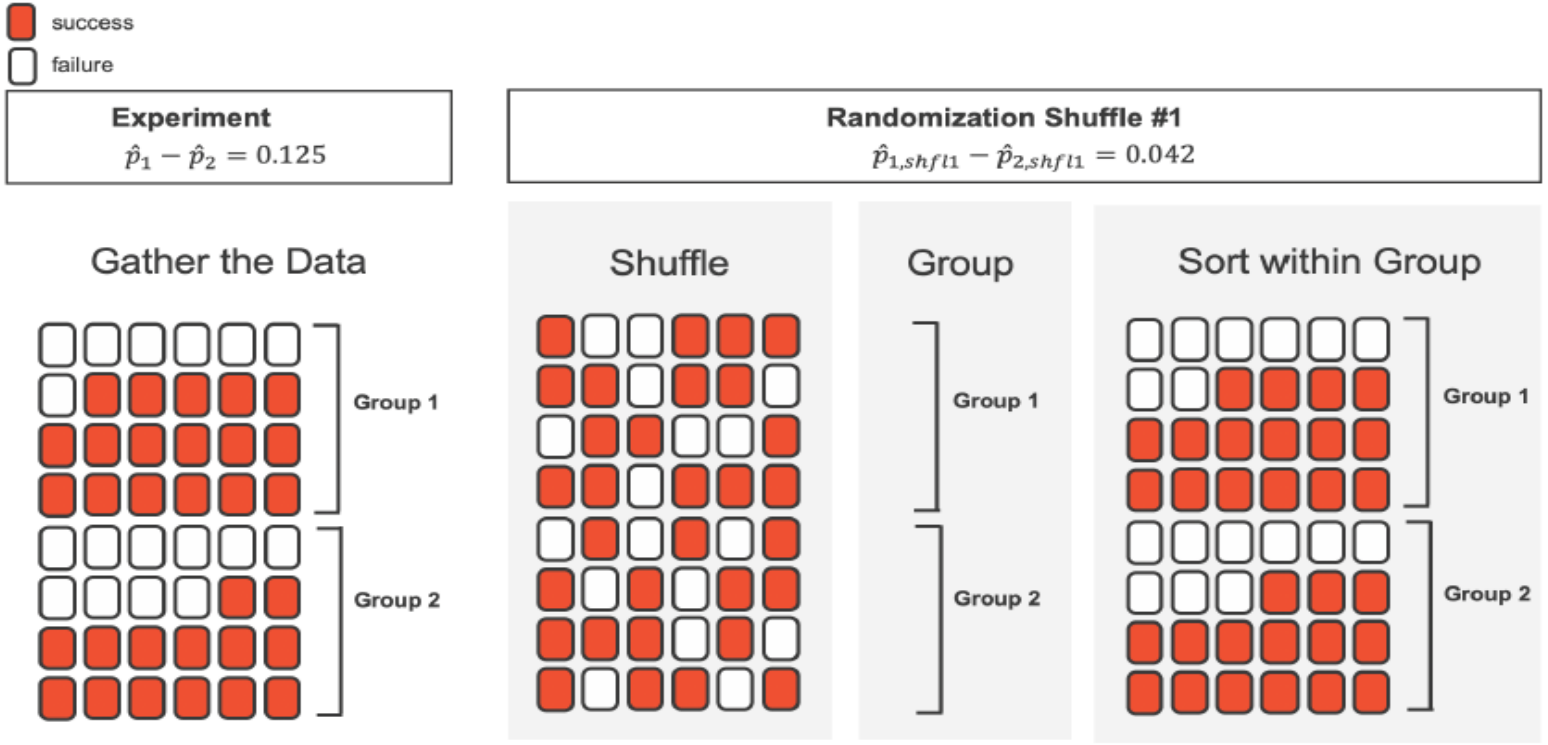


Figure 11.8: An example of one simulation of the full randomization procedure from a hypothetical dataset as visualized in the first panel. We repeat the steps hundreds or thousands of times.

Randomization testing procedure

- **Frame the research question in terms of hypotheses.** Hypothesis tests are appropriate for research questions that can be summarized in two competing hypotheses. The null hypothesis (H_0) usually represents a skeptical perspective or a perspective of no relationship between the variables. The alternative hypothesis (H_A) usually represents a new view or the existence of a relationship between the variables.
- **Collect data with an observational study or experiment.** If a research question can be formed into two hypotheses, we can collect data to run a hypothesis test. If the research question focuses on associations between variables but does not concern causation, we would use an observational study. If the research question seeks a causal connection between two or more variables, then an experiment should be used.
- **Model the randomness that would occur if the null hypothesis was true.** In the examples above, the variability has been modeled as if the treatment (e.g., sexual identity, opportunity) allocation was independent of the outcome of the study. The computer generated null distribution is the result of many different randomizations and quantifies the variability that would be expected if the null hypothesis was true.
- **Analyze the data.** Choose an analysis technique appropriate for the data and identify the p-value. So far, we have only seen one analysis technique: randomization. Throughout the rest of this textbook, we'll encounter several new methods suitable for many other contexts.
- **Form a conclusion.** Using the p-value from the analysis, determine whether the data provide evidence against the null hypothesis. Also, be sure to write the conclusion in plain language so casual readers can understand the results.

Hypotheses test

Null and alternative hypotheses.

The null hypothesis (H_0) often represents either a skeptical perspective or a claim of "no difference" to be tested.

The alternative hypothesis (H_A) represents an alternative claim under consideration and is often represented by a range of possible values for the value of interest.

If a person makes a somewhat unbelievable claim, we are initially skeptical. However, if there is sufficient evidence that supports the claim, we set aside our skepticism.

Hypotheses test

- **p-value and statistical discernibility**

p-value.

The p-value is the probability of observing data at least as favorable to the alternative hypothesis as our current dataset, if the null hypothesis were true. We typically use a summary statistic of the data, such as a difference in proportions, to help compute the p-value and evaluate the hypotheses. This summary value that is used to compute the p-value is often called the test statistic.

When the p-value is small, i.e., less than a previously set threshold, we say the results are **statistically discernible**. This means the data provide such strong evidence against H_0 that we reject the null hypothesis in favor of the alternative hypothesis.[8] The threshold is called the **discernibility level** and often represented by α (the Greek letter *alpha*). [9] The value of α represents how rare an event needs to be in order for the null hypothesis to be rejected. Historically, many fields have set $\alpha = 0.05$, if the null hypothesis is to be rejected. The value of α can vary depending on the the field or the application.

[8] Many texts use the phrase "statistically significant" instead of "statistically discernible". We have chosen to use "discernible" to indicate that a precise statistical event has happened, as opposed to a notable effect which may or may not fit the statistical definition of discernible or significant.

[9] Here, too, we have chosen "discernibility level" instead of "significance level" which you will see in some texts.

Hypotheses test

- **p-value and statistical discernibility**

Statistical discernibility.

We say that the data provide **statistically discernible** evidence against the null hypothesis if the p-value is less than some predetermined threshold (e.g., 0.01, 0.05, 0.1).

What's so special about 0.05?

We often use a threshold of 0.05 to determine whether a result is statistically discernible. But why 0.05? Maybe we should use a bigger number, or maybe a smaller number. If you're a little puzzled, that probably means you're reading with a critical eye – good job! We've made a video to help clarify *why 0.05*:

https://www.openintro.org/book/stat/why05/

Sometimes it's also a good idea to deviate from the standard.

Confidence intervals with bootstrapping

Confidence intervals with bootstrapping

- **We expand on idea of using a sample proportion to estimate a population proportion. That is we create what is called confidence interval, which is a range of plausible values where we may find the true population value.**
- **The process for creating confidence interval is based on understainding how a statistics (here sample proportion) varies around the parameter (here the population proportion) when many diferent statistics are calculated from many different samples.**

Confidence intervals with bootstrapping

If we could, we would measure the variability of the statistics by repeatedly taking sample data from the population and compute the sample proportion. Then we could do it again. And again. And so on until we have a good sense of the variability of our original estimate.

When the variability across the samples is large, we would assume that the original statistic is possibly far from the true population parameter of interest (and the interval estimate will be wide). When the variability across the samples is small, we expect the sample statistic to be close to the true parameter of interest (and the interval estimate will be narrow).

The ideal world where sampling data is free or extremely cheap is almost never the case, and taking repeated samples from a population is usually impossible. So, instead of using a "resample from the population" approach, bootstrapping uses a "resample from the sample" approach.

Randomization vs Bootstrapping

- **Randomization is a suitable technique for evaluating whether a difference in sample proportions is due to a chance.**

Randomization tests are best suited for modeling experiments where the treatment (explanator variable) has been randomly assigned to the observational units and there is an attempt to answer simple yes/no research question.

For example, consider the following research questions that can be well assessed with a randomization test:

- Does this vaccine make it less likely that a person will get malaria?
- Does drinking caffeine affect how quickly a person can tap their finger?
- Can we predict whether candidate A will win the upcoming election?

Randomization vs Bootstrapping

- **We are now interested in a diferent approach to understanding population parameter.**

Instead, of testing a claim, the goal now is to estimate the unknown value of a population

For example,

- How much less likely am I to get malaria if I get the vaccine?
- How much faster (or slower) can a person tap their finger, on average, if they drink caffeine first?
- What proportion of the vote will go to candidate A?

Here, we explore the situation where the focus is on a single proportion, and we introduce a new simulation method: bootstrapping.

Bootstrapping

Bootstrapping is best suited for modeling studies where the data have been generated through random sampling from a population. As with randomization tests, our goal with bootstrapping is to understand variability of a statistic. Unlike randomization tests (which modeled how the statistic would change if the treatment had been allocated differently), the bootstrap will model how a statistic varies from one sample to another taken from the population. This will provide information about how different the statistic is from the parameter of interest.

Quantifying the variability of a statistic from sample to sample is a hard problem. Fortunately, sometimes the mathematical theory for how a statistic varies (across different samples) is well-known; this is the case for the sample proportion

However, some statistics do not have simple theory for how they vary, and bootstrapping provi a computational approach for providing interval estimates for almost any population parameter. In

Bootstrapping: case study

People providing an organ for donation sometimes seek the help of a special medical consultant. These consultants assist the patient in all aspects of the surgery, with the goal of reducing the possibility of complications during the medical procedure and recovery. Patients might choose a consultant based in part on the historical complication rate of the consultant's clients.

One consultant tried to attract patients by noting the average complication rate for liver donor surgeries in the US is about 10%, but her clients have had only 3 complications in the 62 liver donor surgeries she has facilitated. She claims this is strong evidence that her work meaningfully contributes to reducing complications (and therefore she should be hired!).

EXAMPLE

We will let p represent the true complication rate for liver donors working with this consultant. (The "true" complication rate will be referred to as the **parameter**.) We estimate p using the data, and label the estimate $\hat{p}$.

The sample proportion for the complication rate is 3 complications divided by the 62 surgeries the consultant has worked on: $\hat{p} = 3/62 = 0.048$.

Bootstrapping: case study

EXAMPLE

Is it possible to assess the consultant's claim (that the reduction in complications is due to her work) using the data?

No. The claim is that there is a causal connection, but the data are observational, so we must be on the lookout for confounding variables. For example, maybe patients who can afford a medical consultant can afford better medical care, which can also lead to a lower complication rate. While it is not possible to assess the causal claim, it is still possible to understand the

Parameter.

A parameter is the "true" value of interest.

We typically estimate the parameter using a point estimate from a sample of data. The point estimate is also known as the statistic.

For example, we estimate the probability p of a complication for a client of the medical consultant by examining the past complications rates of her clients:

$$\hat{p} = 3/62 = 0.048 \text{ is used to estimate } p$$

Bootstrapping: case study

- Variability of the statistics

In the medical consultant case study, the parameter is p, the true probability of a complication for a client of the medical consultant. There is no reason to believe that p is exactly $\hat{p} = 3/62$, but there is also no reason to believe that p is particularly far from $\hat{p} = 3/62$. By sampling with replacement from the dataset (a process called bootstrapping), the variability of the possible $\hat{p}$ values can be approximated.

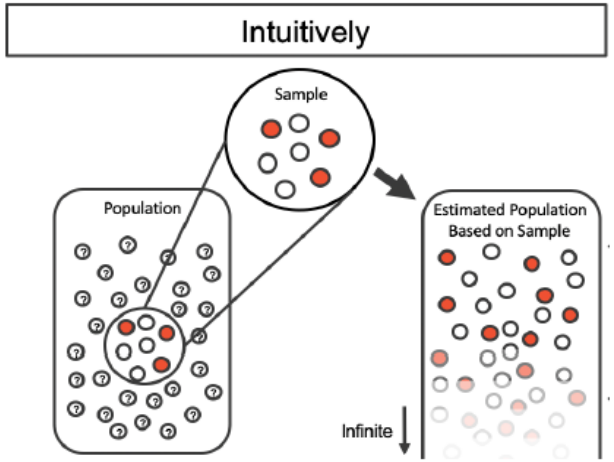


Figure 12.1: The unknown population is estimated using the observed sample data. Note that we can use the sample to create an estimated or bootstrapped population from which to sample. The observed data include three red and four white marbles, so the estimated population contains 3/7 red marbles and 4/7 white marbles.

Bootstrapping: case study

- Variability of the statistics

By taking repeated samples from the estimated population, the variability from sample to sample can be observed. In Figure 12.2 the repeated bootstrap samples are obviously different both from each other and from the original population. Recall that the bootstrap samples were taken from the same (estimated) population, and so the differences are due entirely to natural variability in the sampling procedure.

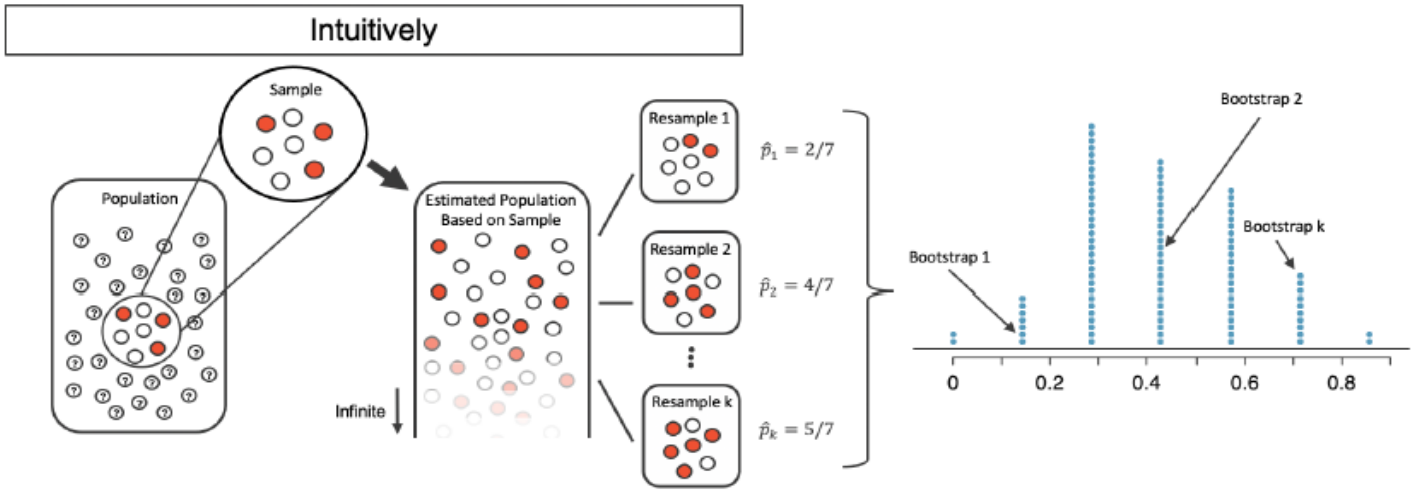
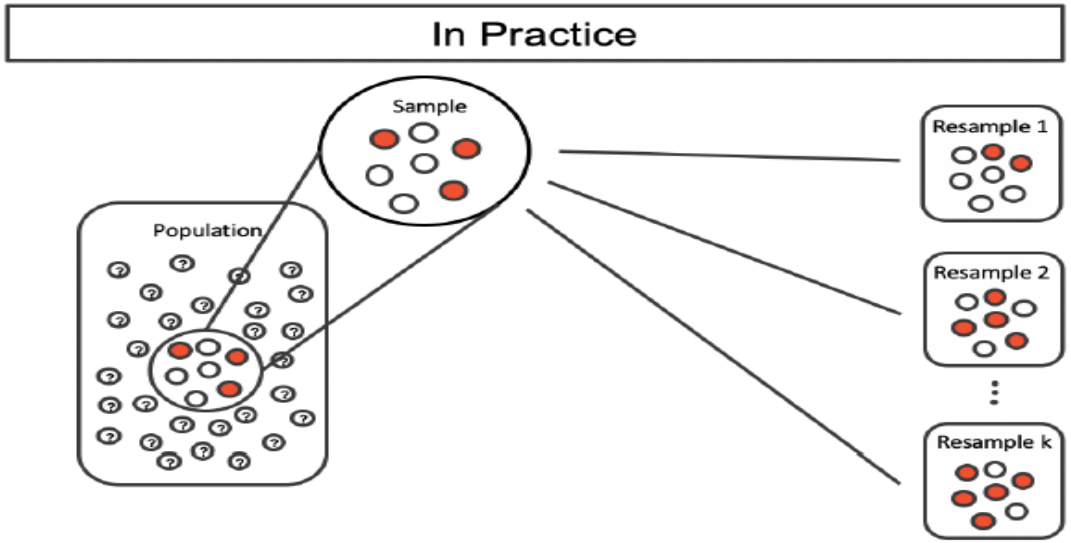


Figure 12.3: The bootstrapped proportion is estimated for each bootstrap sample. The resulting bootstrap distribution (dotplot) provides a measure for how the proportions vary from sample to sample

Bootstrapping: case study

It turns out that in practice, it is very difficult for computers to work with an infinite population (with the same proportional breakdown as in the sample). However, there is a physical and computational method which produces an equivalent bootstrap distribution of the sample proportion in a computationally efficient manner.

Consider the observed data to be a bag of marbles 3 of which are success (red) and 4 of which are failures (white). By drawing the marbles out of the bag with replacement, we depict the exact same sampling **process** as was done with the infinitely large estimated population.



Bootstrapping: case study

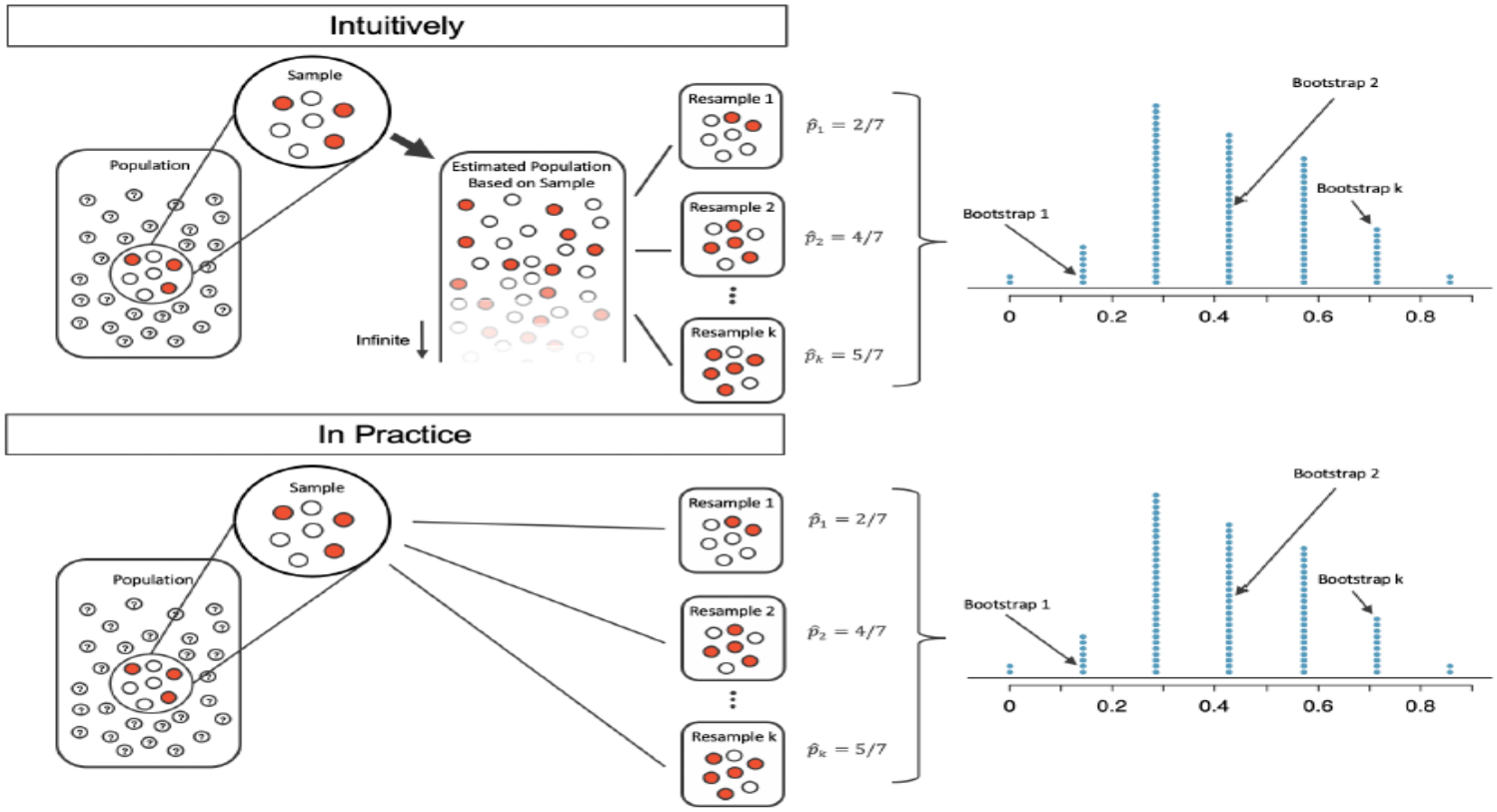


Figure 12.5: A comparison of the process of sampling from the estimate infinite population and resampling with replacement from the original sample. Note that the dotplot of bootstrapped proportions is the same because the process by which the statistics were estimated is equivalent.

Bootstrapping: case study

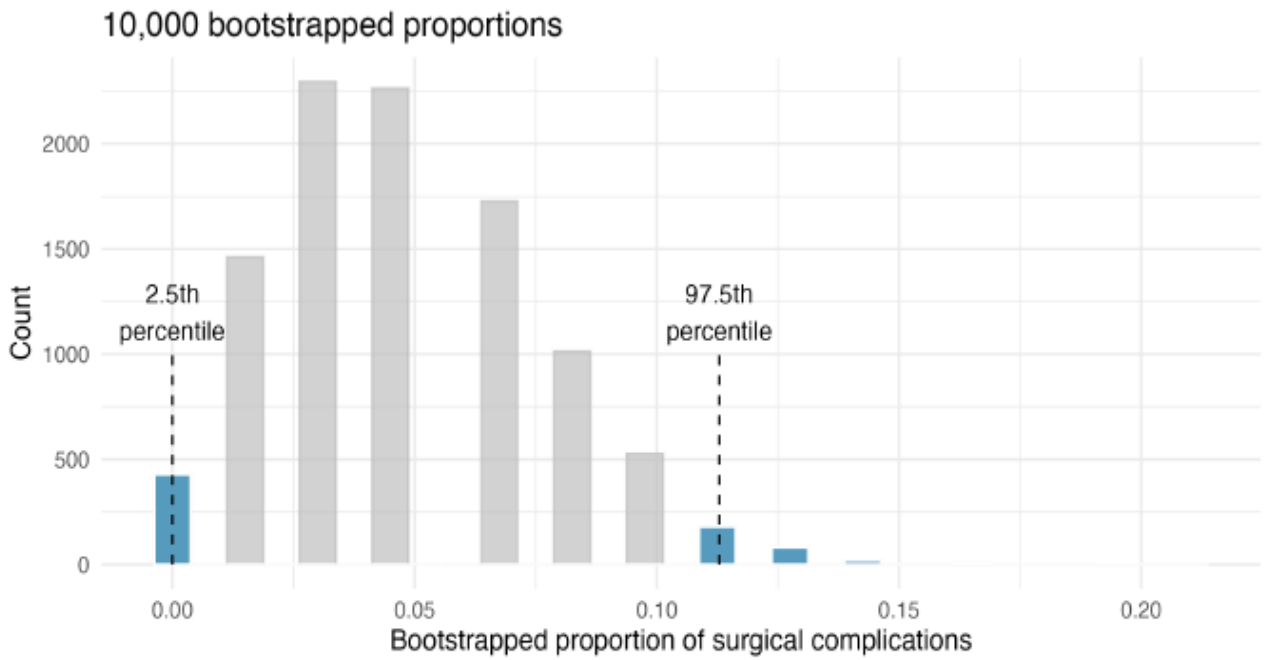


Figure 12.6: The original medical consultant data is bootstrapped 10,000 times. Each simulation creates a sample from the original data where the probability of a complication is $\hat{p} = 3/62$. The bootstrap 2.5 percentile proportion is 0 and the 97.5 percentile is 0.113. The result is: we are confident that, in the population, the true probability of a complication is between 0% and 11.3%.

Bootstrapping: case study

EXAMPLE

The original claim was that the consultant's true rate of complication was under the national rate of 10%. Does the interval estimate of 0% to 11.3% for the true probability of complication indicate that the surgical consultant has a lower rate of complications than the national average? Explain.

No. Because the interval overlaps 10%, it might be that the consultant's work is associated with a lower risk of complications, or it might be that the consultant's work is associated with a higher risk (i.e., greater than 10%) of complications! Additionally, as previously mentioned, because this is an observational study, even if an association can be measured, there is no evidence that the consultant's work is the cause of the complication rate (being higher or lower).

Bootstrapping: case study

Convidence intervals

A point estimate provides a single plausible value for a parameter. However, a point estimate is rarely perfect; usually there is some error in the estimate. In addition to supplying a point estimate of a parameter, a next logical step would be to provide a plausible *range of values* for the parameter.

As we saw above, a **bootstrap sample** is a sample of the original sample. In the case of the medical complications data, we proceed as follows:

- Randomly sample one observation from the 62 patients (replace the marble back into the bag so as to keep the population constant).
- Randomly sample a second observation from the 62 patients. Because we sample with replacement (i.e., we do not actually remove the marbles from the bag), there is a 1-in-62 chance that the second observation will be the same one sampled in the first step!
- Keep going one sampled observation at a time ...
- Randomly sample the 62nd observation from the 62 patients.

Bootstrap sampling is often called **sampling with replacement**.

A bootstrap sample behaves similarly to how an actual sample from a population would behave, and we compute the point estimate of interest (here, compute $\hat{p}_{boot}$).

Based on theory that is beyond this text, we know that the bootstrap proportions $\hat{p}_{boot}$ vary around $\hat{p}$ similarly to how different sample proportions (i.e., values of $\hat{p}$) vary around the true parameter p. Therefore, an interval estimate for p can be produced using the $\hat{p}_{boot}$ values themselves.

Bootstrapping: case study

Convidence intervals

Based on theory that is beyond this text, we know that the bootstrap proportions $\hat{p}_{boot}$ vary around $\hat{p}$ similarly to how different sample proportions (i.e., values of $\hat{p}$) vary around the true parameter p. Therefore, an interval estimate for p can be produced using the $\hat{p}_{boot}$ values themselves.

95% bootstrap percentile confidence interval for a parameter p.

The 95% bootstrap confidence interval for the parameter p can be obtained directly using the ordered $\hat{p}_{boot}$ values.

Consider the sorted $\hat{p}_{boot}$ values. Call the 2.5% bootstrapped proportion value "lower", and call the 97.5% bootstrapped proportion value "upper".

The 95% confidence interval is given by: (lower, upper)

Boostrap process and confidence interval

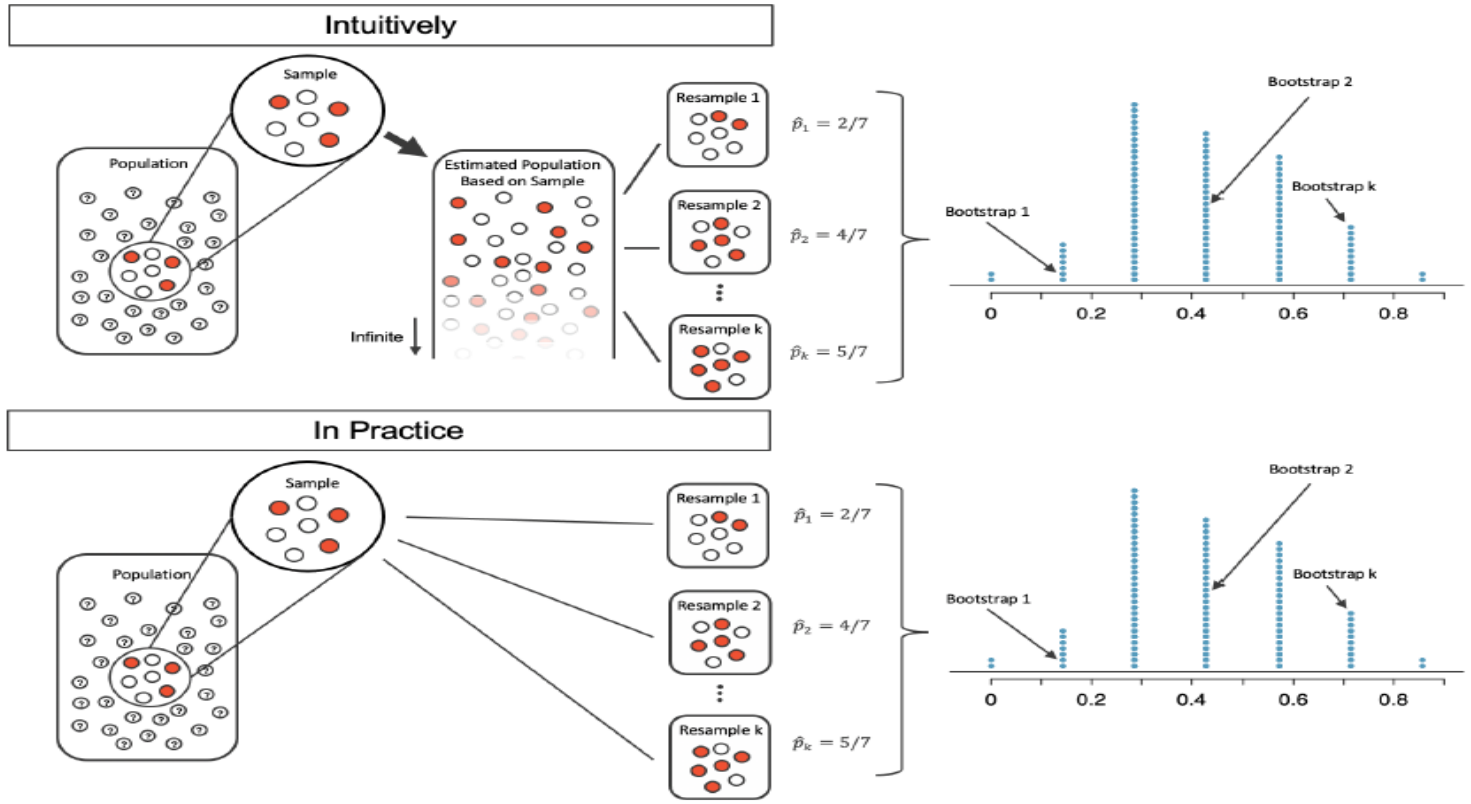


Figure 12.8: We will use sampling with replacement to measure the variability of the statistic of interest (here the proportion). Sampling with replacement is a computational tool which is equivalent to using the sample as a way of estimating an infinitely large population from which to sample.

Boostrap process and confidence interval

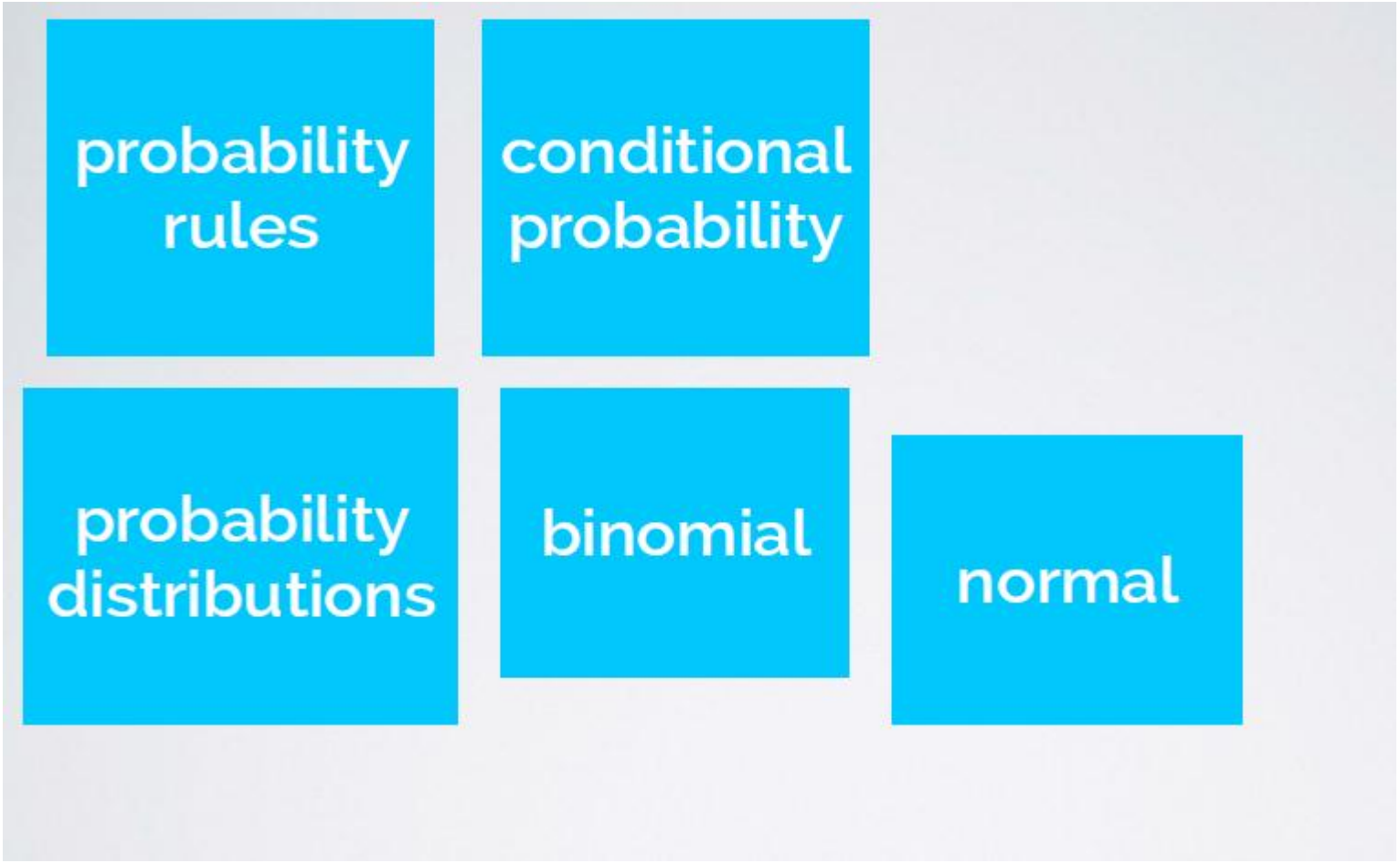
- **Frame the research question in terms of a parameter to estimate.** Confidence Intervals are appropriate for research questions that aim to estimate a number from the population (called a parameter).
- **Collect data with an observational study or experiment.** If a research question can be formed as a query about the parameter, we can collect data to calculate a statistic which is the best guess we have for the value of the parameter. However, we know that the statistic won't be exactly equal to the parameter due to natural variability.
- **Model the randomness by using the data values as a proxy for the population.** In order to assess how far the statistic might be from the parameter, we take repeated resamples from the dataset to measure the variability in bootstrapped statistics. The variability of the bootstrapped statistics around the observed statistic (a quantity which can be measured through computational technique) should be approximately the same as the variability of many observed sample statistics around the parameter (a quantity which is very difficult to measure because in real life we only get exactly one sample).
- **Create the interval.** After choosing a particular confidence level, use the variability of the bootstrapped statistics to create an interval estimate which will hope to capture the true parameter. While the interval estimate associated with the particular sample at hand may or may not capture the parameter, the researcher knows that over their lifetime, the confidence level will determine the percentage of their research confidence intervals that do capture the true parameter.
- **Form a conclusion.** Using the confidence interval from the analysis, report on the interval estimate for the parameter of interest. Also, be sure to write the conclusion in plain language so casual readers can understand the results.

Inference with mathematical models

Inference with mathematial models

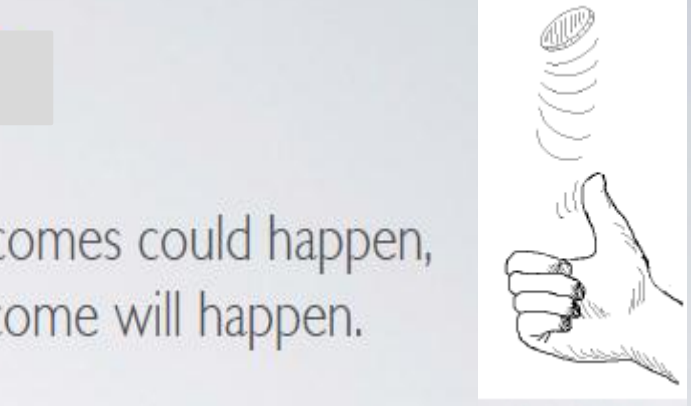
- **So far questions about population parameters were addressed using computational techniques.**
  - **With randomization tests, the data were permuted assuming the null hypothesis.**
  - **With bootstrapping, the data were resampled in order to measure variability.**
- **In many cases (indeed with sample proportions), the variability of the statistics can be described by the computational method or by a mathematical fomulas.**

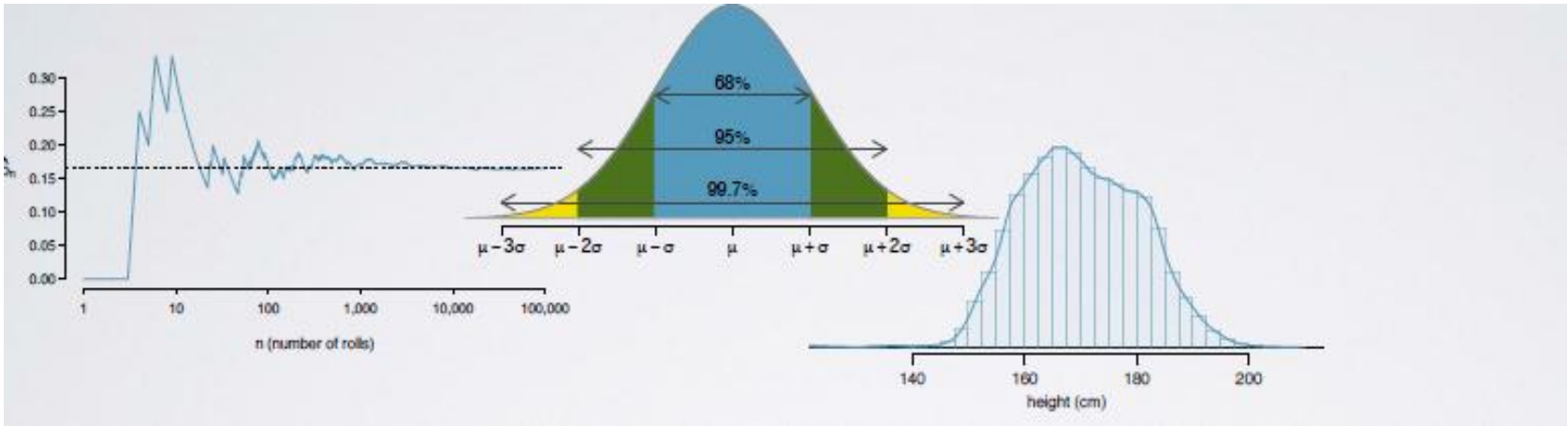
Probability and distributions



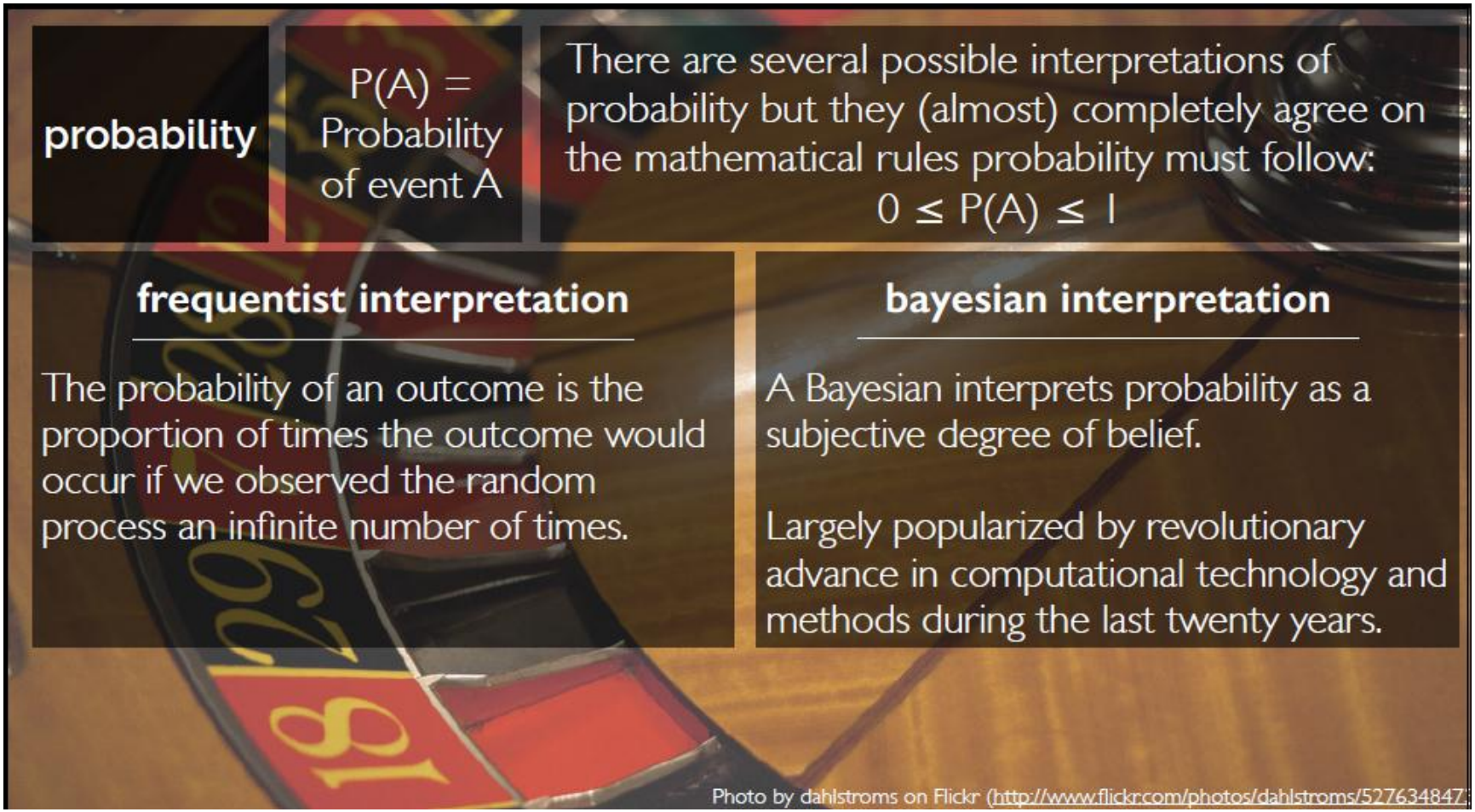
Random process

In a random process we know what outcomes could happen, but we don't know which particular outcome will happen.





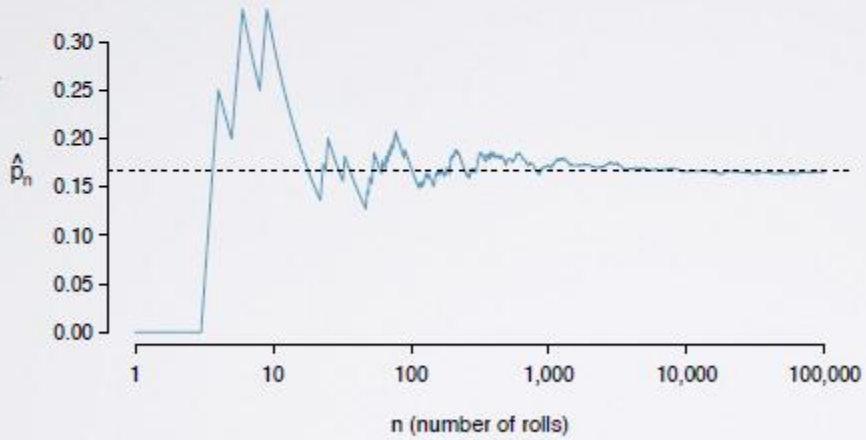
Probability



Law of Large Numbers

law of large numbers states that as more observations are collected, the proportion of occurrences with a particular outcome converges to the probability of that outcome.

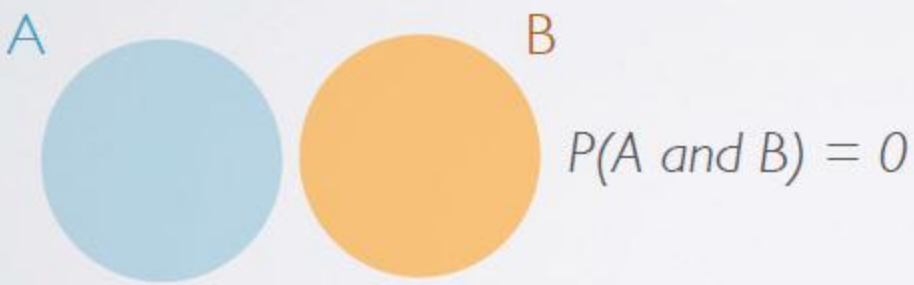
examples



Disjoint (mutually exclusive)

disjoint (mutually exclusive) events cannot happen at the same time.

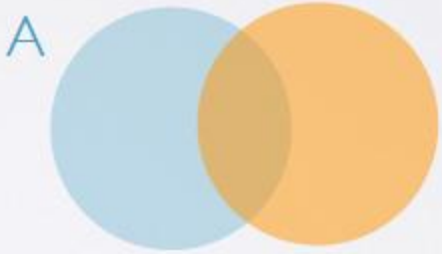
- the outcome of a single coin toss cannot be a head and a tail.
- a student can't both fail and pass a class.
- a single card drawn from a deck cannot be an ace and a queen.



$P(A \text{ and } B) = 0$

non-disjoint events can happen at the same time.

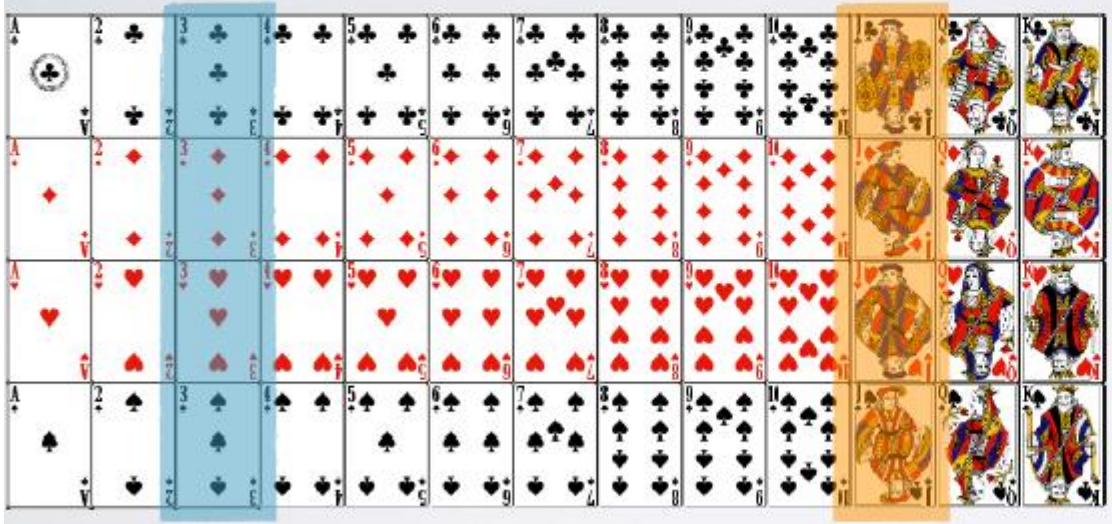
- a student can get an A in Stats and A in Econ in the same semester.



$P(A \text{ and } B) \neq 0$

Union of disjoint events

What is the probability of drawing a Jack or a three from a well shuffled full deck of cards?

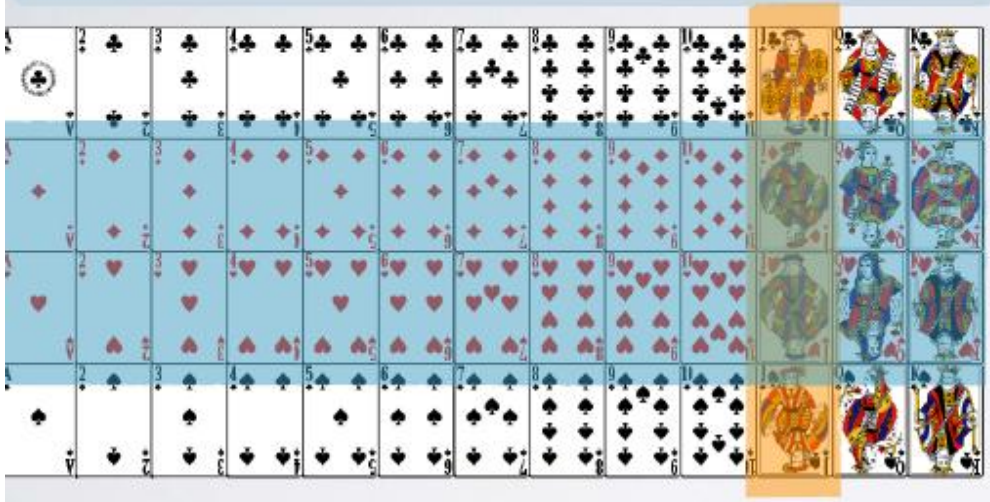


$$P(J \text{ or } 3)$$
$$= P(J) + P(3)$$
$$= (4/52) + (4/52)$$
$$\approx 0.154$$

For disjoint events A and B,
$P(A \text{ or } B) = P(A) + P(B)$

Union of ono-disjoint events

What is the probability of drawing a Jack or a red card from a well shuffled full deck of cards?



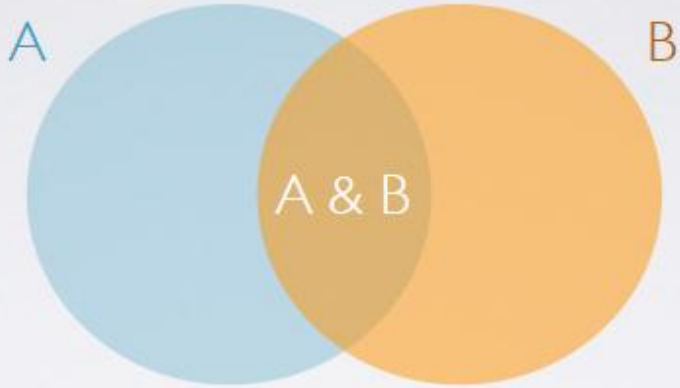
$$P(J \text{ or red}) = P(J) + P(\text{red}) - P(J \text{ and red})$$
$$= (4/52) + (26/52) - (2/52)$$
$$\approx 0.538$$

For non-disjoint events A and B,
P(A or B) = P(A) + P(B) - P(A and B)

General addition rule

General addition rule:

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



Note: When A and B are disjoint, $P(A \text{ and } B) = 0$, so the formula simplifies to $P(A \text{ or } B) = P(A) + P(B)$.

Sample space

a sample space is a collection of all possible outcomes of a trial.

A couple has two kids, what is the sample space for the sex of these kids? For simplicity assume that sex can only be male or female.

$$S = \{MM, FF, FM, MF\}$$

Probability distributions

a probability distribution lists all possible outcomes in the sample space, and the probabilities with which they occur.

one toss	head	tail
probability	0.5	0.5

two tosses	head - head	tail - tail	head - tail	tail - head
probability	0.25	0.25	0.25	0.25

rules

1. the events listed must be disjoint
2. each probability must be between 0 and 1
3. the probabilities must total 1

Complementary events

complementary events are two mutually exclusive events whose probabilities that add up to 1.

one toss	head	tail
probability	0.5	0.5

two tosses	head - head	tail - tail	head - tail	tail - head
probability	0.25	0.25	0.25	0.25

Disjoint vs complementary

Do the sum of probabilities of two disjoint outcomes always add up to 1?

Not necessarily, there may be more than 2 outcomes in the sample space.

Do the sum of probabilities of two complementary outcomes always add up to 1?

Yes, that's the definition of complementary.

Independence

two processes are independent if knowing the outcome of one provides no useful information about the outcome of the other.

1st toss 2nd toss

P(H) = 0.5 P(T) = 0.5

outcomes of two tosses of a coin are independent

1st draw 2nd draw

P(A) = 3/51 P(J) = 4/51

outcomes of two draws from a deck of cards (without replacement) are dependent

Image sources:

Coin: http://commons.wikimedia.org/wiki/File:1913_Eliasberg_Liberty_Head_Nickel.png

Card: Open Clip Art Library (http://openclipart.org/cgi-bin/navigate/recreation/games/cards/white)

Independence

Checking for independence:

$P(A \mid B) = P(A)$, then A and B are independent.

given

Independence

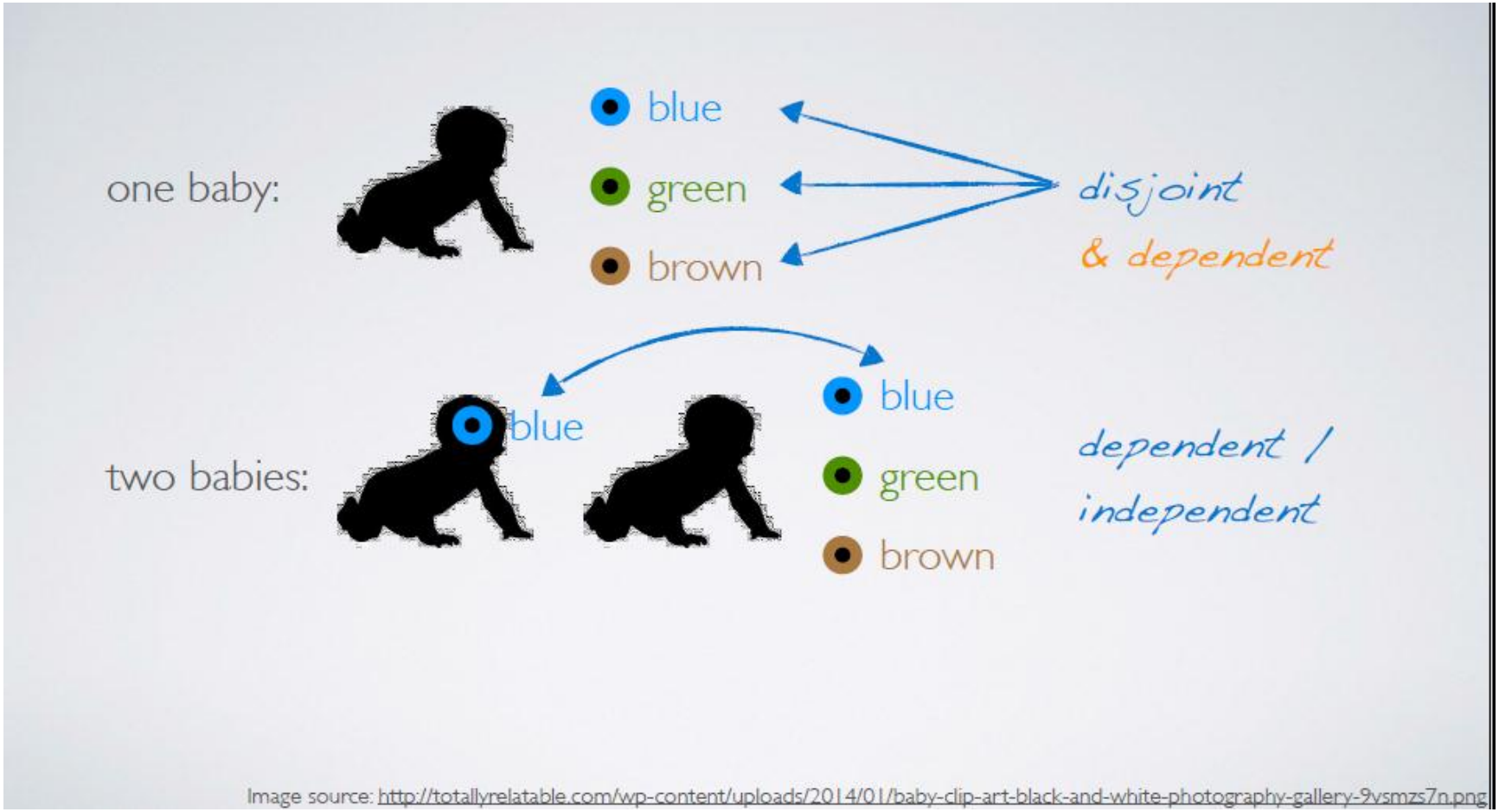
two events that are
disjoint
(mutually exclusive)
cannot happen
at the same time

$P(A \text{ and } B) = 0$

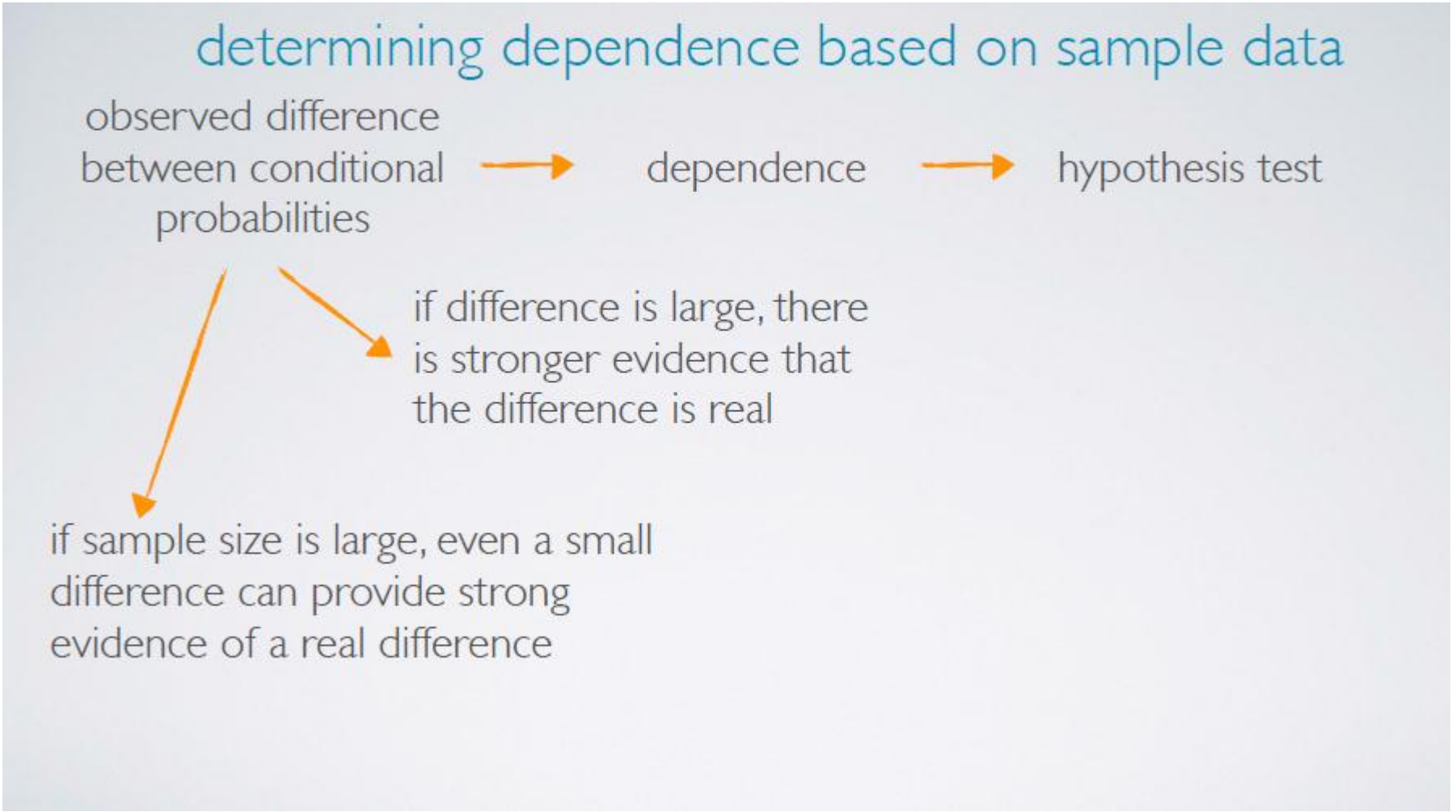
two processes are
independent
if knowing the outcome
of one
provides no useful
information about the
outcome of the other

$P(A \mid B) = P(A)$

Independence



Determining dependence



Determining dependence

Product rule for independent events:

If A and B are independent, $P(A \text{ and } B) = P(A) \times P(B)$

You toss a coin twice, what is the probability of getting two tails in a row?

P(two tails in a row) =

= P(T on the 1st toss) × P(T on the 2nd toss)

= (1/2) × (1/2)

= 1/4

Note: If $A_1, A_2, \ldots, A_k$ are independent, $P(A_1 \text{ and } A_2 \text{ and } \ldots A_k) = P(A_1) \times P(A_2) \times \ldots \times P(A_k)$

Example: probability

- sample spaces
- disjoint, complementary, and independent events
- addition rule for unions of events
- multiplication rule for joint probabilities for independent events

Example

The World Values Survey is an ongoing worldwide survey that polls the world population about perceptions of life, work, family, politics, etc.

The most recent phase of the survey that polled 77,882 people from 57 countries estimates that a 36.2% of the world's population agree with the statement "Men should have more right to a job than women."

The survey also estimates that 13.8% of people have a university degree or higher, and that 3.6% of people fit both criteria.

$P(\text{agree}) = 0.362$

$P(\text{uni. degree}) = 0.138$

$P(\text{agree } \& \text{ uni. degree}) = 0.036$

Survey: http://www.worldvaluessurvey.org/

Example

(1) Are agreeing with the statement "Men should have more right to a job than women" and having a university degree or higher disjoint events?

$P(\text{agree}) = 0.362$

$P(\text{uni. degree}) = 0.138$

$P(\text{agree} \,\&\, \text{uni. degree}) = 0.036 \neq 0 \longrightarrow$ not disjoint

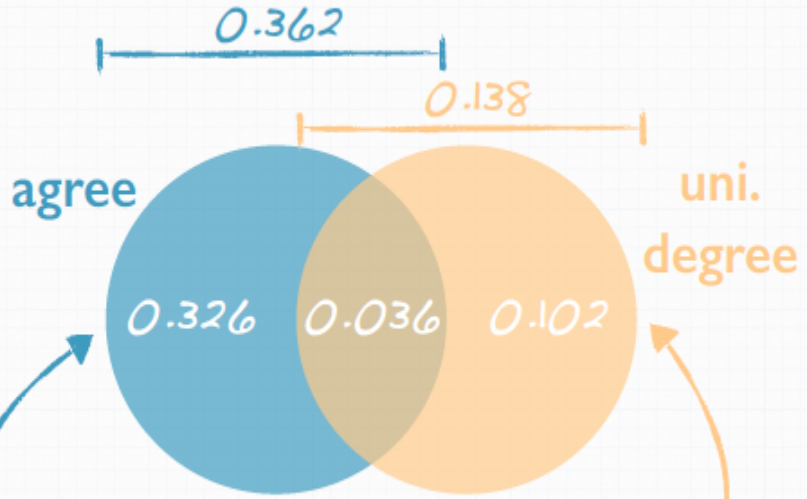
Example

(2) Draw a Venn diagram summarizing the variables and their associated probabilities.

$P(\text{agree}) = 0.362$

$P(\text{uni. degree}) = 0.138$

$P(\text{agree} \ \& \ \text{uni. degree}) = 0.036$

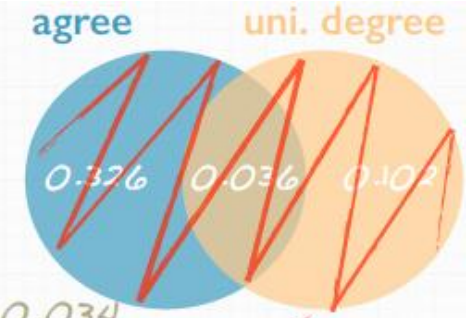


$0.362 - 0.036 = 0.326$

$0.138 - 0.036 = 0.102$

Example

(3) What is the probability that a randomly drawn person has a university degree or higher or agrees with the statement about men having more right to a job than women?



$P(\text{agree}) = 0.362$

$P(\text{uni. degree}) = 0.138$

$P(\text{agree \& uni. degree}) = 0.034$

General addition rule:

P(A or B) = P(A) + P(B) - P(A and B)

$P(\text{agree or uni. degree})$

$= P(\text{agree}) + P(\text{uni. degree}) - P(\text{agree \& uni. degree})$

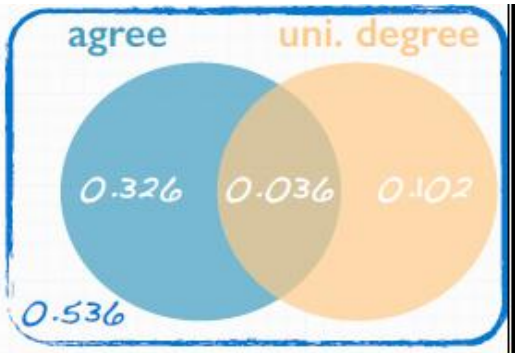
$= 0.362 + 0.138 - 0.036$

$= 0.464$

$0.326 + 0.036 + 0.102 = 0.464$

Example

(4) What percent of the world population do not have a university degree and disagree with the statement about men having more right to a job than women?



$P(\text{agree}) = 0.362$

$P(\text{uni. degree}) = 0.138$

$P(\text{agree \& uni. degree}) = 0.036$

$P(\text{agree or uni. degree}) = 0.464$

$P(\text{neither agree nor uni. degree})$

$= 1 - P(\text{agree or uni. degree})$

$= 1 - 0.464 = 0.536$

Example

(5) Does it appear that the event that someone agrees with the statement is independent of the event that they have a university degree or higher?

$P(\text{agree}) = 0.362$

$P(\text{uni. degree}) = 0.138$

$P(\text{agree \& uni. degree}) = 0.036$

Product rule for independent events:

If A and B are independent, P(A and B) = P(A) × P(B)

$P(\text{agree \& uni. degree}) \stackrel{?}{=} P(\text{agree}) \times P(\text{uni. degree})$

$0.036 \stackrel{?}{=} 0.362 \times 0.138$

$0.036 \neq 0.05 \longrightarrow$ not independent

Example

(6) What is the probability that at least 1 in 5 randomly selected people agree with the statement about men having more right to a job than women?

$P(\text{agree}) = 0.362$

$S = \{0, 1, 2, 3, 4, 5\} \longrightarrow S = \{0, \text{at least } 1\}$

$P(\text{at least 1 agree}) = 1 - P(\text{none agree})$

$= 1 - P(D\ D\ D\ D\ D)$

$= 1 - 0.638^5$

$= 1 - 0.106 = 0.894$

$P(\text{disagree}) = 1 - P(\text{agree}) = 1 - 0.362 = 0.638$

Conditional probability

study

ADOLESCENTS' UNDERSTANDING OF SOCIAL CLASS

study examining teens' beliefs about social class

sample: 48 working class and 50 upper middle class 16-year-olds

study design:

- "objective" assignment to social class based on self-reported measures of both parents' occupation and education, and household income
- "subjective" association based on survey questions

Study reference: Goodman, Elizabeth, et al. "Adolescents' understanding of social class: a comparison of white upper middle class and working class youth." Journal of adolescent health 27.2 (2000): 80-83.

Conditional probability

results:		objective social class position		
		working class	upper middle class	Total
subjective social class identity	poor	0	0	0
	working class	8	0	8
	middle class	32	13	45
	upper middle class	8	37	45
	upper class	0	0	0
	Total	48	50	98

Marginal probability

marginal

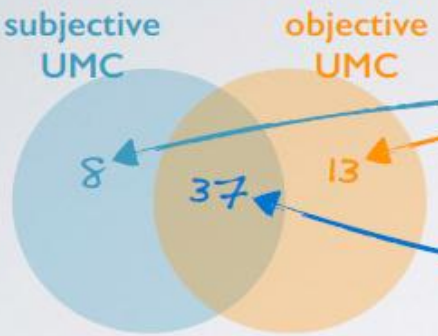
		objective social class position		
		working class	upper middle class	Total
subjective social class identity	poor	0	0	0
	working class	8	0	8
	middle class	32	13	45
	upper middle class	8	37	45
	upper class	0	0	0
	Total	48	50	98

What is the probability that a student's objective social class position is upper middle class?

P(obj UMC) = 50 / 98 ≈ 0.51

Joint probability

joint



		objective social class position		
		working class	upper middle class	Total
subjective social class identity	poor	0	0	0
	working class	8	0	8
	middle class	32	13	45
	upper middle	8	37	45
	upper class	0	0	0
	Total	48	50	98

What is the probability that a student's objective position *and* subjective identity are both upper middle class?

P(obj UMC & subj UMC)

= 37 / 98 ≈ 0.38

Conditional probability

conditional

		objective social class position		
		working class	upper middle class	Total
subjective social class identity	poor	0	0	0
	working class	8	0	8
	middle class	32	13	45
	upper middle	8	37	45
	upper class	0	0	0
	Total	48	50	98

What is the probability that a student who is objectively in the working class associates with upper middle class?

$P(\text{subj UMC} \mid \text{obj WC}) = 8 / 48 \approx 0.17$

Conditional probability

Bayes' theorem:

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

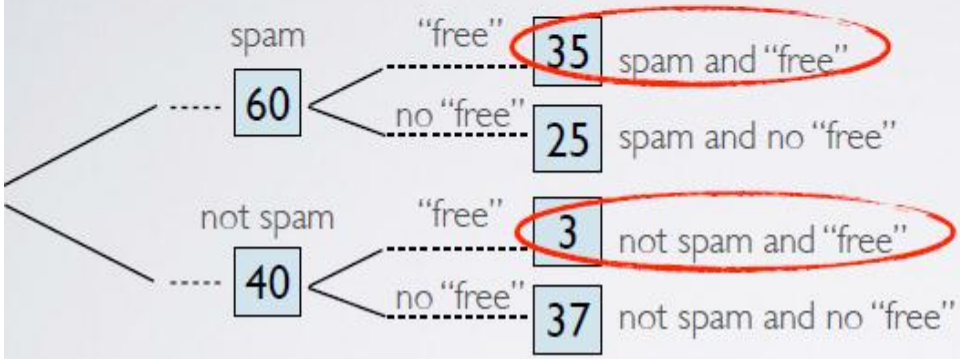
		objective social class position		
		working class	upper middle class	Total
subjective social class identity	poor	0	0	0
	working class	8	0	8
	middle class	32	13	45
	upper middle	8	37	45
	upper class	0	0	0
	Total	48	50	98

$$P(\text{subj UMC} \mid \text{obj WC}) = \frac{P(\text{subj UMC \& obj WC})}{P(\text{obj WC})} = \frac{8/98}{48/98} = 8/48 \approx 0.17$$

Probability trees

$P(A \mid B) \rightarrow P(B \mid A)$

You have 100 emails in your inbox: 60 are spam, 40 are not. Of the 60 spam emails, 35 contain the word "free". Of the rest, 3 contain the word "free". If an email contains the word "free", what is the probability that it is spam?



$$P(\text{spam} \mid \text{"free"}) = \frac{35}{35 + 3} = 0.92$$

Probability trees

As of 2009, Swaziland had the highest HIV prevalence in the world. 25.9% of this country's population is infected with HIV. The ELISA test is one of the first and most accurate tests for HIV. For those who carry HIV, the ELISA test is 99.7% accurate. For those who do not carry HIV, the test is 92.6% accurate. If an individual from Swaziland has tested positive, what is the probability that he carries HIV?

$P(H/IV) = 0.259$

$P(+ \mid H/IV) = 0.997$ $P(- \mid no\ H/IV) = 0.926$

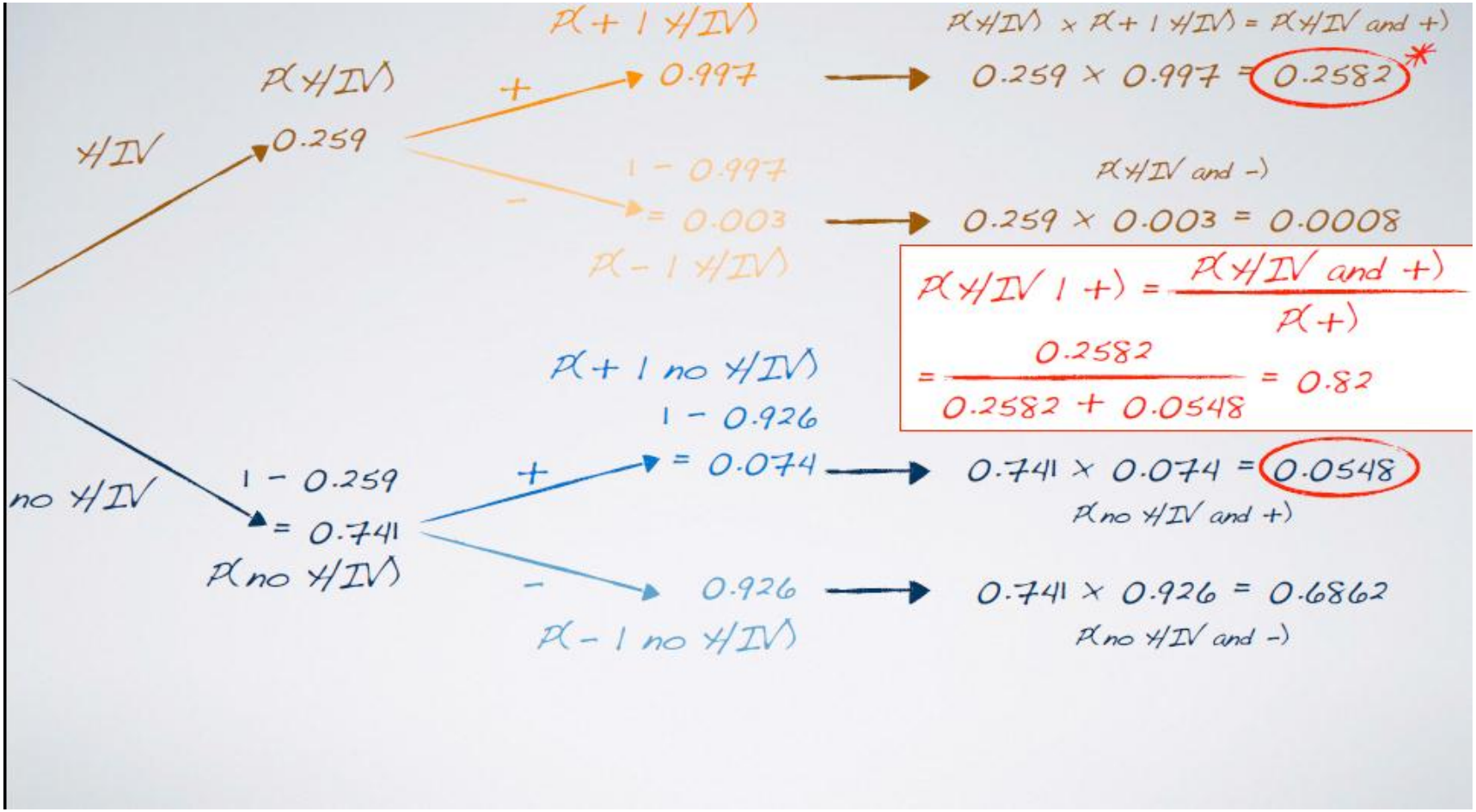
tree diagram!

$P(H/IV \mid +) = ?$

Image source: http://en.wikipedia.org/wiki/File:Location_Swaziland_AU_Africa.svg

Data source: CIA Factbook, Country Comparison: HIV/AIDS - Adult Prevalence Rate https://www.cia.gov/library/publications/the-world-factbook/rankorder/2155rank.html

Probability trees



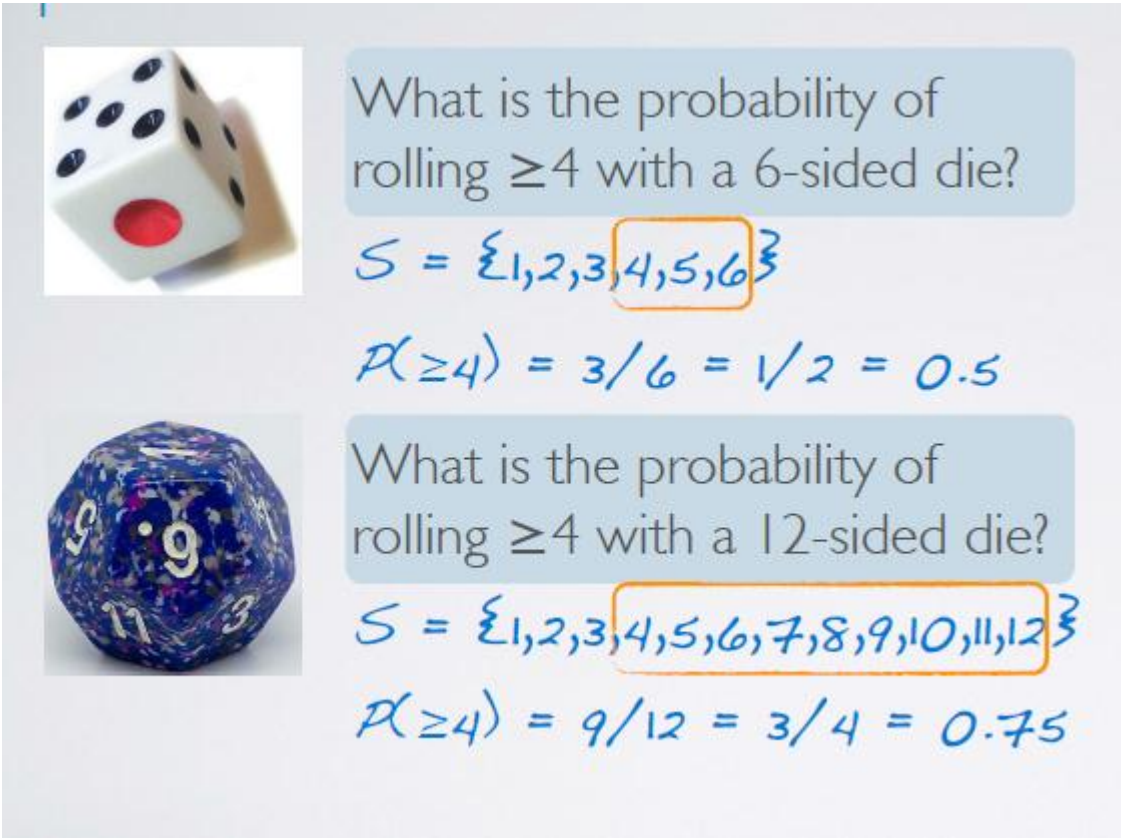
Probability trees

If an individual from Swaziland has tested positive, what is the probability that he carries HIV?

$$P(\text{HIV} \mid +) = 0.82$$

There is an 82% chance that an individual from Swaziland who has tested positive actually carries HIV.

Bayesian inference



Bayesian inference

"good die"

Say you're playing a game where the goal is to roll ≥ 4. If you could get your pick, which die would you prefer to play this game with?

(a)

$P(\geq 4) = 0.5$

good die

(b)

$P(\geq 4) = 0.75$

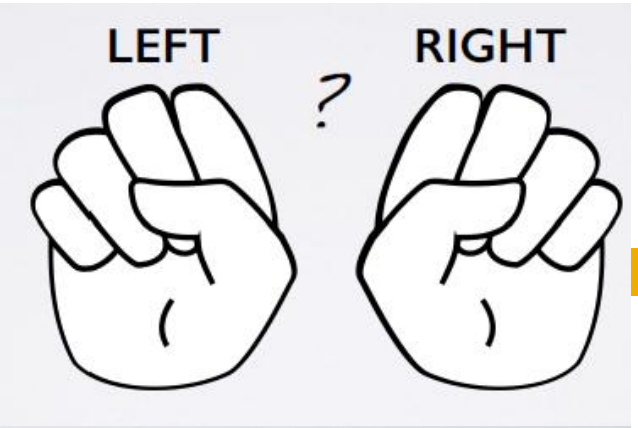
Bayesian inference

rules

hypotheses and decisions

		Truth	
		Right good, Left bad	Right bad, Left good
Decision	pick Right	You win the game!	You lose :(
	pick Left	You lose :(	You win the game!

Bayesian inference



before you collect data

Before we collect any data, you have no idea if I am holding the good die (12-sided) on the right hand or the left hand. Then, what are the probabilities associated with the following hypotheses?

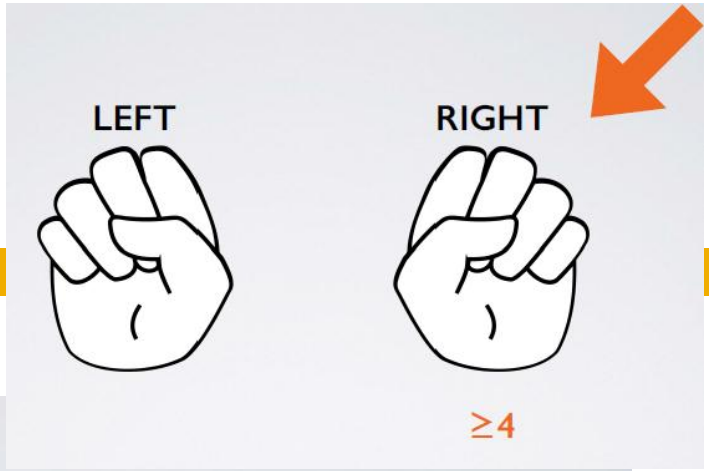
H_1: good die on the Right (bad die on the Left)

H_2: good die on the Left (bad die on the Right)

	$P(H_1$: good die on the Right)	$P(H_2$: good die on the Left)
(a)	0.33	0.67
(b)	0.5	0.5
(c)	0	1
(d)	0.25	0.75

→ prior

Bayesian inference



after you see the data

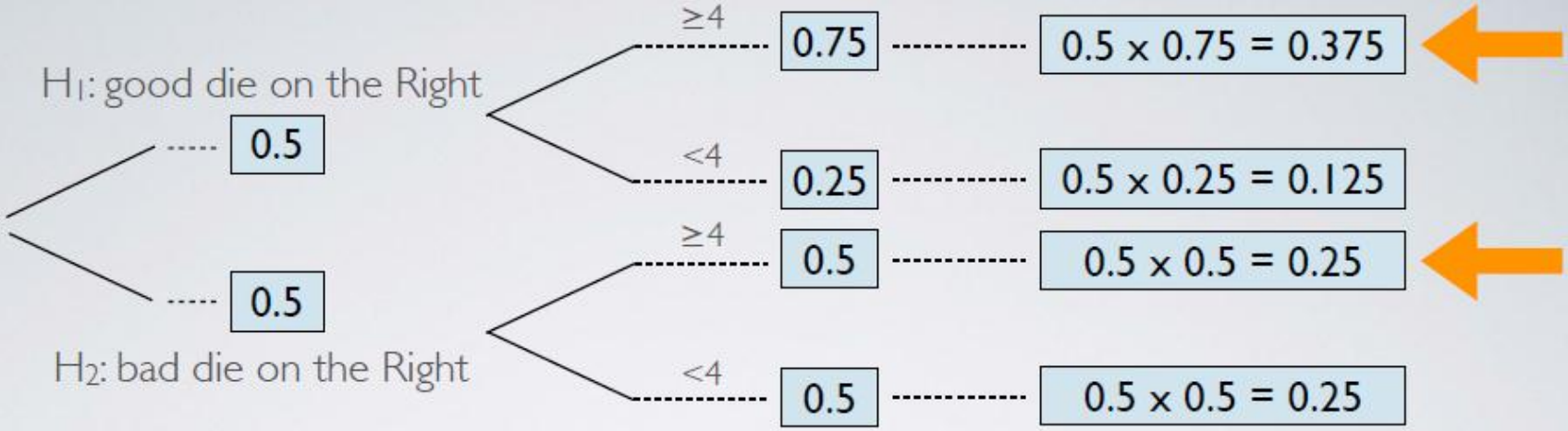
You chose the right hand, and you won (rolled a number ≥4). Having observed this data point how, if at all, do the probabilities you assign to the same set of hypotheses change?

H_1: good die on the Right (bad die on the Left)

H_2: good die on the Left (bad die on the Right)

	$P(H_1$: good die on the Right)	$P(H_2$: good die on the Left)
(a)	0.5	0.5
(b)	more than 0.5	less than 0.5
(c)	less than 0.5	more than 0.5

Bayesian inference



$P(H_1$: good die on the Right | you rolled ≥4 with the die on the Right) =

$$= \frac{P(\text{good Right \& } \geq 4 \text{ Right})}{P(\geq 4 \text{ Right})} = \frac{0.375}{0.375 + 0.25} = 0.6$$

Bayesian inference

posterior

- The probability we just calculated is also called the posterior probability.

  $P(H_1$: good die on the Right | you rolled ≥ 4 with the die on the Right)
- Posterior probability is generally defined as P(hypothesis | data).
- It tells us the probability of a hypothesis we set forth, given the data we just observed.
- It depends on both the prior probability we set and the observed data.
- This is different than what we calculated at the end of the randomization test on gender discrimination – the probability of observed or more extreme data given the null hypothesis being true, i.e. P(data | hypothesis), also called a p-value.

Bayesian inference

updating the prior

- In the Bayesian approach, we evaluate claims iteratively as we collect more data.
- In the next iteration (roll) we get to take advantage of what we learned from the data.
- In other words, we update our prior with our posterior probability from the previous iteration.

updated:

$P(H_1$: good die on the Right)	$P(H_2$: good die on the Left)
0.6	0.4

Bayesian inference

recap

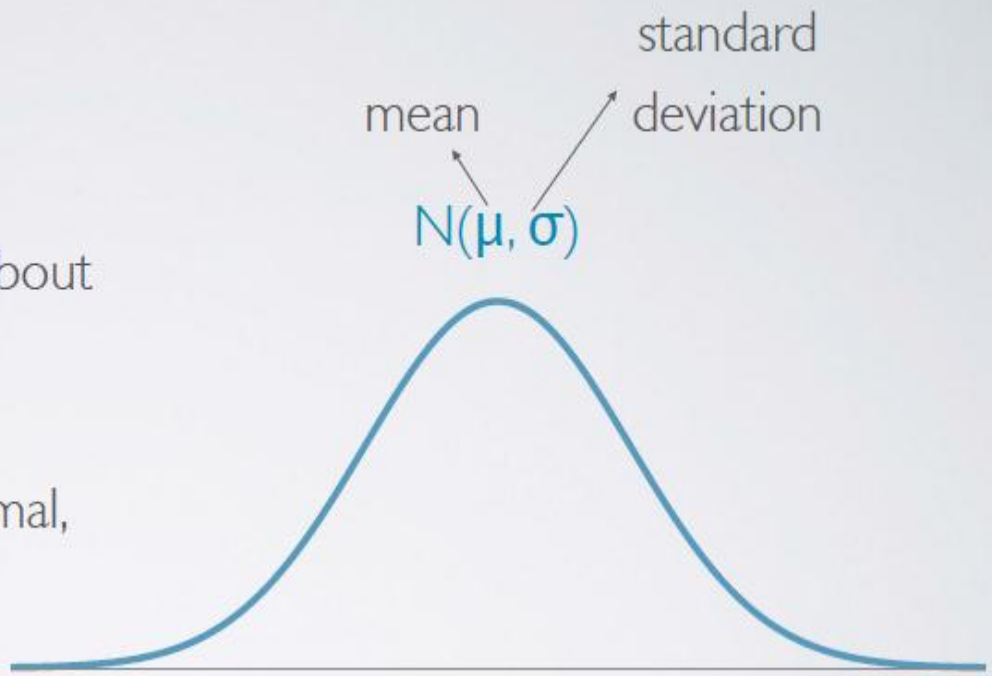
- Take advantage of prior information, like a previously published study or a physical model.
- Naturally integrate data as you collect it, and update your priors.
- Avoid the counter-intuitive definition of a p-value:

  P(observed or more extreme outcome | H0 is true)
- Instead base decisions on the posterior probability:

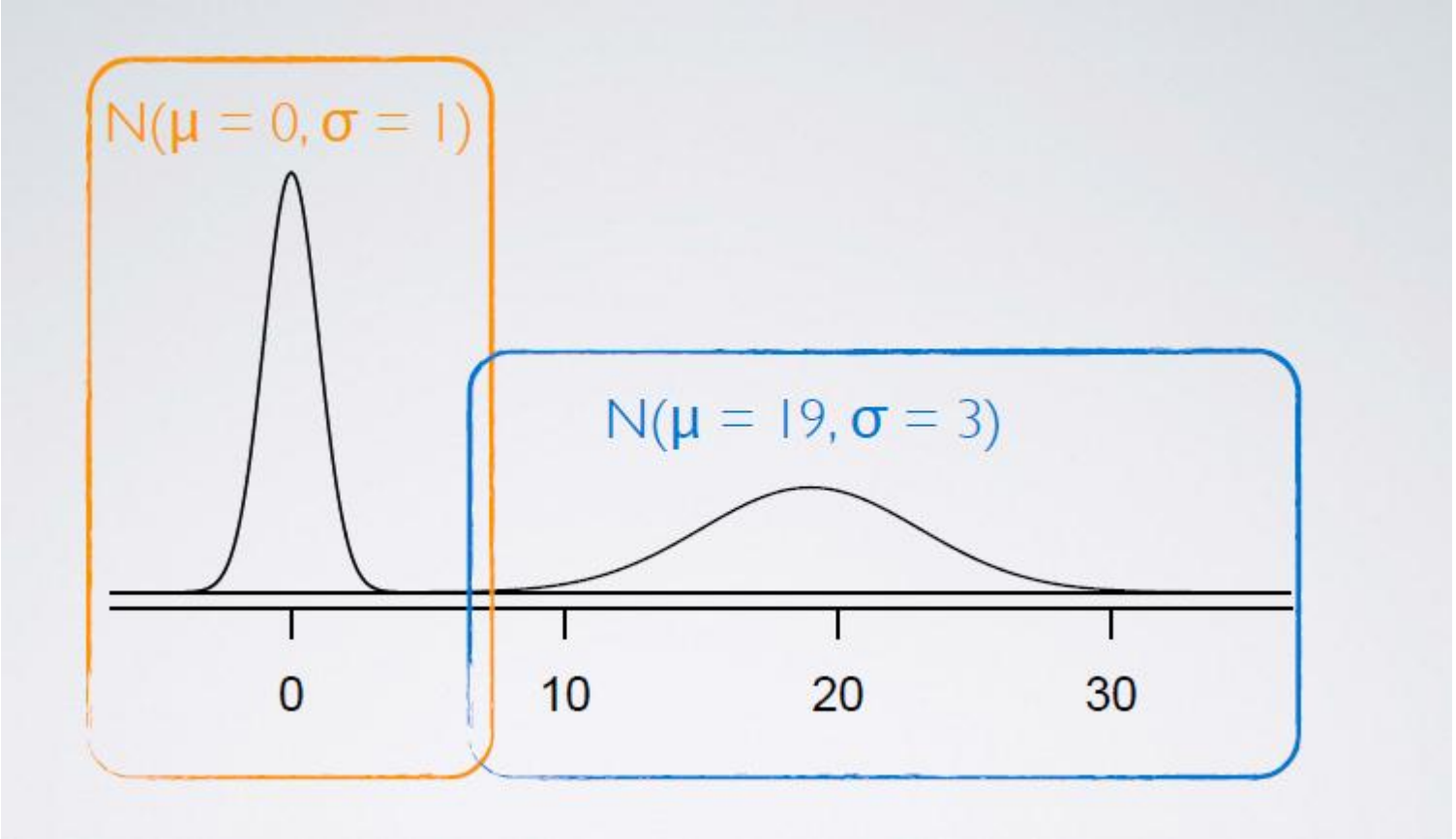
  P(hypothesis is true | observed data)
- A good prior helps, a bad prior hurts, but the prior matters less the more data you have.
- More advanced Bayesian techniques offer flexibility not present in Frequentist models.

Normal distribution

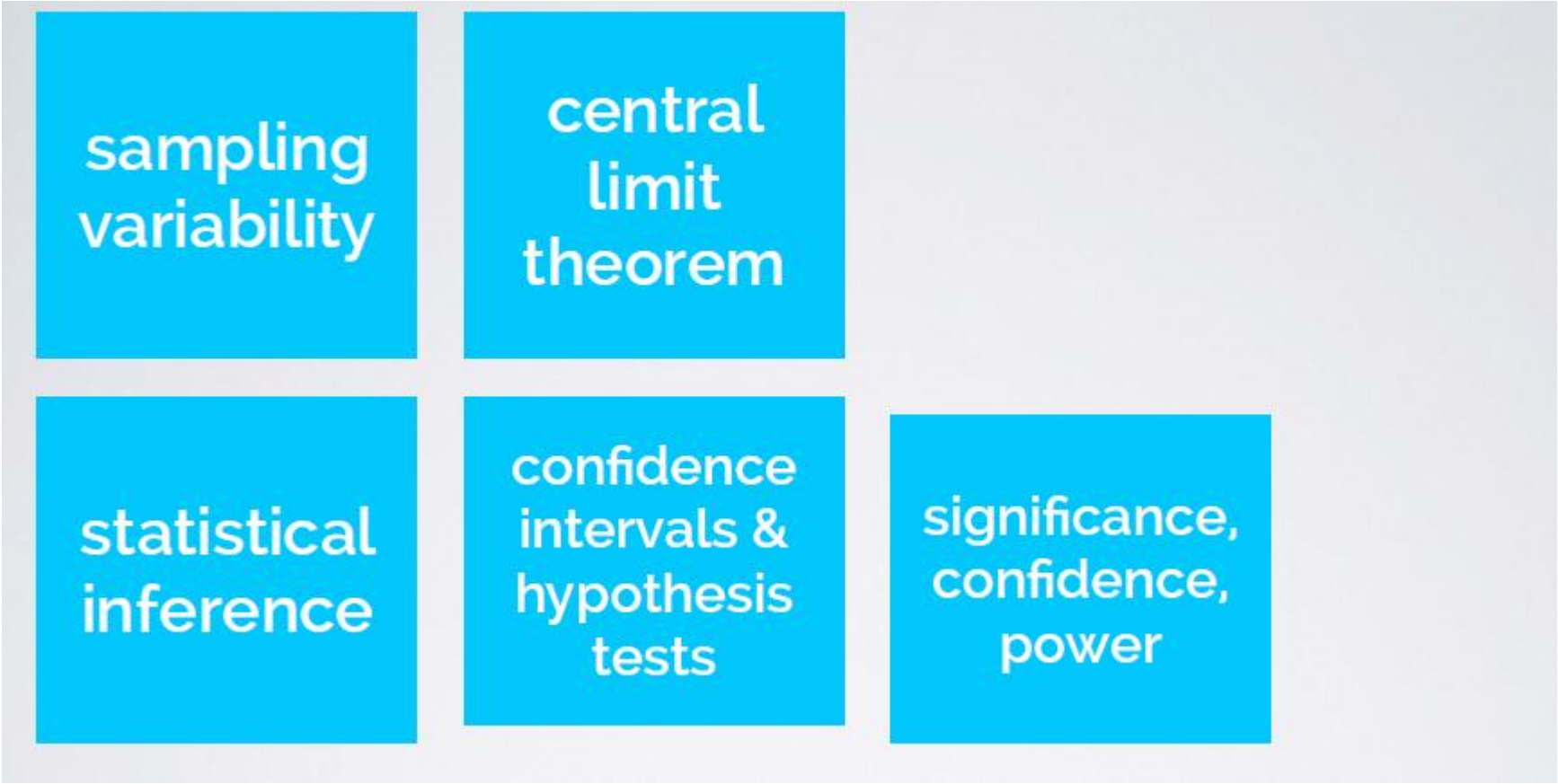
- unimodal and symmetric
  - bell curve
- follows very strict guidelines about how variably the data are distributed around the mean
- many variables are nearly normal, but none are exactly normal



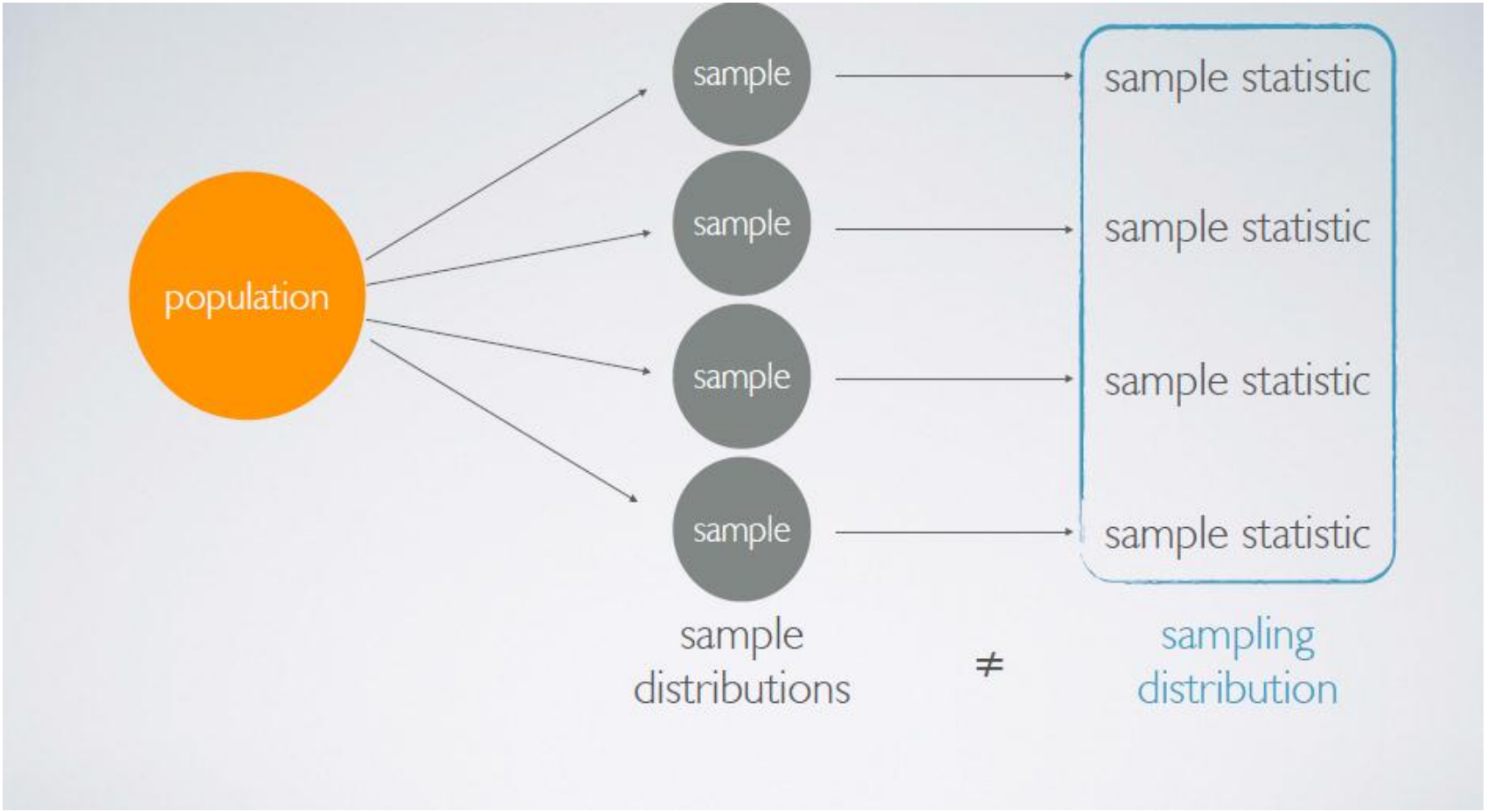
Normal distribution



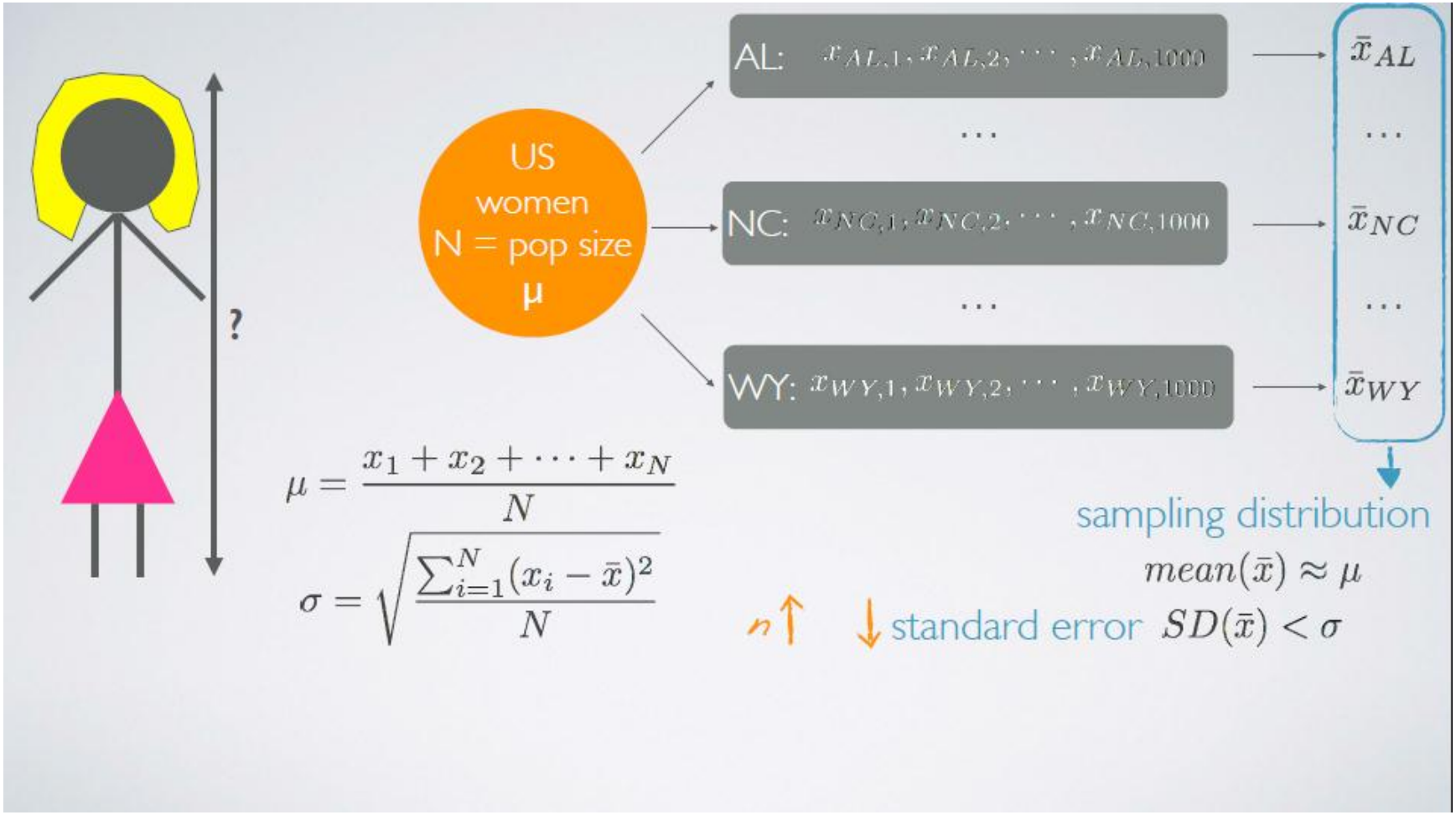
Mathematical models



Sampling distribution



Sampling distribution



Central Limit Theorem

Central Limit Theorem (CLT): The distribution of sample statistics is nearly normal, centered at the population mean, and with a standard deviation equal to the population standard deviation divided by square root of the sample size.

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$

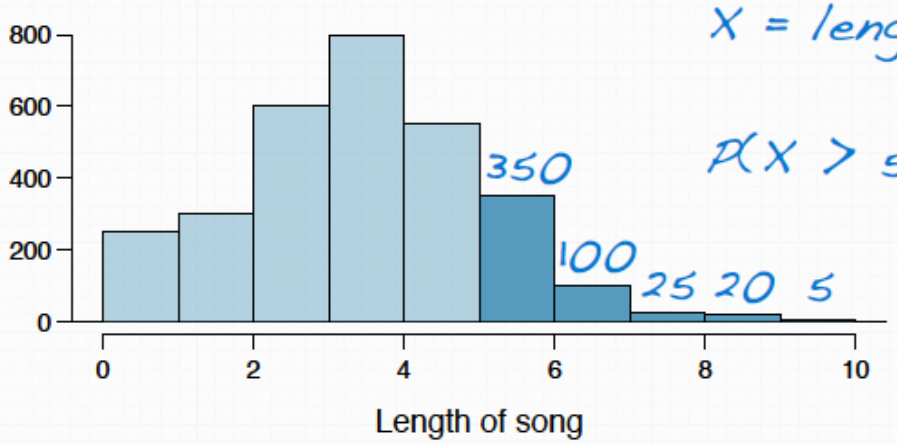
shape center spread

Conditions for the CLT:

1. ***Independence:*** Sampled observations must be independent.
   - random sample/assignment
   - if sampling without replacement, n < 10% of population
2. ***Sample size/skew:*** Either the population distribution is normal, or if the population distribution is skewed, the sample size is large (rule of thumb: n > 30).

Example

Suppose my iPod has 3,000 songs. The histogram below shows the distribution of the lengths of these songs. We also know that, for this iPod, the mean length is 3.45 minutes and the standard deviation is 1.63 minutes. Calculate the probability that a randomly selected song lasts more than 5 minutes.



X = length of one song

$$P(X > 5) = \frac{350 + 100 + 25 + 20 + 5}{3000}$$

$$= 500 / 3000$$

$$\approx 0.17$$

Example

I'm about to take a trip to visit my parents and the drive is 6 hours. I make a random playlist of 100 songs. What is the probability that my playlist lasts the entire drive?

6 hours = 360 minutes

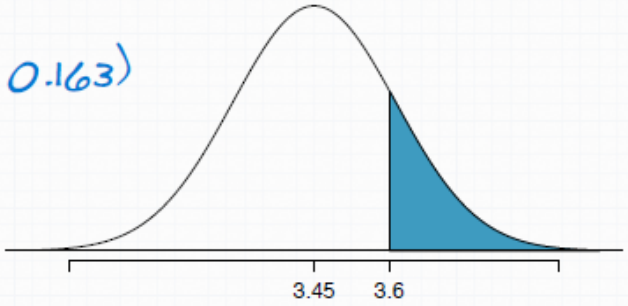
$P(X_1 + X_2 + \ldots + X_{100} > 360 \text{ min}) = ?$

$P(\bar{X} > 3.6) = ?$

$\bar{X} \sim N(\text{mean} = \mu = 3.45, SE = \frac{\sigma}{\sqrt{n}} = \frac{1.63}{\sqrt{100}} = 0.163)$

$Z = \frac{3.6 - 3.45}{0.163} = 0.92$

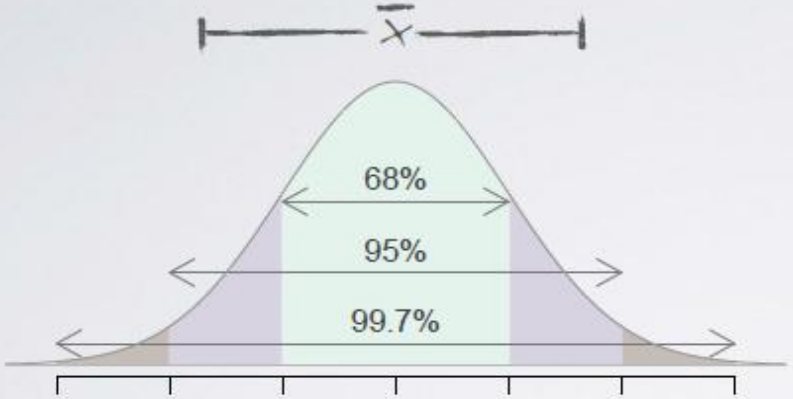
$P(Z > 0.92) = 0.179$



Confidence interval

Central Limit Theorem (CLT):

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$



approximate 95% CI: $\bar{x} \pm 2SE$

margin of error (ME)

Confidence interval

Confidence interval for a population mean: Computed as the sample mean plus/minus a margin of error (critical value corresponding to the middle XX% of the normal distribution times the standard error of the sampling distribution).

$$\bar{x} \pm z^\star \frac{s}{\sqrt{n}}$$

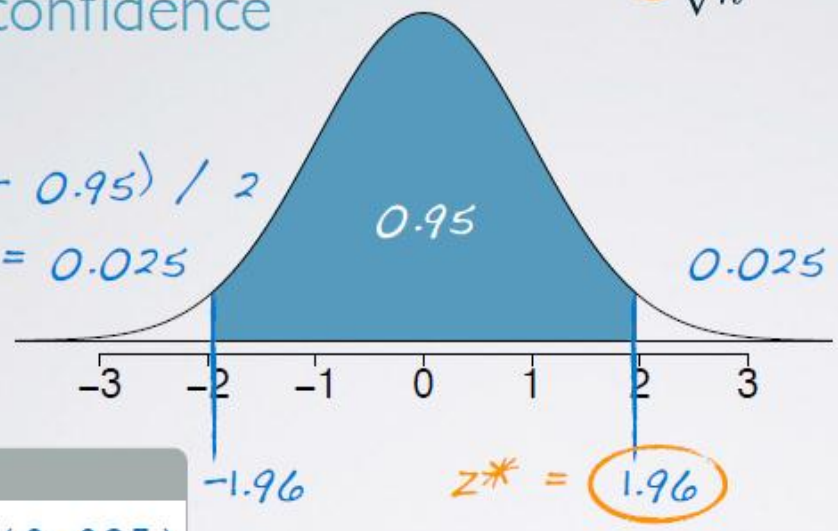
Conditions for this confidence interval:

1. ***Independence:*** Sampled observations must be independent.
   - random sample/assignment
   - if sampling without replacement, n < 10% of population
2. ***Sample size/skew:*** n ≥ 30, larger if the population distribution is very skewed.

Confidence interval

finding the critical value
95% confidence

$$\bar{x} \pm z^\star \frac{s}{\sqrt{n}}$$



R

```
> qnorm(0.025)
[1] -1.96
```

Second decimal place					
0.07	0.06	0.05	0.04	0.00	z
0.0003	0.0003	0.0003	0.0003	0.0003	−3.4
0.0004	0.0004	0.0004	0.0004	0.0005	−3.3
0.0005	0.0006	0.0006	0.0006	0.0007	−3.2
0.0008	0.0008	0.0008	0.0008	0.0010	−3.1
0.0011	0.0011	0.0011	0.0012	0.0013	−3.0
0.0015	0.0015	0.0016	0.0016	0.0019	−2.9
0.0021	0.0021	0.0022	0.0023	0.0026	−2.8
0.0028	0.0029	0.0030	0.0031	0.0035	−2.7
0.0038	0.0039	0.0040	0.0041	0.0047	−2.6
0.0051	0.0052	0.0054	0.0055	0.0062	−2.5
0.0068	0.0069	0.0071	0.0073	0.0082	−2.4
0.0089	0.0091	0.0094	0.0096	0.0107	−2.3
0.0116	0.0119	0.0122	0.0125	0.0139	−2.2
0.0150	0.0154	0.0158	0.0162	0.0179	−2.1
0.0192	0.0197	0.0202	0.0207	0.0228	−2.0
0.0244	0.0250	0.0256	0.0262	0.0287	−1.9
0.0307	0.0314	0.0322	0.0329	0.0359	−1.8

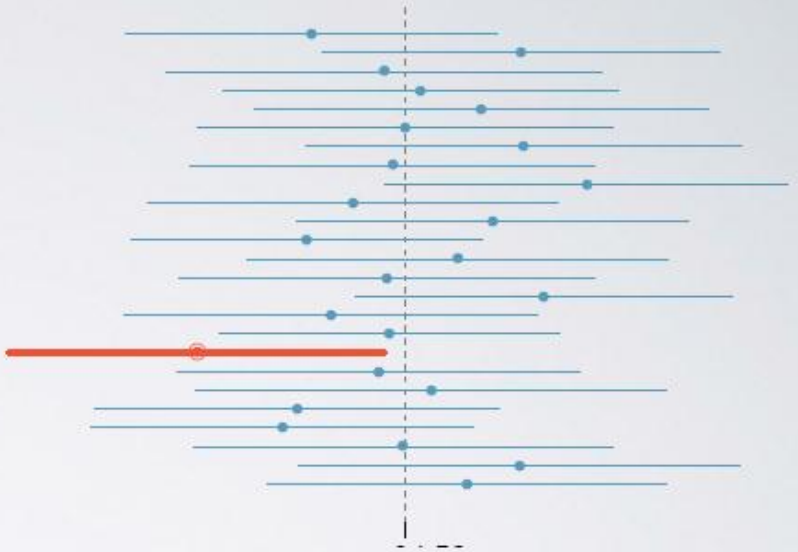
Confidence level

confidence level

- Suppose we took many samples and built a confidence interval from each sample using the equation

$$point\ estimate \pm 1.96 \times SE$$

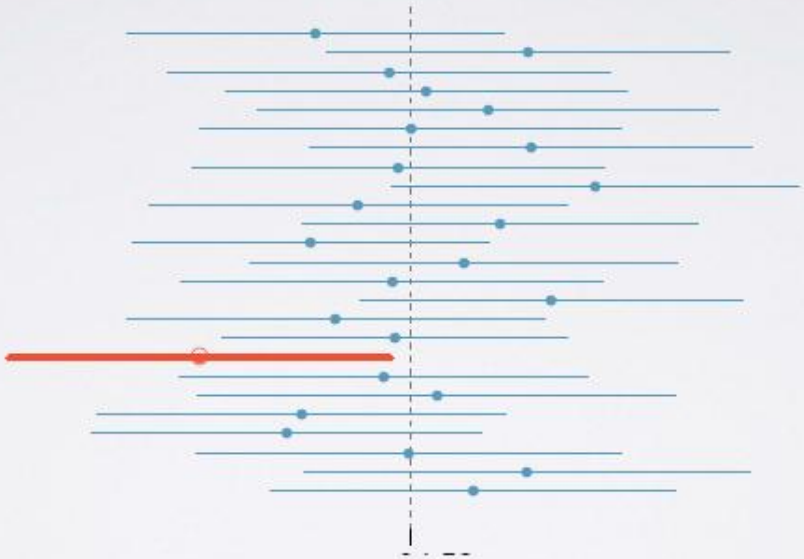
- Then about 95% of those intervals would contain the true population mean (μ).
- Commonly used confidence levels in practice are 90%, 95%, 98%, and 99%.



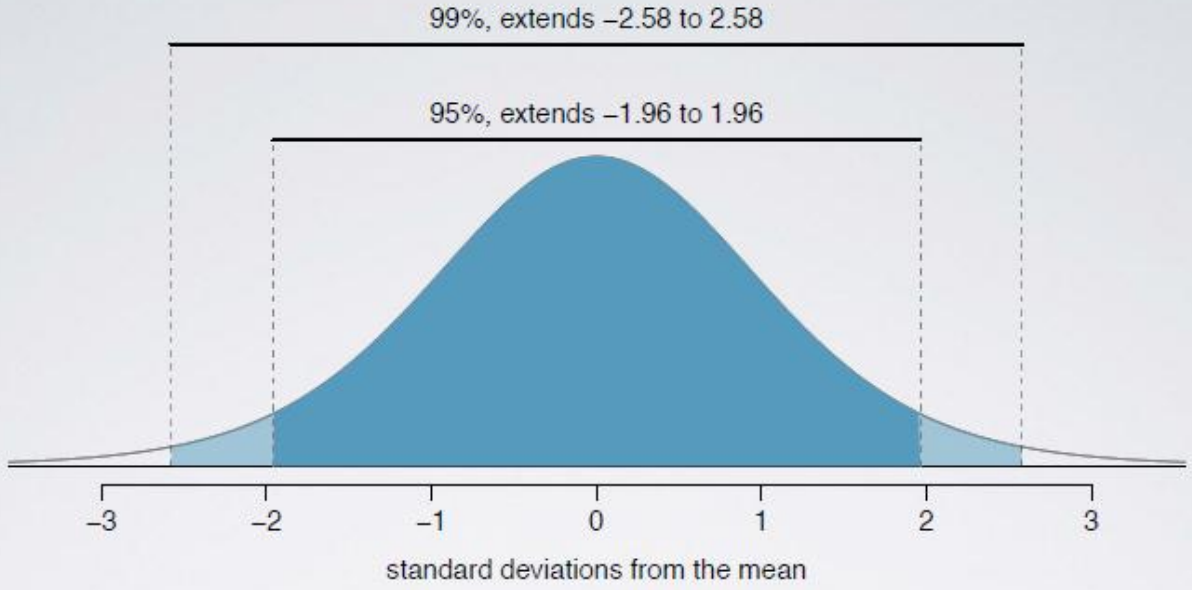
24 / 25 = 0.96

Confidence level

If we want to be very certain that we capture the population parameter, should we use a wider interval or a narrower interval?

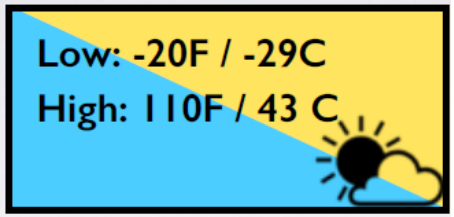


Confidence level



CL ↑ width ↑ accuracy ↑

precision ↓



Confidence level

Required sample size

backtracking to n for a given ME

given a target margin of error, confidence level, and information on the variability of the sample (or the population), we can determine the required sample size to achieve the desired margin of error.

$$ME = z^\star \frac{s}{\sqrt{n}} \quad \rightarrow n = \left(\frac{z^\star s}{ME}\right)^2$$

Examples: Confidence interval

The General Social Survey asks: "For how many days during the past 30 days was your mental health, which includes stress, depression, and problems with emotions, not good?" Based on responses from 1,151 US residents, the survey reported a 95% confidence interval of 3.40 to 4.24 days in 2010. Interpret this interval in context of the data.

We are 95% confident that Americans on average have 3.40 to 4.24 bad mental health days per month.

Examples: Confidence interval

The General Social Survey asks: "For how many days during the past 30 days was your mental health, which includes stress, depression, and problems with emotions, not good?" Based on responses from 1,151 US residents, the survey reported a 95% confidence interval of 3.40 to 4.24 days in 2010.

In this context, what does a 95% confidence level mean?

95% of random samples of 1,151 Americans will yield CIs that capture the true population mean of number of bad mental health days per month.

Examples: Confidence interval

The General Social Survey asks: "For how many days during the past 30 days was your mental health, which includes stress, depression, and problems with emotions, not good?" Based on responses from 1,151 US residents, the survey reported a 95% confidence interval of 3.40 to 4.24 days in 2010.

Suppose the researchers think a 99% confidence level would be more appropriate for this interval. Will this new interval be narrower or wider than the 95% confidence interval?

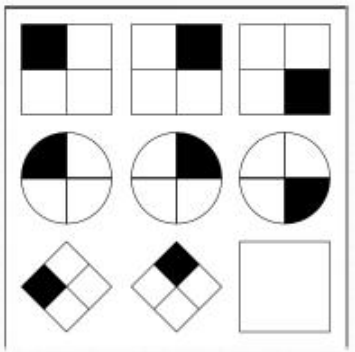
As CL increases so does the width of the confidence interval, so wider.

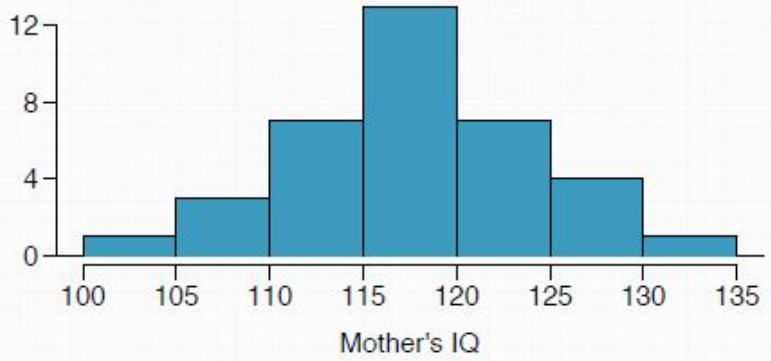
Hypothesis testing framework

- We start with a null hypothesis (H_0) that represents the status quo.
- We also have an alternative hypothesis (H_A) that represents our research question, i.e. what we're testing for.
- We conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation or theoretical methods — methods that rely on the CLT
- If the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, we stick with the null hypothesis. If they do, then we reject the null hypothesis in favor of the alternative.

Example

Researchers investigating characteristics of gifted children collected data from schools in a large city on a random sample of thirty-six children who were identified as gifted children soon after they reached the age of four. In this study, along with variables on the children, the researchers also collected data on their mothers' IQ scores. The histogram shows the distribution of these data, and also provided are some sample statistics.





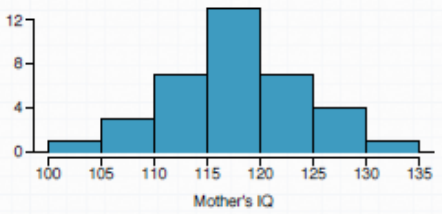
n	36
min	101
mean	118.2
sd	6.5
max	131

Example

Perform a hypothesis test to evaluate if these data provide convincing evidence of a difference between the average IQ score of mothers of gifted children and the average IQ score for the population at large, which is 100. Use a significance level of 0.01.

1. Set the hypotheses μ = average IQ score of mothers of gifted children

$H_0: \mu = 100$ $H_A: \mu \neq 100$



n	36
min	101
mean	118.2
sd	6.5
max	131

2. Calculate the point estimate

$\bar{x} = 118.2$

3. Check conditions

1. random & 36 < 10% of all gifted children -> independence
2. n > 30 & sample not skewed -> nearly normal sampling distribution

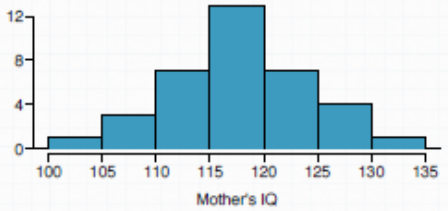
Example

$H_0: \mu = 100$

$H_A: \mu \neq 100$

$\bar{x} = 118.2$

$\bar{X} \sim N(\mu = 100, SE = \frac{s}{\sqrt{n}} = \frac{6.5}{\sqrt{36}} \approx 1.083)$

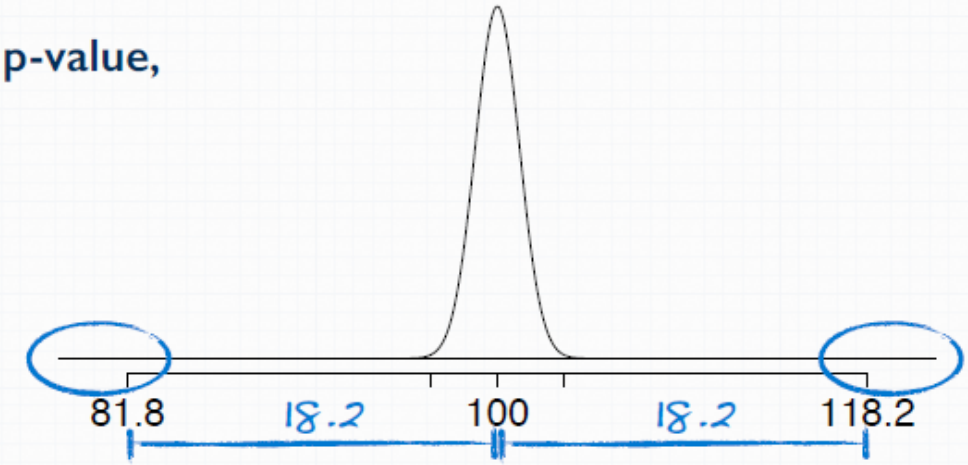


n	36
min	101
mean	118.2
sd	6.5
max	131

4. Draw sampling distribution, shade p-value, calculate test statistic

$$Z = \frac{118.2 - 100}{1.083} = 16.8$$

p-value ≈ 0



Example

5. **Make a decision, and interpret it in context of the research question**

p-value is very low -> strong evidence against the null

We reject the null hypothesis and conclude that the data provide convincing evidence of a difference between the average IQ score of mothers of gifted children and the average IQ score for the population at large.

Inference for other estimators

nearly normal sampling distributions

sample mean	$\bar{x}$
difference between sample means	$\bar{x}_1 - \bar{x}_2$
sample proportion	$\hat{p}$
difference between sample proportions	$\hat{p}_1 - \hat{p}_2$

Inference for other estimators

unbiased estimator

An important assumption about point estimates is that they are unbiased, i.e. the sampling distribution of the estimate is centered at the true population parameter it estimates.

- That is, an unbiased estimate does not naturally over or underestimate the parameter, it provides a "good" estimate.
- The sample mean is an example of an unbiased point estimate, as well as others we just listed.

Inference for other estimators

confidence intervals
for nearly normal point estimates

$$point\ estimate \pm z^\star \times SE$$

Process of using normal model

- **Frame the research question.** The mathematical model can be applied to both the hypothesis testing and the confidence interval framework. Make sure that your research question is being addressed by the most appropriate inference procedure.
- **Collect data with an observational study or experiment.** To address the research question, collect data on the variables of interest. Note that your data may be a random sample from a population or may be part of a randomized experiment.
- **Model the randomness of the statistic.** In many cases, the normal distribution will be an excellent model for the randomness associated with the statistic of interest. The Central Limit Theorem tells us that if the sample size is large enough, sample averages (which can be calculated as either a proportion or a sample mean) will be approximately normally distributed when describing how the statistics change from sample to sample.
- **Calculate the variability of the statistic.** Using formulas, come up with the standard deviation (or more typically, an estimate of the standard deviation called the standard error) of the statistic. The SE of the statistic will give information on how far the observed statistic is from the null hypothesized value (if performing a hypothesis test) or from the unknown population parameter (if creating a confidence interval).
- **Use the normal distribution to quantify the variability.** The normal distribution will provide a probability which measures how likely it is for your observed and hypothesized (or observed and unknown) parameter to differ by the amount measured. The unusualness (or not) of the discrepancy will form the conclusion to the research question.
- **Form a conclusion.** Using the p-value or the confidence interval from the analysis, report on the research question of interest. Also, be sure to write the conclusion in plain language so casual readers can understand the results.

Decision errors

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✔	Type I error
	H_A true	Type 2 error	✔

- Type I error is rejecting H_0 when H_0 is true.
- Type 2 error is failing to reject H_0 when H_A is true.
- We (almost) never know if H_0 or H_A is true, but we need to consider all possibilities.

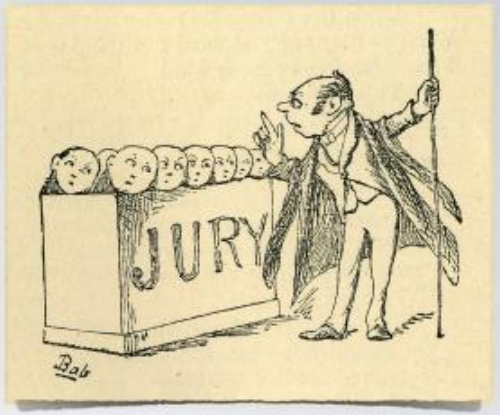
Decision errors

hypothesis test as a trial

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

H_0 : Defendant is innocent

H_A : Defendant is guilty



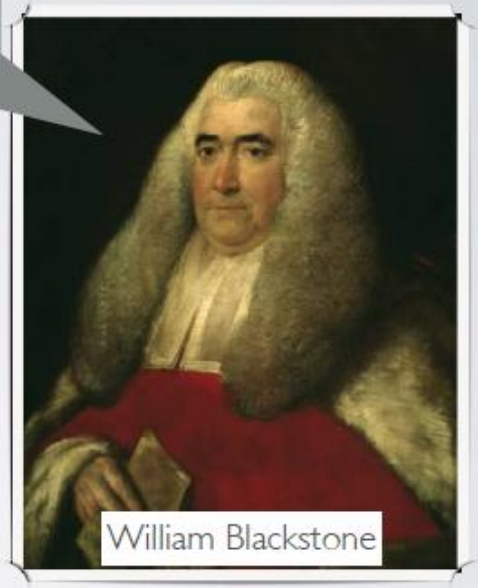
Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty — Type 2 error
- Declaring the defendant guilty when they are actually innocent — Type 1 error

Jury: http://upload.wikimedia.org/wikipedia/commons/5/5d/Trial_by_Jury_Usher.jpg

Decision errors

"better that ten guilty persons escape than that one innocent suffer"



Which error is the worst error to make?

- Type 2 : Declaring the defendant innocent when they are actually guilty
- Type 1 : Declaring the defendant guilty when they are actually innocent

William Blackstone: http://en.wikipedia.org/wiki/File:SirWilliamBlackstone.jpg

Decision errors

type I error rate

- We reject H_0 when the p-value is less than 0.05 ($\alpha = 0.05$).
- This means that, for those cases where H_0 is actually true, we do not want to incorrectly reject it more than 5% of those times.
- In other words, when using a 5% significance level there is about 5% chance of making a Type I error if the null hypothesis is true.

$$P(\text{Type I error} \mid H_0 \text{ true}) = \alpha$$

- This is why we prefer small values of α – increasing α increases the Type I error rate.

Decision errors

If Type 1 Error is dangerous or especially costly, choose a small significance level (e.g. 0.01).

Goal: we want to be very cautious about rejecting H_0, so we demand very strong evidence favoring H_A before we would do so.

choosing α

If a Type 2 Error is relatively more dangerous or much more costly, choose a higher significance level (e.g. 0.10).

Goal: we want to be cautious about failing to reject H_0 when the null is actually false.

Scale: http://commons.wikimedia.org/wiki/File:US_Department_of_Justice_Scales_Of_Justice.svg

Decision errors

goal: keep α and β low

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	$1 - \alpha$	Type I error, α
	H_A true	Type 2 error, β	$1 - \beta$

- Type I error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level).
- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β.
- Power of a test is the probability of correctly rejecting H_0, and the probability of doing so is $1 - \beta$

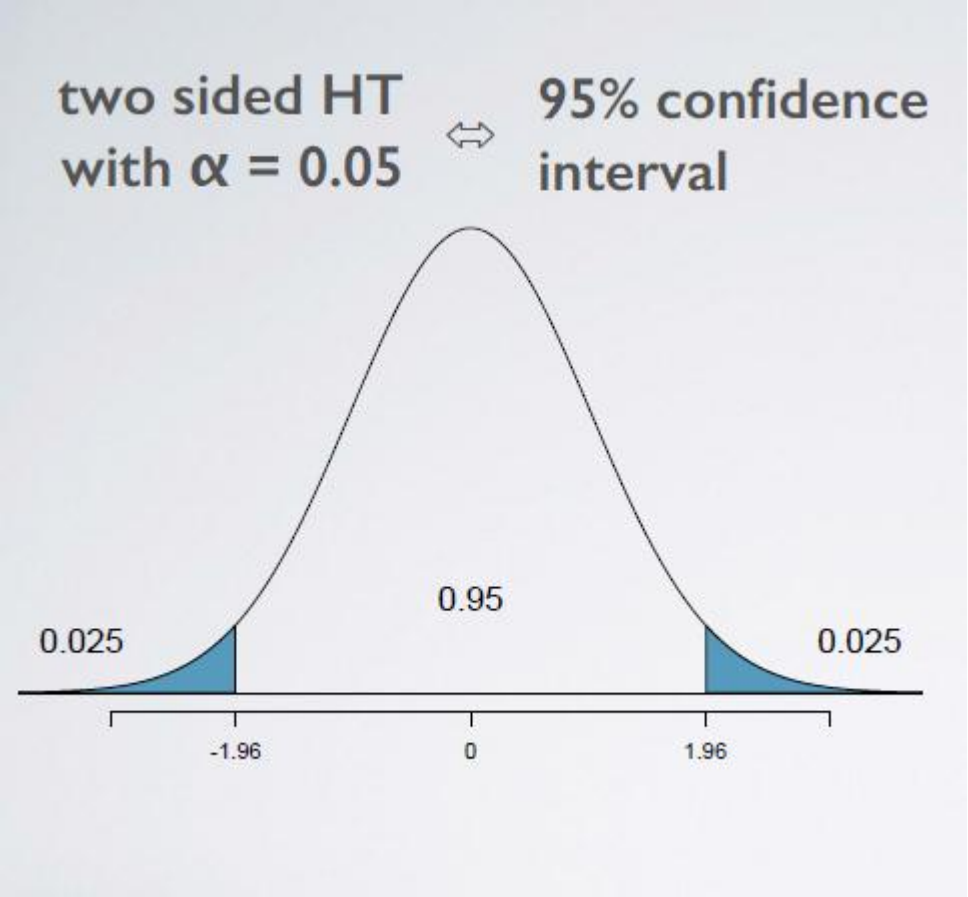
Decision errors

type 2 error rate

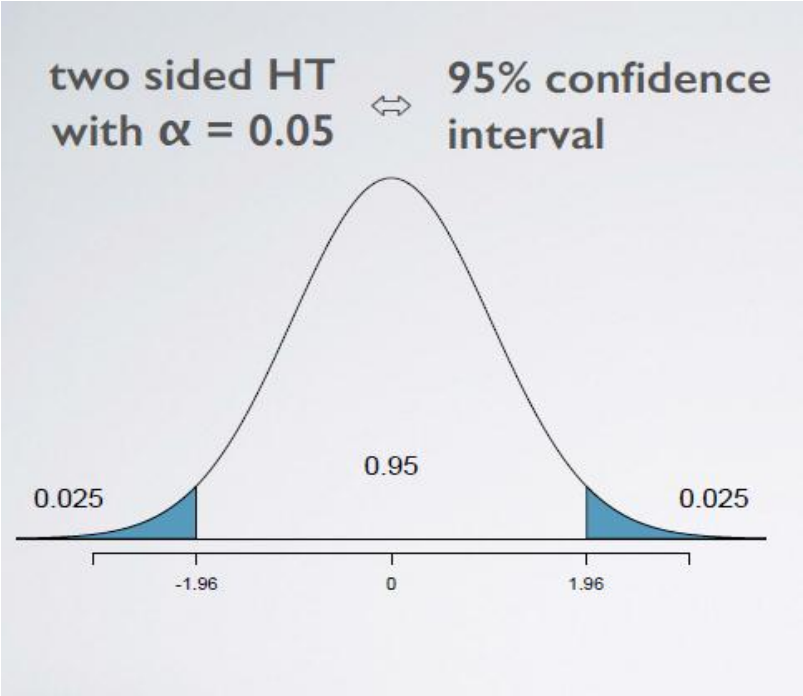
If the alternative hypothesis is actually true, what is the chance that we make a Type 2 Error, i.e. we fail to reject the null hypothesis even when we should reject it?

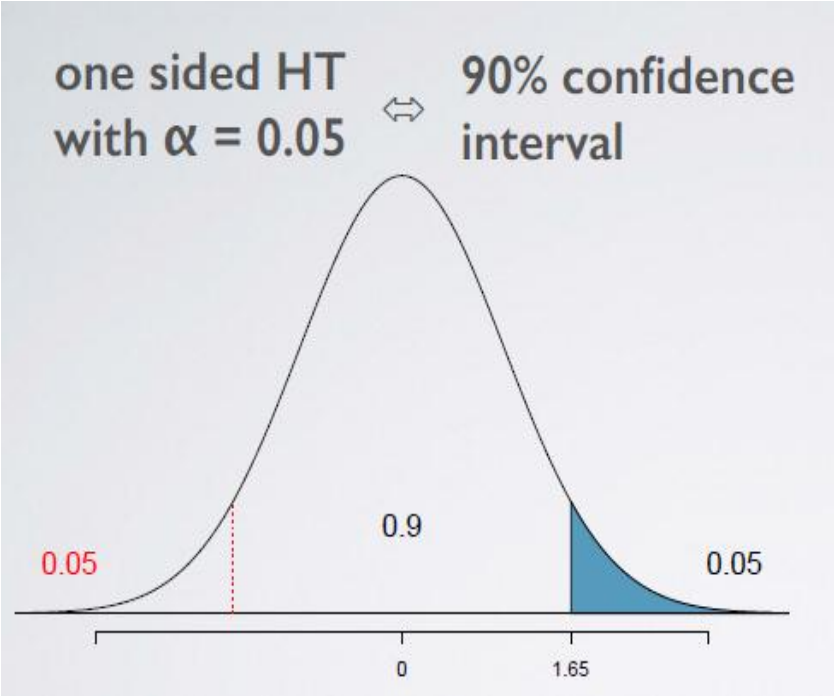
- The answer is not obvious.
- If the true population average is very close to the null value, it will be difficult to detect a difference (and reject H_0).
- If the true population average is very different from the null value, it will be easier to detect a difference.
- Clearly, β depends on the effect size (δ), difference between point estimate and null value.

Significance vs confidence level



Significance vs confidence level





Significance vs confidence level

agreement of CI and HT

- A two sided hypothesis with threshold of α is equivalent to a confidence interval with CL = 1 − α.
- A one sided hypothesis with threshold of α is equivalent to a confidence interval with CL = 1 − (2 × α).
- If H_0 is rejected, a confidence interval that agrees with the result of the hypothesis test should not include the null value.
- If H_0 is failed to be rejected, a confidence interval that agrees with the result of the hypothesis test should include the null value.

Summary on inferential statistical methods

Table 15.1: Summary and comparison of randomization, bootstrapping, and mathematical models as inferential statistical methods.

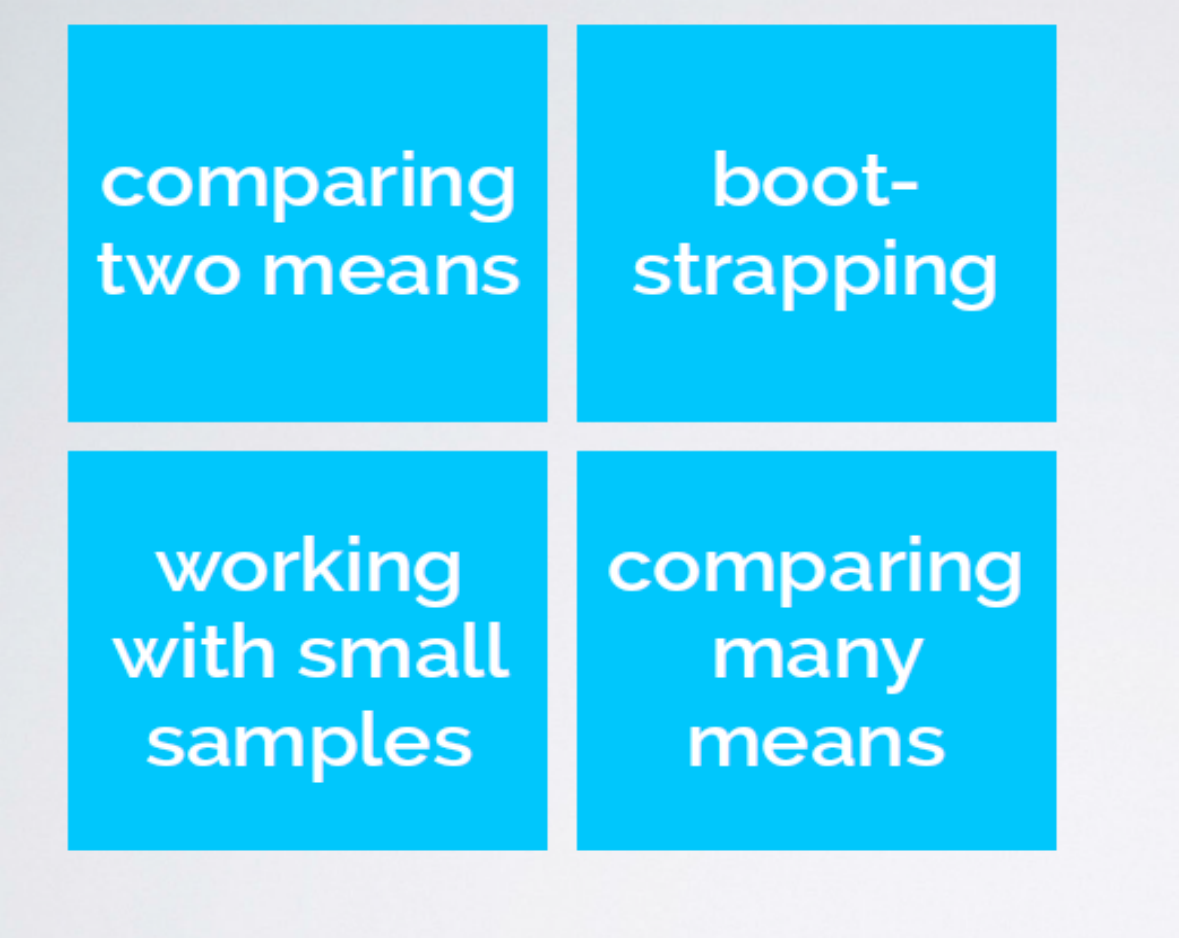
	Answer		
Question	Randomization	Bootstrapping	Mathematical models
What does it do?	Shuffles the explanatory variable to mimic the natural variability found in a randomized experiment	Resamples (with replacement) from the observed data to mimic the sampling variability found by collecting data from a population	Uses theory (primarily the Central Limit Theorem) to describe the hypothetical variability resulting from either repeated randomized experiments or random samples
What is the random process described?	Randomized experiment	Random sampling from a population	Randomized experiment or random sampling

Summary on inferential statistical methods

Question	Answer		
	Randomization	Bootstrapping	Mathematical models
What other random processes can be approximated?	Can also be used to describe random sampling in an observational model	Can also be used to describe random allocation in an experiment	Can also be used to describe random sampling in an observational model or random allocation in an experiment
What is it best for?	Hypothesis testing (can also be used for confidence intervals, but not covered in this text)	Confidence intervals (can also be used for bootstrap hypothesis testing for one proportion as well)	Quick analyses through, for example, calculating a Z score
What physical object represents the simulation process?	Shuffling cards	Pulling marbles from a bag with replacement	Not applicable

Statistical inference

Inference for numerical variables



Hypothesis testing for paired data

high school and beyond

200 observations were randomly sampled from the High School and Beyond survey. The same students took a reading and writing test. At a first glance, how are the distributions of reading and writing scores similar? How are they different?

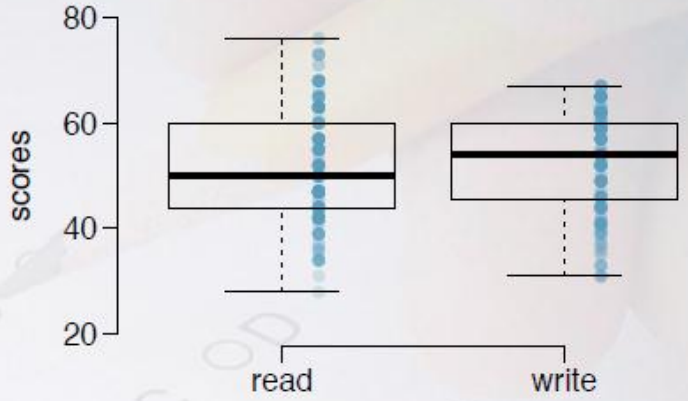


Photo by Alberto G. http://www.flickr.com/photos/albertogp123/5843577306/ (CC BY 2.0)

Hypothesis testing for paired data

Given that the same students took the reading and the writing tests, are the reading and writing scores of each student independent of each other?

	ID	read	write
1	70	57	52
2	86	44	33
3	141	63	44
4	172	47	52
…	…	…	…
200	137	63	65

Hypothesis testing for paired data

analyzing paired data

- When two sets of observations have this special correspondence (not independent), they are said to be paired.
- To analyze paired data, it is often useful to look at the difference in outcomes of each pair of observations:

  $$\text{diff} = \text{read} - \text{write}$$

- It is important that we always subtract using a consistent order.

	ID	read	write	diff
1	70	57	52	5
2	86	44	33	11
3	141	63	44	19
4	172	47	52	-5
…	…	…	…	…
200	137	63	65	-2

Hypothesis testing for paired data

parameter of interest

Average difference between the reading and writing scores of **all** high school students.

μ_{diff}

point estimate

Average difference between the reading and writing scores of **sampled** high school students.

$\bar{x}_{diff}$

Hypothesis testing for paired data

If in fact there was no difference between the scores on the reading and writing exams, what would you expect the average difference to be?

	ID	read	write	diff
1	70	57	52	5
2	86	44	33	11
3	141	63	44	19
4	172	47	52	-5
…	…	…	…	…
200	137	63	65	-2

$$\bar{x}_{diff} = -0.545$$
$$s_{diff} = 8.887$$
$$n_{diff} = 200$$

Hypothesis testing for paired data

hypotheses for paired means

$H_0 : \mu_{diff} = 0$ There is no difference between the average reading and writing scores.

$H_A : \mu_{diff} \neq 0$ There is a difference between the average reading and writing scores.

Hypothesis testing for paired data

nothing new!

one numerical variable

diff
5
11
19
-5
...
-2

hypothesis about the mean

$$H_0 : \mu_{diff} = 0$$
$$H_A : \mu_{diff} \neq 0$$

Hypothesis testing for paired data

Hypothesis testing for a ~~single mean~~ difference between paired means:

1. Set the hypotheses: $H_0 : \mu_{diff} = null\ value$
   $H_A : \mu_{diff} < or > or \neq null\ value$
2. Calculate the point estimate: $\bar{x}_{diff}$
3. Check conditions:
   1. ***Independence:*** Sampled observations must be independent (random sample/assignment & if sampling without replacement, n_{diff} < 10% of population)
   2. ***Sample size/skew:*** $n_{diff} \geq 30$, larger if the population distribution is very skewed.
4. Draw sampling distribution, shade p-value, calculate test statistic

$$Z = \frac{x_{diff} - \mu_{diff}}{SE_{\bar{x}_{diff}}}$$

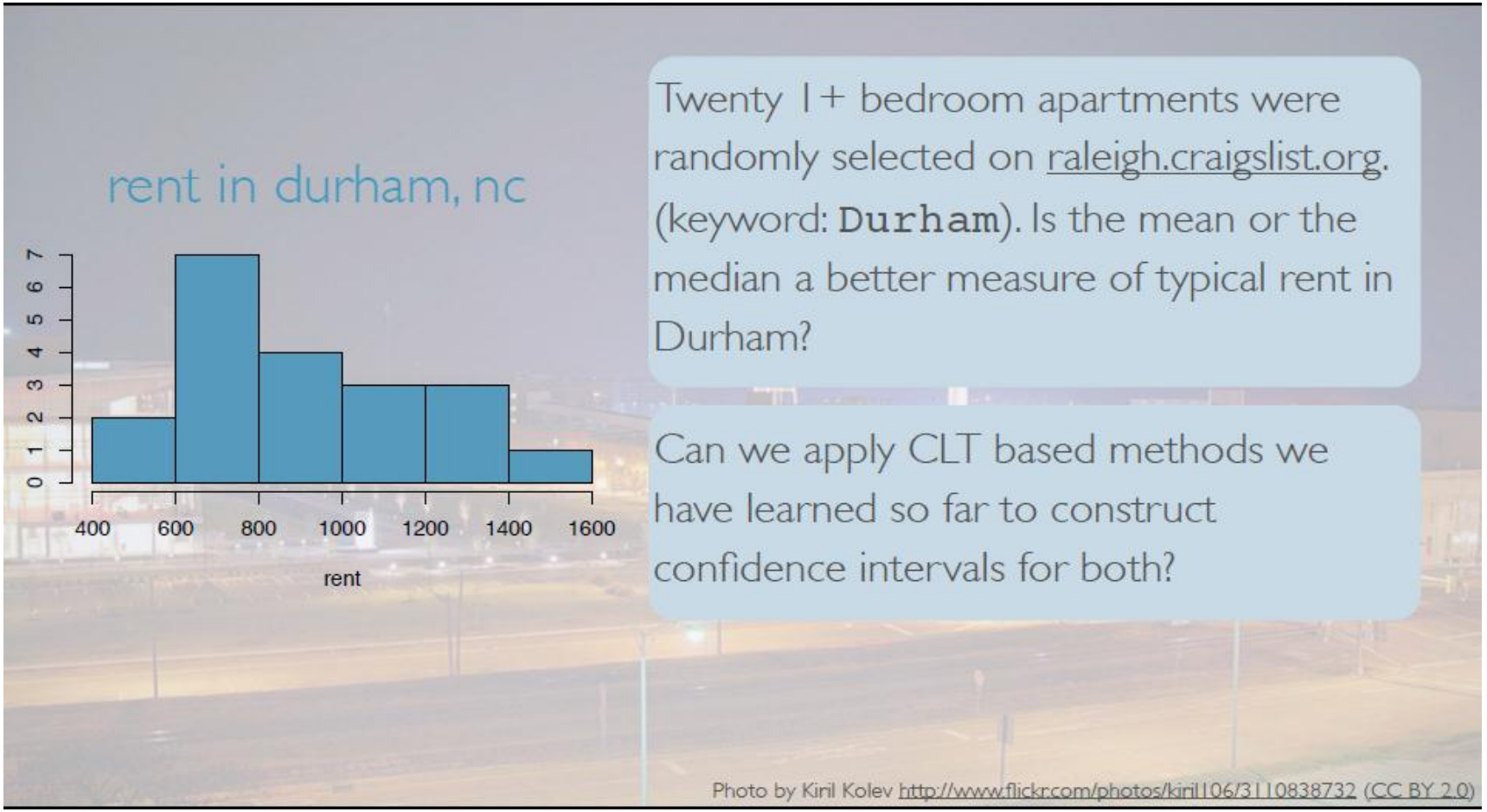
5. Make a decision, and interpret it in context of the research question:

Hypothesis testing for paired data

summary

- paired data (2 vars.) → differences (1 var.)
- most often $H_0 : \mu_{diff} = 0$
- same individuals: pre-post studies, repeated measures, etc.
- different (but dependent) individuals: twins, partners, etc.

Bootstrapping



Bootstrapping

- An alternative approach to constructing confidence intervals is bootstrapping.
- This term comes from the phrase *"pulling oneself up by one's bootstraps"*, which is a metaphor for accomplishing an impossible task without any outside help.
- In this case the ~~im~~possible task is estimating a population parameter, and we'll accomplish it using data from only the given sample.

Boots: http://openclipart.org/detail/26401/-by--26401

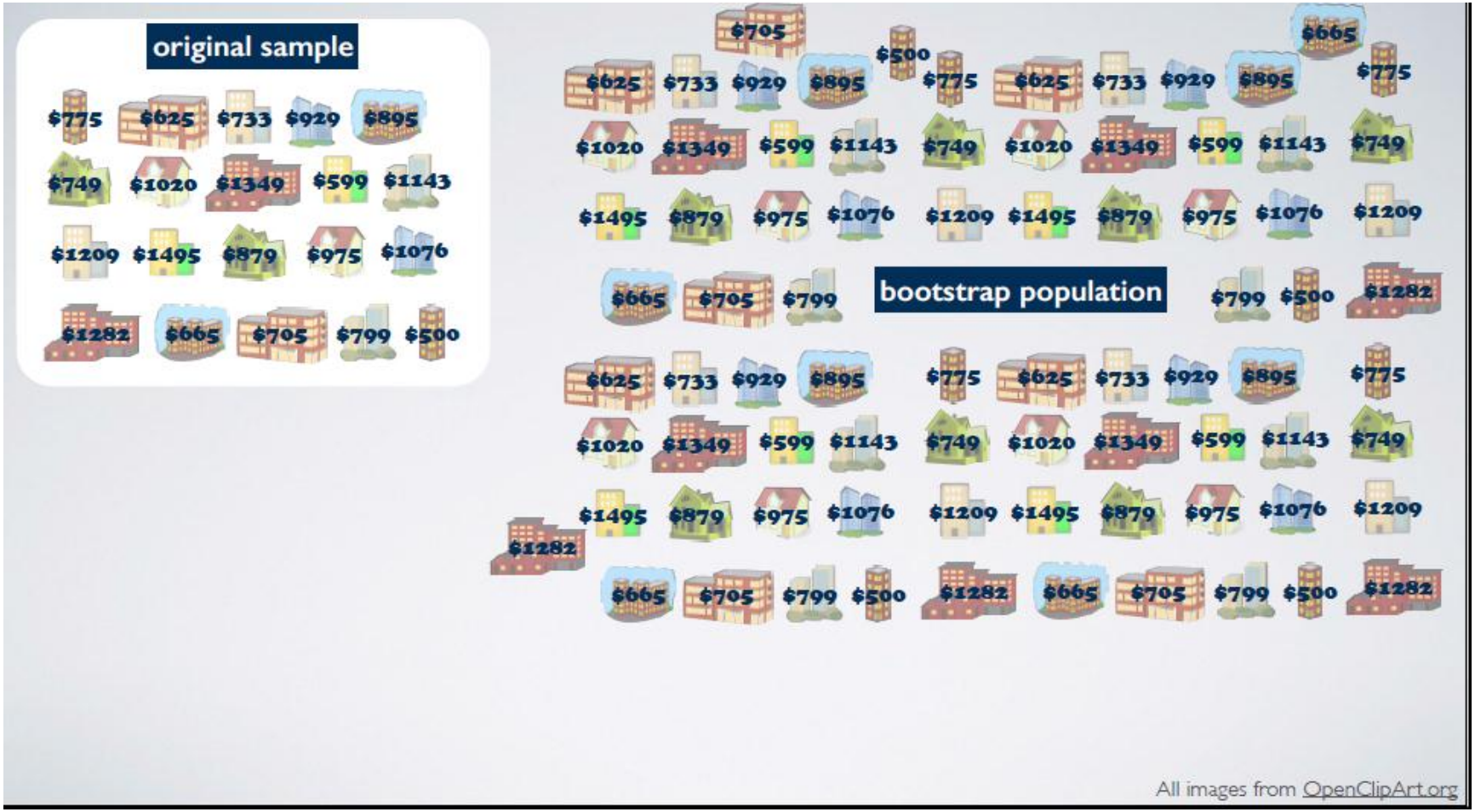
Bootstrapping

Bootstrapping

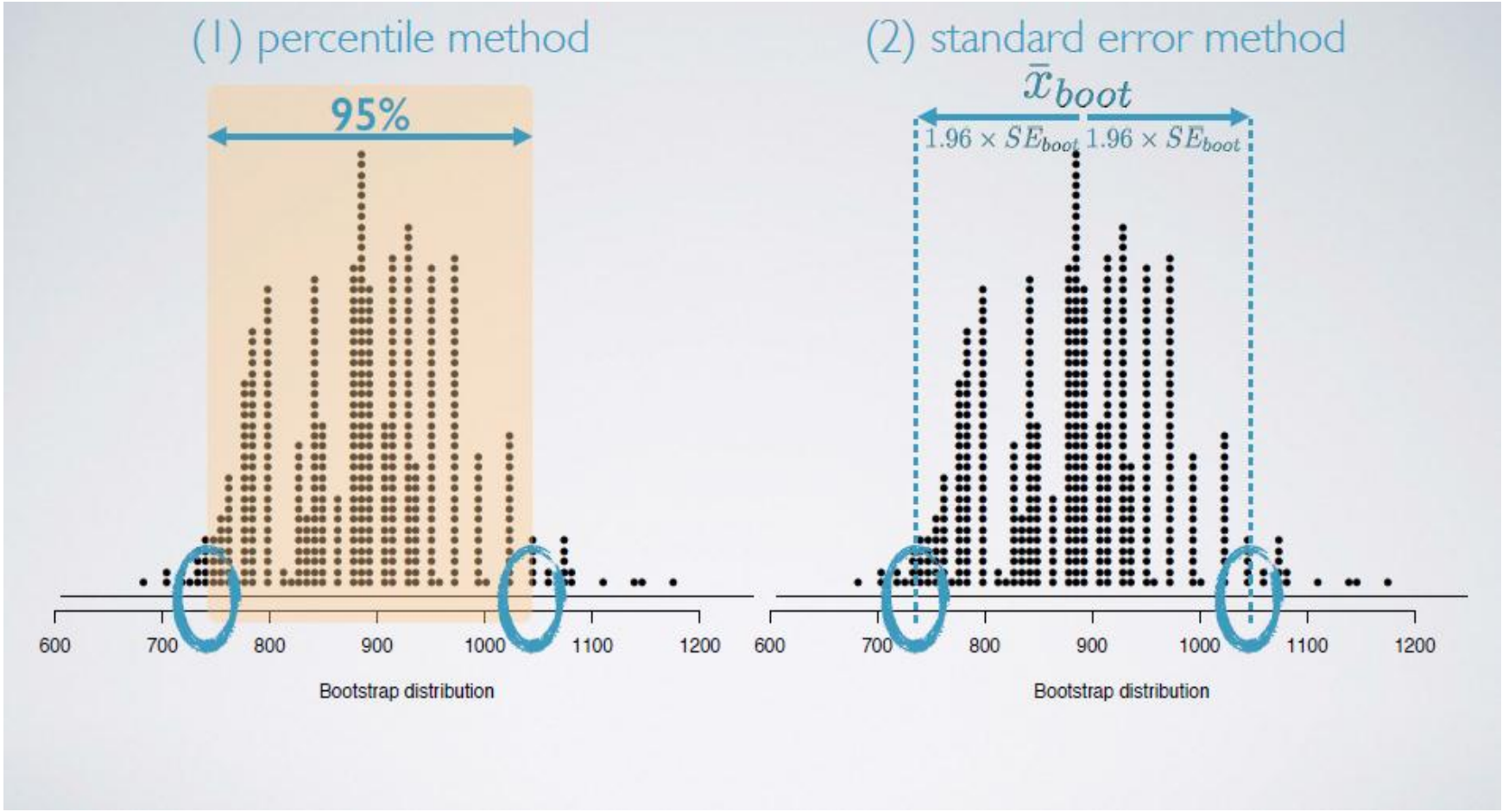
bootstrapping scheme

(1) take a bootstrap sample - a random sample taken **with replacement** from the original sample, of the same size as the original sample
(2) calculate the bootstrap statistic - a statistic such as mean, median, proportion, etc. computed on the bootstrap samples
(3) repeat steps (1) and (2) many times to create a bootstrap distribution - a distribution of bootstrap statistics

Bootstrapping



Bootstrapping



Bootstrapping limitations

- Not as rigid conditions as CLT based methods.
- However if the bootstrap distribution is extremely skewed or sparse, the bootstrap interval might be unreliable.
- A representative sample is required for generalizability. If the sample is biased, the estimates resulting from this sample will also be biased.

Bootstrapping vs sampling distribution

- Sampling distribution created using sampling (with replacement) from the population.
- Bootstrap distribution created using sampling (with replacement) from the sample.
- Both are distributions of sample statistics.

t distribution

- Student's t
- William Gosset (1876 - 1937)
- "Head Experimental Brewer" at the Guinness brewing company

Gosset: http://commons.wikimedia.org/wiki/File:William_Sealy_Gosset.jpg

review:
what purpose does a large sample serve?

As long as observations are independent, and the population distribution is not extremely skewed, a large sample would ensure that...

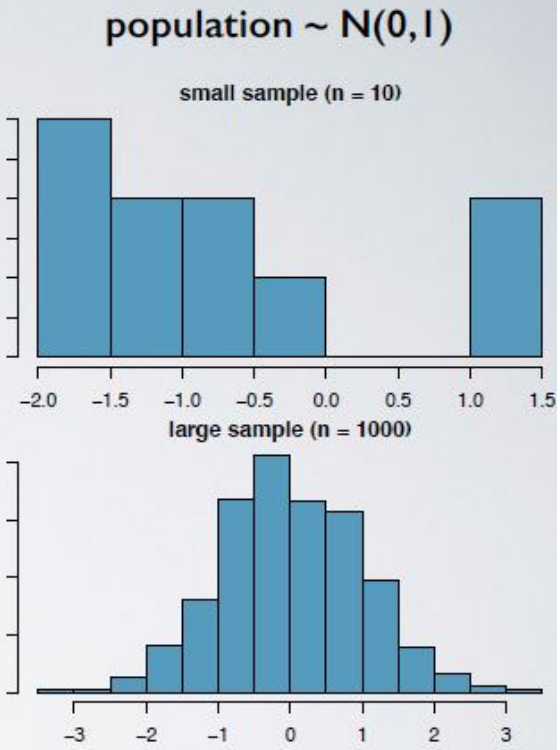
- the sampling distribution of the mean is nearly normal
- the estimate of the standard error is reliable: $\frac{s}{\sqrt{n}}$

Photo by Kheel Center, Cornell University on Flickr http://www.flickr.com/photos/kheelcenter/5279081507/ (CC BY 2.0)

t distribution

review:
normality of sampling distributions

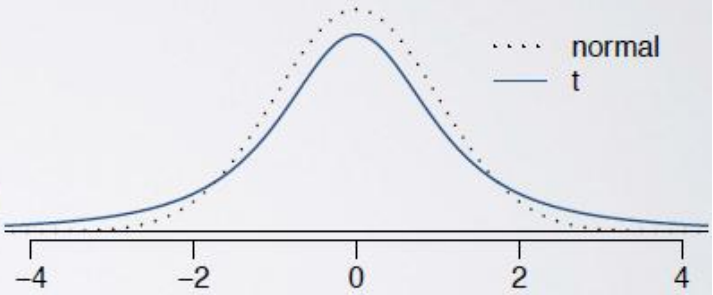
- CLT: sampling distributions are nearly normal as long as the population distribution is nearly normal, for **any** sample size.
- Helpful special case, but difficult to verify normality in small data sets.
- Careful with the normality condition for small samples: don't just examine the sample, also think about where the data come from.
  - *"Would I expect this distribution to be symmetric, and am I confident that outliers are rare?"*



t distribution

t distribution

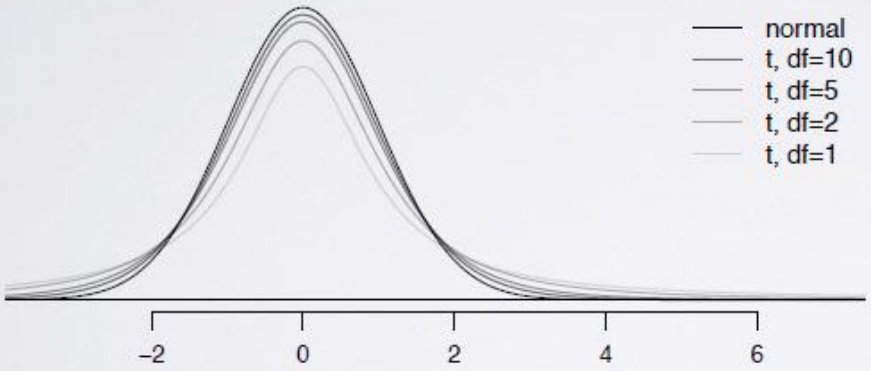
- n is small & σ unknown (almost always), use the t distribution to address the uncertainty of the standard error estimate
- bell shaped but thicker tails than the normal
  - observations more likely to fall beyond 2 SDs from the mean
  - extra thick tails helpful for mitigating the effect of a less reliable estimate for the standard error of the sampling distribution



t distribution

t distribution

- always centered at 0 (like the standard normal)
- has one parameter: degrees of freedom (df) - determines thickness of tails
  - remember, the normal distribution has two parameters: mean and SD



What happens to the shape of the t-distribution as degrees of freedom increases?

approaches the normal dist.

t distribution

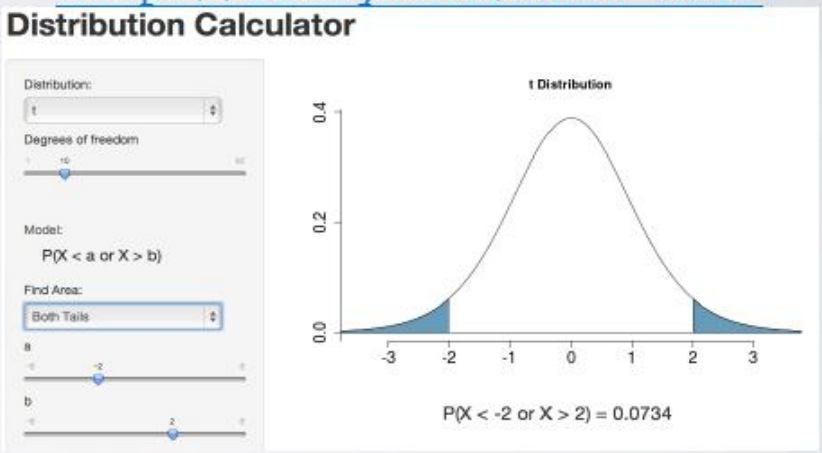
t statistic

- for inference on a mean where
  - σ unknown
  - $n < 30$
- calculated the same way

$$T = \frac{obs - null}{SE}$$

- p-value (same definition)
  - one or two tail area, based on H_A
  - using R, applet, or table

http://bitly.com/dist_calc



Inference for a small sample mean

PLAYING A COMPUTER GAME DURING LUNCH AFFECTS FULLNESS, MEMORY FOR LUNCH, AND LATER SNACK INTAKE

distraction and recall of food consumed and snacking

sample: 44 patients: 22 men and 22 women

study design:

- randomized into two groups:

(1) play solitaire while eating - "win as many games as possible"

(2) eat lunch without distractions

- both groups provided same amount of lunch
- offered biscuits to snack on after lunch

biscuit intake	$\bar{x}$	s	n
solitaire	52.1 g	45.1 g	22
no distraction	27.1 g	26.4 g	22

Study reference: Oldham-Cooper, Rose E., et al. "Playing a computer game during lunch affects fullness, memory for lunch, and later snack intake." The American journal of clinical nutrition 93.2 (2011): 308-313.

Inference for a small sample mean

estimating the mean (based on a small sample)

point estimate ± margin of error

$$\bar{x} \pm t^\star_{df} SE_{\bar{x}}$$

$$\bar{x} \pm t^\star_{df} \frac{s}{\sqrt{n}}$$

$$\bar{x} \pm t^\star_{n-1} \frac{s}{\sqrt{n}}$$

Degrees of freedom for t statistic for inference on one sample mean

$$df = n - 1$$

Inference for a small sample mean

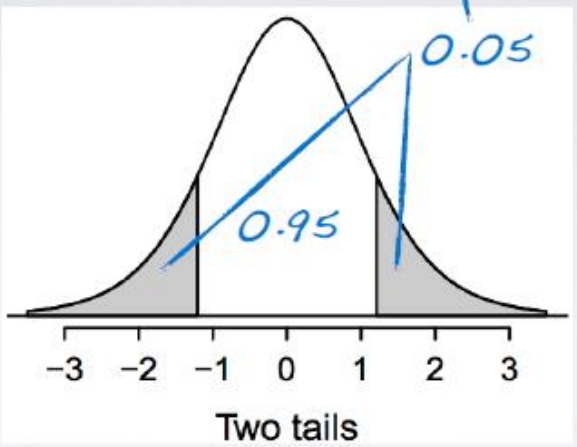
finding the critical t score

using the table

1. determine df

$df = 22 - 1 = 21$

2. find corresponding tail area for desired confidence level



	one tail	0.100	0.050	0.025	0.010	0.005
	two tails	0.200	0.100	0.050	0.020	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	4	1.53	2.13	2.78	3.75	4.60
	5	1.48	2.02	2.57	3.36	4.03
	6	1.44	1.94	2.45	3.14	3.71
	7	1.41	1.89	2.36	3.00	3.50
	8	1.40	1.86	2.31	2.90	3.36
	9	1.38	1.83	2.26	2.82	3.25
	10	1.37	1.81	2.23	2.76	3.17
	11	1.36	1.80	2.20	2.72	3.11
	12	1.36	1.78	2.18	2.68	3.05
	13	1.35	1.77	2.16	2.65	3.01
	14	1.35	1.76	2.14	2.62	2.98
	15	1.34	1.75	2.13	2.60	2.95
	16	1.34	1.75	2.12	2.58	2.92
	17	1.33	1.74	2.11	2.57	2.90
	18	1.33	1.73	2.10	2.55	2.88
	19	1.33	1.73	2.09	2.54	2.86
	20	1.33	1.72	2.09	2.53	2.85
	21	1.32	1.72	2.08	2.52	2.83
	22	1.32	1.72	2.07	2.51	2.82
	23	1.32	1.71	2.07	2.50	2.81
	24	1.32	1.71	2.06	2.49	2.80
	25	1.32	1.71	2.06	2.49	2.79
	26	1.31	1.71	2.06	2.48	2.78
	27	1.31	1.70	2.05	2.47	2.77

Inference for comparing two small sample means

PLAYING A COMPUTER GAME DURING LUNCH AFFECTS FULLNESS, MEMORY FOR LUNCH, AND LATER SNACK INTAKE

distraction and recall of food consumed and snacking

sample: 44 patients: 22 men and 22 women

study design:

- randomized into two groups:

(1) play solitaire while eating - "win as many games as possible"

(2) eat lunch without distractions

- both groups provided same amount of lunch
- offered biscuits to snack on after lunch

biscuit intake	$\bar{x}$	s	n
solitaire	52.1 g	45.1 g	22
no distraction	27.1 g	26.4 g	22

Study reference: Oldham-Cooper, Rose E., et al. "Playing a computer game during lunch affects fullness, memory for lunch, and later snack intake." The American journal of clinical nutrition 93.2 (2011): 308-313.

Inference for comparing two small sample means

comparing means based on small samples

confidence interval

point estimate ± margin of error

$$(\bar{x}_1 - \bar{x}_2) \pm t^\star_{df} SE_{(\bar{x}_1 - \bar{x}_2)}$$

hypothesis test

$$T_{df} = \frac{obs - null}{SE}$$

$$T_{df} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE_{(\bar{x}_1 - \bar{x}_2)}}$$

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

DF for t statistic for inference on difference of two means

$$df = min(n_1 - 1, n_2 - 1)$$

Comparing more than two means

vocabulary score and class

from the 2010 GSS

	wordsum	8
1	6	middle class
2	9	working class
3	6	working class
4	5	working class
5	6	working class
6	6	working class
…	…	…
795	9	middle class

10 question vocabulary test (scores range from 0 to 10)

self identified social class (lower, working, middle, upper)

Comparing more than two means

vocabulary score

Choose a word from a list of provided options that comes closest to the meaning of the first word provided in capital letters.

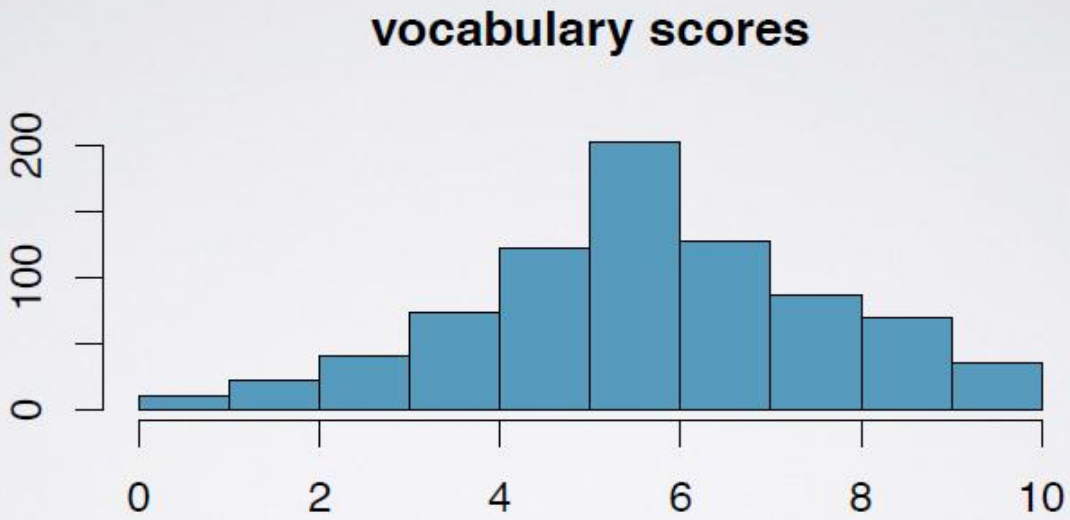
`wordsum`

1. SPACE (school, noon, captain, room, board, don't know)
2. BROADEN (efface, make level, elapse, embroider, widen, don't know)
3. EMANATE (populate, free, prominent, rival, come, don't know)
4. EDIBLE (auspicious, eligible, fit to eat, sagacious, able to speak, don't know)
5. ANIMOSITY (hatred, animation, disobedience, diversity, friendship, don't know)
6. PACT (puissance, remonstrance, agreement, skillet, pressure, don't know)
7. CLOISTERED (miniature, bunched, arched, malady, secluded, don't know)
8. CAPRICE (value, a star, grimace, whim, inducement, don't know)
9. ACCUSTOM (disappoint, customary, encounter, get used to, business, don't know)
10. ALLUSION (reference, dream, eulogy, illusion, aria, don't know)

Comparing more than two means

vocabulary score

wordsum



Comparing more than two means

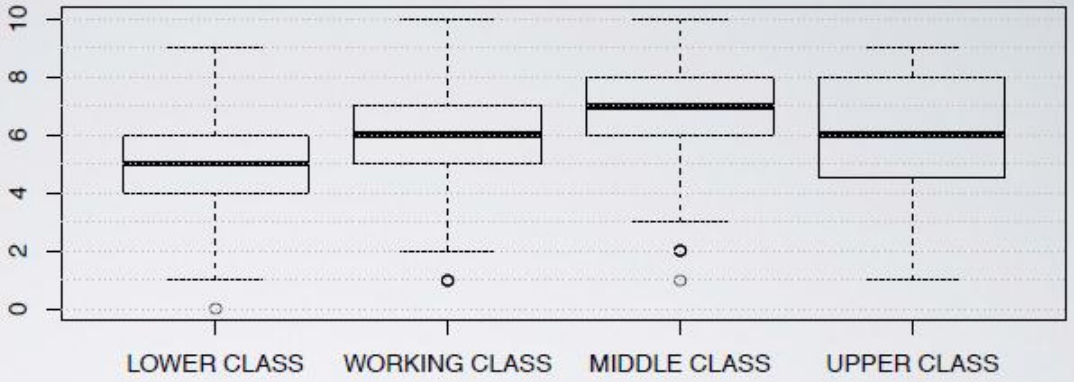
self identified social class

`class`

If you were asked to use one of four names for your social class, which would you say you belong in: the lower class, the working class, the middle class, or the upper class?

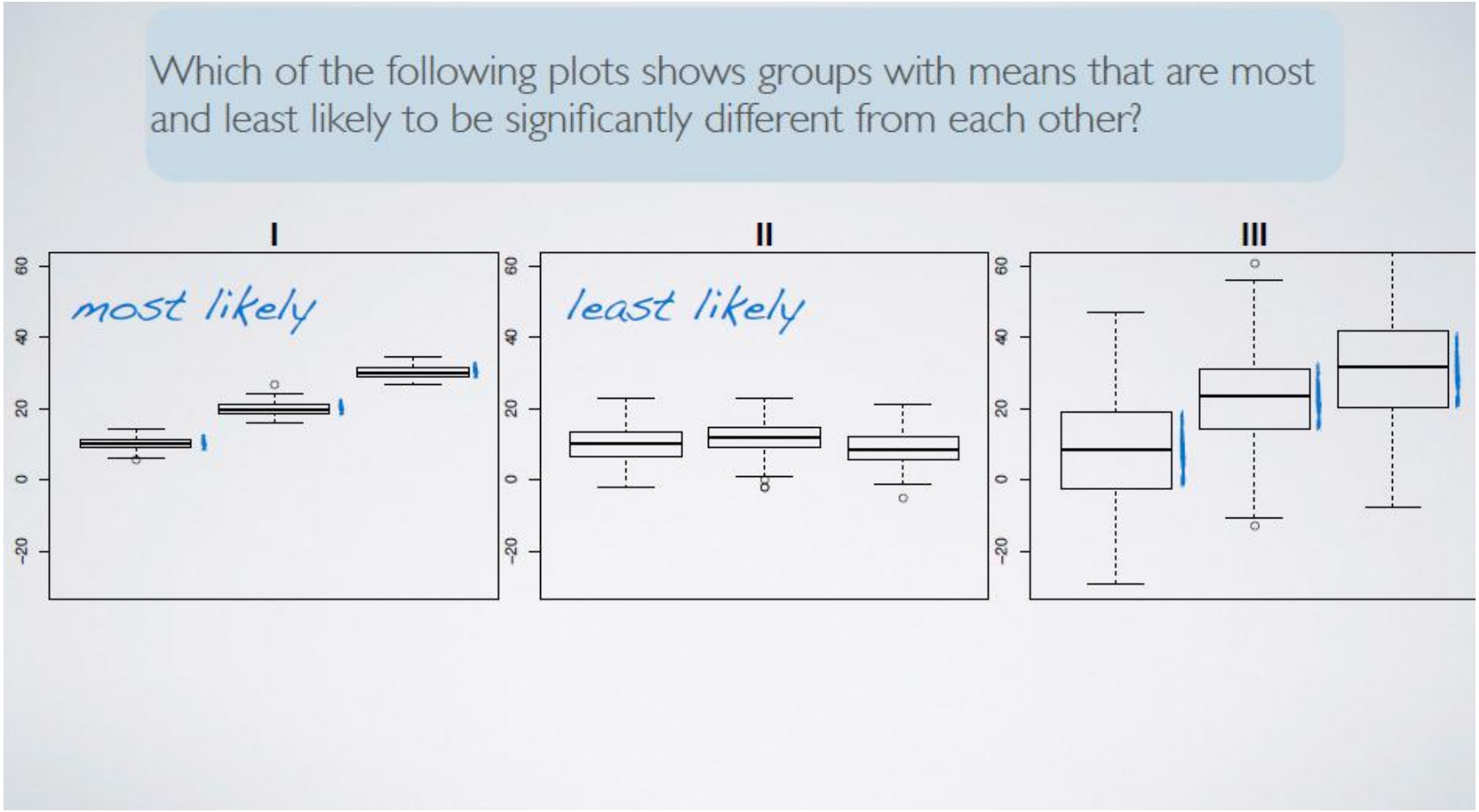
Comparing more than two means

exploratory analysis



	n	mean	sd
lower class	41	5.07	2.24
working class	407	5.75	1.87
middle class	331	6.76	1.89
upper class	16	6.19	2.34
overall	795	6.14	1.98

Comparing more than two means



Comparing more than two means

Is there a difference between the average vocabulary scores of Americans from different (self reported) classes?

- To compare means of 2 groups we use a Z or a T statistic.
- To compare means of 3+ groups we use a new test called analysis of variance (ANOVA) and a new statistic called F.

ANOVA

Comparing more than two means

anova

H_0: The mean outcome is the same across all categories

$$\mu_1 = \mu_2 = \cdots = \mu_k$$

H_A: At least one pair of means are different from each other

μ_i : mean of the outcome for observations in category *i*

k : number of groups

Comparing more than two means

z / t test

Compare means from **two** groups: are so far apart that the observed difference cannot reasonably be attributed to sampling variability?

$$H_0 : \mu_1 = \mu_2$$

anova

Compare means from **more than two** groups: are they so far apart that the observed differences cannot all reasonably be attributed to sampling variability?

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

Comparing more than two means

z / t test

Compute a test statistic (a ratio).

$$z/t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE_{(\bar{x}_1 - \bar{x}_2)}}$$

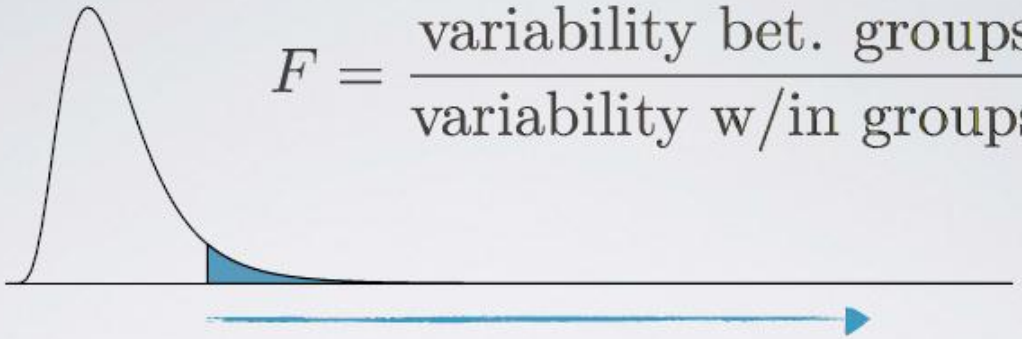
anova

Compute a test statistic (a ratio).

$$F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}}$$

- Large test statistics lead to small p-values.
- If the p-value is small enough H_0 is rejected, and we conclude that the data provide evidence of a difference in the population means.

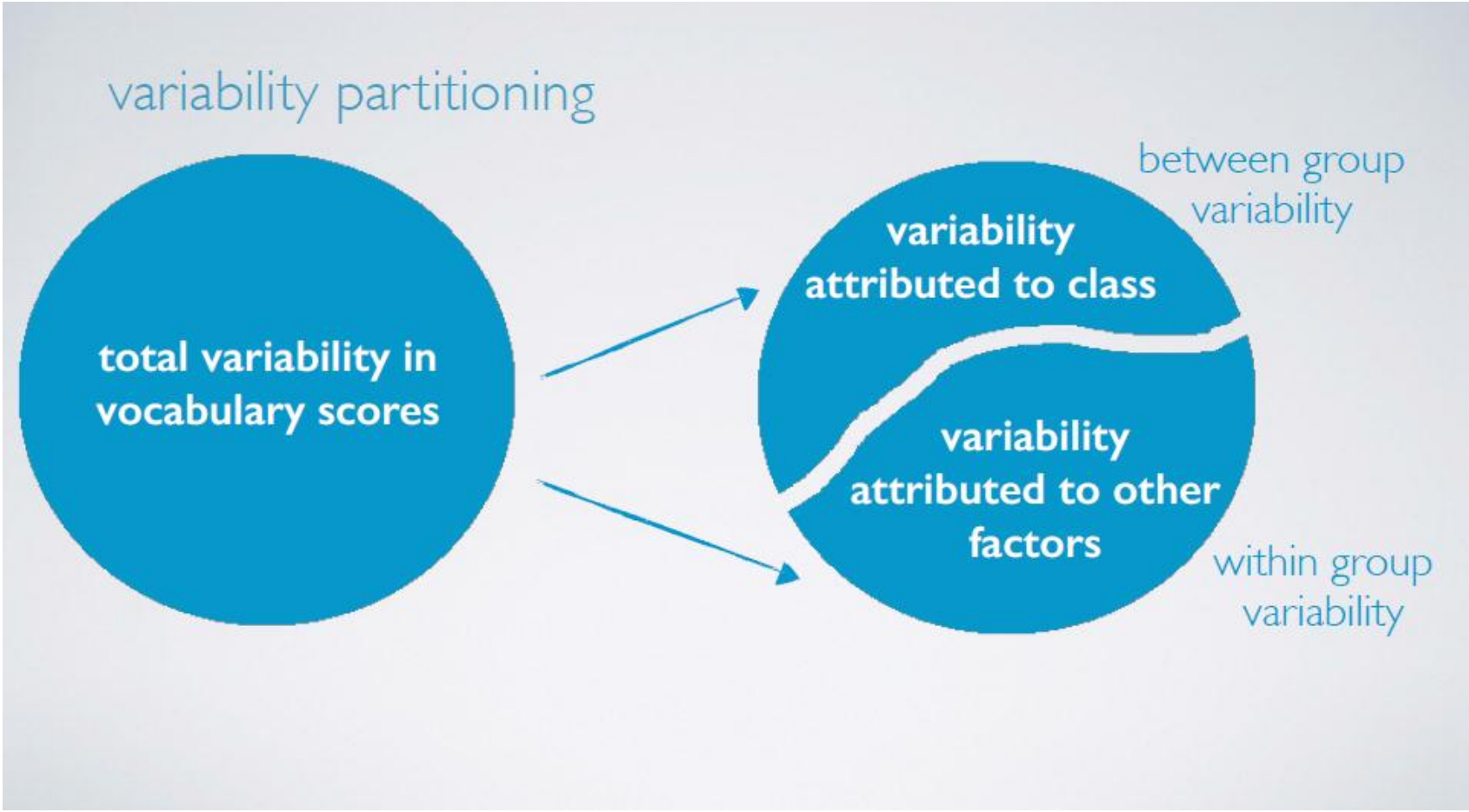
Comparing more than two means



$$F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}} \; \begin{matrix} ? \\ ? \end{matrix}$$

- In order to be able to reject H_0, we need a small p-value, which requires a large F statistic.
- In order to obtain a large F statistic, variability between sample means needs to be greater than variability within sample means.

ANOVA


variability partitioning
total variability in vocabulary scores
between group variability
variability attributed to class
variability attributed to other factors
within group variability

ANOVA

vocabulary score and class

	wordsum	class
1	6	middle class
2	9	working class
3	6	working class
4	5	working class
5	6	working class
6	6	working class
…	…	…
795	9	middle class

	n	mean	sd
lower class	41	5.07	2.24
working class	407	5.75	1.87
middle class	331	6.76	1.89
upper class	16	6.19	2.34
overall	795	6.14	1.98

H_0: The mean outcome is the same across all categories

$$\mu_1 = \mu_2 = \mu_3 = \mu_4$$

H_A: At least one pair of means are different from each other

ANOVA

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class	3	236.56	78.855	21.735	<0.0001
Error	Residuals	791	2869.80	3.628		
	Total	794	3106.36			

ANOVA

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class		236.56			
Error	Residuals		2869.80			
	Total		3106.36			

sum of squares total (SST)

- measures the **total variability** in the response variable
- calculated very similarly to variance (except not scaled by the sample size)

ANOVA

Sum of squares total (SST):

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

y_i : value of the response variable for each observation

$\bar{y}$: grand mean of the response variable

	wordsum	class
1	6	middle class
2	9	working class
3	6	working class
…	…	…
795	9	middle class

	n	mean	sd
overall	795	6.14	1.98

$SST = (6-6.14)^2$
$+ (9-6.14)^2$
$+ (6-6.14)^2$
$+ \cdots$
$+ (9-6.14)^2 = 3106.36$

ANOVA

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class		236.56			
Error	Residuals		2869.80			
	Total		3106.36			

sum of squares groups (SSG)

- measures the variability **between groups**
- **explained variability:** deviation of group mean from overall mean, weighted by sample size

ANOVA

Sum of squares group (SSG):

$$SSG = \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2$$

n_j : number of observations in group j

$\bar{y}_j$: mean of the response variable for group j

$\bar{y}$: grand mean of the response variable

	n	mean	sd
lower class	41	5.07	2.24
working class	407	5.75	1.87
middle class	331	6.76	1.89
upper class	16	6.19	2.34
overall	795	6.14	1.98

$$SSG = (41 \times (5.07 - 6.14)^2) + (407 \times (5.75 - 6.14)^2) + (331 \times (6.76 - 6.14)^2) + (16 \times (6.19 - 6.14)^2) \approx 236.56$$

ANOVA

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class		236.56			
Error	Residuals		2869.8			
	Total		3106.36			

Sum of squares error (SSE):

$$SSE = SST - SSG$$

sum of squares error (SSE)

- measures the variability **within groups**
- **unexplained variability:** unexplained by the group variable, due to other reasons

3106.36 - 236.56 = 2869.8

ANOVA

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class		236.56	?		
Error	Residuals		2869.8	?		
	Total		3106.36	?		

- now we need a way to get from these measures of total variability to average variability
- scaling by a measure that incorporates sample sizes and number of groups → degrees of freedom

degrees of freedom

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class	3	236.56			
Error	Residuals	791	2869.80			
	Total	794	3106.36			

Degrees of freedom associated with ANOVA:

- total: $df_T = n - 1$ → $795 - 1 = 794$
- group: $df_G = k - 1$ → $4 - 1 = 3$
- error: $df_E = df_T - df_G$ → $794 - 3 = 791$

mean square error

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class	3	236.56	78.855		
Error	Residuals	791	2869.80	3.628		
	Total	794	3106.36			

Mean squares: Average variability between and within groups, calculated as the total variability (sum of squares) scaled by the associated degrees of freedom.

- group: $MSG = SSG/df_G$ → 236.56 / 3 ≈ 78.855
- error: $MSE = SSE/df_E$ → 2869.8 / 791 ≈ 3.628

F statistic

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class	3	236.56	78.855	21.735	
Error	Residuals	791	2869.80	3.628		
	Total	794	3106.36			

F statistic: Ratio of the between group and within group variability:

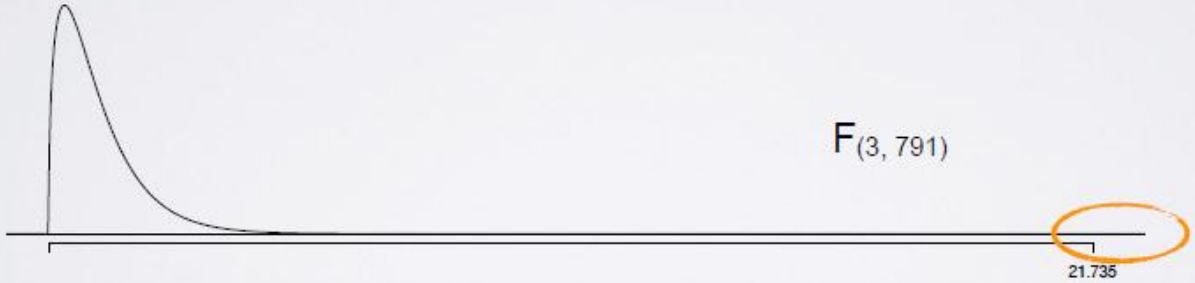
$$F = \frac{MSG}{MSE}$$

$$\frac{78.855}{3.628} \approx 21.735$$

p-value

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class	3	236.56	78.855	21.735	<0.0001
Error	Residuals	791	2869.80	3.628		
	Total	794	3106.36			

- p-value is the probability of at least as large a ratio between the "between" and "within" group variabilities if in fact the means of all groups are equal
- area under the F curve, with degrees of freedom df_G and df_E, above the observed F statistic.



conclusion

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class	3	236.56	78.855	21.735	<0.0001
Error	Residuals	791	2869.80	3.628		
	Total	794	3106.36			

- If p-value is small (less than α), reject H_0.
  - The data provide convincing evidence that at least one pair of population means are different from each other (but we can't tell which one).
- If p-value is large, fail to reject H_0.
  - The data do not provide convincing evidence that one pair of population means are different from each other, the observed differences in sample means are attributable to sampling variability (or chance).

Conditions for ANOVA

Conditions for ANOVA

1. ***Independence:***
   - ✓ **within groups:** sampled observations must be independent
   - ✓ **between groups:** the groups must be independent of each other (non-paired)
2. ***Approximate normality:*** distributions should be nearly normal within each group
3. ***Equal variance:*** groups should have roughly equal variability

Conditions for ANOVA

(1) independence

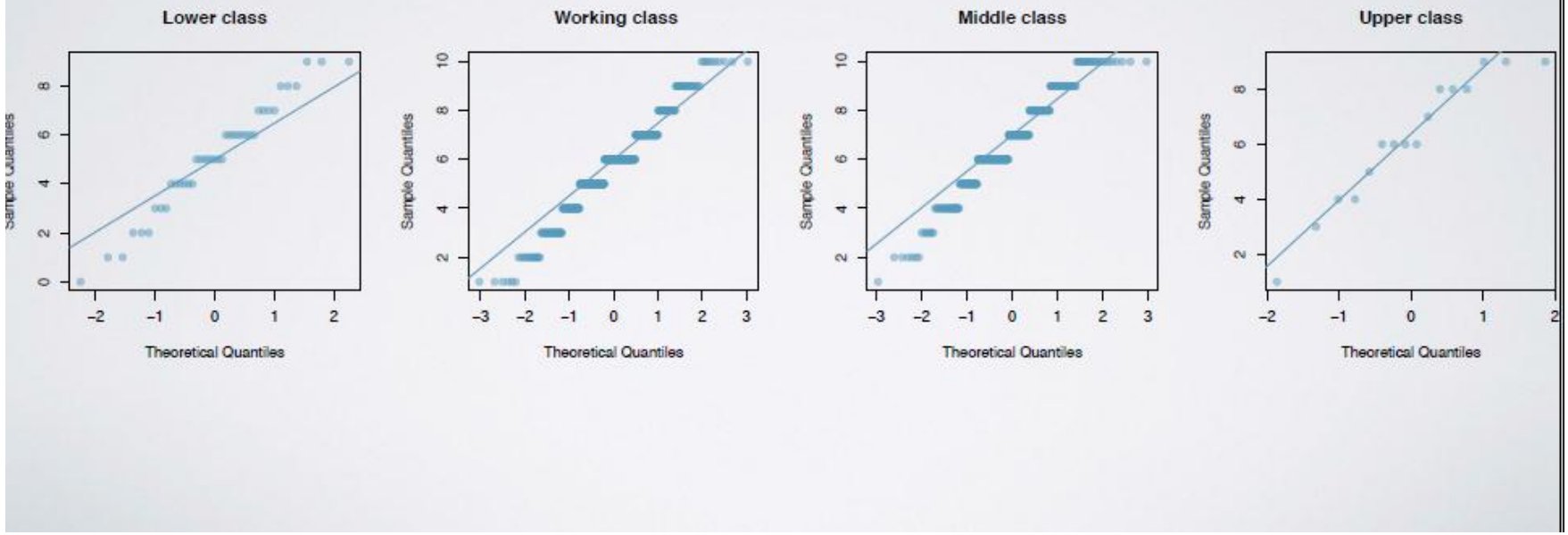
sampled observations must be independent of each other

- random sample / assignment
- each n_j less than 10% of respective population
- carefully consider whether the groups may be independent (e.g. no pairing) → repeated measures anova
- always important, but sometimes difficult to check

Conditions for ANOVA

(2) approximately normal

- distribution of response variable within each group should be approximately normal
- especially important when sample sizes are small

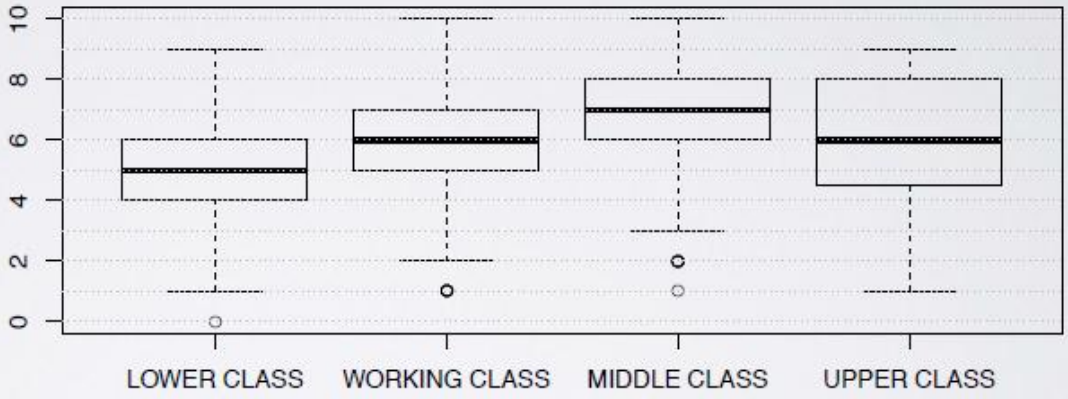


Conditions for ANOVA

(3) constant variance

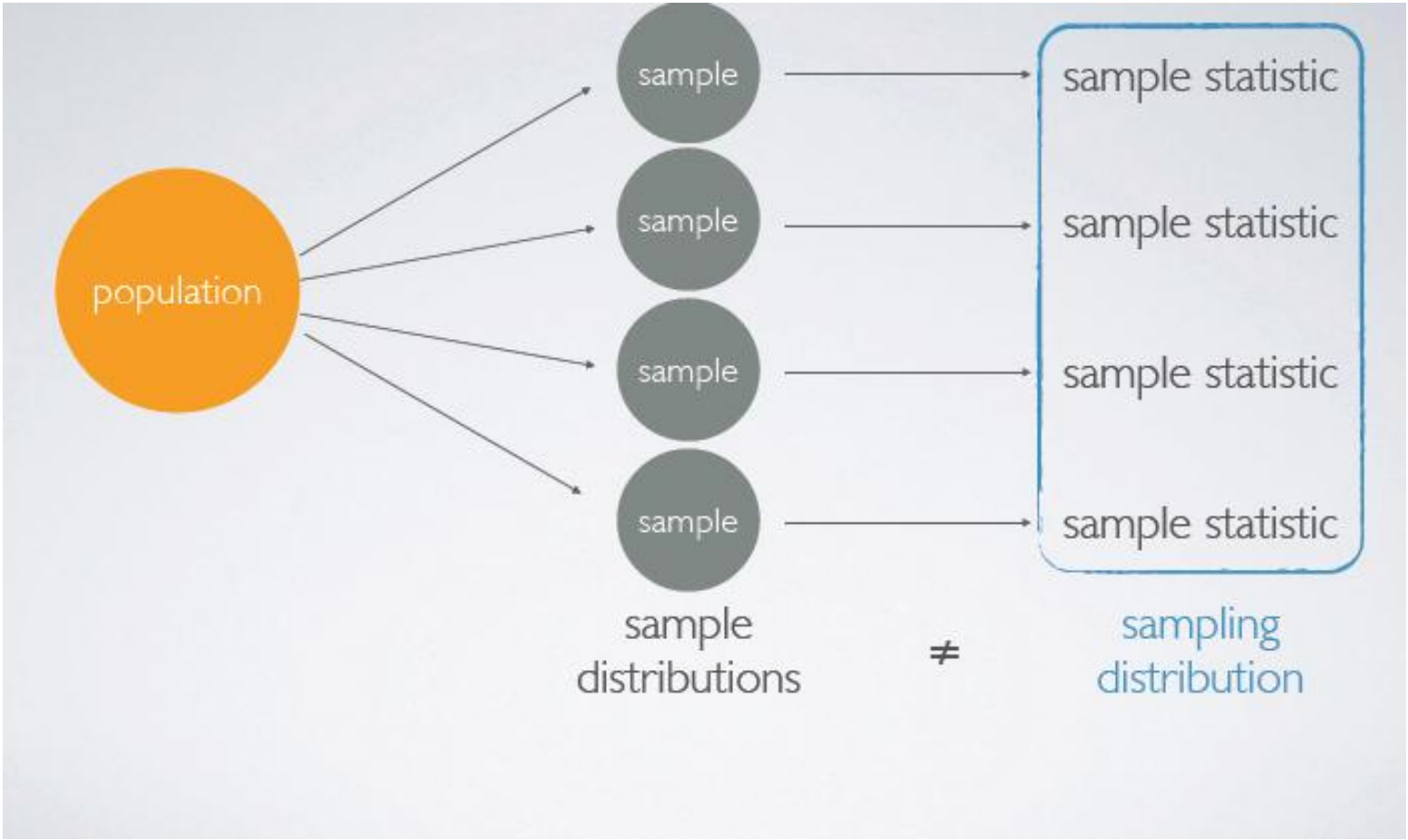
- variability should be consistent across groups: homoscedastic groups
- especially important when sample sizes differ between groups

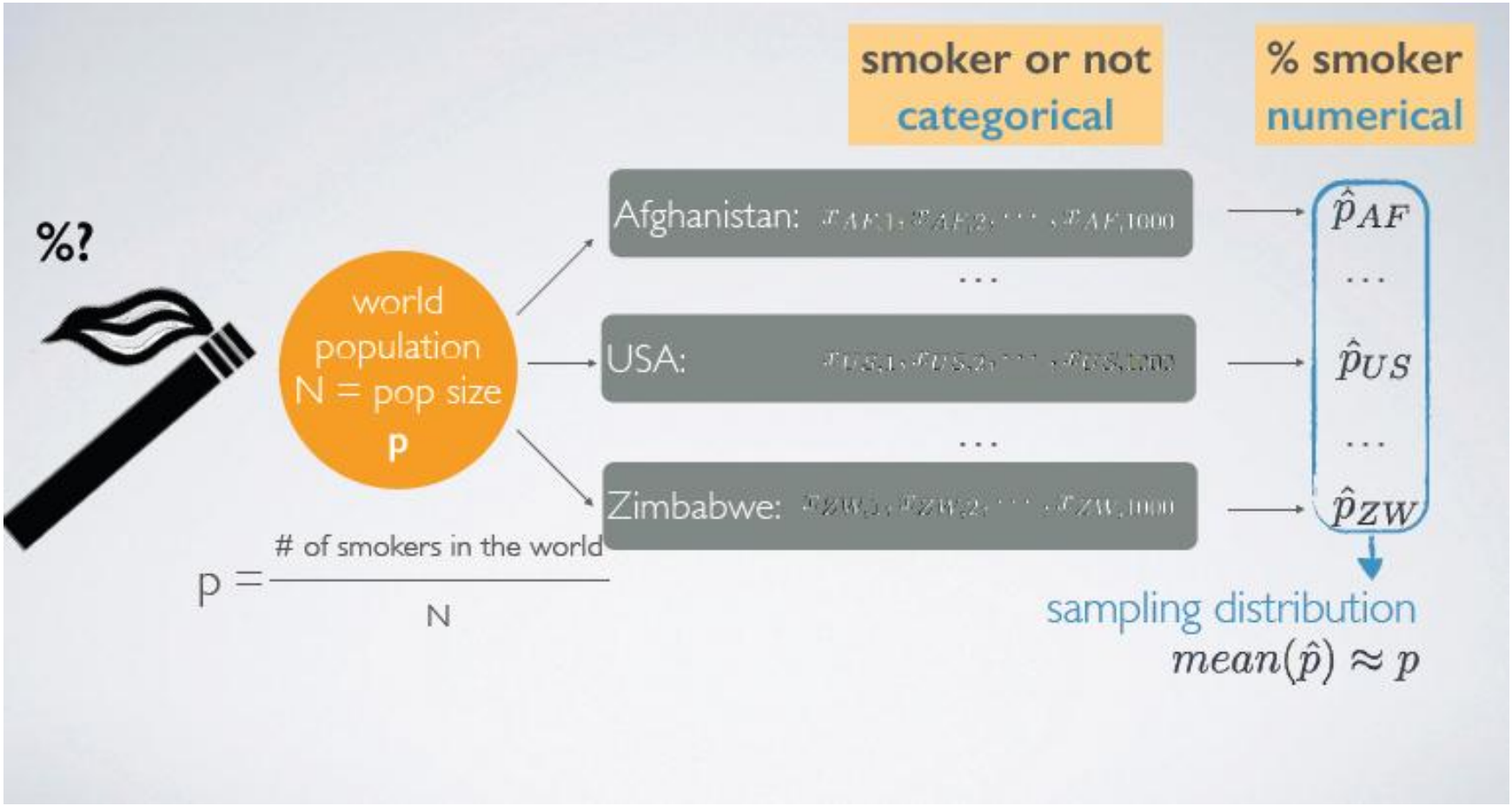
	n	sd
lower class	41	2.24
working class	407	1.87
middle class	331	1.89
upper class	16	2.34
overall	795	1.98



Inference for categorical variables

Sampling variability & CLT for proportions




smoker or not
categorical
% smoker
numerical
%?
world
population
N = pop size
p
Afghanistan:
USA:
Zimbabwe:
$\hat{p}_{AF}$
$\hat{p}_{US}$
$\hat{p}_{ZW}$
$p = \frac{\text{\# of smokers in the world}}{N}$
sampling distribution
$mean(\hat{p}) \approx p$

CLT for proportions: The distribution of sample proportions is nearly normal, centered at the population proportion, and with a standard error inversely proportional to the sample size.

$$\hat{p} \sim N\left(mean = p, SE = \sqrt{\frac{p(1-p)}{n}}\right)$$

shape center spread

Conditions for the CLT:

1. ***Independence:*** Sampled observations must be independent.
   - random sample/assignment
   - if sampling without replacement, n < 10% of population
2. ***Sample size/skew:*** There should be at least 10 successes and 10 failures in the sample: $np \geq 10$ and $n(1-p) \geq 10$.

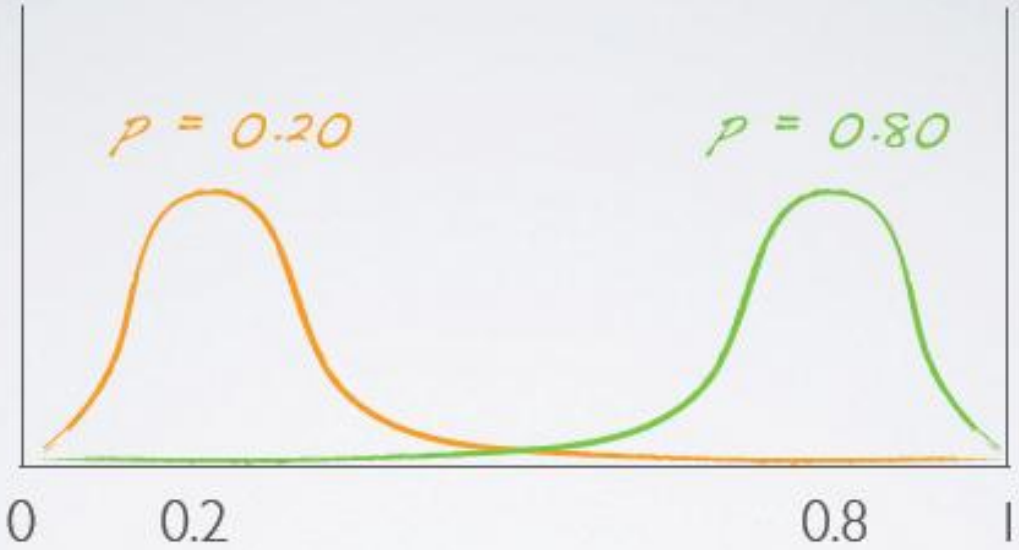
if p unknown, use $\hat{p}$

What if

if the success-failure condition is not met:

- the center of the sampling distribution will still be around the true population proportion
- the spread of the sampling distribution can still be approximated using the same formula for the standard error
- the shape of the distribution will depend on whether the true population proportion is closer to 0 or closer to 1

shape of the sampling distribution



Hypothesis testing for a proportion

Hypothesis testing for a single proportion:

1. Set the hypotheses: $H_0 : p = null\ value$
   $H_A : p <\ or\ >\ or\ \neq\ null\ value$
2. Calculate the point estimate: $\hat{p}$
3. Check conditions:
   1. ***Independence:*** Sampled observations must be independent (random sample/assignment & if sampling without replacement, n < 10% of population)
   2. ***Sample size/skew:*** $np \geq 10$ and $n(1-p) \geq 10$
4. Draw sampling distribution, shade p-value, calculate test statistic $Z = \frac{\hat{p} - p}{SE}, \quad SE = \sqrt{\frac{p(1-p)}{n}}$
5. Make a decision, and interpret it in context of the research question:
   - If p-value < α, reject H_0; the data provide convincing evidence for H_A.
   - If p-value > α, fail to reject H_0 the data *do not* provide convincing evidence for H_A.

$\hat{p}$ vs. p

	confidence interval	hypothesis test
success-failure condition	$n\hat{p} \geq 10$ $n(1-\hat{p}) \geq 10$	$np \geq 10$ $n(1-p) \geq 10$
standard error	$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$SE = \sqrt{\frac{p(1-p)}{n}}$

Estimating diference between two proportions

estimating the difference between two proportions

point estimate ± margin of error

$$(\hat{p}_1 - \hat{p}_2) \pm z^\star SE_{(\hat{p}_1 - \hat{p}_2)}$$

Standard error for difference between two proportions, for calculating a confidence interval:

$$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Estimating diference between two proportions

Conditions for inference for comparing two independent proportions:

1. ***Independence:***
   - ✓ **within groups:** sampled observations must be independent within each group
     - random sample/assignment
     - if sampling without replacement, n < 10% of population
   - ✓ **between groups:** the two groups must be independent of each other (non-paired)
2. ***Sample size/skew:*** Each sample should meet the success-failure condition:
   - ✓ $n_1p_1 \geq 10$ and $n_1(1-p_1) \geq 10$
   - ✓ $n_2p_2 \geq 10$ and $n_2(1-p_2) \geq 10$

Hypothesis tests for comparing two proportions

A SurveyUSA poll asked respondents whether any of their children have ever been the victim of bullying. Also recorded on this survey was the gender of the respondent (the parent). Below is the distribution of responses by gender of the respondent.

	Male	Female
Yes	34	61
No	52	61
Not sure	4	0
Total	90	122
$\hat{p}$	0.38	0.50

34 / 90 61 / 122

$H_0: p_{male} - p_{female} = 0$

$H_A: p_{male} - p_{female} \neq 0$

✔ check conditions

✔ calculate test statistic & p-value

Link to poll: http://www.surveyusa.com/client/PollReport.aspx?g=1823ef50-44c7-4d2a-9efc-ead711b4ad9c

Image by Eddie~S: http://en.wikipedia.org/wiki/File:Bully_Free_Zone.jpg (CC BY 2.0)

flashback to working with one proportion: $\hat{p}$ vs. p

	observed confidence interval	expected hypothesis test
success-failure condition	$n\hat{p} \geq 10$ $n(1-\hat{p}) \geq 10$	$np \geq 10$ $n(1-p) \geq 10$
standard error	$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$SE = \sqrt{\frac{p(1-p)}{n}}$

working with two proportions: $\hat{p}$ vs. p

	observed confidence interval	expected hypothesis test
success-failure condition	$n_1\hat{p}_1 \geq 10$ $\quad n_2\hat{p}_2 \geq 10$ $n_1(1-\hat{p}_1) \geq 10$ $\quad n_2(1-\hat{p}_2) \geq 10$	$H_0 : p_1 = p_2$
standard error	$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	

INFERENTIAL MODELING

Inferential modeling

Inference for linear regression with a single predictor

Inference for linear regression with a multiple predictor

Inference for logistic regression

08/01, 15/01, 22/01 2025

Inference for linear regresion with a a single predictor

Case study: sandwitch store

- **Consider o hypothetical population of all the sandwitch stores of a particular chain.**

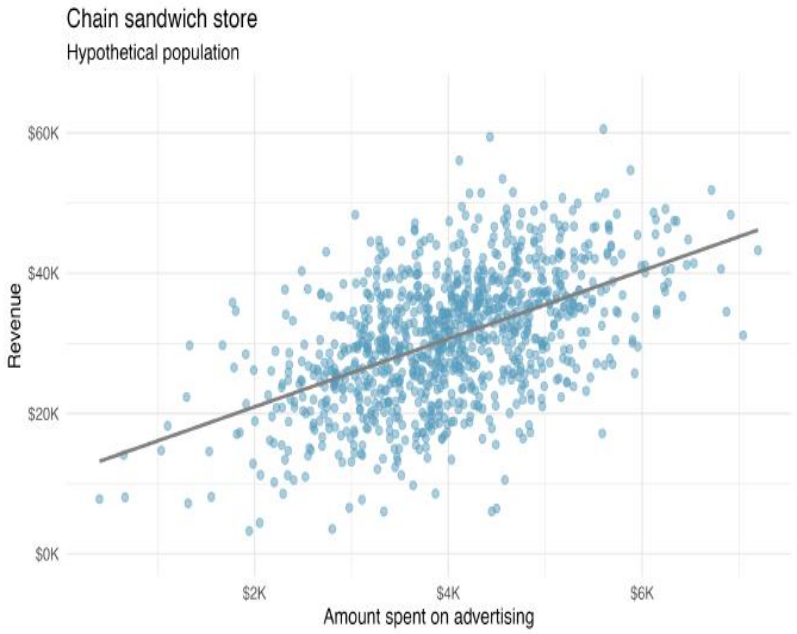


Figure 24.1: Revenue as a linear model of advertising dollars for a population of sandwich stores, in thousands of dollars.

Population model:

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

True population model:

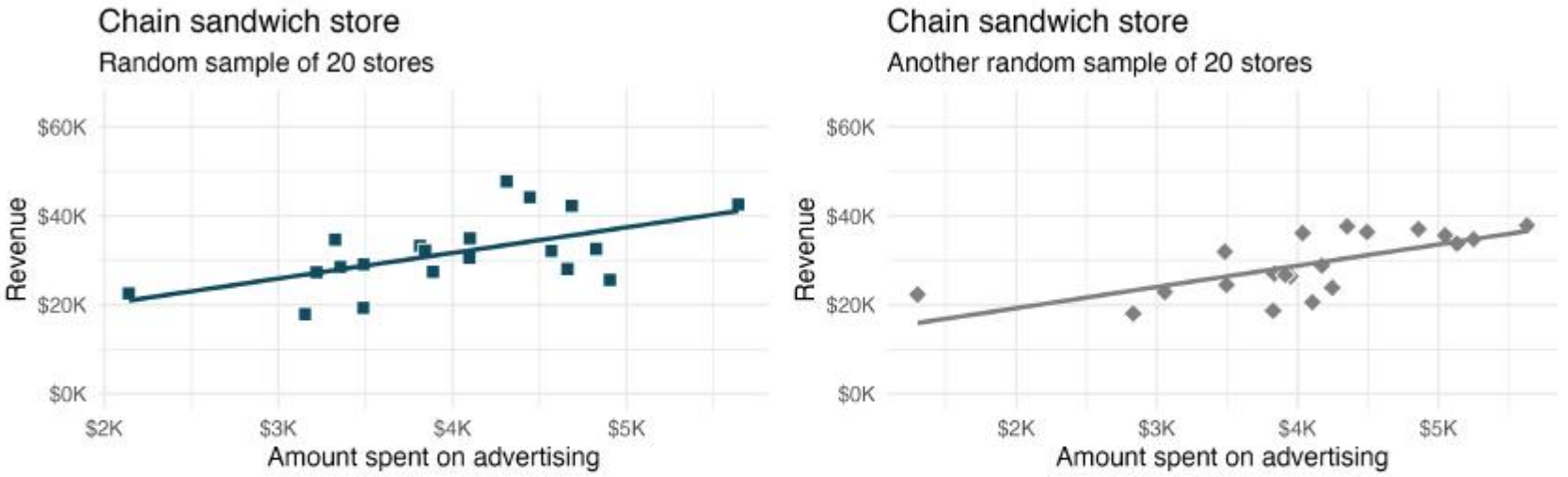
$$\texttt{expected revenue} = 11.23 + 4.8 \times \texttt{advertising}.$$

Case study: sandwitch store

- **Variability of the statistics**

Least square regression fit:

$$\hat{y} = b_0 + b_1 x.$$



(a) First sample.

(b) Second sample.

Figure 24.2: Two random samples of 20 stores from the entire population. A linear trend between advertising and revenue is observed in both.

Case study: sandwitch store

- **Variability of the statistics**

Least square regression fit:

$$\hat{y} = b_0 + b_1 x.$$

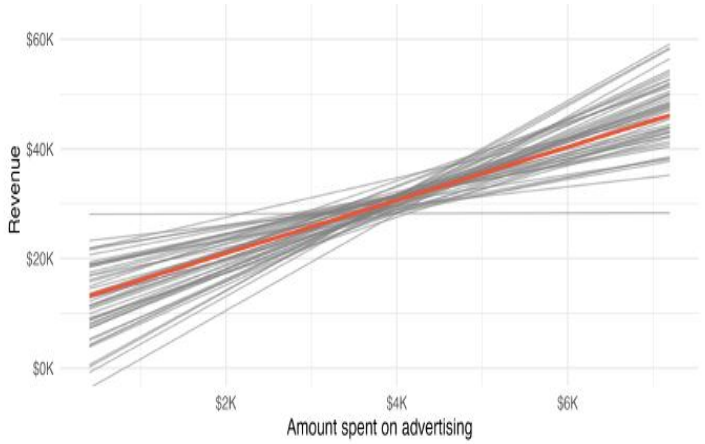


Figure 24.4: If repeated samples of size 20 are taken from the entire population, each linear model will be slightly different. The red line provides the linear fit to the entire population.

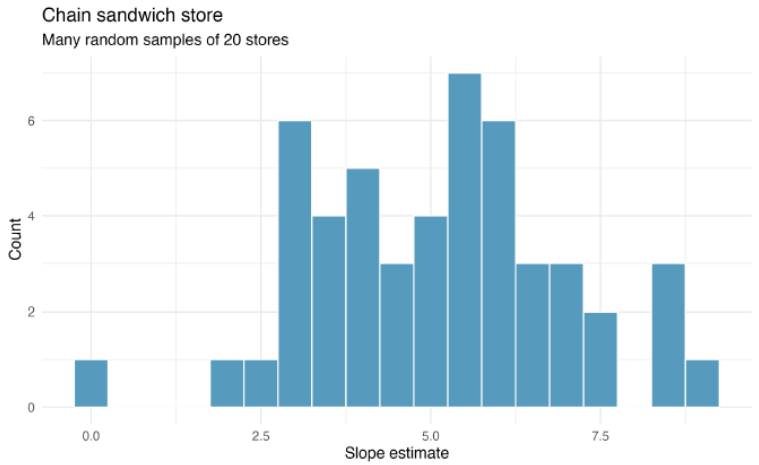


Figure 24.5: Variability of slope estimates from many different samples of stores, each of size 20.

Case study: sandwitch store

Recall, the example described in this introduction is hypothetical. That is, we created an entire population in order demonstrate how the slope of a line would vary from sample to sample. The tools in this textbook are designed to evaluate only one single sample of data. With actual studies, we do not have repeated samples, so we are not able to use repeated samples to visualize the variability in slopes. We have seen variability in samples throughout this text, so it should not come as a surprise that different samples will produce different linear models. However, it is nice to visually consider the linear models produced by different slopes. Additionally, as with measuring the variability of previous statistics (e.g., $\overline{X}_1 - \overline{X}_2$ or $\hat{p}_1 - \hat{p}_2$), the histogram of the sample statistics can provide information related to inferential considerations.

In the following the distribution (i.e., histogram) of b_1 (the estimated slope coefficient) will be constructed in three ways

- by randomizing (permuting) the response variable
- bootstrap the data by taking random samples of size n from the original dataset
- use mathematical tools to describe the variability using the t-distribution

Randomization test for the slope

Consider data on 100 randomly selected births gathered originally from the US Department of Health and Human Services. Some of the variables are plotted in Figure 24.6.

The scientific research interest at hand will be in determining the linear relationship between weight of baby at birth (in lbs) and number of weeks of gestation. The dataset is quite rich and deserves exploring, but for this example, we will focus only on the weight of the baby.

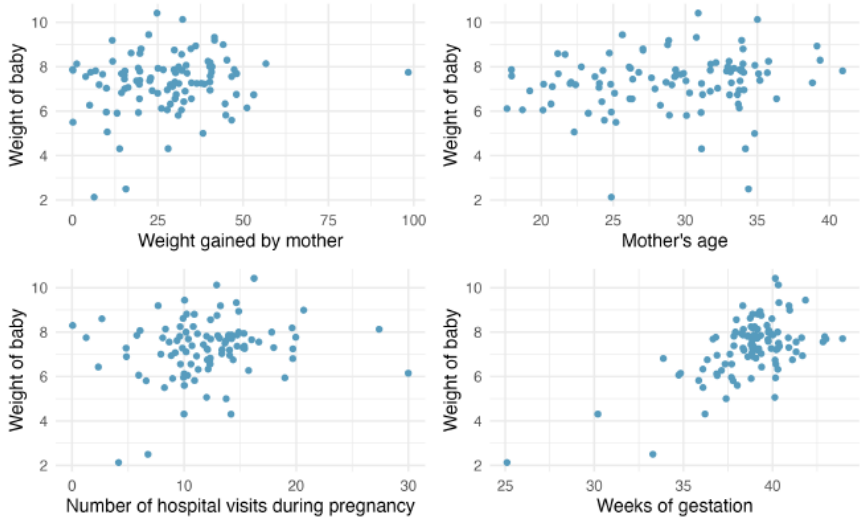


Figure 24.6: Weight of baby at birth (in lbs) as plotted by four other birth variables (mother's weight gain, mother's age, number of hospital visits, and weeks gestation).

Randomization test for the slope

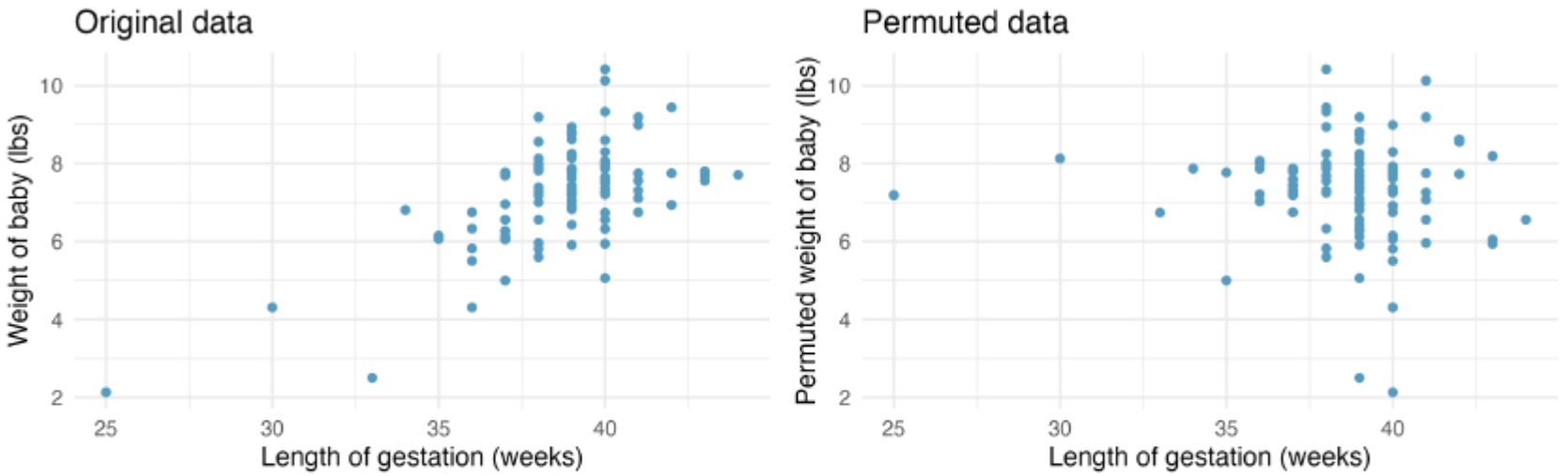
The relevant hypotheses for the linear model setting can be written in terms of the population slope parameter. Here the population refers to a larger population of births in the US.

- $H_0 : \beta_1 = 0$, there is no linear relationship between `weight` and `weeks`.
- $H_A : \beta_1 \neq 0$, there is some linear relationship between `weight` and `weeks`.

Recall that for the randomization test, we permute one variable to eliminate any existing relationship between the variables. That is, we set the null hypothesis to be true, and we measure the natural variability in the data due to sampling but **not** due to variables being correlated. Figure 24.7a shows the observed data and Figure 24.7b shows one permutation of the `weight` variable. The careful observer can see that each of the observed values for `weight` (and for `weeks`) exist in both the original data plot as well as the permuted `weight` plot, but the `weight` and `weeks` gestation are no longer matched for a given birth. That is, each `weight` value is randomly assigned to a new `weeks` gestation.

By repeatedly permuting the response variable, any pattern in the linear model that is observed is due only to random chance (and not an underlying relationship). The randomization test compares the slopes calculated from the permuted response variable with the observed slope. If the observed slope is inconsistent with the slopes from permuting, we can conclude that there is some underlying relationship (and that the slope is not merely due to random chance).

Randomization test for the slope



(a) Original data.

(b) Permuted data.

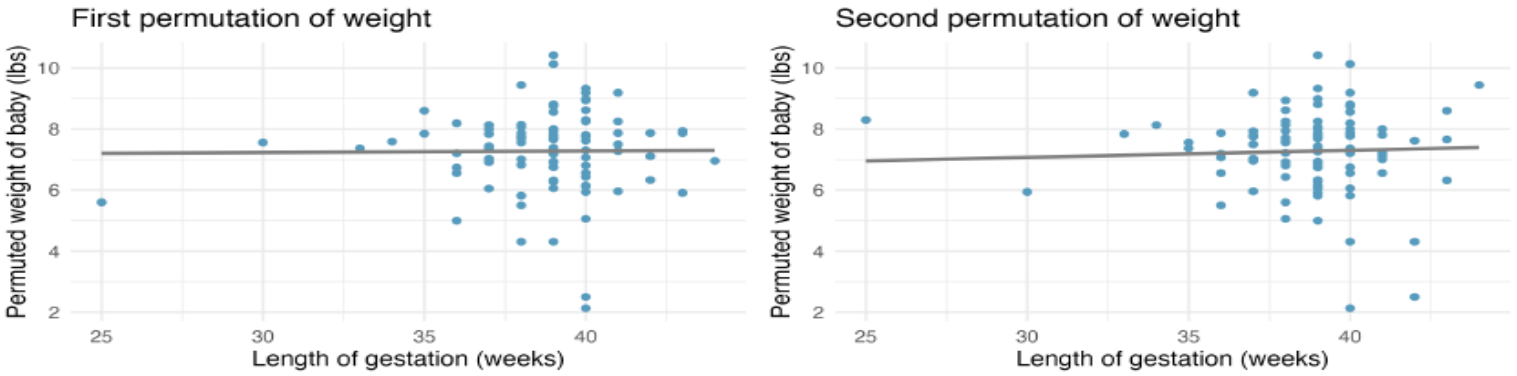
Figure 24.7: Permutation removes the linear relationship between `weight` and `weeks`. Repeated permutations allow for quantifying the variability in the slope under the condition that there is no linear relationship (i.e., that the null hypothesis is true).

Slope in original data: +0.35

Randomization test for the slope

Variability of statistics

After permuting the data, the least squares estimate of the line can be computed. Repeated permutations and slope calculations describe the variability in the line (i.e., in the slope) due only to the natural variability and not due to a relationship between `weight` and `weeks` gestation. Figure 24.8 shows two different permutations of `weight` and the resulting linear models.



(a) First permutation. (b) Second permutation.

Figure 24.8: Two permutations of `weight` with slightly different least squares regression lines.

As you can see, sometimes the slope of the permuted data is positive, sometimes it is negative. Because the randomization happens under the condition of no underlying relationship (because the response variable is completely mixed with the explanatory variable), we expect to see the center of the randomized slope distribution to be zero.

Randomization test for the slope

- Observed statistics vs null statistics

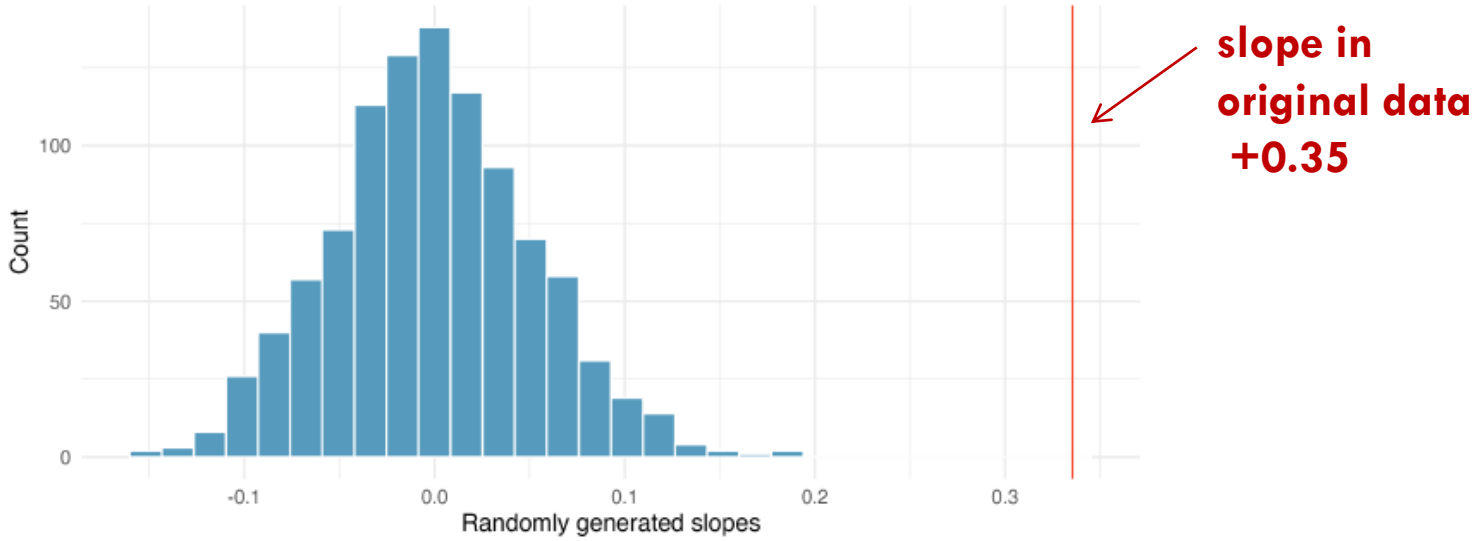


Figure 24.9: Histogram of slopes given different permutations of the `weight` variable. The vertical red line is at the observed value of the slope, 0.335.

Natural variability of the slopes would produce estimated between – 0.15 and +0.15. We reject the null hypothesis. We believe that the slope observed in the original data is not due to natural variability and indeed there is a linear relationship.

Bootstrap confidence interval for slope

- Observed data:

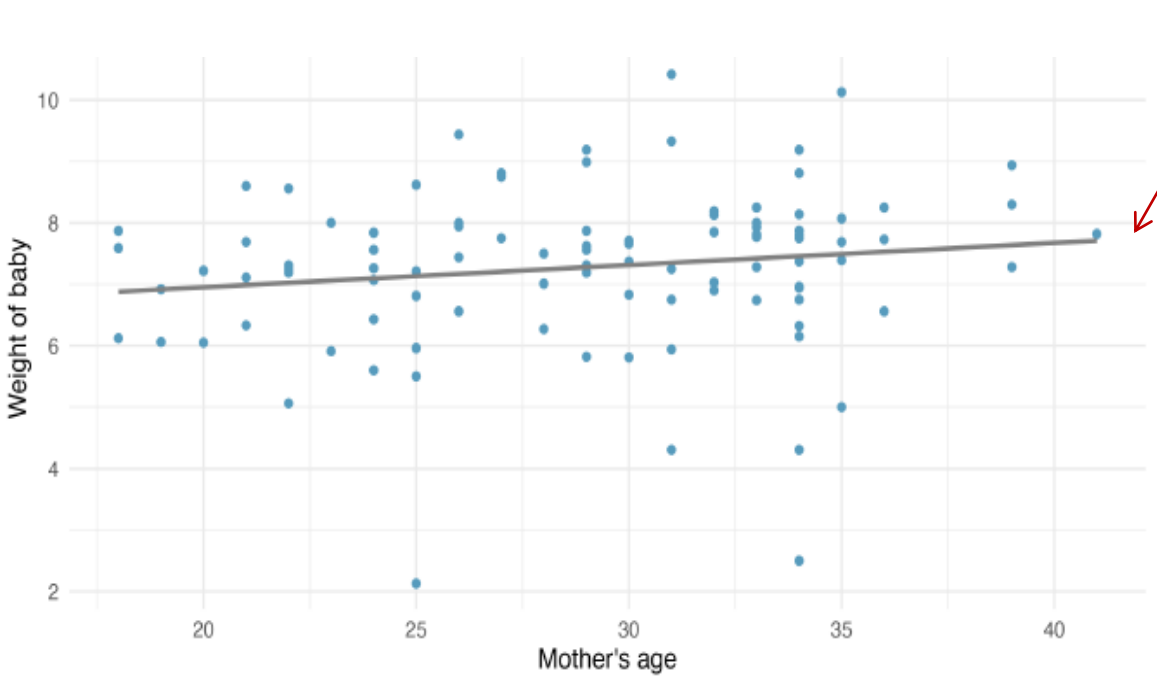
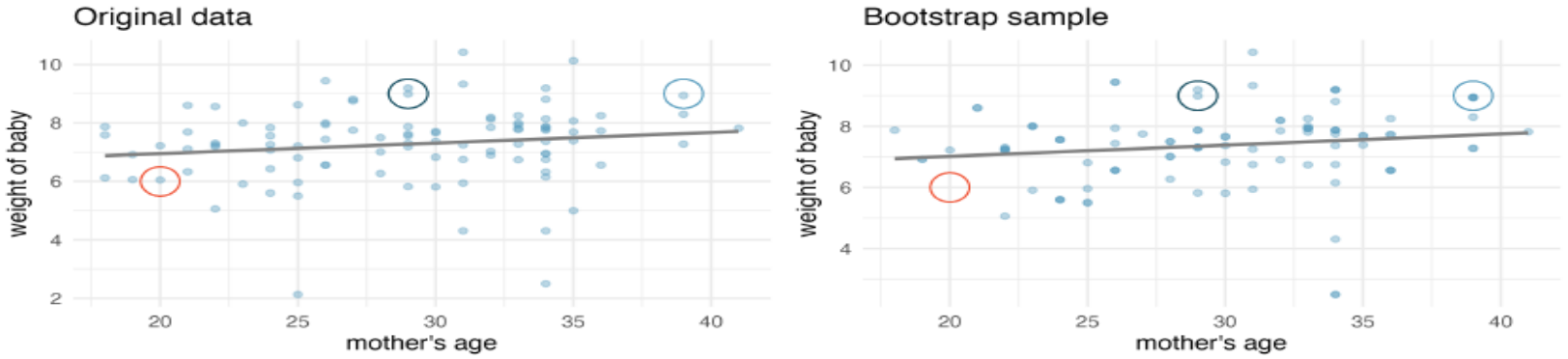


Figure 24.10: Using the original data, the weight of baby as a linear model of mother's age. Notice that the relationship between mother's age and weight of baby is not as strong as the relationship we saw previously between weeks gestation and weight of baby.

Bootstrap confidence interval for slope

Variability of the statistics

Because the focus here is *not* on a null distribution, we sample with replacement $n = 100$ observations from the original dataset. Recall that with bootstrapping the resample always has the same number of observations as the original dataset in order to mimic the process of taking a sample from the population. When sampling in the linear model case, consider each observation to be a single dot. If the dot is resampled, both the `weight` and the `mage` measurement are observed. The measurements are linked to the dot (i.e., to the birth in the sample).



(a) Original data.

(b) Bootstrapped data.

Figure 24.11: Original and one bootstrap sample of the births data. It is difficult to differentiate between the two plots, as (within a single bootstrap sample) the observations which have been resampled twice are plotted as points on top of one another. The red circles represent points in the original data which were not included in the bootstrap sample. The blue circles represent a data point that was repeatedly resampled (and is therefore darker) in the bootstrap sample. The green circles represent a particular structure to the data which is observed in both the original and bootstrap samples.

Bootstrap confidence interval for slope

- Variability of the statistics

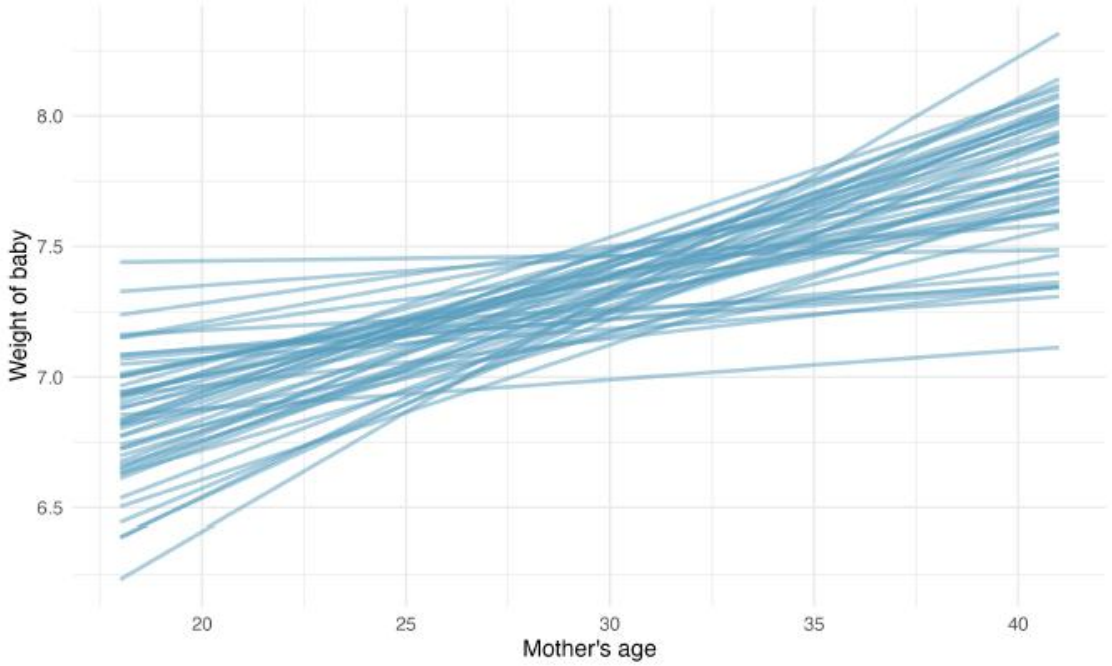


Figure 24.12: Repeated bootstrap resamples of size 100 are taken from the original data. Each of the bootstrapped linear models is slightly different.

Bootstrap confidence interval for slope

- Confidence level (95%): (-0.01, 0.081)

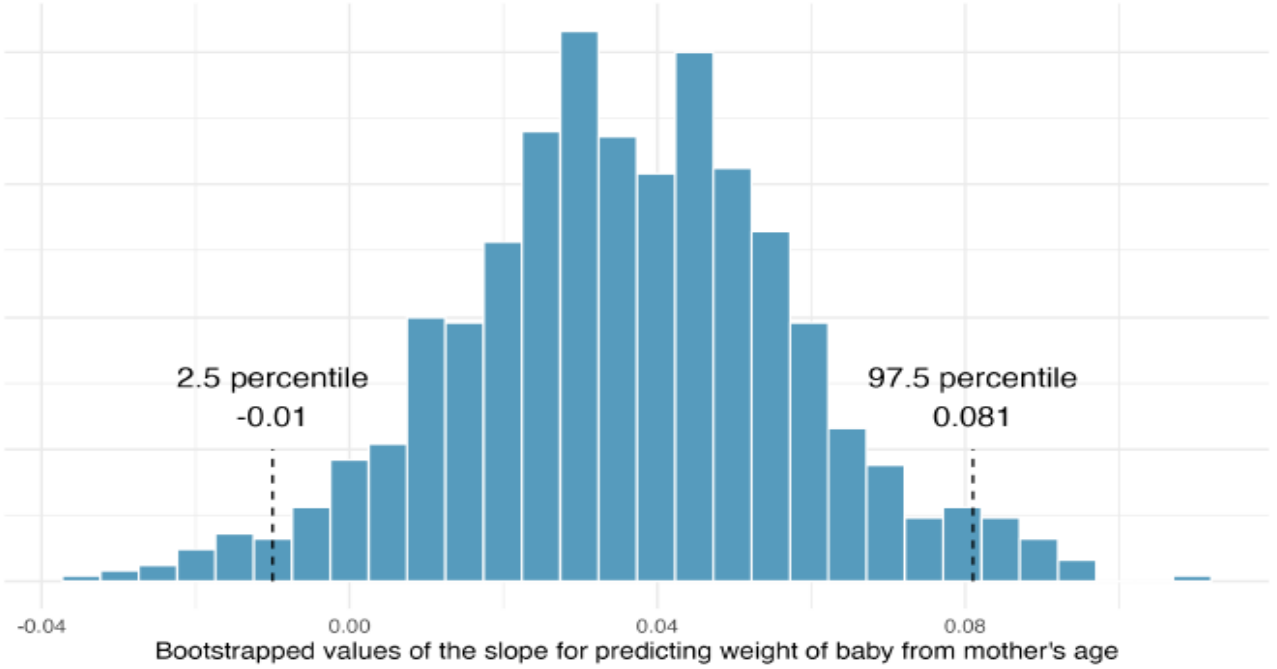


Figure 24.13: The original births data on baby's weight and mother's age is bootstrapped 1,000 times. The histogram provides a sense for the variability of the slope of the linear model from sample to sample.

Bootstrap confidence interval for slope

- Confidence level (95%): (-0.01, 0.081)

EXAMPLE

Using Figure 24.13, calculate the bootstrap estimate for the standard error of the slope. Using the bootstrap standard error, find a 95% bootstrap SE confidence interval for the true population slope, and interpret the interval in context.

Notice that most of the bootstrapped slopes fall between -0.01 and +0.08 (a range of 0.09). Using the empirical rule (that with bell-shaped distributions, most observations are within two standard errors of the center), the standard error of the slopes is approximately 0.0225. The critical value for a 95% confidence interval is $z^\star = 1.96$ which leads to a confidence interval of $b_1 \pm 1.96 \cdot SE \rightarrow 0.036 \pm 1.96 \cdot 0.0225 \rightarrow (-0.0081, 0.0801)$. The bootstrap SE confidence interval is almost identical to the bootstrap percentile interval. In context, we are 95% confident that for the model describing the population of births, predicting weight of baby from mother's age, a one unit increase in `mage` (in years) is associated with an increase in predicted average baby `weight` of between -0.0081 and 0.0801 pounds.

Mathematical model for testing the slope

When certain technical conditions apply, it is convenient to use mathematical approximations to test and estimate the slope parameter.

The approximations will build on the t-distribution

The mathematical model is often correct and is usually easy to implement computationally.

Mathematical model for testing the slope

- Observed data

Midterm elections and unemployment

Elections for members of the United States House of Representatives occur every two years, coinciding every four years with the U.S. Presidential election. The set of House elections occurring during the middle of a Presidential term are called midterm elections. In America's two-party system (the vast majority of House members through history have been either Republicans or Democrats), one political theory suggests the higher the unemployment rate, the worse the President's party will do in the midterm elections. In 2020 there were 232 Democrats, 198 Republicans, and 1 Libertarian in the House.

To assess the validity of the claim related to unemployment and voting patterns, we can compile historical data and look for a connection. We consider every midterm election from 1898 to 2018, with the exception of those elections during the Great Depression. The House of Representatives is made up of 435 voting members.

Mathematical model for testing the slope

- **Observed data**

Regression line

$$\text{percent change in House seats for President's party} = -7.36 - 0.89 \times (\text{unemployment rate})$$

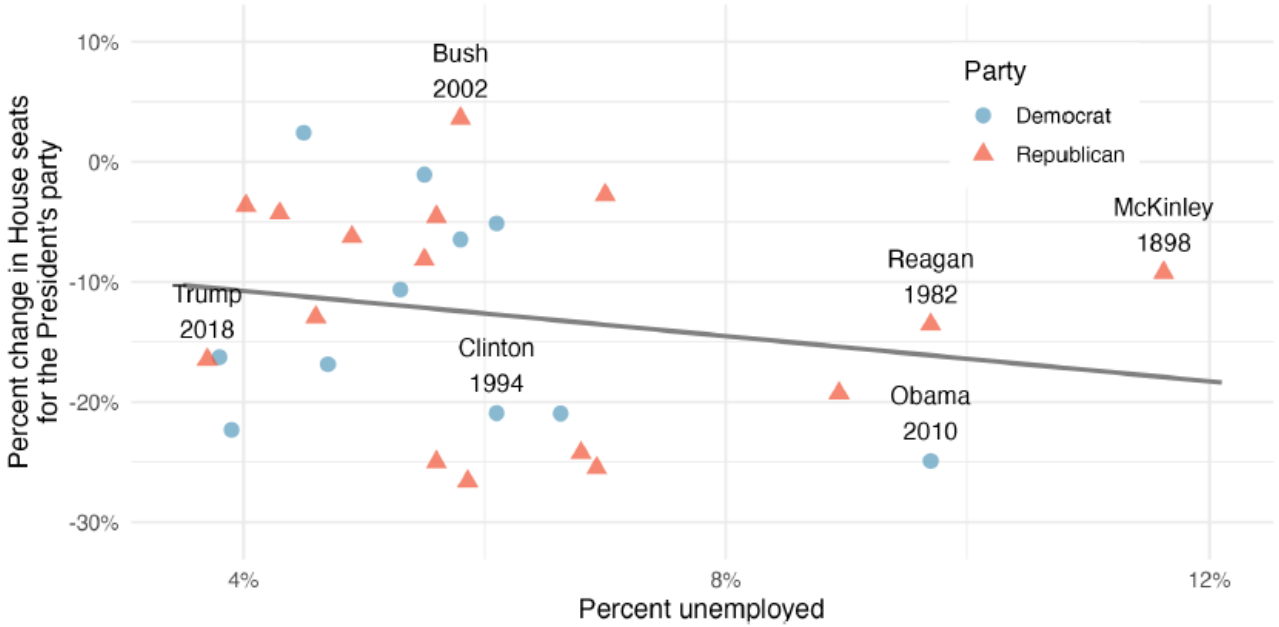


Figure 24.14: The percent change in House seats for the President's party in each election from 1898 to 2010 plotted against the unemployment rate. The two points for the Great Depression have been removed, and a least squares regression line has been fit to the data.

Mathematical model for testing the slope

There is a negative slope in the line shown in Figure 24.14. However, this slope (and the y-intercept) are only estimates of the parameter values. We might wonder, is this convincing evidence that the "true" linear model has a negative slope? That is, do the data provide strong evidence that the political theory is accurate, where the unemployment rate is a useful predictor of the midterm election? We can frame this investigation into a statistical hypothesis test:

- H_0: $\beta_1 = 0$. The true linear model has slope zero.
- H_A: $\beta_1 \neq 0$. The true linear model has a slope different than zero. The unemployment is predictive of whether the President's party wins or loses seats in the House of Representatives.

We would reject H_0 in favor of H_A if the data provide strong evidence that the true slope parameter is different than zero. To assess the hypotheses, we identify a standard error for the estimate, compute an appropriate test statistic, and identify the p-value.

Mathematical model for testing the slope

Table 24.3: Output from statistical software for the regression line modeling the midterm election losses for the President's party as a response to unemployment.

term	estimate	std.error	statistic	p.value
(Intercept)	-7.36	5.16	-1.43	0.16
unemp	-0.89	0.83	-1.07	0.30

The entries in the first column represent the least squares estimates, b_0 and b_1, and the values in the second column correspond to the standard errors of each estimate. Using the estimates, we could write the equation for the least square regression line as

$$\hat{y} = -7.36 - 0.89x$$

where $\hat{y}$ in this case represents the predicted change in the number of seats for the president's party, and x represents the unemployment rate.

Mathematical model for testing the slope

Table 24.3: Output from statistical software for the regression line modeling the midterm election losses for the President's party as a response to unemployment.

term	estimate	std.error	statistic	p.value
(Intercept)	-7.36	5.16	-1.43	0.16
unemp	-0.89	0.83	-1.07	0.30

is very similar. In the hypotheses we consider, the null value for the slope is 0, so we can compute the test statistic using the T score formula:

$$T = \frac{\text{estimate} - \text{null value}}{\text{SE}} = \frac{-0.89 - 0}{0.835} = -1.07$$

The T score we calculated corresponds to the third column of Table 24.3.

Mathematical model for testing the slope

Table 24.3: Output from statistical software for the regression line modeling the midterm election losses for the President's party as a response to unemployment.

term	estimate	std.error	statistic	p.value
(Intercept)	-7.36	5.16	-1.43	0.16
unemp	-0.89	0.83	-1.07	0.30

EXAMPLE

Use Table 24.3 to determine the p-value for the hypothesis test.

The last column of the table gives the p-value for the two-sided hypothesis test for the coefficient of the unemployment rate 0.2961 That is, the data do not provide convincing evidence that a higher unemployment rate has any correspondence with smaller or larger losses for the President's party in the House of Representatives in midterm elections. If there was no linear relationship between the two variables (i.e., if $\beta_1 = 0$), then we would expect to see linear models as or more extreme that the observed model roughly 30% of the time.

Mathematical model for testing the slope

- Observed statistics vs null statistics

As the final step in a mathematical hypothesis test for the slope, we use the information provided to make a conclusion about whether the data could have come from a population where the true slope was zero (i.e., $\beta_1 = 0$). Before evaluating the formal hypothesis claim, sometimes it is important to check your intuition. Based on everything we have seen in the examples above describing the variability of a line from sample to sample, ask yourself if the linear relationship given by the data could have come from a population in which the slope was truly zero.

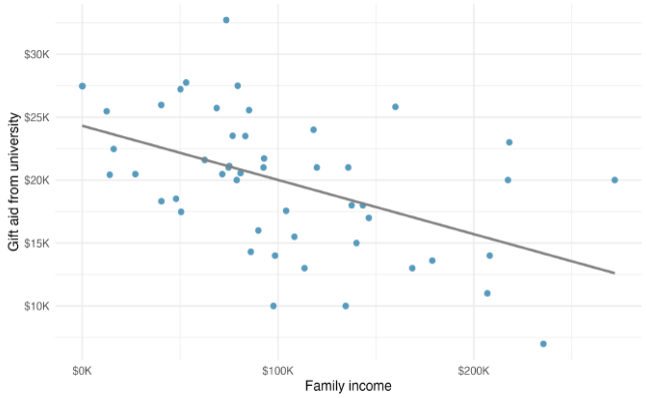


Figure 7.13: Gift aid and family income for a random sample of 50 first-year students from Elmhurst College.

EXAMPLE

Examine Figure 7.13, which relates the Elmhurst College aid and student family income. Are you convinced that the slope is discernibly different from zero? That is, do you think a formal hypothesis test would reject the claim that the true slope of the line should be zero?

While the relationship between the variables is not perfect, there is an evident decreasing trend in the data. Such a distinct trend suggests that the hypothesis test will reject the null claim that the slope is zero.

Mathematical model for testing the slope

Table 24.4: Summary of least squares fit for the Elmhurst College data, where we are predicting the gift aid by the university based on the family income of students.

term	estimate	std.error	statistic	p.value
(Intercept)	24319.33	1291.45	18.83	<0.0001
family_income	-0.04	0.01	-3.98	2e-04

GUIDED PRACTICE

Table 24.4 shows statistical software output from fitting the least squares regression line shown in Figure 7.13. Use the output to formally evaluate the following hypotheses.[2]

- H_0: The true coefficient for family income is zero.
- H_A: The true coefficient for family income is not zero.

Inference for regression.

We usually rely on statistical software to identify point estimates, standard errors, test statistics, and p-values in practice. However, be aware that software will not generally check whether the method is appropriate, meaning we must still verify conditions are met. See Section 24.6.

Mathematical model for testing the slope

Confidence intervals for coefficients.

Confidence intervals for model coefficients (e.g., the intercept or the slope) can be computed using the t-distribution:

$$b_i \pm t^\star_{df} \times SE_{b_i}$$

where $t^\star_{df}$ is the appropriate $t^\star$ cutoff corresponding to the confidence level with the model's degrees of freedom, $df = n - 2$.

EXAMPLE

Compute the 95% confidence interval for the coefficient using the regression output from Table 24.4.

The point estimate is -0.0431 and the standard error is $SE = 0.0108$. When constructing a confidence interval for a model coefficient, we generally use a t-distribution. The degrees of freedom for the distribution are noted in the regression output, $df = 48$, allowing us to identify $t^\star_{48} = 2.01$ for use in the confidence interval.

We can now construct the confidence interval in the usual way:

$$\begin{aligned} &\text{point estimate} \pm t^\star_{48} \times SE \\ &-0.0431 \pm 2.01 \times 0.0108 \\ &(-0.0648, -0.0214) \end{aligned}$$

We are 95% confident that for an additional one unit (i.e., \$1000 increase) in family income, the university's gift aid is predicted to decrease on average by \$21.40 to \$64.80.

Mathematical model for testing the slope

Technical conditions for mathematical model

- **Linearity.** The data should show a linear trend. If there is a nonlinear trend (e.g., first panel of Figure 24.15) an advanced regression method from another book or later course should be applied.
- **Independent observations.** Be cautious about applying regression to data that are sequential observations in time such as a stock price each day. Such data may have an underlying structure that should be considered in a different type of model and analysis. An example of a dataset where successive observations are not independent is shown in the fourth panel of Figure 24.15.
- **Nearly normal residuals.** Generally, the residuals should be nearly normal. When this condition is found to be unreasonable, it is often because of outliers or concerns about influential points, An example of a residual that would be potentially concerning is shown in the second panel of Figure 24.15 A strategy for dealing with outliers is to present two analyses: one with the outlier and one without the outlier.
- **Constant or equal variability.** The variability of points around the least squares line remains roughly constant. An example of non-constant variability is shown in the third panel of Figure 24.15, which represents the most common pattern observed when this condition fails: the variability of y is larger when x is larger.

Mathematical model for testing the slope

- Technical conditions for mathematical model

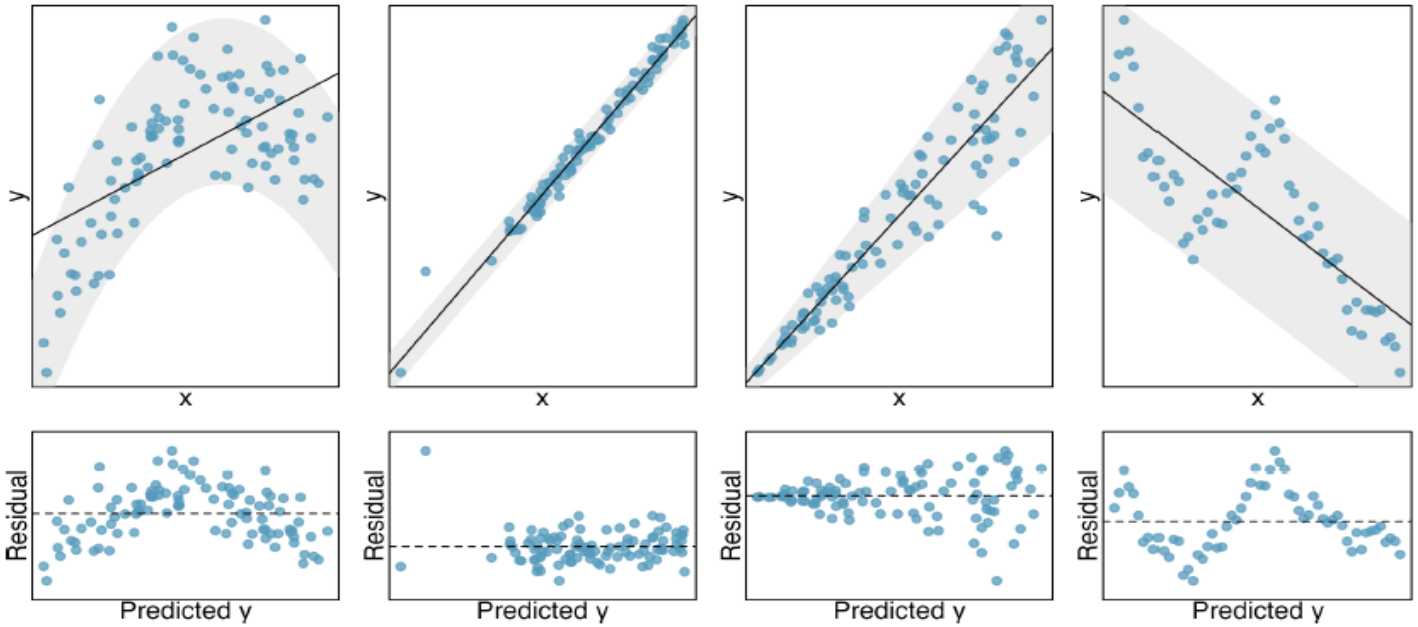


Figure 24.15: Four examples showing when the methods in this chapter are insufficient to apply a linear model to the data. The top set of graphs represents the x and y relationship. The bottom set of graphs is a residual plot. First panel – linearity fails. Second panel – there are outliers, most especially one point that is very far away from the line. Third panel – the variability of the errors is related to the value of x. Fourth panel – a time series dataset is shown, where successive observations are highly correlated.

Inference for a linear regression with multiple predictor

Mathematical model

Now, our goal is to create a model where `interest_rate` can be predicted using the variables `debt_to_income`, `term`, and `credit_checks`. As you learned in Chapter 8, least squares can be used to find the coefficient estimates for the linear model. The unknown population model can be written as:

$$\begin{aligned} E[\texttt{interest\_rate}] = \beta_0 &+ \beta_1 \times \texttt{debt\_to\_income} \\ &+ \beta_2 \times \texttt{term} \\ &+ \beta_3 \times \texttt{credit\_checks} \end{aligned}$$

Table 25.1: Summary of a linear model for predicting interest rate based on `debt_to_income`, `term`, and `credit_checks`. Each of the variables has its own coefficient estimate as well as a p-value.

term	estimate	std.error	statistic	p.value
(Intercept)	4.31	0.20	22.1	<0.0001
debt_to_income	0.04	0.00	13.3	<0.0001
term	0.16	0.00	37.9	<0.0001
credit_checks	0.25	0.02	12.8	<0.0001

The estimated equation for the regression model may be written as a model with three predictor variables:

$$\widehat{\texttt{interest\_rate}} = 4.31 + 0.041 \times \texttt{debt\_to\_income} + 0.16 \times \texttt{term} + 0.25 \times \texttt{credit\_checks}$$

Not only does Table 25.1 provide the estimates for the coefficients, it also provides information on the inference analysis (i.e., hypothesis testing)

Mathematical model

hypothesis test for a linear model with **one predictor**[1] can be written as:

if only one predictor, $H_0 : \beta_1 = 0$.

That is, if the true population slope is zero, the p-value measures how likely it would be to select data which produced the observed slope (b_1) value.

With **multiple predictors**, the hypothesis is similar, however, it is now conditioned on each of the other variables remaining in the model.

if multiple predictors, $H_0 : \beta_i = 0$ given other variables in the model

Using the example above and focusing on each of the variable p-values (here we won't discuss the p-value associated with the intercept), we can write out the three different hypotheses:

- $H_0 : \beta_1 = 0$, given `term` and `credit_checks` are included in the model
- $H_0 : \beta_2 = 0$, given `debt_to_income` and `credit_checks` are included in the model
- $H_0 : \beta_3 = 0$, given `debt_to_income` and `term` are included in the model

The very low p-values from the software output tell us that each of the variables acts as an important predictor in the model, despite the inclusion of the other two.

Mathematical model

Table 25.1: Summary of a linear model for predicting interest rate based on `debt_to_income`, `term`, and `credit_checks`. Each of the variables has its own coefficient estimate as well as a p-value.

term	estimate	std.error	statistic	p.value
`(Intercept)`	4.31	0.20	22.1	<0.0001
`debt_to_income`	0.04	0.00	13.3	<0.0001
`term`	0.16	0.00	37.9	<0.0001
`credit_checks`	0.25	0.02	12.8	<0.0001

The low p-value says that it would be extremely unlikely to see data that produce a coefficient on `debt_to_income` as large as 0.041 if the true relationship between `debt_to_income`and `interest_rate` was non-existent (i.e., if $\beta_1 = 0$) and the model also included `term` and `credit_checks`. You might have thought that the value 0.041 is a small number (i.e., close to zero), but in the units of the problem, 0.041 turns out to be far away from zero, it's all about context! The p-values on `term` and on `credit_checks` are interpreted similarly.

Sometimes a set of predictor variables can impact the model in unusual ways, often due to the predictor variables themselves being correlated.

Hypothesis tests

hypothesis test for a linear model with **one predictor**[1] can be written as:

if only one predictor, $H_0 : \beta_1 = 0$.

That is, if the true population slope is zero, the p-value measures how likely it would be to select data which produced the observed slope (b_1) value.

With **multiple predictors**, the hypothesis is similar, however, it is now conditioned on each of the other variables remaining in the model.

if multiple predictors, $H_0 : \beta_i = 0$ given other variables in the model

Using the example above and focusing on each of the variable p-values (here we won't discuss the p-value associated with the intercept), we can write out the three different hypotheses:

- $H_0 : \beta_1 = 0$, given `term` and `credit_checks` are included in the model
- $H_0 : \beta_2 = 0$, given `debt_to_income` and `credit_checks` are included in the model
- $H_0 : \beta_3 = 0$, given `debt_to_income` and `term` are included in the model

The very low p-values from the software output tell us that each of the variables acts as an important predictor in the model, despite the inclusion of the other two. Consider the p-value on $H_0 : \beta_1 = 0$.

Inference for logistic regression

Conditions for logistic regression

- **For logistic regression:**

it is imperative that the response variable is binary. Additionally, the key technical condition for logistic regression has to do with the relationship between the predictor variables (x_i values) and the probability the outcome will be a success. It turns out, the relationship is a specific functional form called a logit function, where $\text{logit}(p) = \log_e(\frac{p}{1-p})$. The function may feel complicated, and memorizing the formula of the logit is not necessary for understanding logistic regression. What you do need to remember is that the probability of the outcome being a success is a function of a linear combination of the explanatory variables.

Logistic regression conditions.

There are two key conditions for fitting a logistic regression model:

1. Each outcome Y_i is independent of the other outcomes.
2. Each predictor x_i is linearly related to $\text{logit}(p_i)$ if all other predictors are held constant.

Spam or not spam

Consider the `email` data which describes email characteristics which can be used to predict whether a particular incoming email is spam (unsolicited bulk email). Without reading every incoming message, it might be nice to have an automated way to identify spam emails. Which of the variables describing each email are important for predicting the status of the email?

The first logistic regression model condition — independence of the outcomes — is reasonable if we can assume that the emails that arrive in an inbox within a few months are independent of each other with respect to whether they're spam or not.

The second condition of the logistic regression model is not easily checked without a fairly sizable amount of data. Luckily, we have 3921 emails in the dataset! Let's first visualize these data by plotting the true classification of the emails against the model's fitted probabilities, as shown in Figure 26.1.

Email dataset

Table 26.1: Variables and their descriptions for the `email` dataset. Many of the variables are indicator variables, meaning they take the value 1 if the specified characteristic is present and 0 otherwise.

Variable	Description
`spam`	Indicator for whether the email was spam.
`to_multiple`	Indicator for whether the email was addressed to more than one recipient.
`from`	Whether the message was listed as from anyone (this is usually set by default for regular outgoing email).
`cc`	Number of people cc'ed.
`sent_email`	Indicator for whether the sender had been sent an email in the last 30 days.
`attach`	The number of attached files.
`dollar`	The number of times a dollar sign or the word "dollar" appeared in the email.
`winner`	Indicates whether "winner" appeared in the email.
`format`	Indicates whether the email was written using HTML (e.g., may have included bolding or active links).
`re_subj`	Whether the subject started with "Re:", "RE:", "re:", or "rE:"
`exclaim_subj`	Whether there was an exclamation point in the subject.
`urgent_subj`	Whether the word "urgent" was in the email subject.
`exclaim_mess`	The number of exclamation points in the email message.
`number`	Factor variable saying whether there was no number, a small number (under 1 million), or a big number.

Email dataset

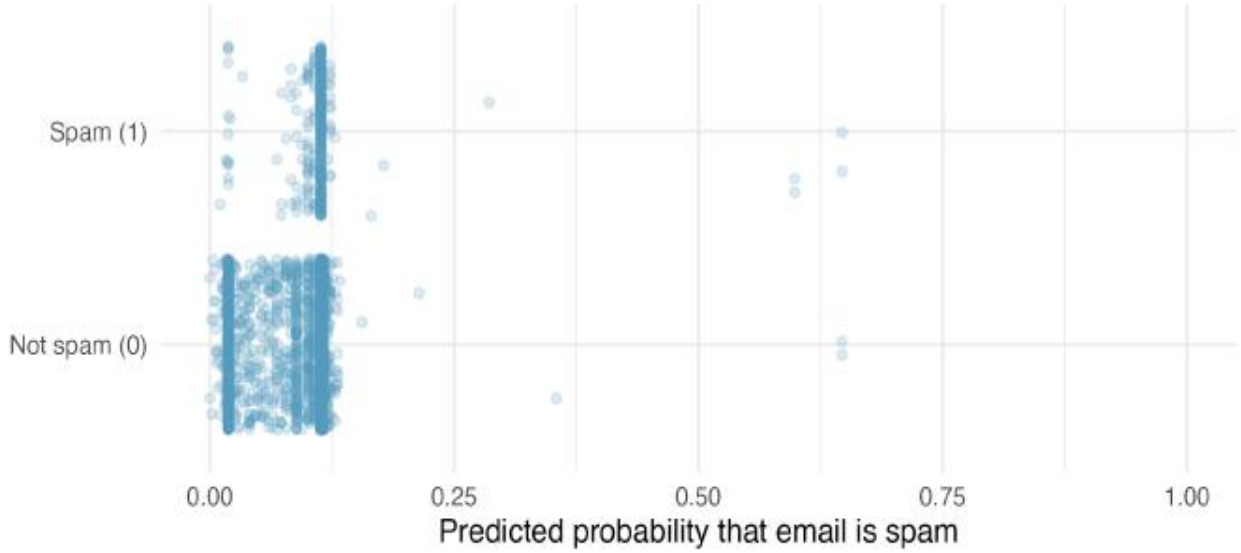


Figure 26.1: The predicted probability that each of the 3921 emails are spam. Points have been jittered so that those with nearly identical values aren't plotted exactly on top of one another.

Quality of the model

We'd like to assess the quality of the model. For example, we might ask: if we look at emails that we modeled as having 10% chance of being spam, do we find out 10% of the actually are spam? We can check this for groups of the data by constructing a plot as follows:

1. Bucket the observations into groups based on their predicted probabilities.
2. Compute the average predicted probability for each group.
3. Compute the observed probability for each group, along with a 95% confidence interval for the true probability of success for those individuals.
4. Plot the observed probabilities (with 95% confidence intervals) against the average predicted probabilities for each group.

Quality of the model

If the model does a good job describing the data, the plotted points should fall close to the line $y = x$, since the predicted probabilities should be similar to the observed probabilities. We can use the confidence intervals to roughly gauge whether anything might be amiss. Such a plot is shown in Figure 26.2.

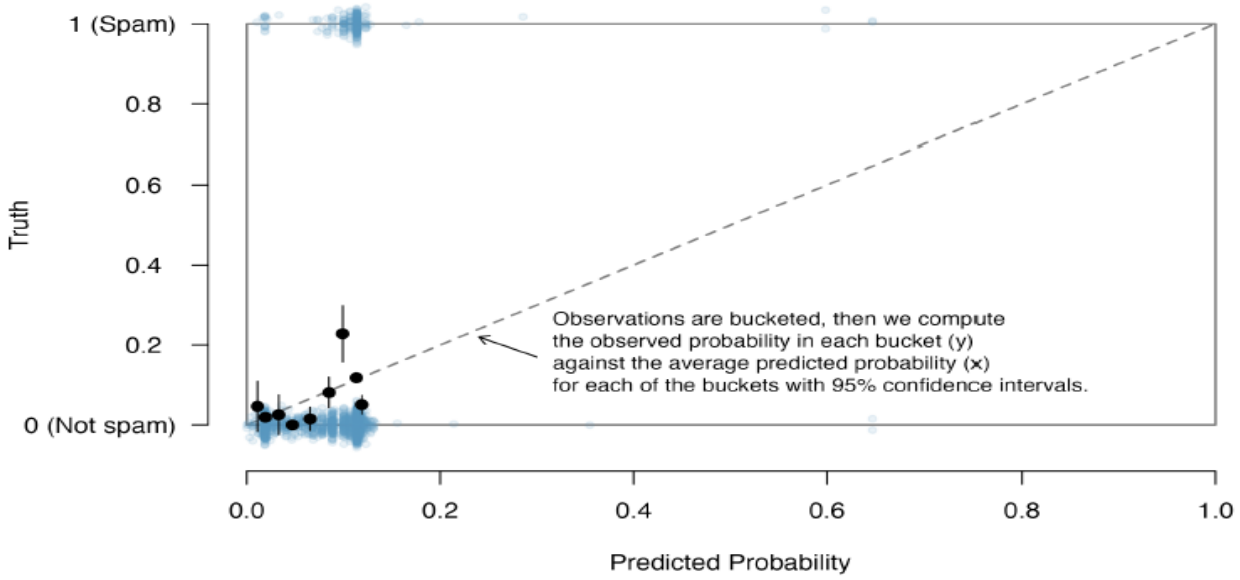


Figure 26.2: A reconfiguration of Figure 26.1. Again, the predicted probabilities are on the x-axis and the truth is on the y-axis for each email. After data have been bucketed into predicted probability groups, the proportion of spam emails (i.e., the observed probability) is given by the black circles. The dashed line is within the confidence bound of the 95% confidence intervals for many of the buckets, suggesting the logistic fit is reasonable.

The larger model

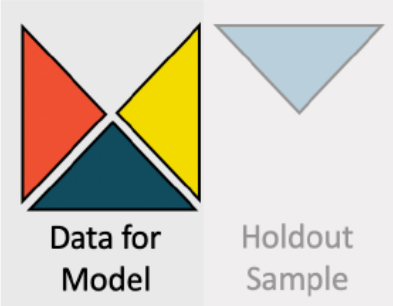
The larger model:

$$\log_e\left(\frac{\hat{p}}{1-\hat{p}}\right) = -0.34 - 2.56 \times \texttt{to\_multiple} + 0.20 \times \texttt{attach} + 1.73 \times \texttt{winner}_{yes}$$
$$- 1.28 \times \texttt{format} - 2.86 \times \texttt{re\_subj} + 0.00 \times \texttt{exclaim\_mess}$$
$$- 1.07 \times \texttt{number}_{small} - 0.42 \times \texttt{number}_{big}$$

Table 26.5: The larger model. Summary of a logistic model for predicting whether an email is spam based on the variables `to_multiple`, `attach`, `winner`, `format`, `re_subj`, `exclaim_mess`, and `number`. Each of the variables has its own coefficient estimate and p-value.

term	estimate	std.error	statistic	p.value
(Intercept)	-0.34	0.11	-3.02	0.0025
to_multiple1	-2.56	0.31	-8.28	<0.0001
attach	0.20	0.06	3.29	0.001
winneryes	1.73	0.33	5.33	<0.0001
format1	-1.28	0.13	-9.80	<0.0001
re_subj1	-2.86	0.37	-7.83	<0.0001
exclaim_mess	0.00	0.00	0.26	0.7925
numbersmall	-1.07	0.14	-7.54	<0.0001
numberbig	-0.42	0.20	-2.10	0.0357

Cross-validation



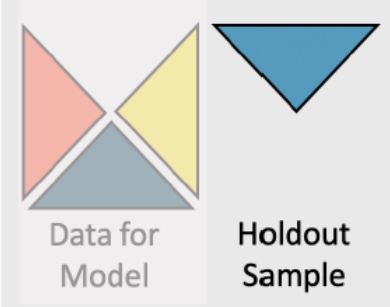


Table 26.6: The larger model. One quarter at a time, the data were removed from the model building, and whether the email was spam (TRUE) or not (FALSE) was predicted. The logistic regression model was fit independently of the removed emails. Now, the variables `to_multiple`, `attach`, `winner`, `format`, `re_subj`, `exclaim_mess`, and `number` are used to predict whether the email is spam. `spamTP` is the proportion of true spam emails that were predicted to be spam. `notspamTP` is the proportion of true not spam emails that were predicted to be not spam.'

fold	count	accuracy	notspamTP	spamTP
1st quarter	980	0.77	0.77	0.71
2nd quarter	981	0.80	0.81	0.70
3rd quarter	979	0.76	0.77	0.65
4th quarter	981	0.78	0.79	0.75