Quantitative Methods

Serena DeStefani - Lecture 1 – 7/6/2020

Introduction

Welcome to Quantitative Methods!

Course Information
LOCATION: Online!

DAY/TIMES: M,T,W,R 2:00PM-4:30PM
(attendance is mandatory)

Textbook
Stats: Data and Models, 4th Edition

Let's introduce ourselves.
Let's look at the Syllabus.

Outline

- Definition of Statistics
- Brief history of Statistics (first part)
- Data and variables
- Categorical Data

THIS IS STATISTICS

Statistics

The Hottest Career of the 21st Century

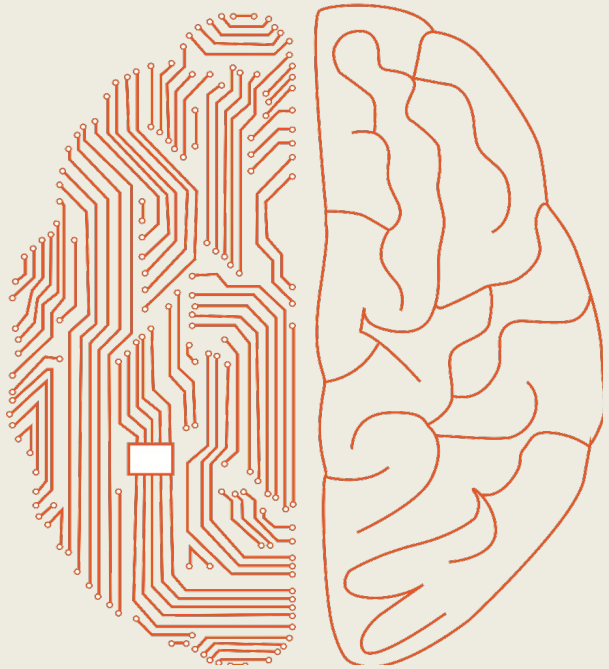
What is statistics?

It’s not what you think it is.

So, what *is* statistics?

Statistics is...

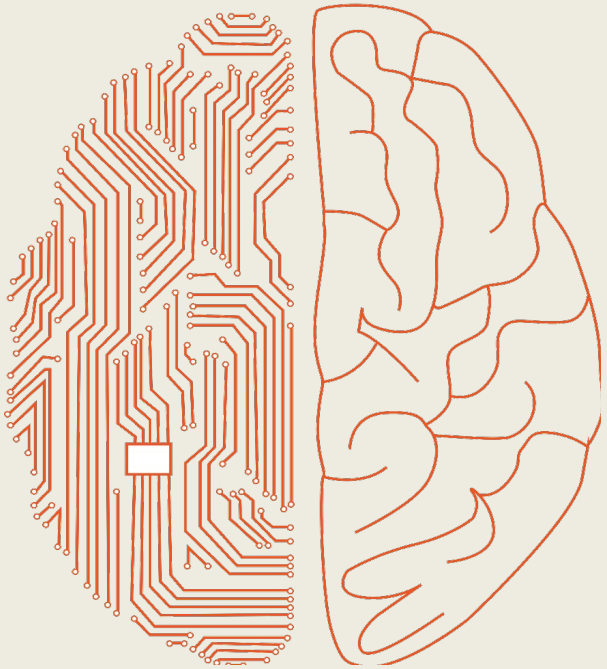
- The **science** of learning from **data**
- It's ***analyzing*** information
- It's creating **models** to capture ***insights***
- It's **solving** complex **problems**



So, what do statisticians *do*?

Statisticians...

- Help companies make sense of the world around us by ***analyzing*** data.
- Use data to **solve** complex ***problems***, in fields like **business**, **medicine**, **public service** and more.



Statisticians help manage the earth's natural resources

Fjallsjökull glacier in Iceland

Statisticians reduce disease and improve medicines

Spiral strand of DNA

Statisticians work for professional sports teams to make strategic draft picks

Jarvis Hayes, Darius Songalia, Mark Madsen and Rashad McCants. Image source: Keith Allison

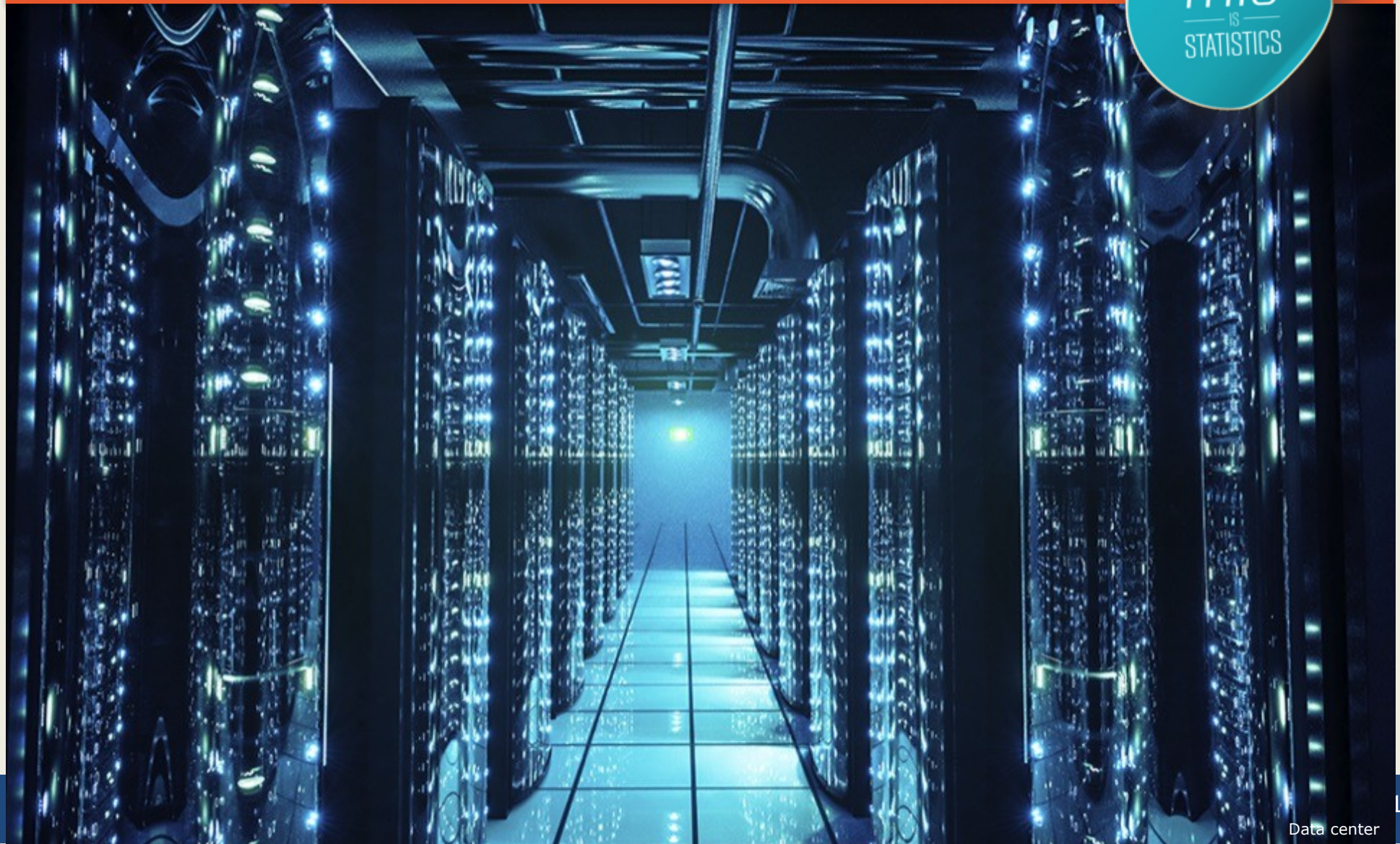
Statisticians improve voter targeting and assess the success of government policies and programs

President Barack Obama campaign rally in Urbandale, Iowa, 2012. Image source: White House

Statisticians help protect human rights in developing countries

Yezidi children in an internally displaced person camp in Sharya, Iraq. Image source: Human Rights Data Analysis Group

Statisticians advance computing through machine learning, speech recognition and artificial intelligence



Data center

Statisticians are crucial to making industry more efficient and financial institutions stronger

Budweiser plant. Image source: Ryan Glenn

Why study statistics?

The opportunity is endless.

What is Statistics?

1. The practice of collecting and collating numerical facts

Goes back to beginning of human civilization.
… in 1749, word *Statistik* introduced by Gottfried Achenwall.
from the latin statisticum
("of the state").

It designated the collection
of data about the state.

The Definition of Statistics

2) The process of reasoning about the data collected

→ inferential statistics:

- parameter estimation

- hypothesis testing

Why do we need statistics?

What is biology about?
What is economics about?
What is sociology about?
…What is statistics about?

Statistics is about *variation.*

Statistics helps us make sense of the data and how the data vary.
Statistics is a collection of conceptual and mathematical tools that allow us to study such variation.

Why is Statistics important for Psychology?

The use of Statistics qualifies Psychology as a science…

- using statistics we can determine whether a psychological hypothesis is true for a **wider population**, or whether a treatment works or not.

Statistical methods provide a unifying force within Psychology.

What's the origin of Statistics?

Starting from the **17th century**,
Statistics (definition #2: the process of reasoning about the data) originated from different fields:

- Demography → statistical summaries
- Astronomy → theory of errors → normal distribution
- Gambling → probability theory
- Agriculture → experimental design

What's the origin of Statistics?

- Let's take a step back…

- In 1687 Newton published *Principia Mathematica*

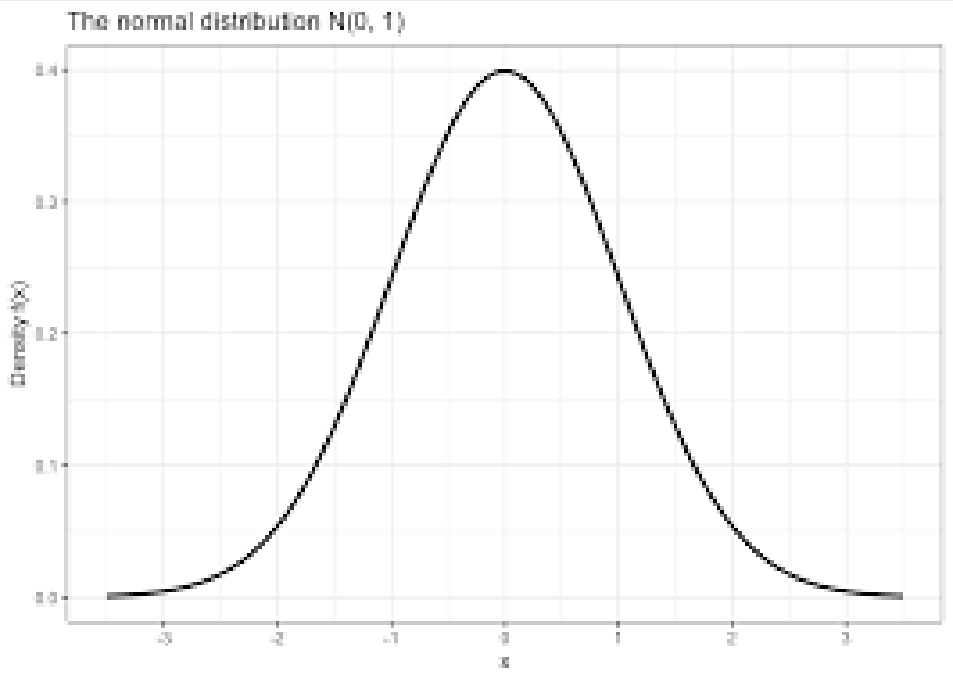
- Laws of motion, gravitation

- Now it's possible to predict mathematically the movement of objects, including stars!

- But what about other complex phenomena?

What's the origin of Statistics?

- At some point it was found that many events follows what we call a normal distribution



- How did they find out?

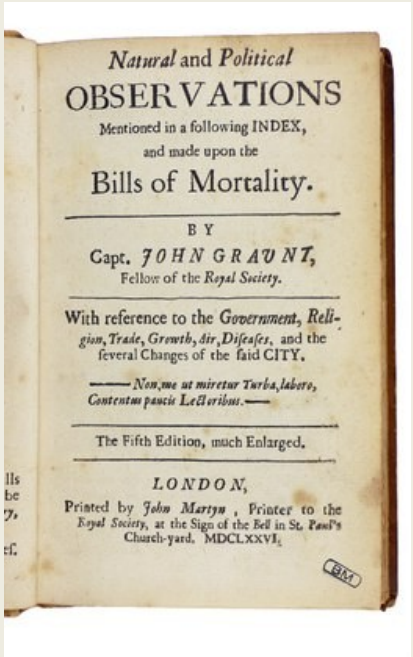
What's the origin of Statistics?

Starting from the **17th century**,
Statistics (definition #2: the process of reasoning about the data) originated from different fields:

- Demography → statistical summaries
- Astronomy → theory of errors → normal distribution
- Gambling → probability theory
- Agriculture → experimental design

Demography

John Graunt, a London haberdasher,
born in 1620, tried to predict
and explain social phenomena from tables
he compiled from the “Bills of Mortality”.



Natural and *Political*

OBSERVATIONS

Mentioned in a following INDEX,
and made upon the

Bills of Mortality.

BY

Capt. *JOHN GRAUNT*,
Fellow of the *Royal Society*.

With reference to the *Government*, *Religion*, *Trade*, *Growth*, *Air*, *Diseases*, and the several Changes of the said CITY.

——*Non, me ut miretur Turba, laboro,*
Contentus paucis Lectoribus.——

The Fifth Edition, much Enlarged.

LONDON,

Printed by *John Martyn*, Printer to the *Royal Society*, at the Sign of the Bell in St. *Paul's* Church-yard. MDCLXXVI.

Astronomy

For science to progress, scientists propose hypotheses to test.

In order to that, they need to **collect data**…

In order to collect data, scientists must make **experimental measurements**

Astronomy

But how to measure the position of the stars?!

Initially, by the naked eye.

Tycho Brahe, a Dane, worked from ~ 1570 - 1601 and built the most accurate naked eye observatory ever.

Scientists used to make one observation only
→ Problem of errors
1720, Roger Cotes: reporting the arithmetic average of **group of observations** decreased the error of the measurement process ?

Astronomy

1755 Thomas Simpson proposed that the **mean** of a series of observations was **a better estimate** of the true quantity of the object to be measured than any single observation, however meticulously obtained.

1755 Bayes, in a comment on Simpson, noted that the mean only made sense as a superior estimator if the deviations from the mean were **symmetric** about it.

Simpson took note and revised his recommendation in 1757: report both the mean (as the “best” estimate) and the scatter of the deviations from the mean.

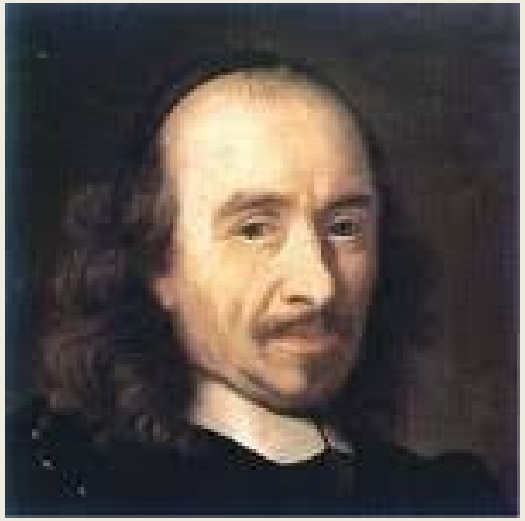
Astronomy

So scientist are now taking about more than one observation and reporting the mean and the deviations from the mean (errors).

But do these errors have a regular distribution???

Gambling

Antoine Gombauld, the Chevalier de Mere, a writer and gambler, consulted his friend,
Blaise Pascal (1623-1662) on how to calculate expected (probable) frequency of gains and losses, and how to divide the stakes fairly if the game was interrupted.



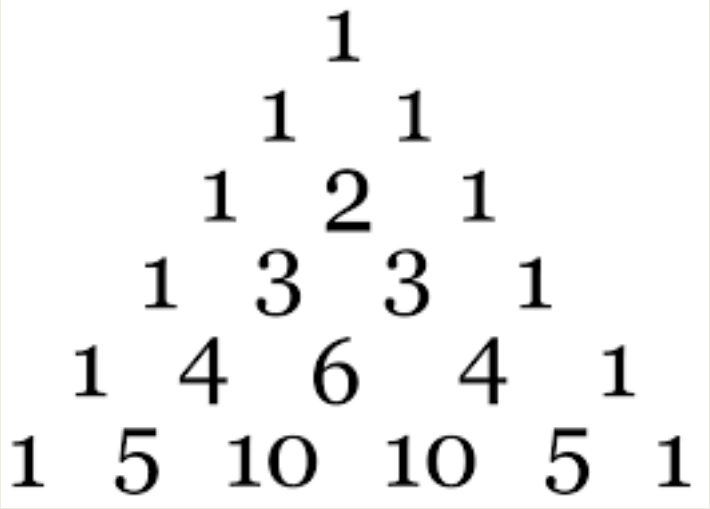
Gambling

Pascal did not know the solution and wrote to his friend Fermat.

Correspondence between Gombauld, Pascal and Fermat: the birth of probability theory

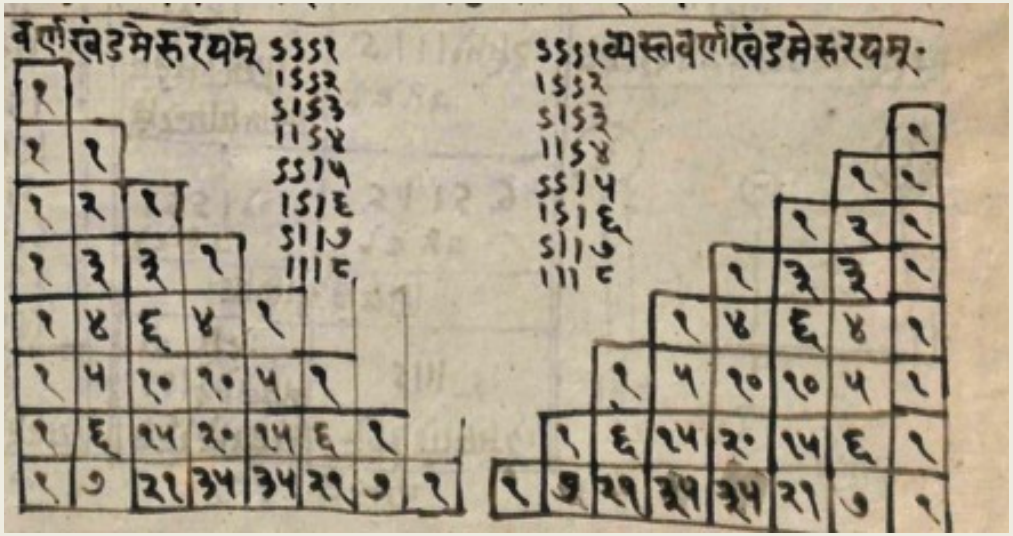
Probability

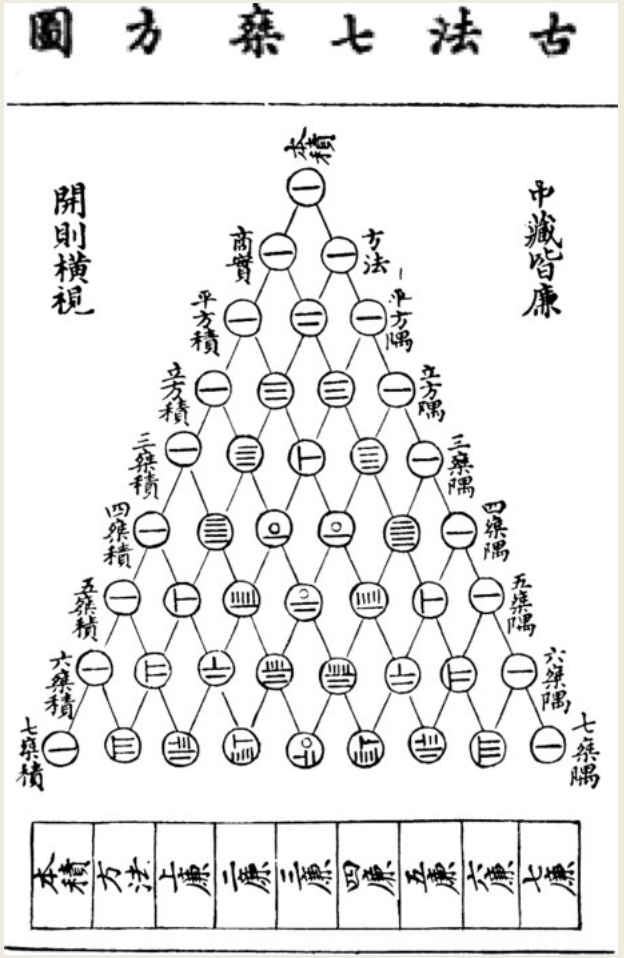
Pascal connected the study of probability with the arithmetic triangle:



The arithmetic triangle

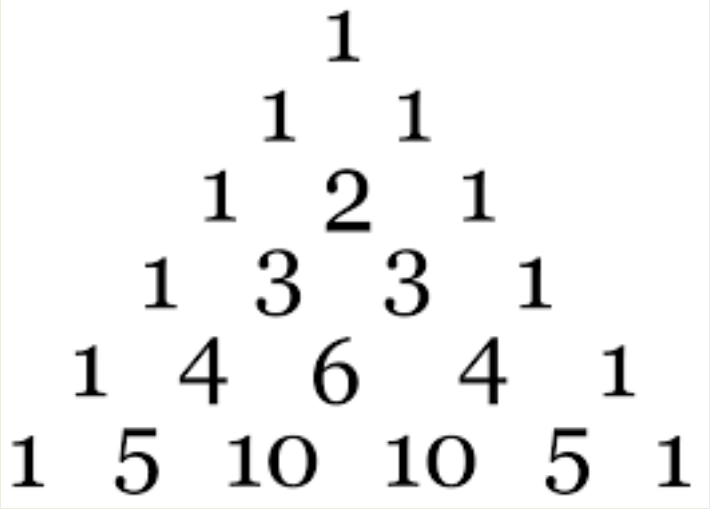
Was already known in India and China





The arithmetic

Pascal connected the study of probability with the arithmetic triangle:



This triangle is linked to the binomial expansion

Binomial expansion

The arithmetic triangle is linked to the binomial expansion

$$(a+b)^0 = 1$$

$$(a+b)^1 = a+b$$

$$(a+b)^2 = a^2 + 2ab + b^2$$

$$(a+b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$$

$$(a+b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$$

$$(a+b)^5 = a^5 + 5a^4b + 10a^3b^2 + 10a^2b^3 + 5ab^4 + b^5$$

$$(a+b)^6 = a^6 + 6a^5b + 15a^4b^2 + 20a^3b^3 + 15a^2b^4 + 6ab^5 + b^6$$

...

Binomial expansion

The arithmetic triangle is linked to the binomial expansion

$$(x+y)^0 = 1 \quad \text{0th row}$$

$$(x+y)^1 = 1x + 1y \quad \text{1st row}$$

$$(x+y)^2 = 1x^2 + 2xy + 1y^2 \quad \text{2nd row}$$

$$(x+y)^3 = 1x^3 + 3x^2y + 3xy^2 + 1y^2 \quad \text{3rd row}$$

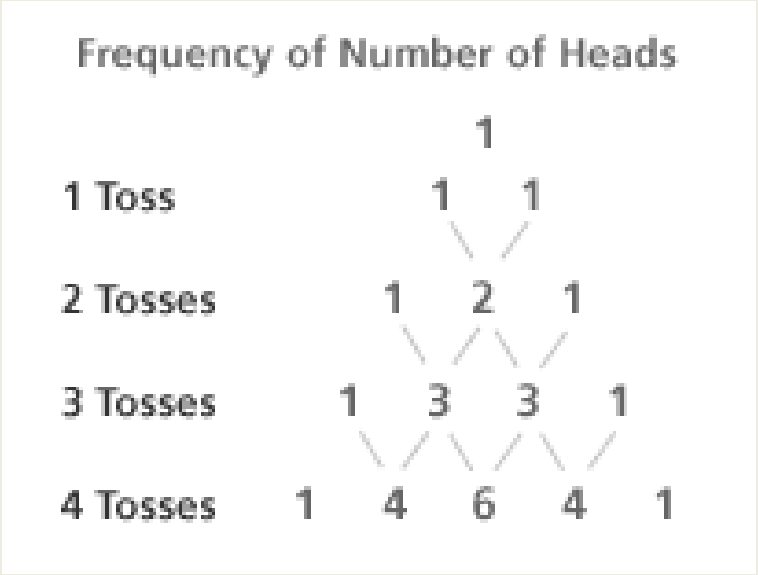
$$(x+y)^4 = 1x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + 1y^4 \quad \text{4th row}$$

$$(x+y)^5 = 1x^5 + 5x^4y + 10x^3y^2 + 10x^2y^3 + 5xy^4 + 1y^5 \quad \text{5th row}$$

Watch video at: https://bit.ly/3dwO969

Binomial expansion and probability

We can use the binomial expansion (and the triangle) to find the probability of some simple events, like tossing a coin

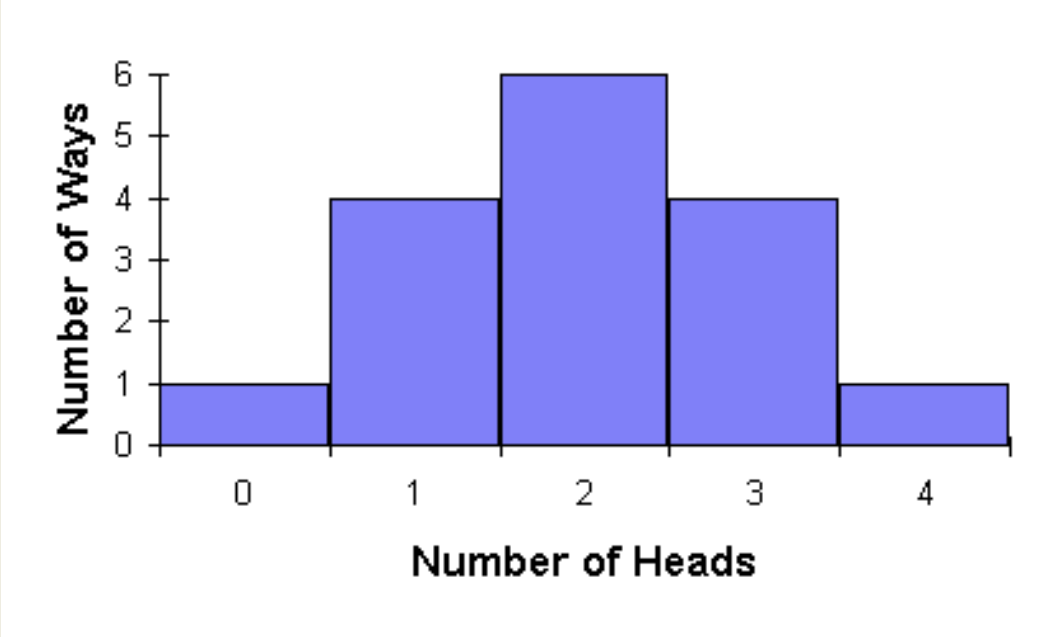


$$(\tfrac{1}{2} + \tfrac{1}{2})^4 = \frac{1}{16} + \frac{4}{16} + \frac{6}{16} + \frac{4}{16} + \frac{1}{16}$$

Binomial expansion and probability

We can plot the frequency of getting heads on an histogram

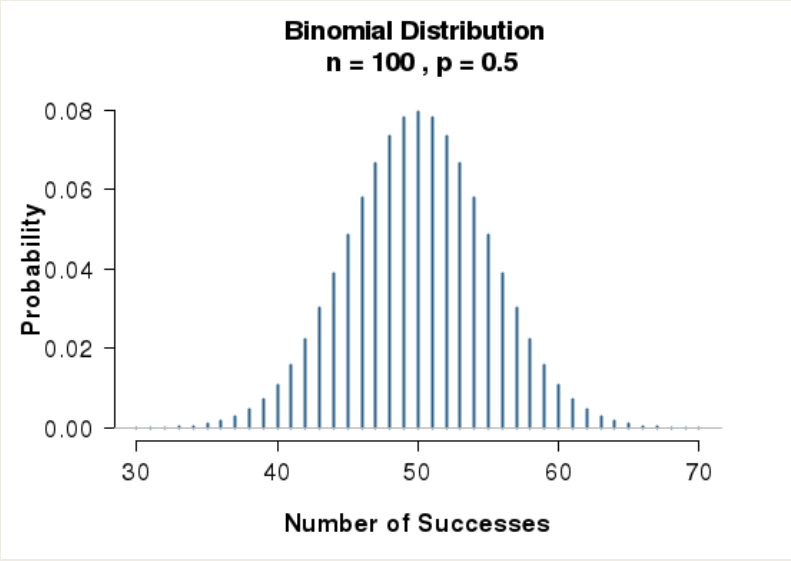
$$(\tfrac{1}{2} + \tfrac{1}{2})^4 = \frac{1}{16} + \frac{4}{16} + \frac{6}{16} + \frac{4}{16} + \frac{1}{16}$$



Binomial expansion and probability

The more coin tosses I make, the more this histogram will resemble a curve:

See simulation at:
https://shiny.rit.albany.edu/stat/binomial/



Binomial expansion and probability

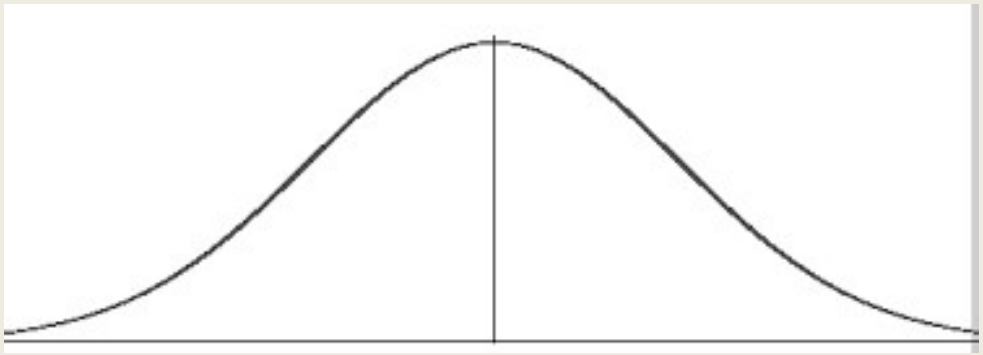
Abraham De Moivre (1667-1754) published in 1738: *The Doctrine of Chances or a Method of Calculating the Probabilities of Events in Play*

In the third edition (1756) he showed a way to approximate the sum of the binomial terms when **n** is very large.

Binomial expansion and probability

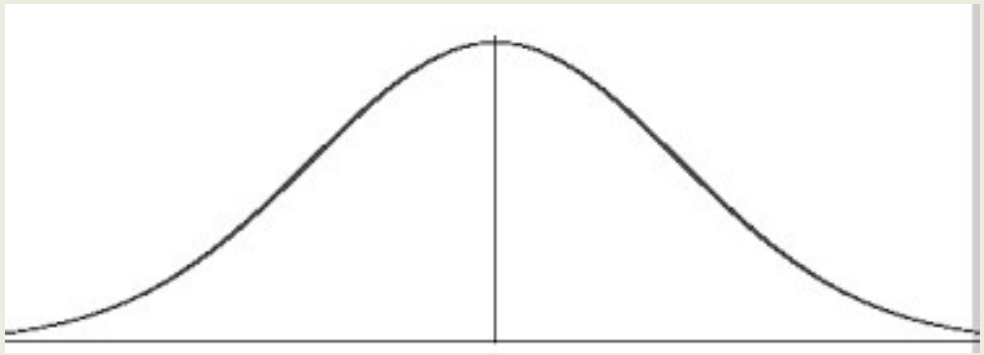
De Moivre (1756) showed a way to approximate the sum of the binomial terms when **n** is very large.

Now you can calculate probabilities for an infinite number of tosses! An if you graph them, you get this curve:



A function for the binomial expansion (large n)

Carl Friedrich Gauss (1777-1855) was the first one to derive a function for this curve (1809)



Normal Probability Density Function

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

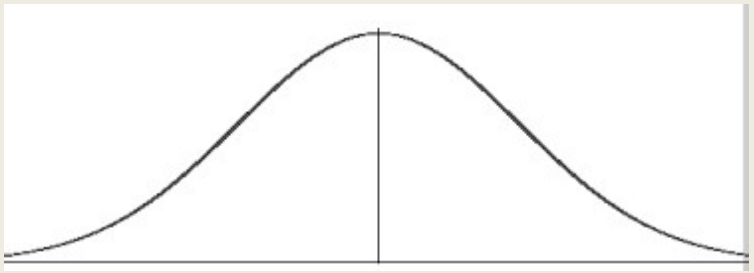
Astronomy

So scientist are now taking about more than one observation and reporting the mean and the deviations from the mean (errors).

But do these errors have a regular distribution???

The Law of Errors

Pierre Laplace (1749-1827) independently derived the formula of the normal distribution in 1812 and understood that this function was the one describing the distribution of errors!
It is called Gauss-Laplace distribution, or Gaussian distribution, or Normal distribution
One of the first applications of the distribution outside of gaming was in the assessment of errors in astronomy.



Normal Probability Density Function

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

The Normal Distribution

Summary:

1654 Gombauld writes to Pascal. Pascal studies the arithmetic triangle and the probability of coin tosses
1756 De Moivre approximates the binomial function
1757 Simpson proposes the mean as optimal measurement, given that the errors are distributed evenly
1809-1812 Gauss and Laplace derive the Normal distribution and the “Law of Errors”

The Normal Distribution

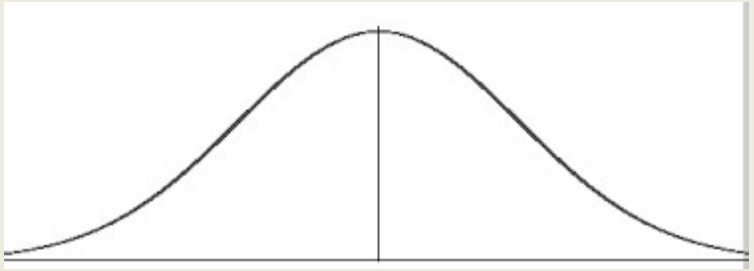
The ready acceptance of the normal distribution as a **law of nature** encouraged its wide application and produced consternation when exceptions were observed.

Are there distribution that are asymmetrical?

The Normal Distribution

It is safe to say that no other theoretical mathematical abstraction as had such an important influence on psychology and the social science.

Using this distribution we can calculate the probabilities of a wide range of events and conduct hypothesis testing.



Normal Probability Density Function

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

Chapter 1

Data

Data

Data

- Any **collection** of numbers, characters, images, or other items that provide **information** about something
- Data **vary**: Surveys and experiments produce a variety of outcomes.

Statistical inference is making a decision or a conclusion based on the data.

Example of Data Uses: Facebook

Facebook collects data about you!

- Personal information: age, gender, education, etc.
- Interests: based on what you “like”

Statistics is used to determine which ads you see.

If you follow an ad, Facebook now has even more data. Your information on Facebook is a goldmine for the company.

Example: Texting While Driving

- Is texting while driving dangerous?
- Texting has grown dramatically in the last five years.
- Driving fatalities have gone down significantly in the last five years.
- Is texting while driving safe?
- How might you decide?

Texting While Driving: University of Utah Study

- Measured reaction times of sober, drunk, and texting drivers in simulated driving emergencies

- Result: Those texting had the *slowest* reactions.

- Is texting while driving safe?

Organizing Data

105-2686834-3759466	Ohio	Nashville	Kansas	10.99	440	N	B00000I5Y6	Katherine H.
105-9318443-4200264	Illinois	Orange County	Boston	16.99	312	Y	B000002BK9	Samuel P.
105-1872500-0198646	Massachusetts	Bad Blood	Chicago	15.98	413	N	B000068ZVQ	Chris G.
103-2628345-9238664	Canada	Let Go	Mammals	11.99	902	N	B0000010AA	Monique D.
002-1663369-6638649	Ohio	Best of Kansas	Kansas	10.99	440	N	B002MXA7Q0	Katherine H.

- Difficult to decipher the data above

Order Number	Name	State/Country	Price	Area Code	Previous Album Download	Gift?	ASIN	New Purchase Artist
105-2686834-3759466	Katherine H.	Ohio	10.99	440	Nashville	N	B0000015Y6	Kansas
105-9318443-4200264	Samuel R	Illinois	16.99	312	Orange County	Y	B000002BK9	Boston
105-1372500-0198646	Chris G.	Massachusetts	15.98	413	Bad Blood	N	B000068ZVQ	Chicago
103-2628345-9238664	Monique D.	Canada	11.99	902	Let Go	N	B0000010AA	Mammals
002-1663369-6638649	Katherine H.	Ohio	10.99	440	Best of Kansas	N	B002MXA7Q0	Kansas

- Presentation can make all the difference.

The “Five W”s

- Who: Describe the individuals who were surveyed.
- What: Determine what is being measured.
- When: When was the research conducted?
- Where: Where was the research conducted?
- Why: What was the purpose of the survey or experiment?
- How: Describe how the survey or experiment was conducted.

Who and What

•Rows correspond to individual **cases**, that may go by different names:

- Respondents: Individuals who answer the survey
- Subjects or Participants: People who are experimented on
- Experimental Units: The object of the experiment when it is not a person
- Records: Rows in a database

Sample and Population

- The goal is to describe the population.
- This is usually impractical or impossible.
- A sample is used to make inferences about the population.
- The sample should be ***representative*** of the population.

Categorical Variables

- Categorical Variable: A variable that tells us what group or category an individual belongs to

- **Synonyms:** nominal and qualitative

- **Examples:** Favorite color, country of birth, area code

- **Drawback of Categorical Variables:** Challenging to analyze with computation

Quantitative Variables

- Quantitative Variable: Contains measured numerical values with measurement units
  - Typically records the ***amount*** or ***degree*** of something
  - **Unit Examples:** ounces, dollars, degrees Fahrenheit

Categorical or Quantitative?

- Amazon knows your age and will use it to present an age-appropriate image customized for you.
- Is *Age* categorical or quantitative?
- Perceived as Child, Teen, Young Adult, Middle Aged, Senior, age is categorical.
- Can also be perceived as quantitative if not categorized into a type.

Identifiers

- Identifier Variable: A variable that is used to uniquely identify the individual. It does not describe the individual.
  - Login ID
  - Customer Number
  - Transaction Number
  - Social Security Number

Ordinal Variables

- Ordinal Variable: A variable that reports order without natural units
  - Four-point Likert Scale: Strongly Disagree, Disagree, Agree, Strongly Agree
  - Olympic Rank: Gold, Silver, Bronze
- Can be treated as quantitative by using the rank number
  - 1 = Strongly Disagree, 2 = Disagree, 3 = Agree, 4 = Strongly Agree

Chapter 2

Displaying and Describing Categorical Data

Displaying and Describing Categorical Data

- Frequency tables
- Bar Charts
- Pie Charts

- Choose the chart that best tells the story of your data.

- Charts often work better when the categories do not overlap.

Frequency Tables

- A frequency table is a table whose first column displays each distinct outcome and second column displays that outcome's frequency.

Class	Count
First	325
Second	285
Third	706
Crew	885

- If there are many distinct outcomes, then combining them into a few categories is recommended.

Relative Frequency Tables

Class	%
First	14.77
Second	12.95
Third	32.08
Crew	40.21

- A relative frequency table is a table whose first column displays each distinct

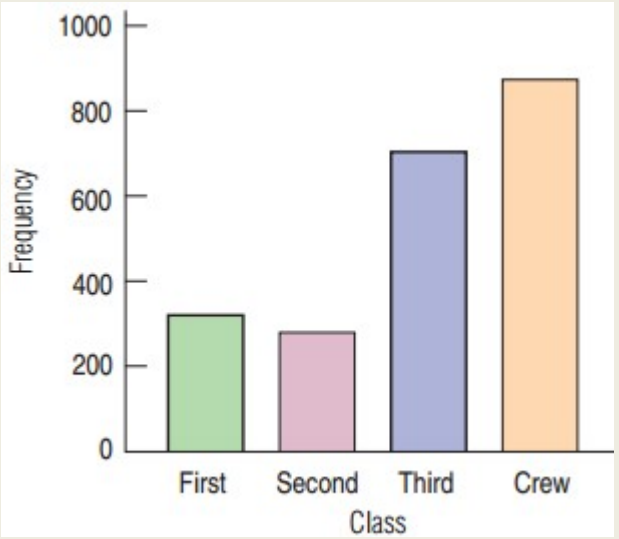
outcome and second column displays that outcome's relative frequency.

- The relative frequency table is similar to

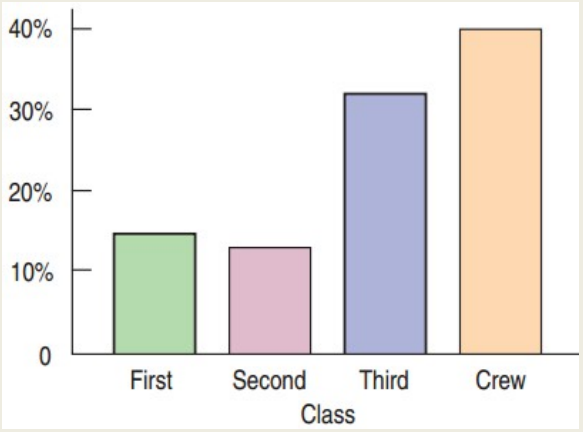
the frequency table, but it displays relative frequencies rather than frequencies.

Bar Charts

- A bar chart displays the frequency or relative frequency of each category.
- All bars must have the same width.
- Good for general audience



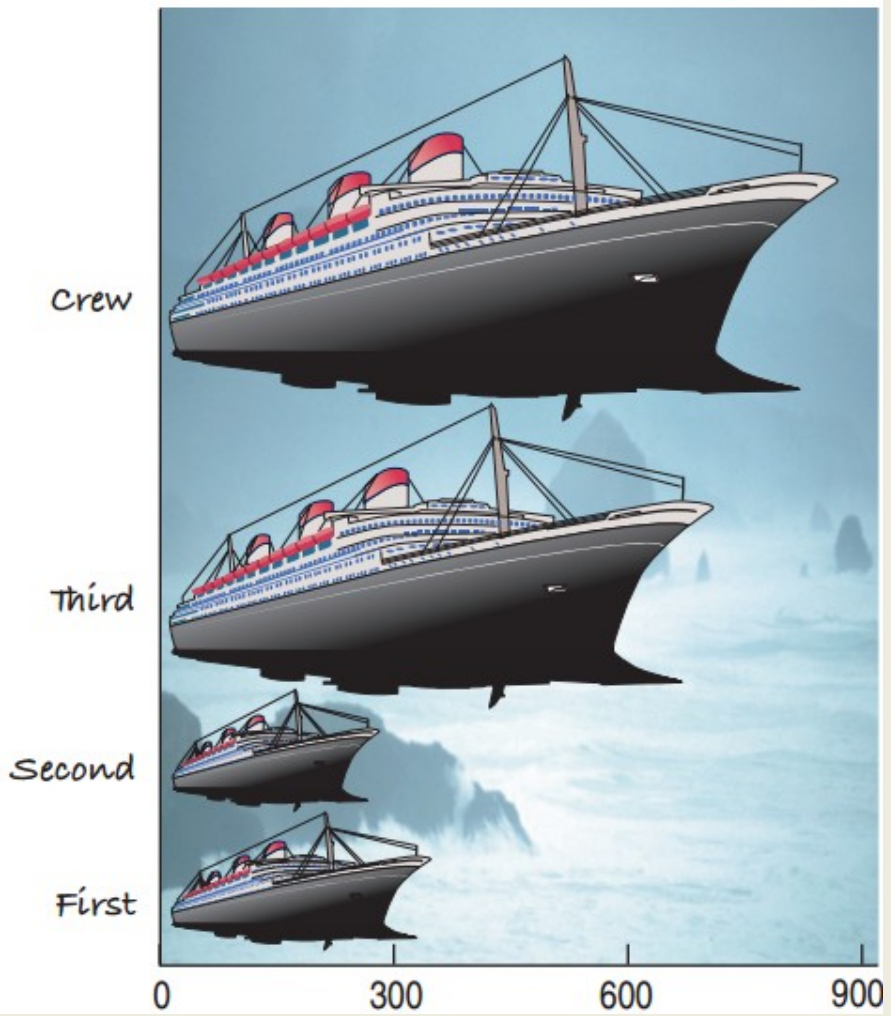
Frequency Bar Chart



Relative Frequency Bar Chart

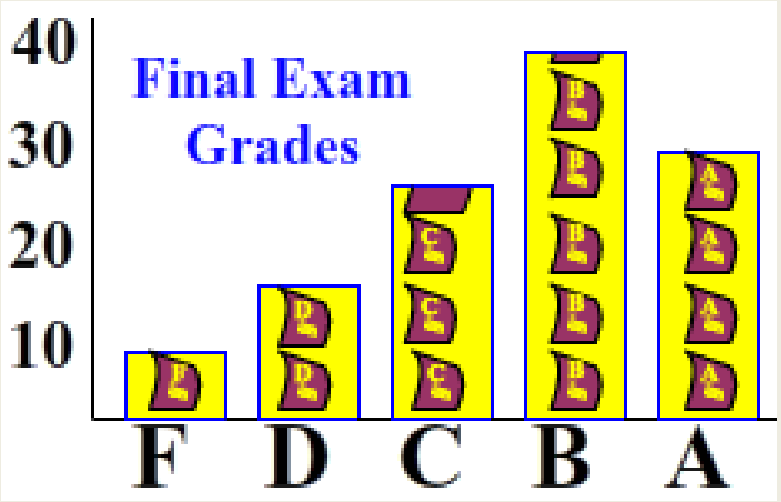
A (wrong) example

- Were most members of the Titanic crew members?
- Three times as many crew members as second-class passengers
- The eyes are tricked by the area being nine times as large for the crew.

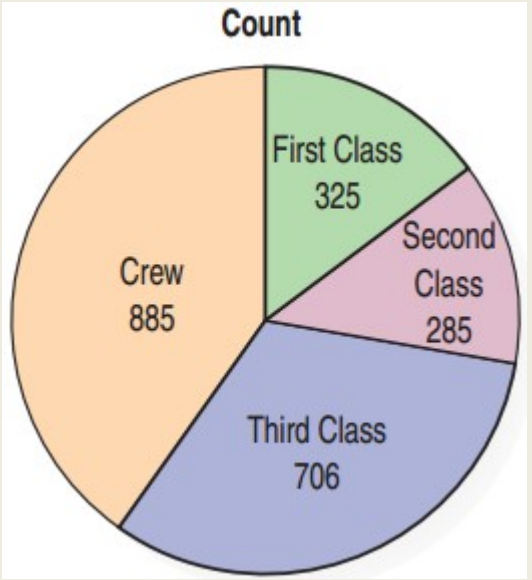


The Area Principle

- The Area Principle: The area occupied by a part of the graph should correspond to the magnitude of the value it represents.
  - Bars should have equal widths in a bar chart.
  - Be cautious when using two-dimensional pictures to exhibit one-dimensional data.



Pie Charts



- A pie chart presents each category as a slice of a circle so that each slice has a size that is proportion to the whole in each category.

- Pie charts help to display the fraction of the whole that each category represents.

- Better not to use a pie chart in science

Contingency Tables

Survival	Class: First	Second	Third	Crew	Total
Alive	203	118	178	212	711
Dead	122	167	528	673	1490
Total	325	285	706	885	2201

- A contingency table is a table that displays **two** categorical variables and their relationships.
- There were 528 third-class ticket holders who died.
- The bottom row represents the totals for class and is called the marginal distribution.
- The right column represents the totals for survival and is also a marginal distribution.

Table of Percents

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	9.2%	5.4%	8.1%	9.6%	32.3%
	Dead	5.5%	7.6%	24.0%	30.6%	67.7%
	Total	14.8%	12.9%	32.1%	40.2%	100%

- A table of percents can be misleading.
- Looking at "Alive", was it better to have a second- or third-class ticket?
  - 8.1% were third-class survivors, 5.4% were second-class survivors.
  - What is wrong with just comparing these percentages?

Conditional Distributions

- A conditional distribution provides the percent of one variable satisfying the conditions of another.

			Class				
			First	Second	Third	Crew	Total
Survival	Alive	Count	203	118	178	212	**711**
		% of Column	62.5%	41.4%	25.2%	24.0%	**32.3%**
	Dead	Count	122	167	528	673	**1490**
		% of Column	37.5%	58.6%	74.8%	76.0%	**67.7%**
	Total	Count	**325**	**285**	**706**	**885**	**2201**
			100%	**100%**	**100%**	**100%**	**100%**

- 25.2% of all third-class ticket holders survived.

- Was it better to have a second- or third-class ticket?

Conditional Distribution: Rows or Columns

- The "Condition" can either be based on rows or columns.

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	203	118	178	212	711
		28.6%	16.6%	25.0%	29.8%	100%
	Dead	122	167	528	673	1490
		8.2%	11.2%	35.4%	45.2%	100%

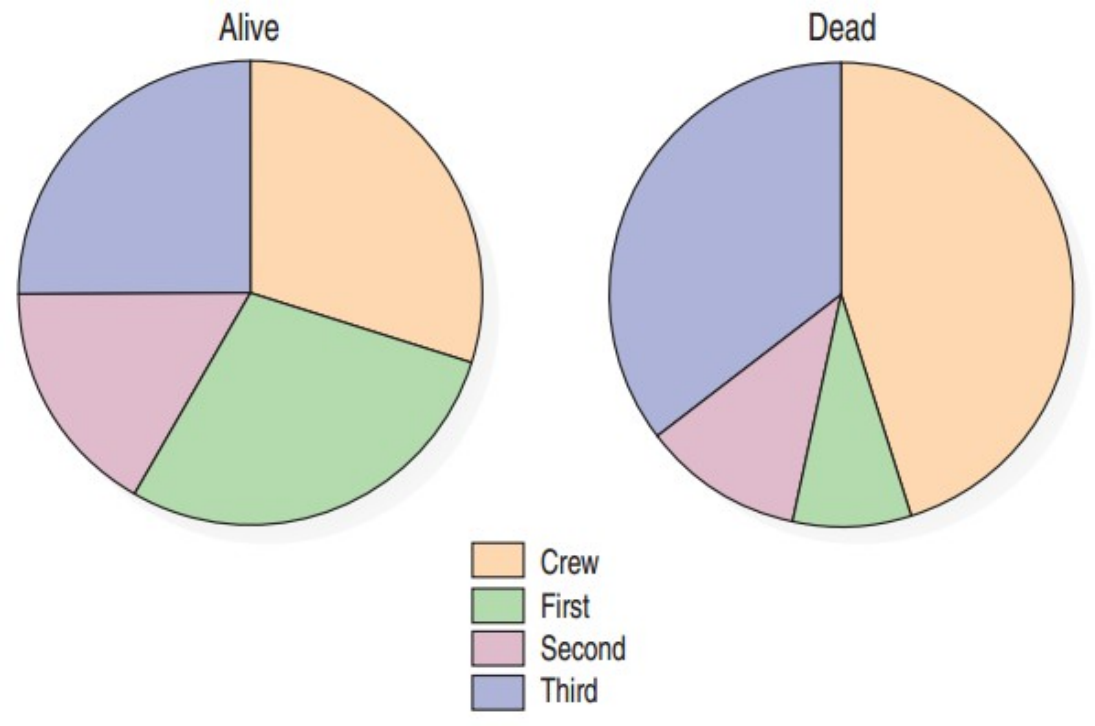
- This table shows that the highest percent of survivors were crew members.

- The highest percent of the dead were also crew members.

Conditional Distributions as Pie Charts

- Pie charts can give a visual representation of the conditional distributions.

- Compare how the first-class ticket holders were represented amongst the survivors vs. the dead.

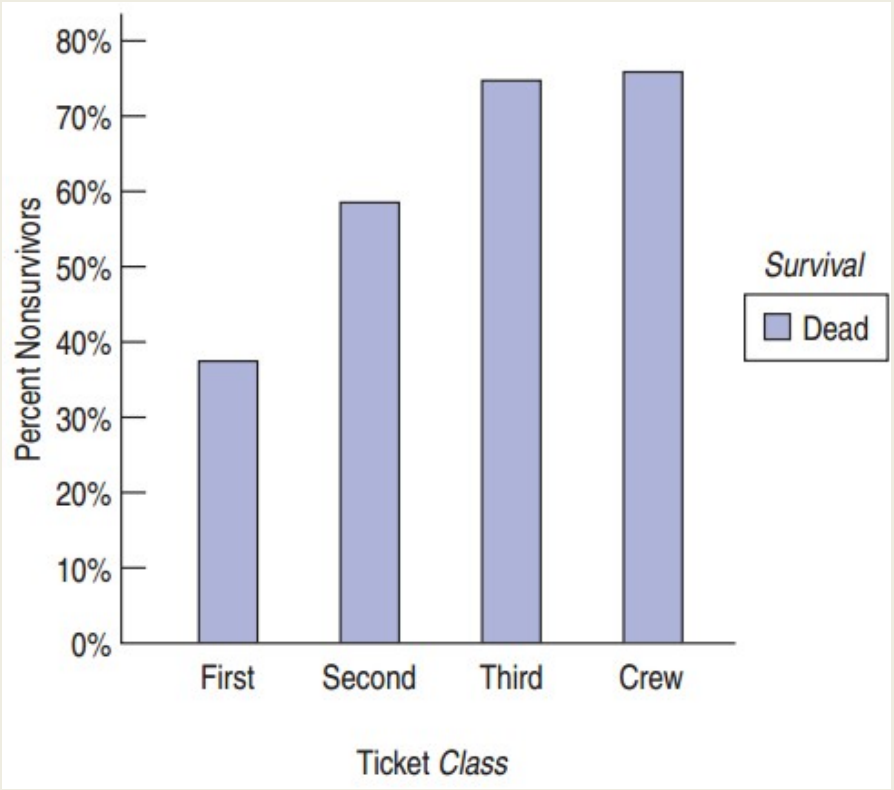


Bar Charts

- Bar charts can also effectively tell the story for conditional distributions.

- Which is best:

Table, Pie chart, or Bar Graph?



Independence

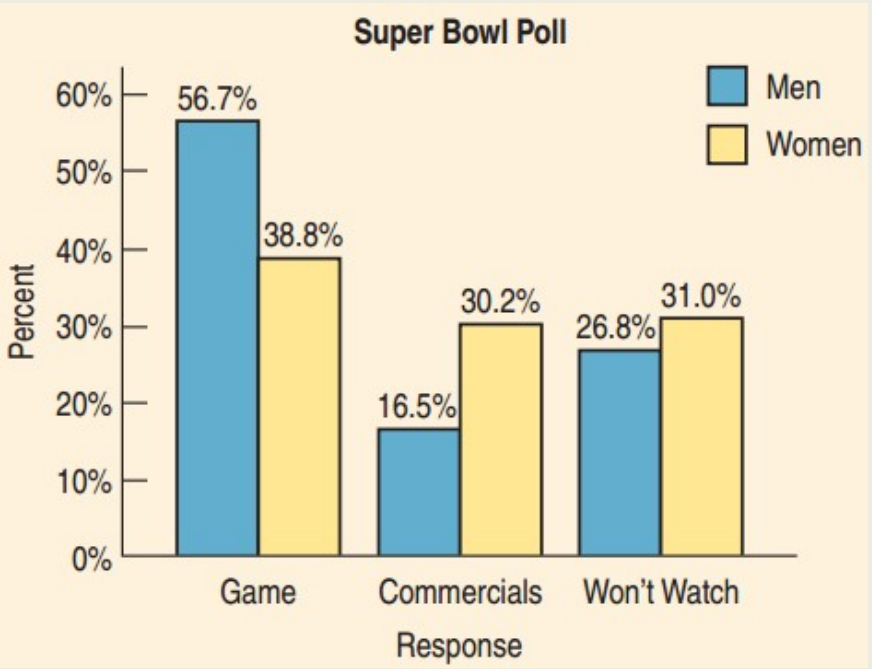
- Independence: The distribution of one variable is the same for all categories of another. There is no association between the two.

- For dependent variables, there is an association between the two variables.

Example

- Is there an association between gender and interest in Super Bowl TV Coverage?
- Large difference for men between watching the game and commercials
- Smaller difference for women
- There is an association between gender and interest.

Response	Sex: Male	Female	Total
Game	279	200	479
Commercials	81	156	237
Won't Watch	132	160	292
Total	492	516	1008



Simpson's Paradox

Definition: An association that holds for all of several groups can reverse direction when the data are combined to form a single group. This reversal is called Simpson's paradox.

History: Simpson's paradox is named after Edward Simpson: he described this paradox in the 1951 paper "The Interpretation of Interaction in Contingency Tables."

Pearson and Yule each observed a similar paradox half a century earlier than Simpson, so Simpson's paradox is sometimes also referred to as the Simpson-Yule effect.

Simpson's Paradox: Example

We want to test two drugs.
We give each drug to a group of people and then count the number of successes (improvements) and failures (no change) for each group.

	Success	Failure	Total
Drug 1	100	100	200
Drug 2	110	80	190

Simpson's Paradox: Example

Let's look at the same result split by gender.

Male:

	Success	Failure	Total
Drug 1	60	20	80
Drug 2	100	50	150

Female:

	Success	Failure	Total
Drug 1	40	80	120
Drug 2	10	30	40

What are the limits of Statistics?

Can data answer every possible questions?

Data will help us remember, but will it let us forget? It will help politicians get elected, but will it help them lead? It will help companies make products addictive, but will it help us get free once we're hooked? It will help advertisers see people as statistics, but will it help us remember those statistics are people? It will help banks prevent credit card fraud, but will it help us stay out of debt? It will help credit card companies predict the impending collapse of a marriage, but will it keep our marriages from falling apart? It will help parents make kids genetically perfect, but will it help us love them regardless? It will help high-frequency traders sell stocks in nanoseconds, but will it help protect markets from feedback loops in their programs? It will help meteorologists predict storms and tornadoes, but will it help us rebuild the homes of survivors? It will help biologists map the migration of fish, but will it keep us from overfishing the oceans? It will help physicists find the "God particle" in a supercollider, but will it help us agree about God? It will help astronomers search for signs of alien life, but will it help us know if aliens are friendly or mean? It will help cardiologists monitor pacemakers with WiFi connections, but will it keep hackers from hacking our hearts? It will help virologists publish the genomes of major diseases, but will it keep terrorists from developing weaponized strains? It will help soldiers kill enemies remotely with drones, but will it help us see war as more than a game? It will help urbanists develop "smart cities," but what will become of our towns? It will help governments map the consumption patterns of cities, but will it help us depend less on consuming? It will help hackers leak evidence of government surveillance, but will we treat those hackers as heroes or thieves? It will help police triangulate the location of gunshots, but will it help us address the underlying causes of violence? It will help educators make excellent standardized tests, but will it help us embrace different standards of excellence? It will help farmers engineer crops to produce bigger yields, but will it keep corporations from patenting our food? It will help search engines know how often people search for "love," but will it help people find it? It will help singles plan a hundred first dates, but will it help them know when they've found the right person? It will help pet owners clone their dogs and their cats, but will it help us love the clones as much as the cloned? It will help neurologists implant chips in our brains, but will it help us turn off the chatter? It will help geneticists sequence our genome, but will it help us understand who we are? It will help us feel connected, but will it help us feel loved? It will help us uncover the facts, but will it help us be wise? It will help us live forever, but will it help us see that life's meaning stems from the fact that it ends? It will help us keep count of everything in our lives, but will it help us understand that not everything that counts in our lives can be counted? It will help us see the world as it is, but will it help us see the world as it could be?