Quantitative Methods

Serena De Stefani – Lecture 9 – 7/24/2017

Announcements

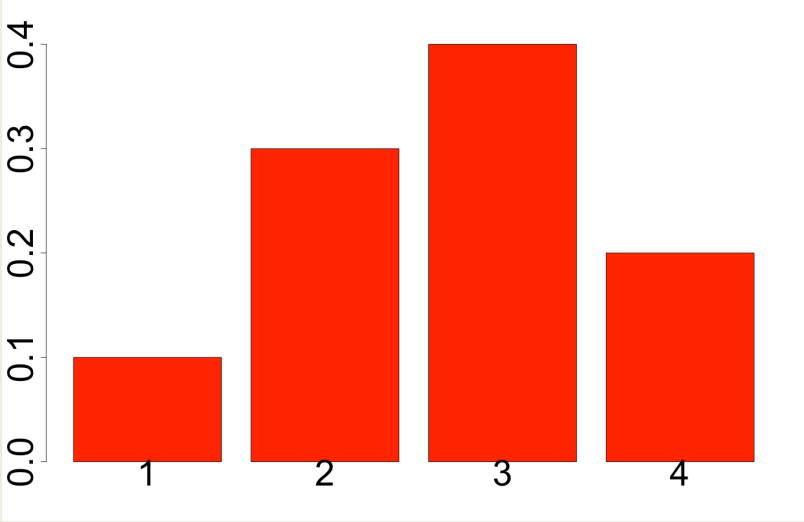
- Review: today and tomorrow
- Thursday: midterm CH 1-14

Summary

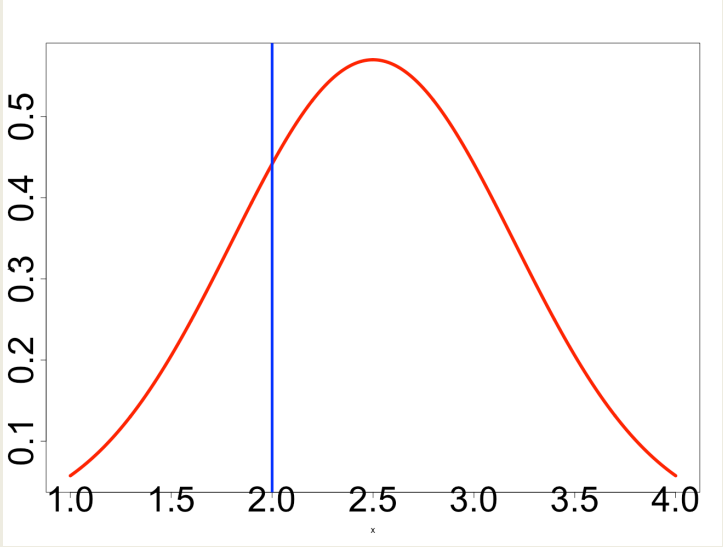
- **Question: is my result likely?** I need a model
- How do we build a model for our data?
- Random variable
- → Probability model (or probability distribution)
- There are many different distribution we can define and use, discrete or continuous.
- Examples:
  - Discrete: geometric or binomial
  - Continuous: normal, uniform, exponential
- Focus on binomial and normal distributions
- Before we look at distributions, we need to define a basic “experiment”, a Bernoulli trial

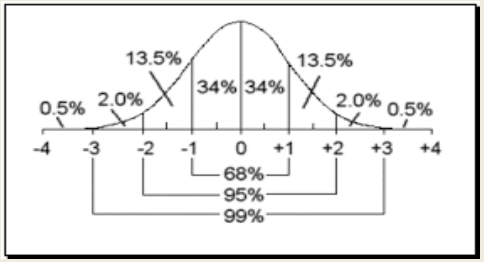
Continuous and Discrete Random Variables

Discrete



Continuous





Review: Sampling Distributions

According to a Gallup poll of 1022 Americans, 57% believe that climate change is due to human activity.

- If many surveys were done of 1022 Americans →calculate sample proportion for each → see how results vary → is this result typical?
- Simulation: I need a MODEL:
  Let's "pretend" the Gallup pol got it exactly right; 57% is the true population proportion

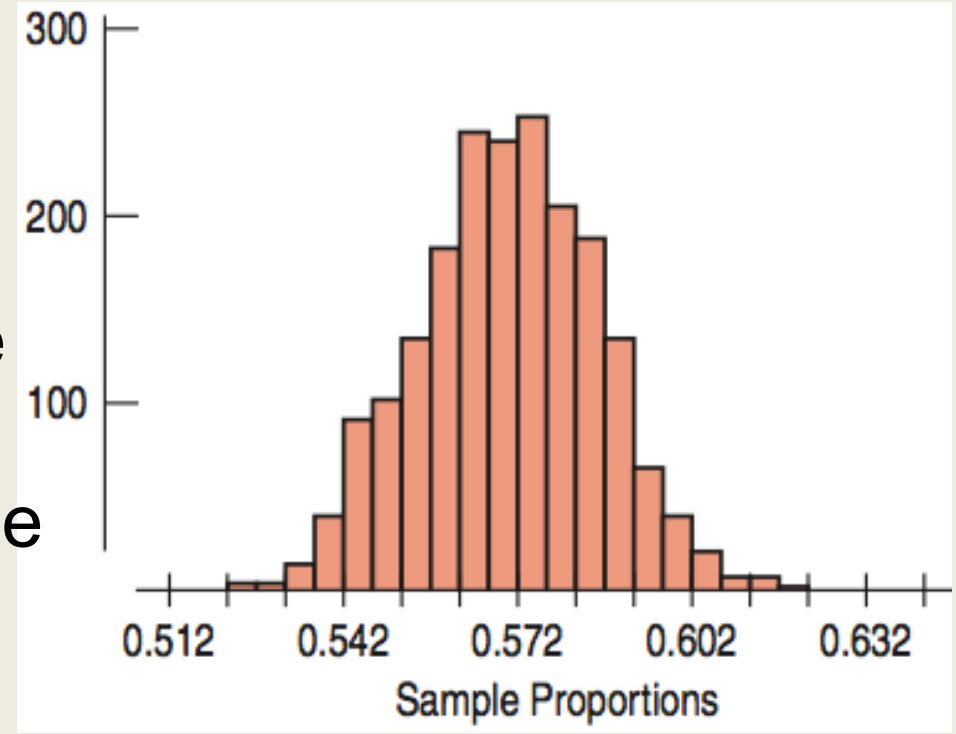
SIMULATION

- Take 10,000 numbers, 5700 are "Yes", 4300 "No"
- Draw a random sample of 1022. Do it 2000 times
- Get 2000 sample proportions. Graph an histogram

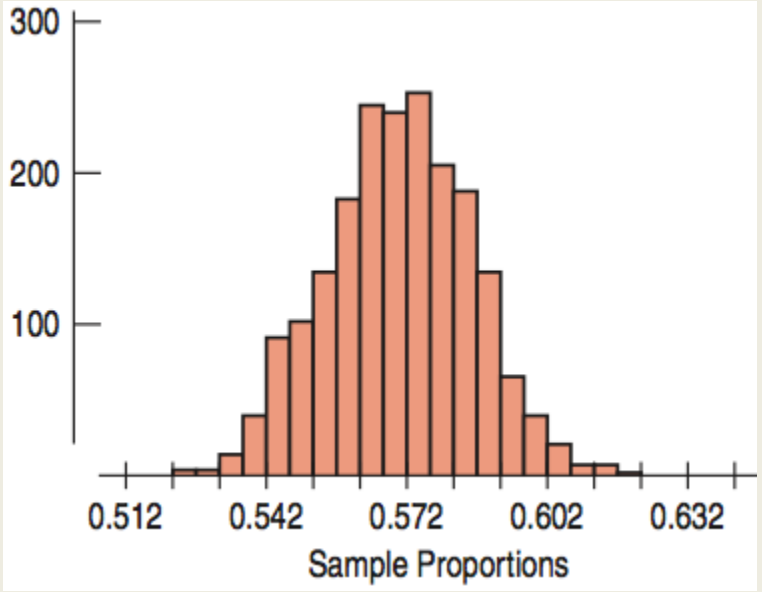
Review: Sampling Distributions

According to a Gallup poll of 1022 Americans, 57% believe that climate change is due to human activity.

- The histogram shows the distribution of a simulation of 2000 sample proportions.
- The distribution of all possible sample proportions from samples with the same sample size is called the sampling distribution.



Sampling Distributions



Sampling Distribution for Proportions

- Symmetric
- Unimodal
- Centered at *p*
- The sampling distribution follows the Normal model.

From One Sample to Many Samples

Distribution of One Sample

- Variable was the answer to the survey question or the result of an experiment.
- Proportion is a fixed value that comes from the one sample.

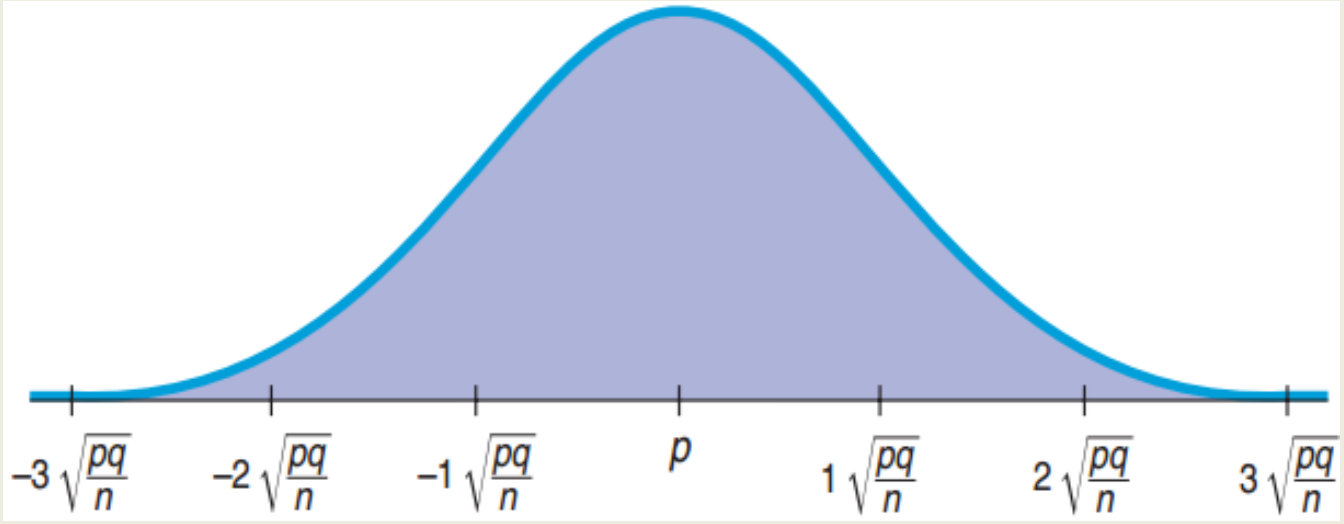
Sampling Distribution

- Variable is the proportion that comes from the entire sample.
- Many proportions that differ from one to another, each coming from a different sample.

Mean and Standard Deviation

Sampling Distribution for Proportions

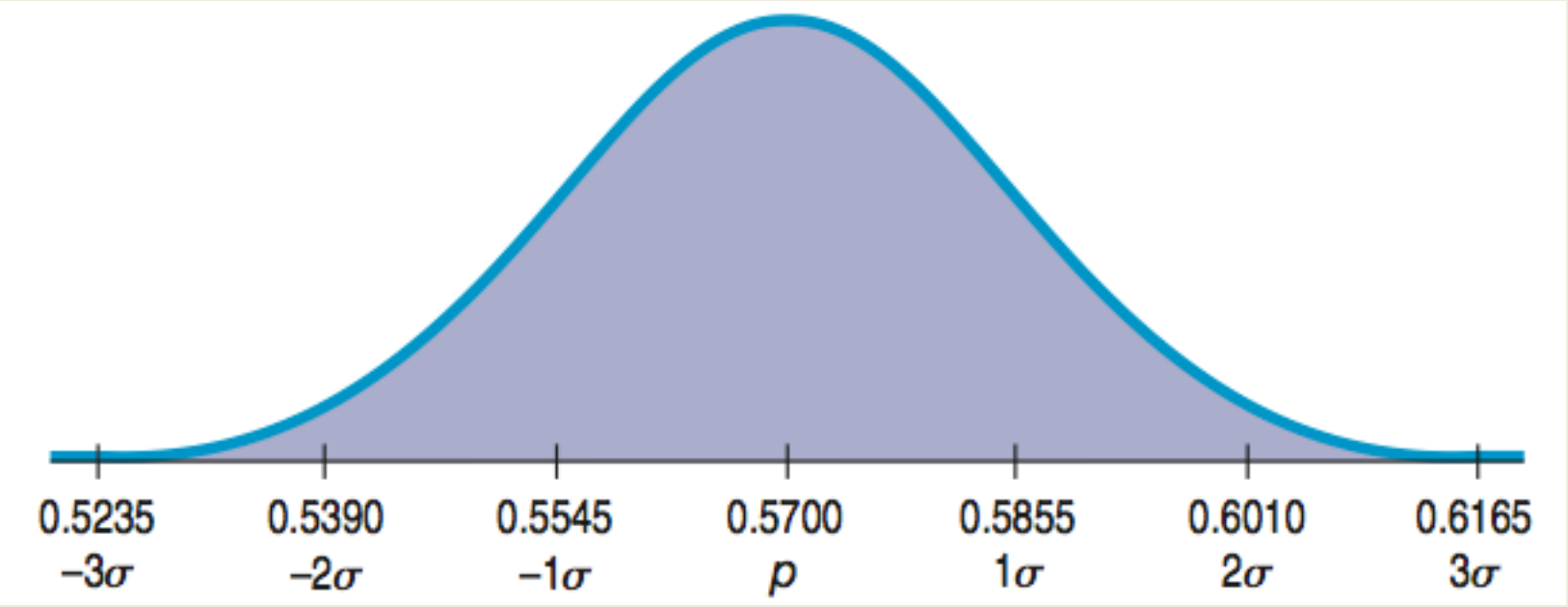
- Mean = p
- $\sigma(\hat{p}) = \dfrac{\sqrt{npq}}{n} = \sqrt{\dfrac{pq}{n}}$
- $N\left(p, \sqrt{\dfrac{pq}{n}}\right)$



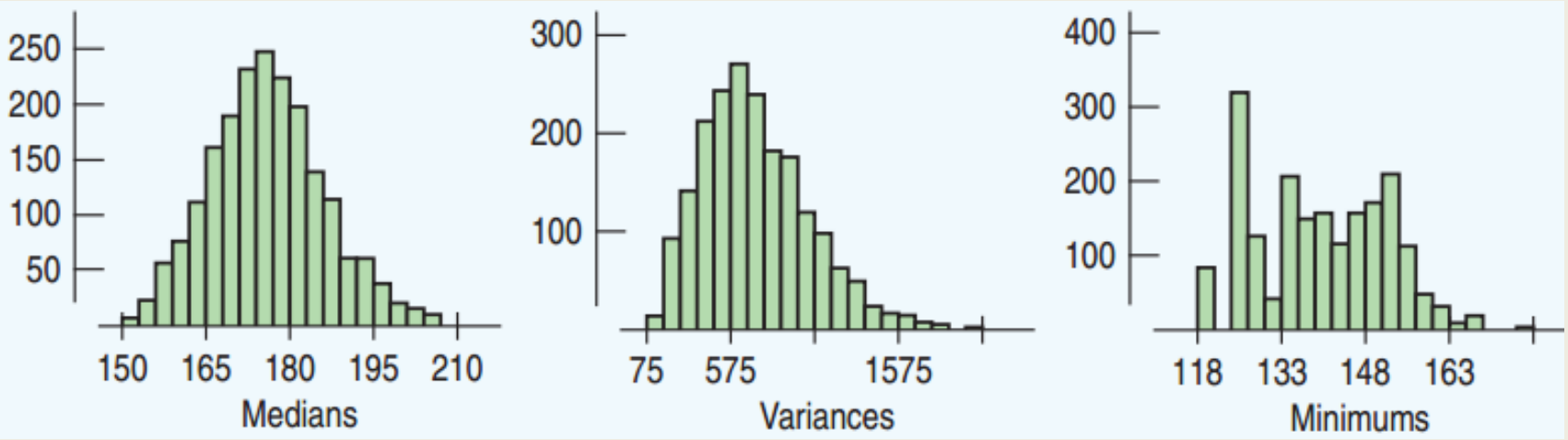
The Normal Model for Climate Change

Population: $p = 0.57$, $n = 1022$. Sampling Distribution:

- Mean = 0.57
- Standard deviation = $SD(\hat{p}) = \sqrt{\frac{(0.57)(0.43)}{1022}} \approx 0.0155$



The Sampling Distribution For Others



- The medians seem to be approximately Normal.
- The variances seem somewhat skewed right.
- The minimums are all over the place.
- In this course, we will focus on the proportions and the means.

The Central Limit Theorem (CLT): sampling distribution for the mean

The mean of a random sample is a random variable whose sampling distribution can be approximated by a Normal model. The larger the sample, the better the approximation will be.

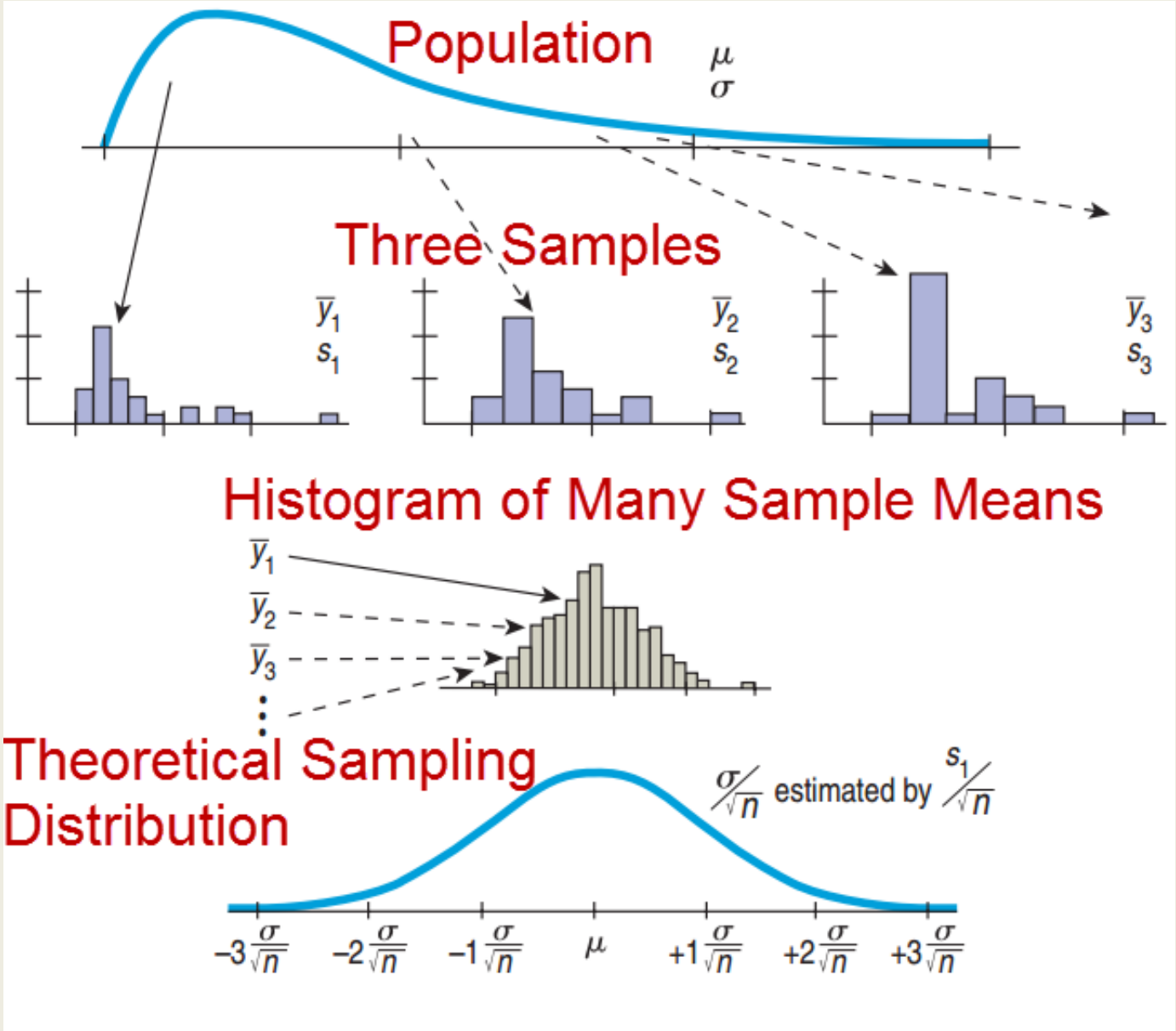
Population Distribution and Sampling Distribution of the Means

Population Distribution		Sampling Distribution for the Means
• Normal	→	Normal (any sample size)
• Uniform	→	Normal (large sample size)
• Bimodal	→	Normal (larger sample size)
• Skewed	→	Normal (larger sample size)

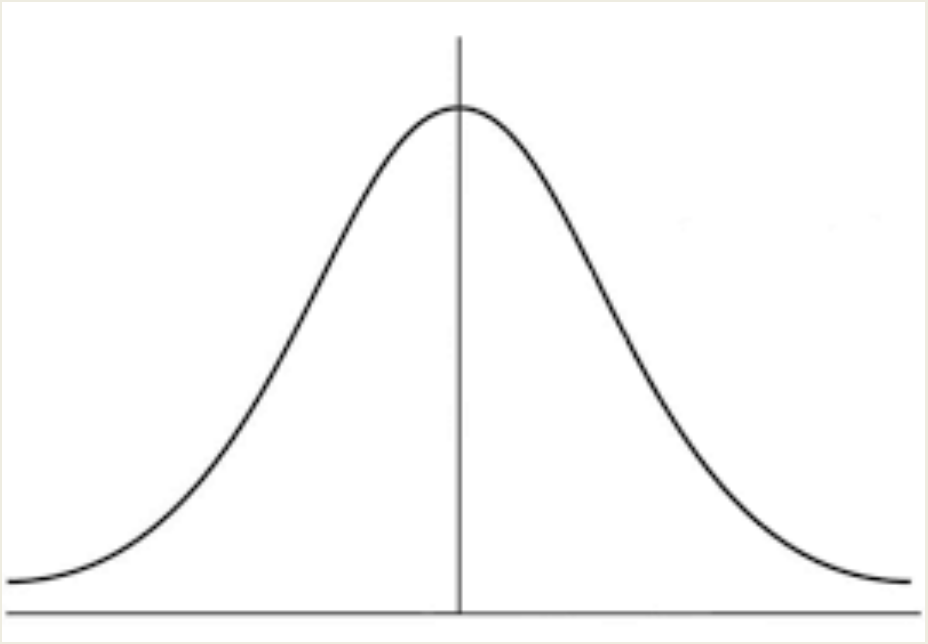
The Sampling Distribution Model for a Mean

When a random sample is drawn from a population with mean μ and standard deviation σ, the sampling distribution has:

- Mean: μ
- Standard Deviation: $\frac{\sigma}{\sqrt{n}}$
- For large sample size, the distribution is approximately normal regardless of the population the random sample comes from.
- The larger the sample size, the closer to Normal.

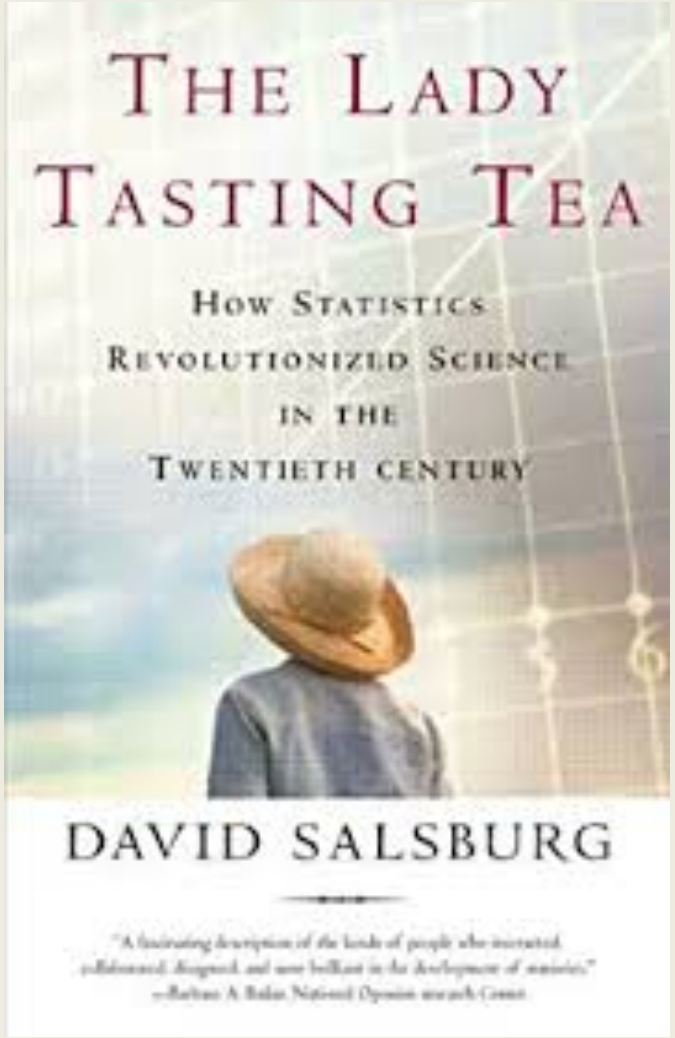

Population
μ
σ
Three Samples
$\bar{y}_1$
s_1
$\bar{y}_2$
s_2
$\bar{y}_3$
s_3
Histogram of Many Sample Means
$\bar{y}_1$
$\bar{y}_2$
$\bar{y}_3$
Theoretical Sampling Distribution
$\frac{\sigma}{\sqrt{n}}$ estimated by $\frac{s_1}{\sqrt{n}}$
$-3\frac{\sigma}{\sqrt{n}}$
$-2\frac{\sigma}{\sqrt{n}}$
$-1\frac{\sigma}{\sqrt{n}}$
μ
$+1\frac{\sigma}{\sqrt{n}}$
$+2\frac{\sigma}{\sqrt{n}}$
$+3\frac{\sigma}{\sqrt{n}}$

The CLT and equality

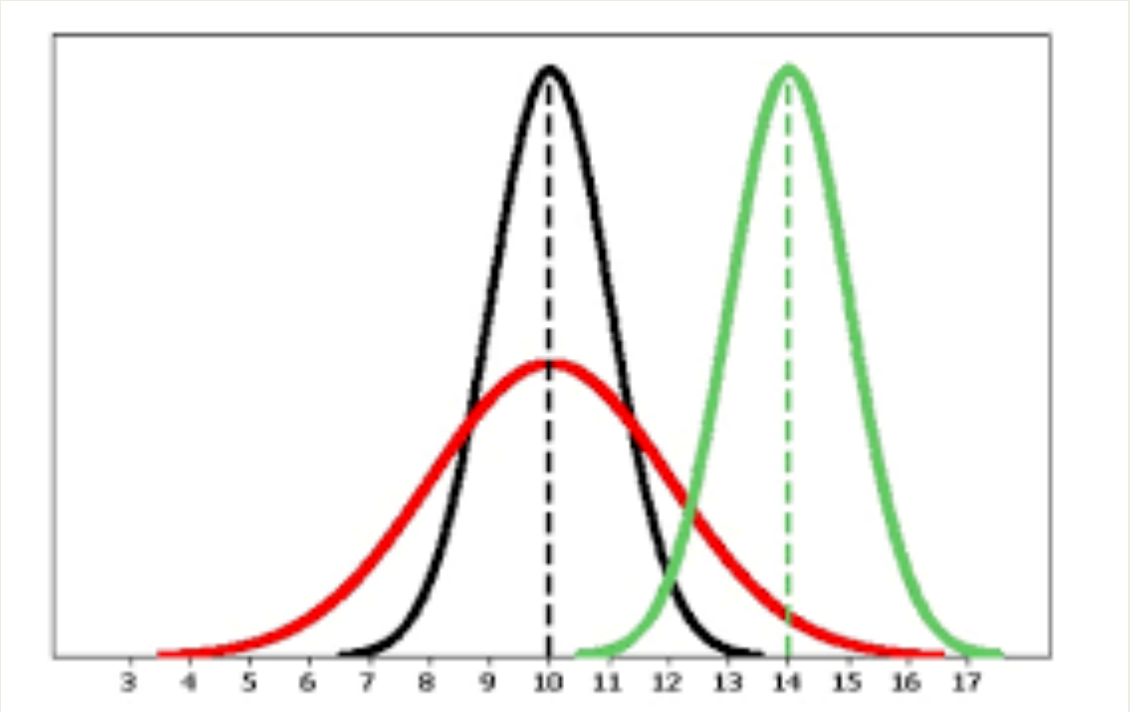


- Average US salaries
- How many hours you worked in a week
- Average artistic level of performers

The CLT and equality



The CLT and equality



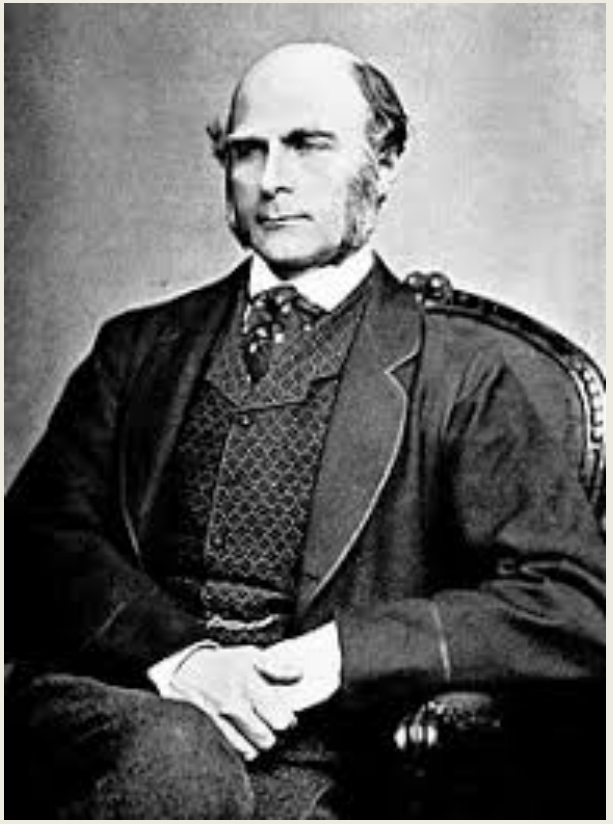
We cannot change the curve, but we can change the shape and position of the curve.
We can change our society.
This is only true for **sample means**.
Statistics cannot tell us anything about an individual.

History of Statistics

Part 2: Galton, Pearson and Eugenics

History of Statistics: Galton

- Sir Francis Galton (1822 – 1911)
- Darwin's half cousin
- Studied variability and hereditary traits
- Problem: how to compare different features? → by their variation
- Idea: features that vary together are "co-related"



History of Statistics: Pearson

- Karl Pearson (1857 – 1936)
- Galton's student
- Studied correlation in the context of variability and hereditary traits
- Came up with the correlation coefficient

Eugenics: Outline

Eugenics: genetic selection of human population
Eugenics in Britain
Eugenics in the United States
Nazi Germany
After WWII
Today?

22

What is eugenics?

Genetic selection of human population by positive or negative means

Positive: selective breeding

Negative: segregation, sterilization, killing

23

Evidence of Eugenics in History

Sparta: phenotypic selection through infanticide

- Weak or undesirable babies were abandoned

Rome:

- Fourth Table of Roman stated that deformed children would be put to death

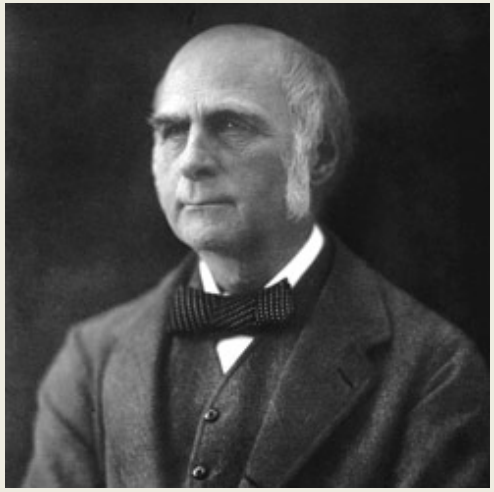
24

Historical Developments

- Christianity
- Statistics
- Evolutionary theory

25

Galton: a second look



Conied the word eugenics

eu “well” + *genes* “born”

“science of improving stock”

“I object to pretensions of natural equality.”

26

Galton: a second look

Inspired by Darwin's *Origin of Species*
He believed societies in favor of spending resources on the weak were in opposition to natural selection
Defined eugenics as: "the science which deals with all influences that improve the inborn qualities of a race; also with those that develop them to the utmost advantage."
His approach to eugenics was based upon social statistics
Initially eugenics was seen as a *positive* movement
Galton was very disappointed when he discovered the regression to the mean: very tall parents generate shorter children

Pearson: a second look

Pearson was also a proponent of eugenics.
Main concern of the first eugenists: the perceived intelligence factors considered to be correlated with the social class.

"*The problem of alien immigration into Great Britain, illustrated by an examination of Russian and Polish Jewish children*"

"…the majority of children examined by Pearson and Moul were recent immigrants, with many actually born abroad. In Pearson's work these children were compared to non-immigrant Gentile children, often without comment as to differing social and economic class status."

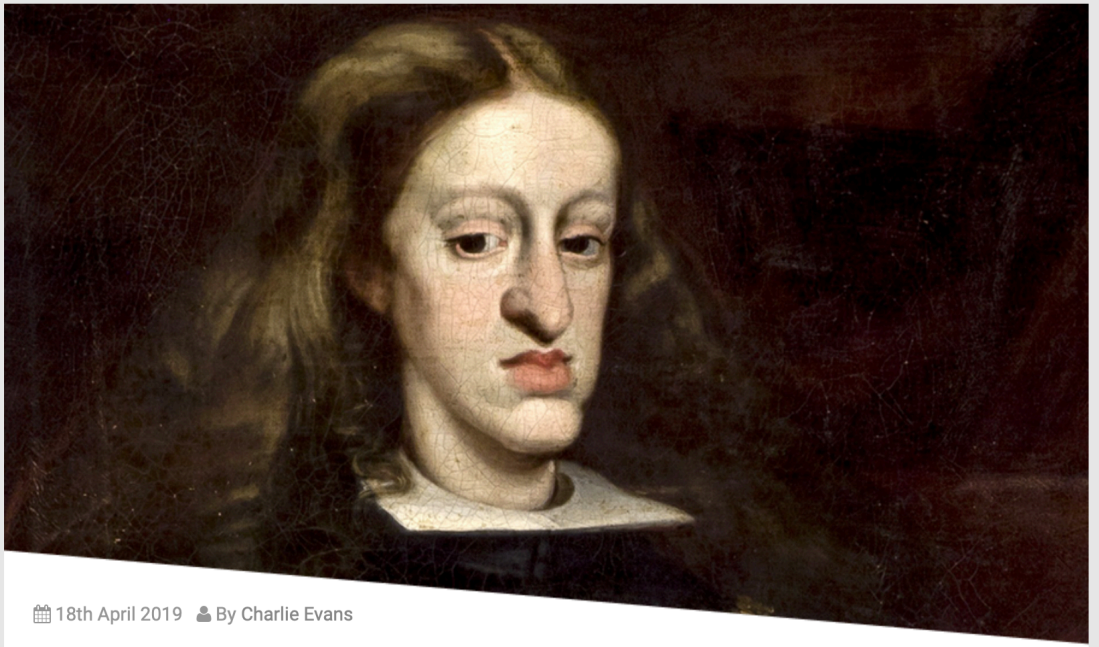
28

Discounting evidence

Inbreeding had brought suffering and havoc in the royal dynasties

Variability is a resource!

https://bit.ly/39iD8oL



18th April 2019 By Charlie Evans

The Dangers Of Royal Inbreeding

From the Spanish Habsburgs to Queen Victoria's grandchildren, how centuries of inbreeding and genetic mutation led Europe's royal families to ruin

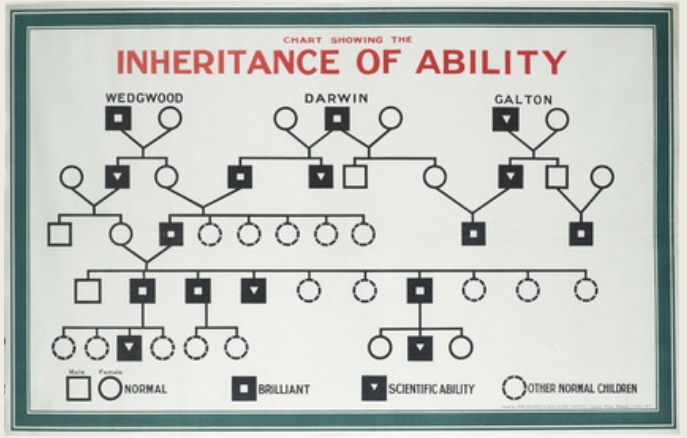
He endured violent convulsions and hallucinations, and his pronounced underbite and engorged tongue meant he was unable to close his teeth together. The malformed jaw made eating and talking nearly impossible, and he suffered uncontrollable
spells of diarrhoea and vomiting.

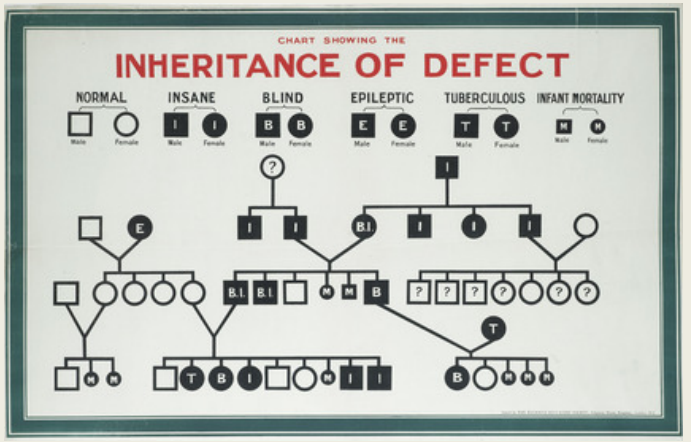
It was rumoured that he was bewitched; his painful and disfigured body the result of witchcraft, a curse, or the ritual consummation of the brains of criminals that he had devoured in hot chocolate drinks. But the truth was just as unsavoury and much closer to home. Charles II of Spain's birth defects were the result of the accumulation of over two centuries of inbreeding.

29

A Larger Cultural Background

Victorian cult of *measurement* (Galton)
Emerging sciences of *heredity* and *statistics*
Longer-term cultures of breeding
Modernist attractions of science of eugenics
Science or pseudo-science?

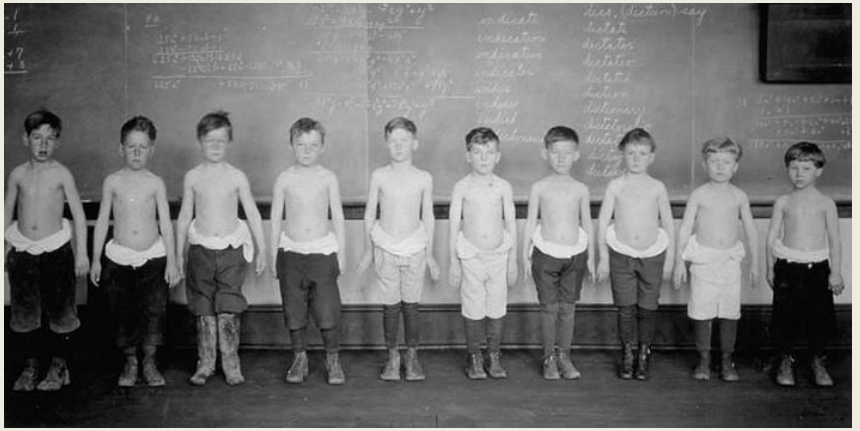




30

New ways of seeing /ranking population

Population problems and differences exposed by rise of mass education, and growing scale of asylums, prisons, workhouses
Emergence of new tools such as the social survey and the psychological test to rank individuals



The contrasting intellectual status of the white versus the negro constituents of the draft appear from table 3. Few residents of the United States probably would have anticipated so great a difference. That the American negro is 90 per cent. illiterate only in part accounts for his inferior intellectual status.

TABLE 3.

INTELLIGENCE OF AMERICAN NEGRO VERSUS WHITE.

Race.	*Number of Cases*	*Percentage making grade*						
		A	B	C+	C	C—	D	D—
Whites	93,973	4·1	8·0	15·0	25·0	23·8	17·1	7·0
Negroes	18,891	0·1	0·6	2·0	5·7	12·9	29·7	49·0
Northern negroes ..	4,705	0·7	2·7	7·2	18·0	25·8	31·2	14·4
Southern negroes ..	6,846	0·1	0·2	0·7	3·4	9·6	29·2	57·0

31

Decline in fertility

From late 19th century a fall in fertility in western nations
Concern in Britain about decline being most marked in middle-classes and among professionals
In France, fear of overall population decline by early 20th century (encourages positive forms of eugenics)
Objections to birth control and to women's groups, birth control advocates, and even sexual liberals

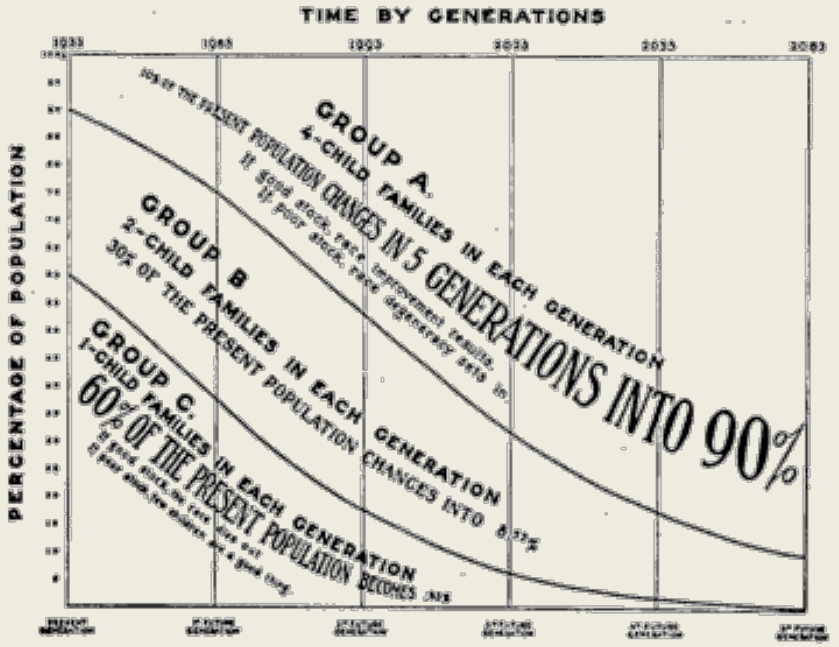


FIG. 65. — Effect of Differential Birth Rates on Future Generations. The "turn over" in population produced by differences in birth rates. It is easily seen from this chart that a group which adopts the one-child family ideal will, five generations hence, have a very small representation in the population, as compared with a group having an average of four children per family. This chart presents in summary form one of the major problems of eugenics: to encourage the better endowed part of the population to avoid a suicidal reduction in birth rate. Courtesy, *Journal of Heredity*.

32

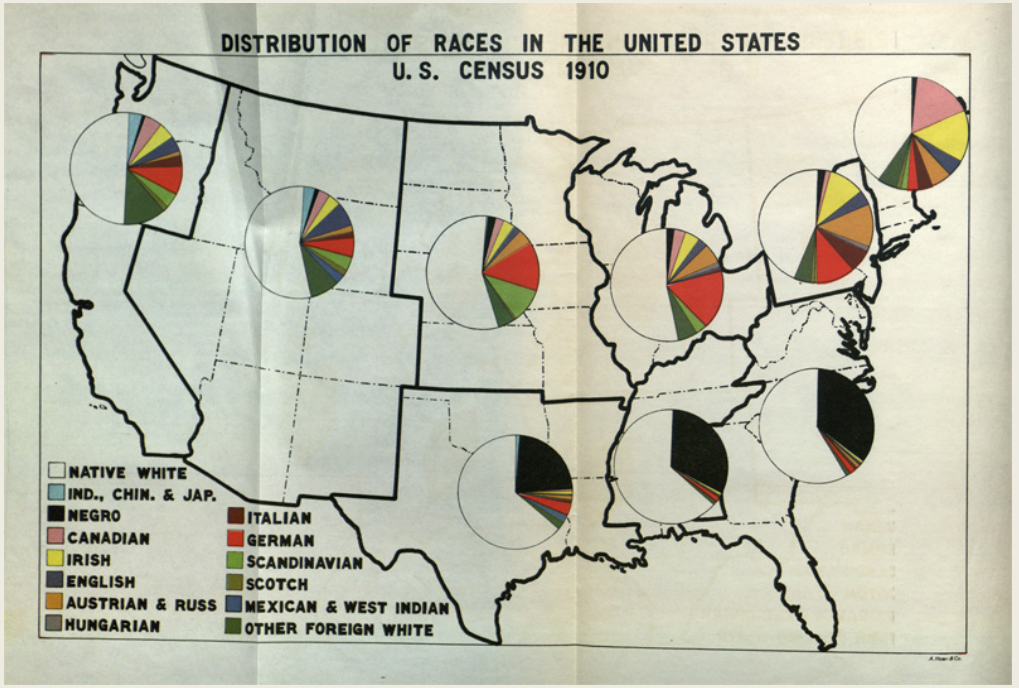
Other voices

- Alfred Russel Wallace (1823-1913)
-
  "Leave heredity alone until we have made the environment of every child from conception to death the best possible for its full and free development, and then we can begin to think about the influences of heredity, which may be small."
-

33

U.S.: Nation building era

Heightened concern about international competition and ‘national efficiency’
Era of nation building and interest in defining the nation (e.g. post WWI)
Immigrant nations concerned about racial mix



34

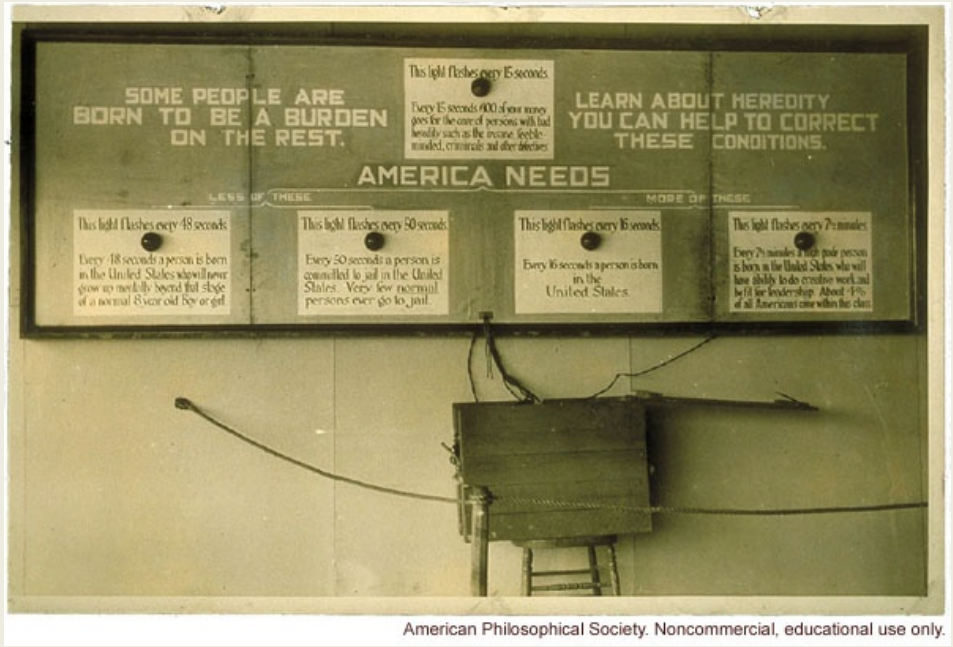
United States

Alexander Graham Bell recommended deaf individuals not be allowed to marry

Thirty states adopted legislation to perform forced sterilization on those deemed “mentally unfit.”

35

The burden of the “feebleminded”

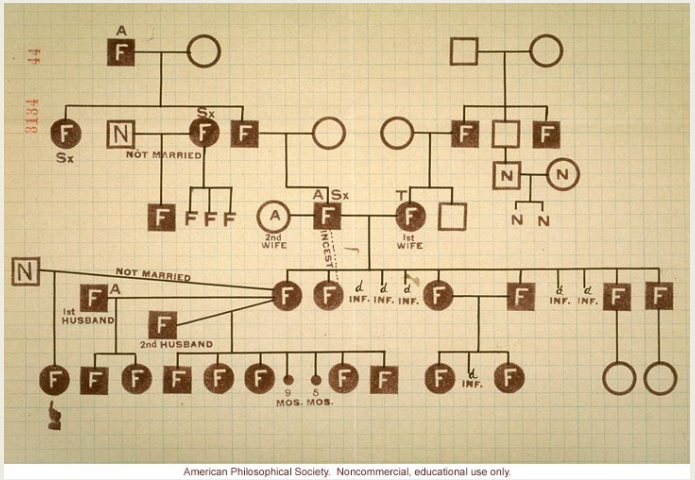


“It is a reproach to our intelligence that we as a people should have to support about half a million insane, feebleminded, epileptic, blind and deaf; 80,000 prisoners and 100,000 paupers at a cost of over 100 million dollars per year.”

-Charles Davenport, founder of the Eugenics Record Office, 1910

36

Compulsory segregation and sterilization



Eugenics Record Office pedigree showing transmission of feeblemindedness over several generations. Such evidence supported state-mandated, coerced sterilization of at least 62,000 Americans in over 30 states, beginning in 1907.

37

Better Babies at the Puyallup Fair, 1910s

38

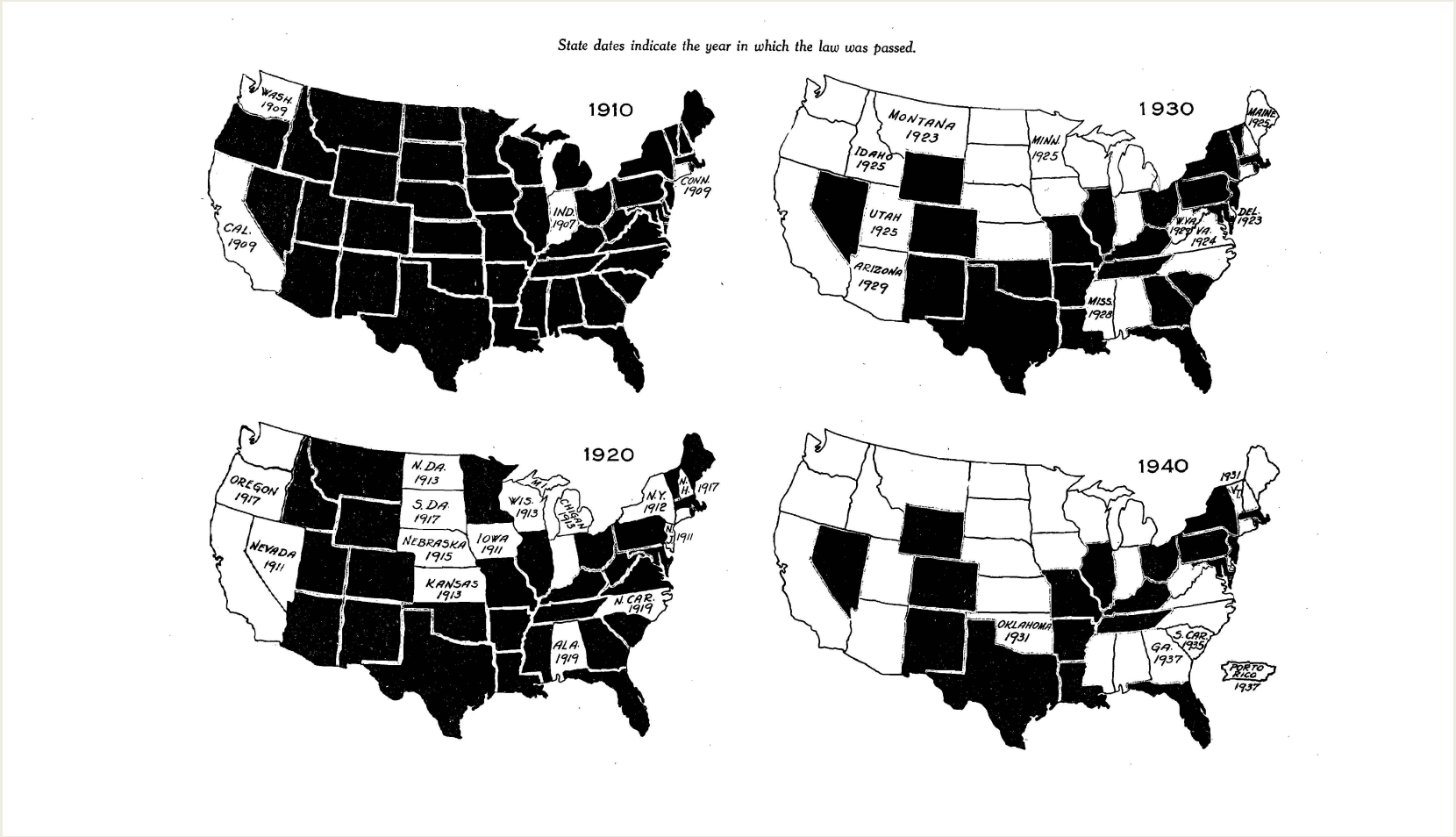
Positive eugenics: Fitter Families contests and educational exhibits

American Philosophical Society. Noncommercial, educational use

American Philosophical Society. Noncommercial, educational use only.

39

Sterilization Laws in the US



40

"1"

Not ordered -

At discharged

REPORT OF SUPERINTENDENT OF

Northern State Hospital

In the matter of [redacted].

To the Institutional Board of Health of Washington:

I hereby report to you [redacted] a female (sex) person, who is feeble-minded, insane, epileptic, an habitual criminal, a moral degenerate or sexual pervert, to-wit: Insane (state which), and who, by reason thereof, is potential to producing offspring who, because of inheritance of inferior or anti-social traits, will probably become a social menace or ward of the state. The said [redacted] is now an inmate of the Northern State Hospital.

Names and addresses of legal guardian, relatives or custodial guardian:

Husband: [redacted], Bellingham, Wash.

Sister: [redacted]; [redacted]

Bellingham, Wash.

REMARKS:

Dated this 1st day of June, 1936.

Superintendent

Northern State Hospital
Institution

All cases to be acted upon by the Board must be reported to the Board quarterly on the first day of January, April, July, or October.

The inmate traits, the mental and physical condition and personal record are as follows:

Normal Disposition, bright and cheerful. Bright in school. After school worked as stenographer, capable. Pleasing personality. Last March she became excited, screamed, was irrational. Religious trend of thought. Became depressed, attempted suicide, and wanted to injure her children. Was disoriented. Memory poor. No hallucinations. Judgment and insight poor. Mal-nourished, asthenic type.

It is recommended that [redacted] Name of Patient

be Sterilized Operation

The object or purpose of operation is to prevent pregnancy.

June Date 1st 1936

Superintendent

Northern State Hospital
Institution

Section 3, Chapter 53, Laws of 1921. PURPOSE AND OBJECTS TO BE SOUGHT - Said investigation, findings and orders of said board shall be made with the purpose in view of securing a betterment of the physical, mental, neural or psychic condition of the person, or to protect society from the acts of such person, or from the menace of procreation by such person, and not in any manner as a punitive measure.

- 2 -

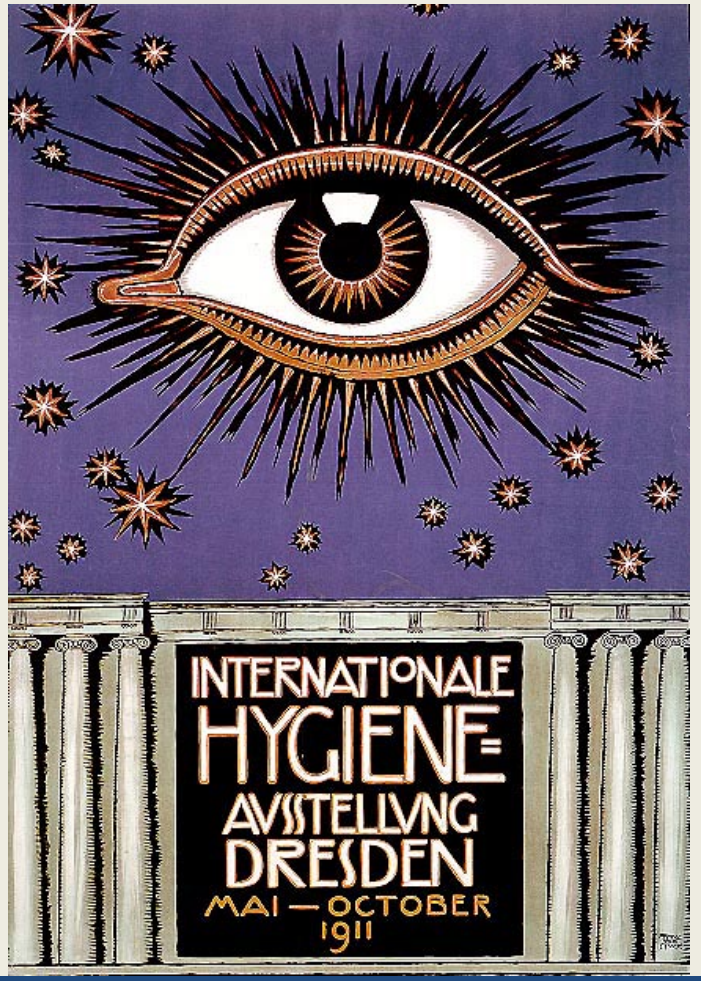
Why Nazi Germany stands out

Importance of ideology of national fitness and purity to politics and culture: ‘the ‘racial state’
Leader in psychiatric and genetic science
Opportunities from a sympathetic state
Degree of economic problems via depression and mobilisation of economy in WWII
Scale of sterilisation policy (375,000)
Use of ‘euthanasia’ for mentally handicapped (Action T4 – 70,000)

42

Yet German eugenics has roots before Nazi era

Sterilisation Law introduced before Nazis come to power
Follows example of USA
Eugenics and an interest in 'race hygiene' well established (1st society 1905)
Germany already a leader in eugenic sciences of psychiatry and genetics, e.g. via Kaiser Wilhelm Institute (and has financial support from Rockefeller Foundation into 1930s)
Deaths of mentally ill in German asylums in First World War (140,000)



43

	U.S. (and Canada)	Nazi Germany
Period	1907-late 1970s	1934-1945
Total	70,000	375,000 (= approx. 1% of adult population of childbearing age) + 70,000 killed
Disproportionately targeted	Disabled; also disenfranchised, poor, women, minorities	Same
Law	State	Federal
Compulsion	Typically formally voluntary (required formal consent)	Compulsory
Reach	Mostly only institutionalized (in a few cases, "extra mural")	Everyone
Text of the law	Varied across states; basic elements similar	Similar to H. Laughlin's model sterilization law
Adjudication	Typically eugenics boards; proceedings public; victims could challenge in civil court	Special "hereditary health courts"; non-public; appeal only to superior health courts
Denunciation by general public (information to commence sterilization proceedings)	No	Yes
Authorities' access to medical and other records of kin groups	Could be extensive	Extensive

44

The post-war reckoning

At post-war Nazi trials the question of medical experiments is subject for prosecution, but eugenics itself in fact attracts little attention
Not until 1970s and 1980s and a new generation of historians does it come into focus

45

After the second World War

Eugenics is scientifically and culturally discreted
It disappears from Universities and the public discourse
But sterilizations continue in US and Scandinavia
(Sweden) till the 70s

46

	% female victims	% mentally ill	% intellectually disabled	total	victims per year/ 100,000 pop. in peak period	sterilization period (length of time)
National	61%	44%	52%	Est. 70,000+		1907-earl.1980s (75 years)
CA	49%	58%	37%	**20,000**	**13**	**1909-60s (55 years)**
MN	78%	18%	**82%**	2,300	5	1925-earl.1960s (35 years)
VA	61%	49%	48%	7,300	**13**	1924-1979s (50 years)
IA	71%	44%	50%	1,910	6	1915-earl.1960s (35 years)
GA	55%	**77%**	22%	3,200	9	1937-1963 (25 years)
NC	**83%**	25%	70%	6,300	7	1929-1974 (45 years)

	% female victims	% mentally ill	% intellectually disabled
NC	**83%**	25%	70%

NC targeted “black welfare queens” in 1950s and 1960s; sterilization law had “extra-mural” component and allowed sterilization of populations not in state institutions

	% female victims	% mentally ill	% intellectually disabled
MN	78%	18%	**82%**

MN: strength of social progressivism with a focus on ‘helping’ intellectually disabled females; sterilization of insane required at least 6 months of continuous institutionalization and consent by individual and next of kin

48

Family Studies in Order of Publication

1. Dugdale, *"The Jukes,"* 1877
   Dugdale, "Hereditary Pauperism," 1877•
2. McCulloch,[7] "The Tribe of Ishmael," 1888•
3. Blackmar, "The Smoky Pilgrims," 1897•
4. Winship, *Jukes-Edwards*, 1900
5. G. Davenport, "Hereditary Crime," 1907•
6. Goddard, *The Kallikak Family*, 1912
   Kite, "Two Brothers," 1912•
7. Danielson and C. Davenport, *The Hill Folk*, 1912•
8. Estabrook and C. Davenport, *The Nam Family*, 1912
9. Kite, "The 'Pineys,'" 1913•
10. Kostir, *The Family of Sam Sixty*, 1916•
11. Finlayson, *The Dack Family*, 1916•
12. Estabrook, *The Jukes in 1915*, 1916
13. Sessions, *The Feeble-Minded in a Rural County of Ohio*, 1918•
14. Rogers and Merrill, *Dwellers in the Vale of Siddem*, 1919•
15. Estabrook and McDougle, *Mongrel Virginians*, 1926

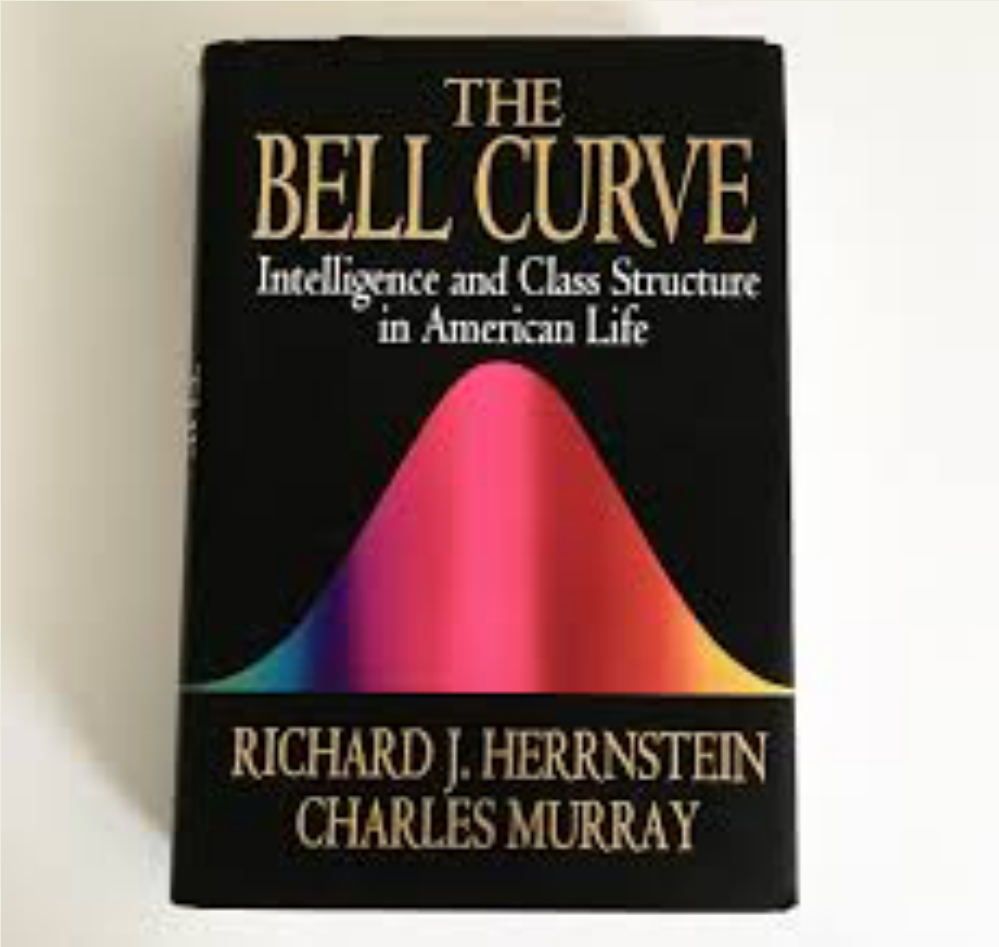
49

Eugenics as a Pseudoscience

Made-up categories
One encompassing explanation (interpretation Vs explanation)
Certainty
Appeal to authority
Propaganda aimed to the public

50

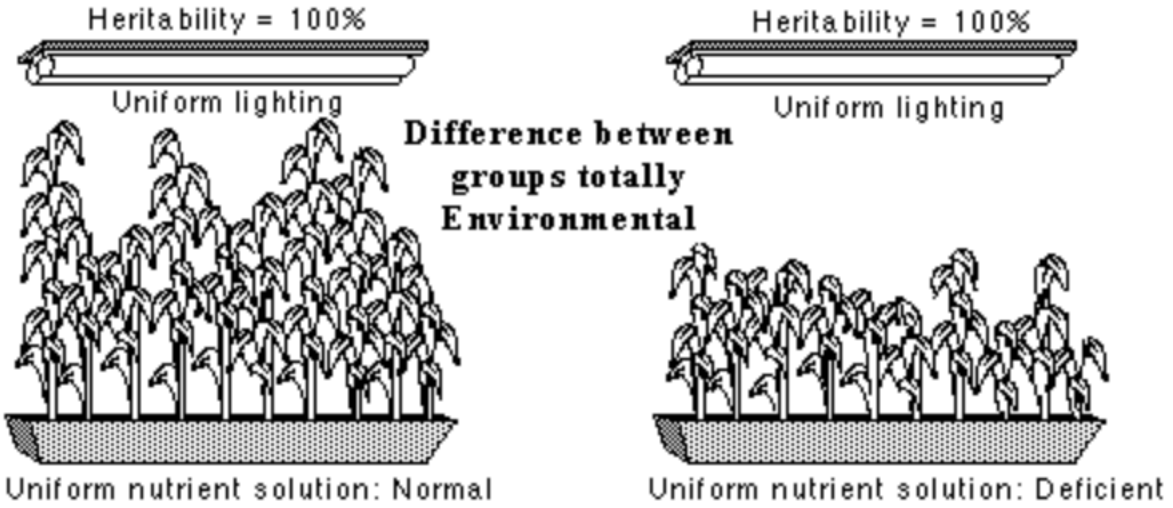
Today?


THE
BELL CURVE
Intelligence and Class Structure
in American Life
RICHARD J. HERRNSTEIN
CHARLES MURRAY

51

Today?

Ned Block (NYU)



https://bit.ly/3hrzUCh

52

Today?

Wallace:

“It is unmitigated humbug to talk about hereditary class distinctions being rooted in Nature. An individual is, of course, a product of nature and nurture, but it is one-tenth the former and nine-tenths the latter"

53

Today?

Abortion is legal for serious fetal anomaly
To prevent life-long suffering for the affected
Euthanasia (for the terminally ill) is illegal but, to some extent, tolerated
COVID-19
Where do we draw a line?

54

*

Midterm

Midterm Thursday during class time
Exam: 40 questions total + formula sheet + z-table
A few questions are worth more than one point
Problems: boxplots, contingency table, z-scores, regression, probability
Best way to study: **review HW and class exercises**
Review lectures slides
Final grades will be curved

Formula sheet

$Range = Max - Min$

$IQR = Q3 - Q1$

Outlier Rule-of-Thumb: $y < Q1 - 1.5 \times IQR$ or $y > Q3 + 1.5 \times IQR$

$$\bar{y} = \frac{\Sigma y}{n}$$

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

$$z = \frac{y - \bar{y}}{s} \text{ (data based)}$$

$$r = \frac{\Sigma z_x z_y}{n - 1}$$

$$\hat{y} = b_0 + b_1 x \qquad \text{where } b_1 = r\frac{s_y}{s_x} \text{ and } b_0 = \bar{y} - b_1\bar{x}$$

$$P(\mathbf{A}) = 1 - P(\mathbf{A}^C)$$

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

$$P(\mathbf{A} \textit{ and } \mathbf{B}) = P(\mathbf{A}) \times P(\mathbf{B} \mid \mathbf{A})$$

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

If **A** and **B** are independent, $P(\mathbf{B} \mid \mathbf{A}) = P(\mathbf{B})$

Chapter 1

Introduction

Why do we need statistics?

Statistics helps us make sense of the data and how the data vary.

Statistics is about *variation.*

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 an unifying force within Psychology.

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.

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.

Variables

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

- Quantitative Variable: Contains measured numerical values with measurement units

- Identifier Variable: A variable that is used to uniquely identify the individual. It does not describe the individual.

- Ordinal Variable: A variable that reports order without natural units

Chapter 2

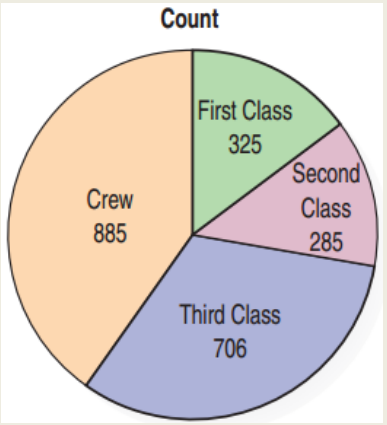
Displaying and Describing Categorical Data

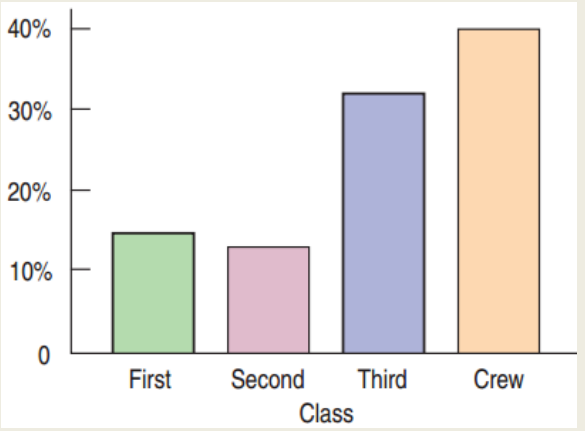
How to Display Categorical Data

- How to display and describe categorical data
- Frequency tables
- Bar Charts
- Pie Charts

Class	Count
First	325
Second	285
Third	706
Crew	885

Class	%
First	14.77
Second	12.95
Third	32.08
Crew	40.21





Comparing Categorical Data

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

- How to compare categorical data:
- Contingency tables
- Marginal distribution

Survival		Class: First	Second	Third	Crew	Total
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%

- Conditional distribution:
  - percent of one variable satisfying the conditions of another
  - can be organized by column or by row

Independence

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

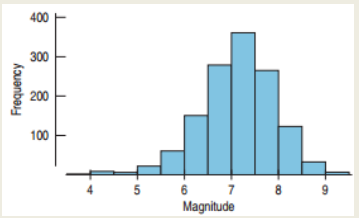
- An association that holds for all of several groups can reverse direction when the data are combined to form a single group. This reversal is called Simpson's paradox.

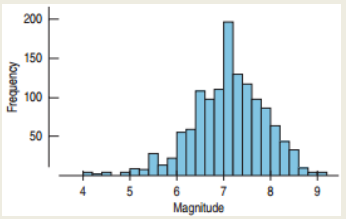
Chapter 3

Displaying and Summarizing Quantitative Data

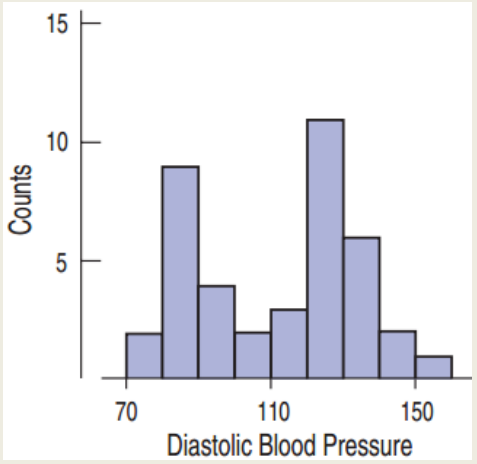
Display of Quantitative Data: Histograms

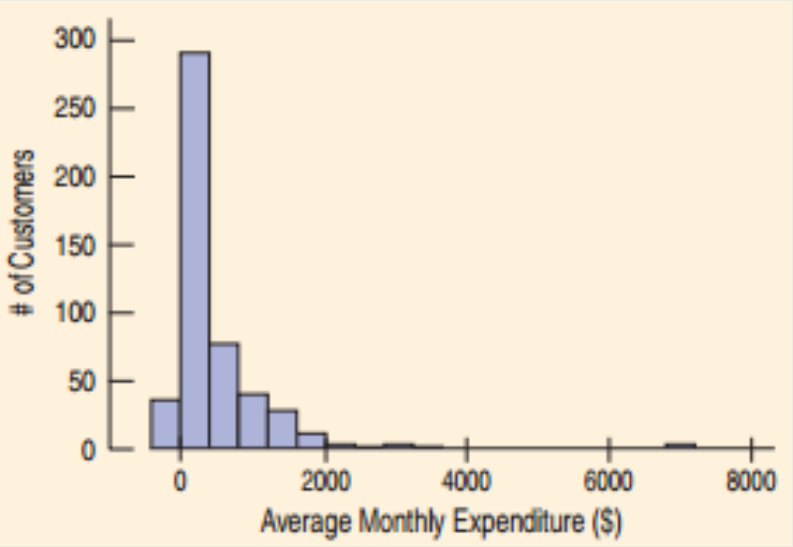
- How to display quantitative data: histograms
  - bin width





  - shape: mode, symmetry, skew
  - outliers

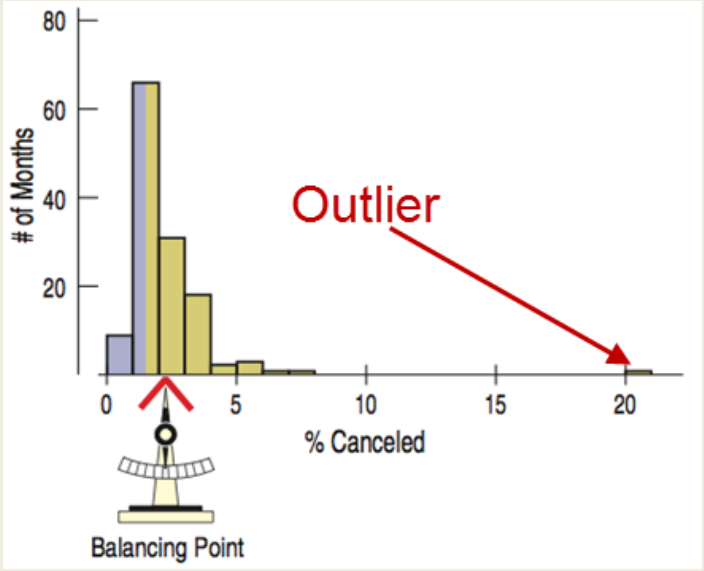




Display of Quantitative Data: Histograms

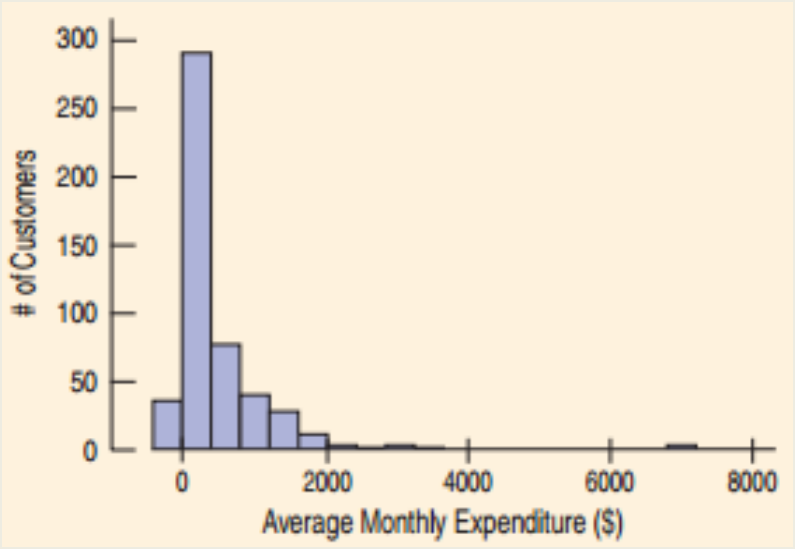
- How to display quantitative data: histograms
  - CENTER: median, mean
  - Median: the center of data values
  - not sensible to outliers
  - Mean: average
  - sensible to outliers

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



Outliers

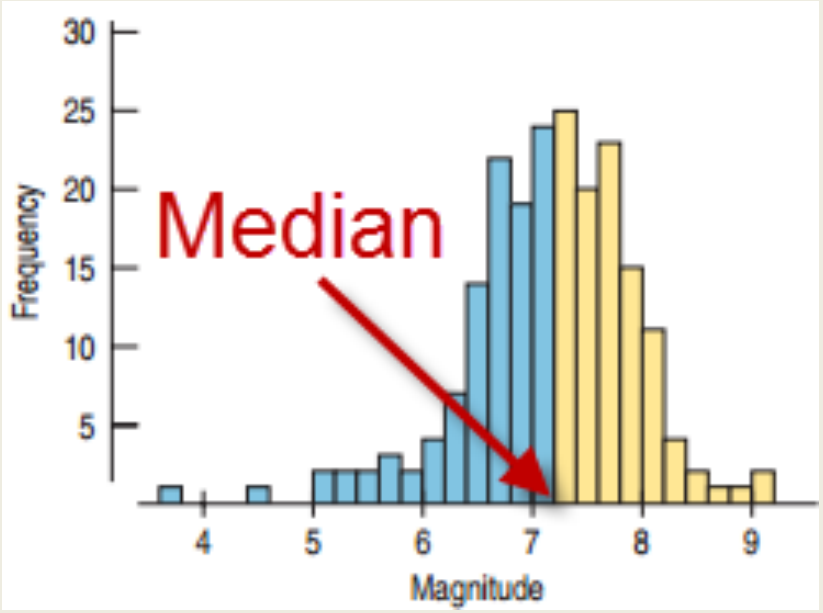
- An Outlier is a data value that is far above or far below the rest of the data values.
- An outlier is sometimes just an error in the data collection.
- An outlier can also be the most important data value.
  - Income of a CEO
  - Temperature of a person with a high fever



Center: the Median

- Median: The center of the data values

- Half of the data values are to the left of the median and half are to the right of the median.



- For symmetric distributions, the median is directly in the middle.

Calculating the Median: Odd Sample Size

- First order the numbers.
- If there are an odd number of numbers, n, the median is at position $\frac{n+1}{2}$.
- Find the median of the numbers: 2, 4, 5, 6, 7, 9, 9.
- $\frac{n+1}{2} = \frac{7+1}{2} = 4$

- The median is the fourth number: 6
- Note that there are 3 numbers to the left of 6 and 3 to the right.

Calculating the Median: Even Sample Size

- First order the numbers.
- If there are an even number of numbers, n, the median is the average of the two middle numbers: $\frac{n}{2}, \frac{n}{2}+1$.
- Find the median of the numbers: 2, 2, 4, 6, 7, 8.
- $\frac{n}{2} = \frac{6}{2} = 3$

- The median is the average of the third and the fourth numbers: $\text{Median} = \frac{4+6}{2} = 5$

The Center of Symmetric Distributions: the Mean

- The Mean is what most people think of as the average.
- Add up all the numbers and divide by the number of numbers.

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

- Recall that $\sum$ means "Add them all."

Mean Vs. Median

- For symmetric distributions, the mean and the median are equal.
  - The balancing point is at the center.

- The tail “pulls” the mean towards it more than it does to the median.

- The mean is more sensitive to outliers than the median.

Spread

- SPREAD: variance, standard deviation
- The variance is a measure of how far the data is spread out from the mean.
- The standard deviation is the square root of the variance
- The sd expresses the average distance form the mean

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

5-Number Summary

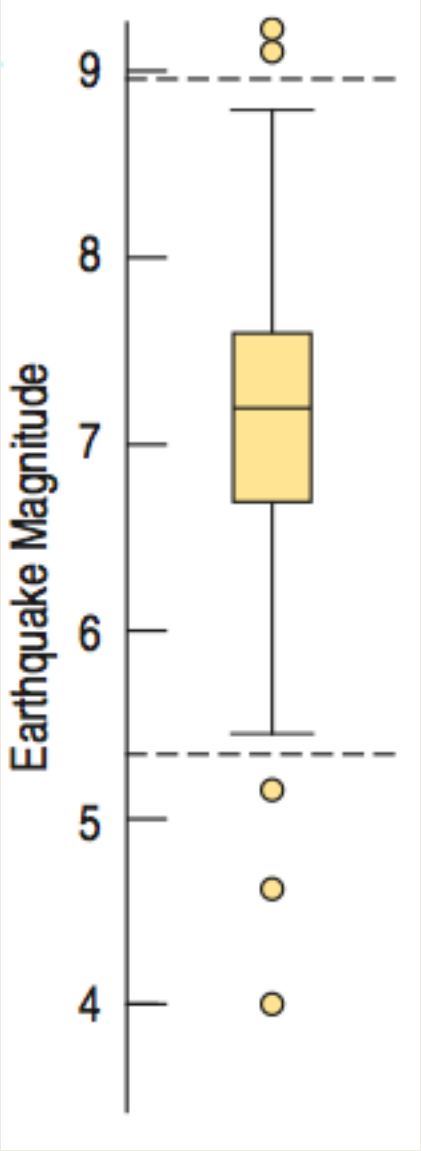
- The 5-Number Summary provides a numerical description of the data. It consists of
  - Minimum
  - First Quartile (Q1)
  - Median
  - Third Quartile (Q3)
  - Maximum

- The list to the right shows the 5-Number Summary for the tsunami data.

Max	9.1
Q3	7.6
Median	7.2
Q1	6.7
Min	4.0

Boxplots

- A Boxplot is a chart that displays the 5-Point Summary and the outliers.
- The Box shows the Interquartile Range.
- The dashed lines are called fences, outside the fences lie the outliers.
- Above and below the box are the whiskers that display the most extreme data values within the fences.
- The line inside the box shows the median.



Finding the Fences

- The lower fence is defined by

  $\text{Lower Fence} = Q1 - 1.5 \times IQR$

- The upper fence is defined by

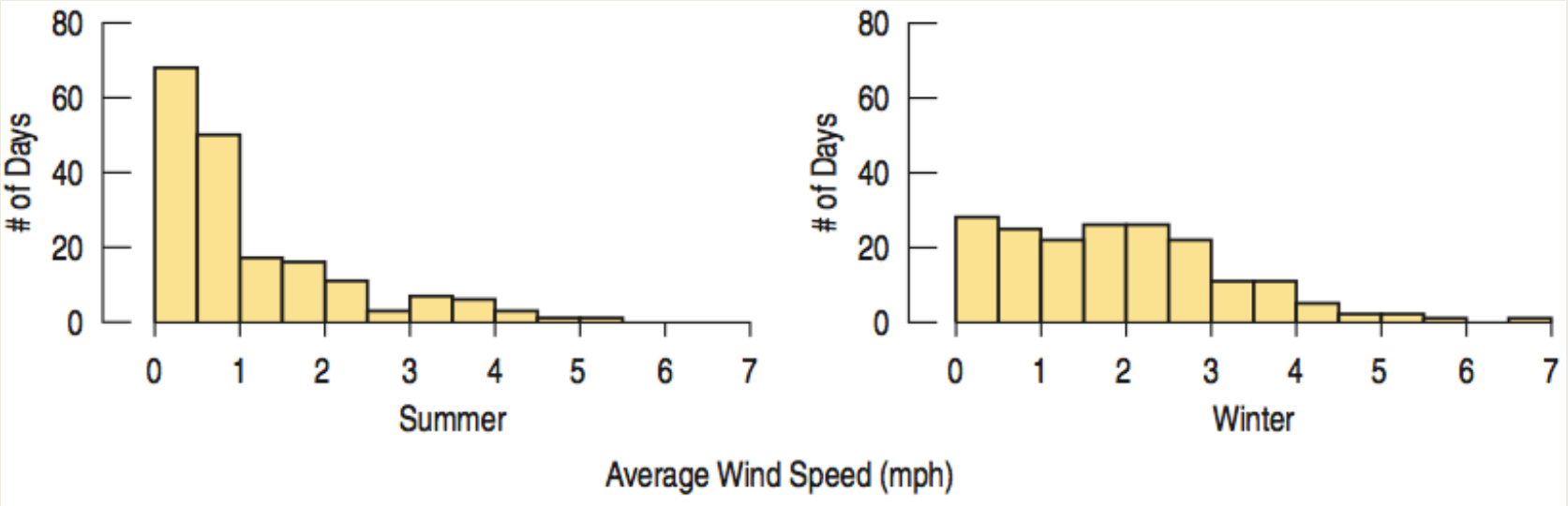
  $\text{Upper Fence} = Q3 + 1.5 \times IQR$

Chapter 4

Understanding and Comparing Distributions

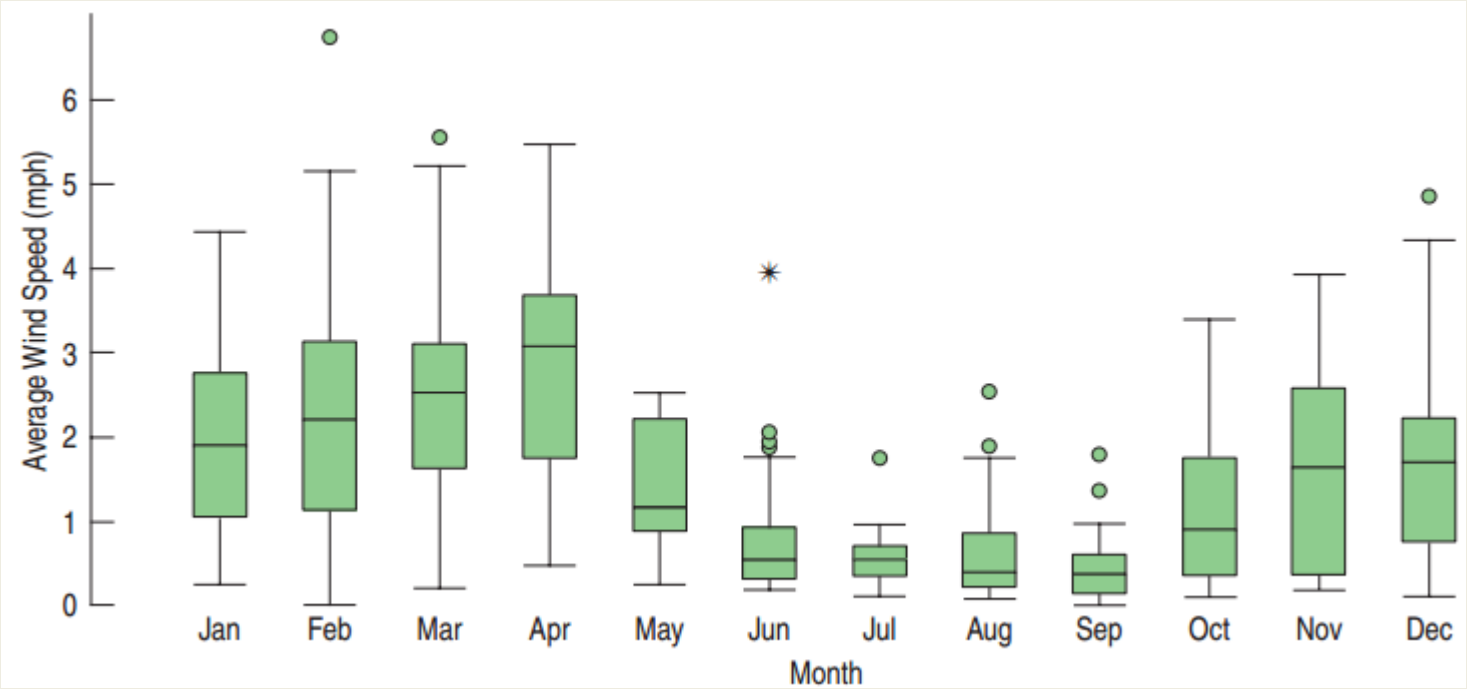
How to compare quantitative data

- How to compare quantitative data: histograms and boxplots



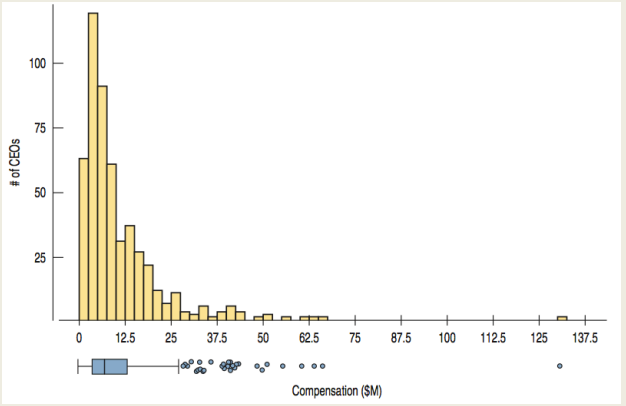
How to compare quantitative data

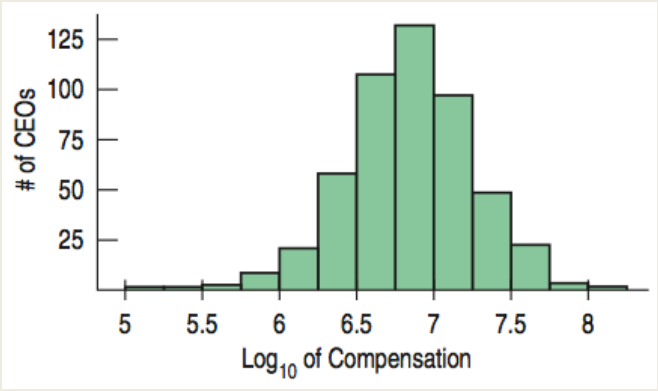
- How to compare quantitative data: histograms and boxplots

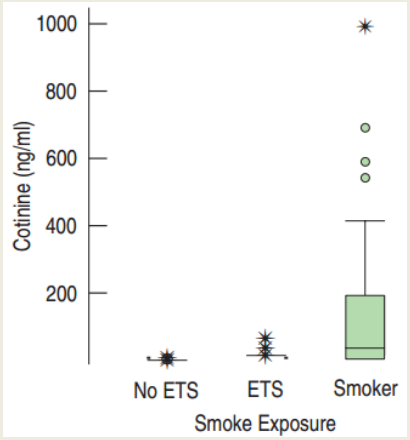


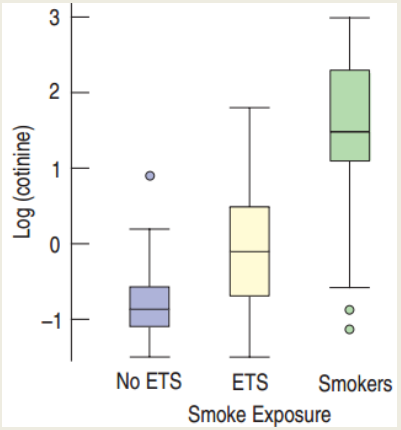
Transformation of the Data

- Taking logarithm of the salaries makes histogram much easier to interpret.









Chapter 5

The Standard Deviation as a Ruler and the Normal Model

How Many Standard Deviations From the Mean?

	Long Jump	200 m
Mean (all contestants)	5.91 m	24.48 s
SD	0.56 m	0.80 s
n	35	36
Chernova	6.54 m	23.67 s
Ennis	6.48 m	22.83 s

- The standard deviation helps us compare: how far from the mean?
- Chernova's long jump was more than 1 standard deviation better than the mean.
- Ennis's winning time in the 200 m was more than 2 standard deviations faster than than the mean.

Is there an even more precise way to calculate these?

The z-Score

- In general, to find the distance between the value and the mean in standard deviations:

1. Subtract the mean from the value.
2. Divide by the standard deviation.

$$z = \frac{y - \bar{y}}{s}$$

- This is called the *z*-score.

The *z*-score

- The *z*-score measures the distance of the value from the mean in standard deviations.
- A positive *z*-score indicates the value is above the mean.
- A negative *z*-score indicates the value is below the mean.
- A small *z*-score indicates the value is close to the mean when compared to the rest of the data values.
- A large *z*-score indicates the value is far from the mean when compared to the rest of the data values.

How Many SDs from Mean?

- Chernova's long jump

$$z = \frac{6.54 - 5.91}{0.56} \approx 1.1$$

- Ennis's 200 m run

$$z = \frac{22.83 - 24.48}{0.80} \approx -2.1$$

	Long Jump	200 m
Mean (all contestants)	5.91 m	24.48 s
SD	0.56 m	0.80 s
n	35	36
Chernova	6.54 m	23.67 s
Ennis	6.48 m	22.83 s

- Ennis's winning time is a little more impressive.
- Judges could assign points based on standard deviations from mean and this system would have a correlation of 0.99 with the one currently used!

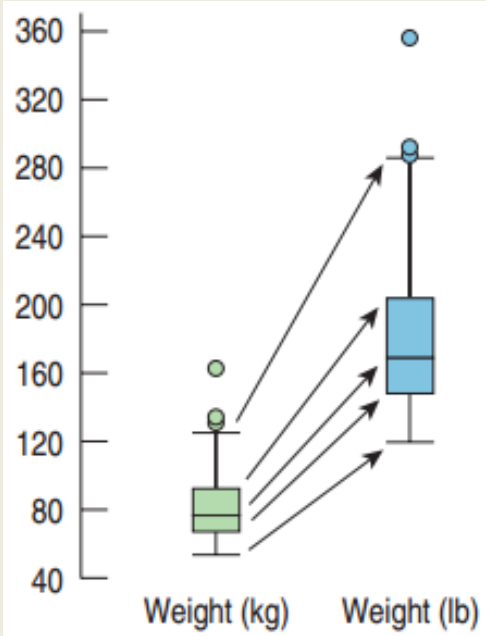
How Many SDs from Mean?

- $-1 < z < 1$: Not uncommon
- $z = \pm 3$: Rare
- $z = 6$: Shouts out for attention!

Shifting

- If the same number is subtracted or added to all data values, then:
  - The measures of the spread – standard deviation, range, and IQR – are all unaffected.
  - The measures of position – mean, median, and mode – are all changed by that number.

Rescaling



	Weight (kg)	Weight (lb)
Min	54.3	119.46
Q1	67.3	148.06
Median	76.85	169.07
Q3	92.3	203.06
Max	161.5	355.30
IQR	25	55
SD	22.27	48.99

- When we multiply (or divide) all the data values by a constant, all measures of position and all measures of spread are multiplied (or divided) by that same constant.