Economics 103 – Statistics for Economists

Francis J. DiTraglia

University of Pennsylvania

Lecture #1 – Introduction

Overview – Population vs. Sample, Probability vs. Statistics

Polling – Sampling vs. Non-sampling Error, Random Sampling

Causality – Observational vs. Experimental Data, RCTs

Racial Discrimination in the Labor Market

Source: Bureau of Labor Statistics

	Oct. 2017	Nov. 2017	Dec. 2017
White:	3.2	3.2	3.4
Black/African American:	7.4	7.3	6.6

Table: Unemployment rate in percentage points for men aged 20 and over in the last quarter of 2017.

The unemployment rate for African Americans has historically been much higher than for whites. What can this information by itself tell us about racial discrimination in the labor market?

This Course: Use Sample to Learn About Population

Population

Complete set of all items that interest investigator

Sample

Observed subset, or portion, of a population

Sample Size

of items in the sample, typically denoted n

Examples...

In Particular: Use Statistic to Learn about Parameter

Parameter

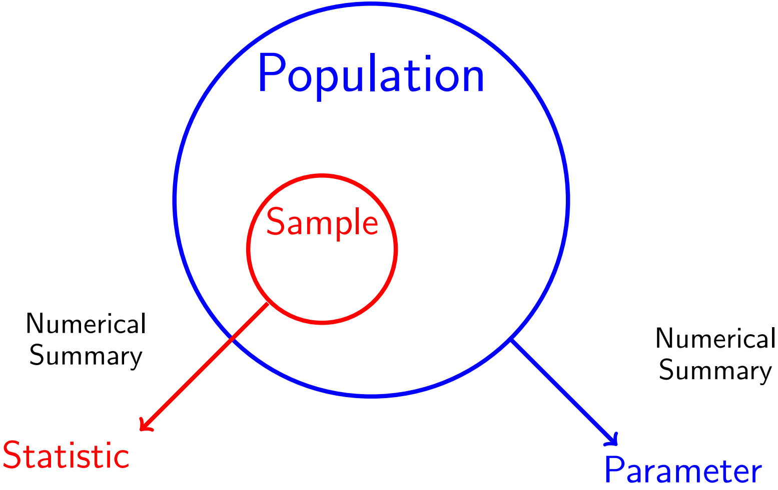
Numerical measure that describes specific characteristic of a population.

Statistic

Numerical measure that describes specific characteristic of sample.

Examples...

Essential Distinction You Must Remember!



This Course

1. Descriptive Statistics: summarize data
   - Summary Statistics
   - Graphics
2. Probability: Population $\rightarrow$ Sample
   - deductive: “safe” argument
     - All ravens are black. Mordecai is a raven, so Mordecai is black.
3. Inferential Statistics: Sample $\rightarrow$ Population
   - inductive: “risky” argument
     - I’ve only every seen black ravens, so all ravens must be black.

Sampling and Nonsampling Error

In statistics we use samples to learn about populations, but samples almost never be *exactly* like the population they are drawn from.

1. Sampling Error
   - *Random* differences between sample and population
   - Cancel out on average
   - Decreases as sample size grows
2. Nonsampling Error
   - *Systematic* differences between sample and population
   - Does *not* cancel out on average
   - Does *not* decrease as sample size grows


NEW COLORED MAP OF POLAND IN THIS ISSUE
Showing the Territorial Changes Wrought by the War
The Literary Digest
(Title Reg. U.S. Pat. Off.)
New York FUNK & WAGNALLS COMPANY London
PUBLIC OPINION New York combined with The LITERARY DIGEST
Vol. 68, No. 8. Whole No. 1609
FEBRUARY 19, 1921
Price 10 Cents

Literary Digest – 1936 Presidential Election Poll

FDR versus Kansas Gov. Alf Landon

Huge Sample

Sent out over 10 million ballots; 2.4 million replies! (Compared to less than 45 million votes cast in actual election)

Prediction

Landslide for Landon: *Landonslide*, if you will.

Spectacularly Mistaken!

FDR versus Kansas Gov. Alf Landon

	Roosevelt	Landon
Literary Digest Prediction:	41%	57%
Actual Result:	61%	37%

What Went Wrong? *Non-sampling Error (aka Bias)*

Source: Squire (1988)

Biased Sample

Some units more likely to be sampled than others.

- Ballots mailed those on auto reg. list and in phone books.

Non-response Bias

Even if sample is unbiased, can't force people to reply.

- Among those who recieved a ballot, Landon supporters were more likely to reply.

In this case, neither effect *alone* was enough to throw off the result but together they did.

Randomize to Get an Unbiased Sample

Simple Random Sample

Each member of population is chosen strictly by chance, so that: (1) selection of one individual doesn't influence selection of any other, (2) each individual is just as likely to be chosen, (3) every possible sample of size n has the same chance of selection.

What about non-response bias? – we'll come back to this...

“Negative Views of Trump’s Transition”

Source: Pew Research Center

Ahead of Donald Trump’s scheduled press conference in New York City on Wednesday, the public continues to give the president-elect low marks for how he is handling the transition process... The latest national survey by Pew Research Center, conducted Jan. 4-9 among 1,502 adults, finds that 39% approve of the job President-elect Trump has done so far explaining his policies and plans for the future to the American people, while a larger share (55%) say they disapprove.

Quantifying Sampling Error

95% Confidence Interval for Poll Based on Random Sample

Margin of Error a.k.a. ME

We report $P \pm \text{ME}$ where $\text{ME} \approx 2\sqrt{P(1-P)/n}$

Trump Transition Approval Rate

$P = 0.39$ and $n = 1502$ so $\text{ME} \approx 0.013$. We'd report 39% plus or minus 1.3% if the poll were based on a simple random sample...

But Pew Reports an ME of 2.9% – more than twice as large as the one we calculated! What's going on here?!

Non-response bias is a huge problem. . .

Source: Pew Research Center

Surveys Face Growing Difficulty Reaching, Persuading Potential Respondents

	1997	2000	2003	2006	2009	2012
	%	%	%	%	%	%
Contact rate (percent of households in which an adult was reached)	90	77	79	73	72	62
Cooperation rate (percent of households contacted that yielded an interview)	43	40	34	31	21	14
Response rate (percent of households sampled that yielded an interview)	36	28	25	21	15	9

PEW RESEARCH CENTER 2012 Methodology Study. Rates computed according to American Association for Public Opinion Research (AAPOR) standard definitions for CON2, COOP3 and RR3. Rates are typical for surveys conducted in each year.

Methodology – "Negative Views of Trump's Transition"

Source: Pew Research Center

The combined landline and cell phone sample are weighted using an iterative technique that matches gender, age, education, race, Hispanic origin and nativity and region to parameters from the 2015 Census Bureaus American Community Survey and population density to parameters from the Decennial Census. The sample also is weighted to match current patterns of telephone status (landline only, cell phone only, or both landline and cell phone), based on extrapolations from the 2016 National Health Interview Survey. The weighting procedure also accounts for the fact that respondents with both landline and cell phones have a greater probability of being included in the combined sample and adjusts for household size among respondents with a landline phone. The margins of error reported and statistical tests of significance are adjusted to account for the surveys design effect, a measure of how much efficiency is lost from the weighting procedures.

Simple Example of Weighting a Survey

Post-stratification

- Women make up 49.6% of the population but suppose they are less likely to respond to your survey than men.
- If women have different opinions of Trump, this will skew the survey.
- Calculate Trump approval rate separately for men P_M vs. women P_W.
- Report $0.496 \times P_W + 0.504 \times P_M$, not the raw approval rate P.

Caveats

- Post-stratification isn't a magic bullet: you have to figure out what factors could skew your poll to adjust for them.
- Calculating the ME is more complicated. Since this is an intro class we'll focus on simple random samples.

Survey to find effect of Polio Vaccine

Ask random sample of parents if they vaccinated their kids or not and if the kids later developed polio. Compare those who were vaccinated to those who weren't.

Would this procedure:

(a) Overstate effectiveness of vaccine

(b) Correctly identify effectiveness of vaccine

(c) Understate effectiveness of vaccine

Confounding

Parents who vaccinate their kids may differ systematically from those who don't in *other ways* that impact child's chance of contracting polio!

Wealth is related to vaccination *and* whether child grows up in a hygenic environment.

Confounder

Factor that influences both outcomes and whether subjects are treated or not. Masks true effect of treatment.

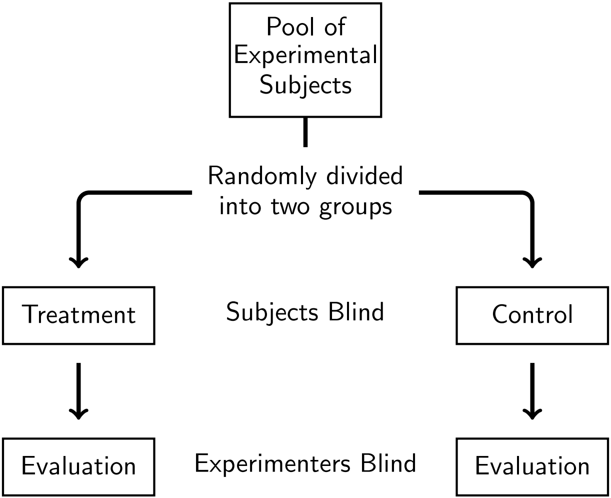
Experiment Using Random Assignment: Randomized Experiment

Treatment Group Gets Vaccine, Control Group Doesn't

Essential Point!

Random assignment *neutralizes* effect of all confounding factors: since groups are initially equal, on average, any difference that emerges must be the treatment effect.

Placebo Effect and Randomized Double Blind Experiment


Pool of Experimental Subjects
Randomly divided into two groups
Treatment
Subjects Blind
Control
Evaluation
Experimenters Blind
Evaluation

Gold Standard: Randomized, Double-blind Experiment

Randomized blind experiments ensure that on average the two groups are initially equal, and continue to be treated equally. Thus a fair comparison is possible.

Randomized, double-blind experiments are considered the "gold standard" for untangling causation.

Sugar Doesn't Make Kids Hyper

http://www.youtube.com/watch?v=mkr9YsmrPAI

Randomization is not always possible, practical, or ethical.

Observational Data

Data that do not come from a randomized experiment.

It much more challenging to untangle cause and effect using observational data because of confounders. But sometimes it's all we have.

Racial Bias in the Labor Market

Bertrand & Mullainathan (2004, American Economic Review)

When faced with observably similar African-American and White applicants, do they [employers] favor the White one? Some argue yes, citing either employer prejudice or employer perception that race signals lower productivity. Others argue that differential treatment by race is a relic of the past . . . Data limitations make it difficult to empirically test these views. Since researchers possess far less data than employers do, White and African-American workers that appear similar to researchers may look very different to employers. So any racial difference in labor market outcomes could just as easily be attributed to differences that are observable to employers but unobservable to researchers.

Racial Bias in the Labor Market: continued . . .

Bertrand & Mullainathan (2004, American Economic Review)

To circumvent this difficulty, we conduct a field experiment . . . We send resumes in response to help-wanted ads in Chicago and Boston newspapers and measure call-back for interview for each sent resume. We experimentally manipulate the perception of race via the name of the ficticious job applicant. We randomly assign very White-sounding names (such as Emily Walsh or Greg Baker) to half the resumes and very African-American-soundsing names (such as Lakisha Washington or Jamal Jones) to the other half.

Racial Bias in the Labor Market: continued ...

Bertrand & Mullainathan (2004, American Economic Review)

Sample	White Names	African-American Names
All sent resumes	9.7	6.5
Females	9.9	6.6
Males	8.9	5.8

Table: % Callback by racial soundingness of names.

Later this semester: if there were no racial bias in callbacks, what is the chance that we would observe such large differences?

Lecture #2 – Summary Statistics Part I

Class Survey

Types of Variables

Frequency, Relative Frequency, & Histograms

Measures of Central Tendency

Measures of Variability / Spread

Class Survey

- Collect some data to analyze later in the semester.
- None of the questions are sensitive and your name will not be linked to your responses. I will post an anonymized version of the dataset on my website.
- The survey is *strictly voluntary* – if you don't want to participate, you don't have to.

Multiple Choice Entry – What is your biological sex?

(a) Male

(b) Female

Numeric Entry – How Many Credits?

How many credits are you taking this semester? Please respond using your remote.

Multiple Choice – Class Standing

What is your class standing at Penn?

(a) Freshman

(b) Sophomore

(c) Junior

(d) Senior

Multiple Choice – What is Your Eye Color?

Please enter your eye color using your remote.

(a) Black

(b) Blue

(c) Brown

(d) Green

(e) Gray

(f) Green

(g) Hazel

(h) Other

How Right-Handed are You?

The sheet in front of you contains a handedness inventory. Please complete it and calculate your handedness score:

$$\frac{\text{Right} - \text{Left}}{\text{Right} + \text{Left}}$$

When finished, enter your score using your remote.

What is your Height in Inches?

Using your remote, please enter your height in inches, rounded to the nearest inch:

$4\text{ft} = 48\text{in}$

$5\text{ft} = 60\text{in}$

$6\text{ft} = 72\text{in}$

$7\text{ft} = 84\text{in}$

What is your Hand Span (in cm)?

On the sheet in front of you is a ruler. Please use it to measure the span of your right hand in centimeters, to the nearest 1/2 cm.

> *Hand Span: the distance from thumb to little finger when your fingers are spread apart*

When ready, enter your measurement using your remote.

We chose (by computer) a random number between 0 and 100. The number selected and assigned to you is written on the slip of paper in front of you. Please do not show your number to anyone else or look at anyone else's number.

Please enter your number now using your remote.

Call your random number X. Do you think that the percentage of countries, among all those in the United Nations, that are in Africa is higher or lower than X?

(a) Higher

(b) Lower

Please answer using your remote.

What is your best estimate of the percentage of countries, among all those that are in the United Nations, that are in Africa?

Please enter your answer using your remote.

Types of Variables

Categorical

Qualitative, assigns each unit to category, number either meaningless or indicates order only

Nominal no order to the categories

Ordinal categories with natural order

Numerical

Quantitative, number meaningful

Discrete Value from discrete set (often count data)

Continuous Value could conceptually be any real number within some range (though *measurements* have finite precision)

Types of Variables – Examples

Categorical (called *Factor* in R)

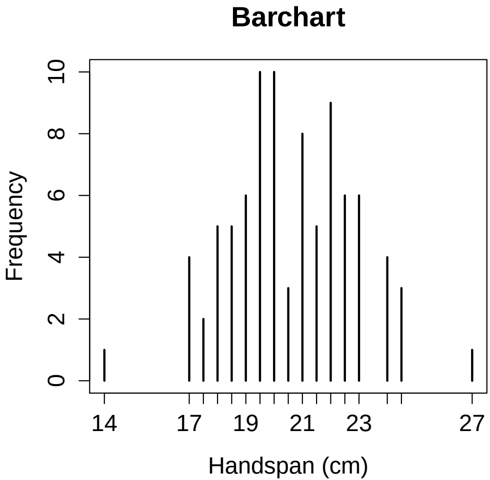
Nominal	eye color, sex
Ordinal	course evaluations ($0 =$ Poor, $1 =$ Fair, $\ldots$)

Numerical

Discrete	credits you are taking this semester
Continuous	height, handspan, handedness (from survey)

Handspan - Frequency and Relative Frequency

cm	Freq.	Rel. Freq.
14.0	1	0.01
17.0	4	0.05
17.5	2	0.02
18.0	5	0.06
18.5	5	0.06
19.0	6	0.07
19.5	10	0.11
20.0	10	0.11
20.5	3	0.03
21.0	8	0.09
21.5	5	0.06
22.0	9	0.10
22.5	6	0.07
23.0	6	0.07
24.0	4	0.05
24.5	3	0.03
27.0	1	0.01
	$n = 89$	1.00



Handspan - Summarize Barchart by "Smoothing"

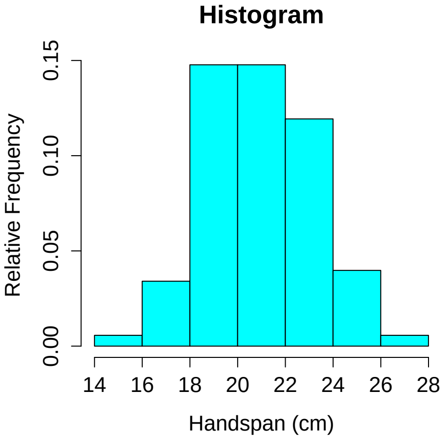
cm	Freq.	Rel. Freq.
14.0	1	0.01
17.0	4	0.05
17.5	2	0.02
18.0	5	0.06
18.5	5	0.06
19.0	6	0.07
19.5	10	0.11
20.0	10	0.11
20.5	3	0.03
21.0	8	0.09
21.5	5	0.06
22.0	9	0.10
22.5	6	0.07
23.0	6	0.07
24.0	4	0.05
24.5	3	0.03
27.0	1	0.01
	$n = 88$	1.00

Group data into non-overlapping bins of equal width:

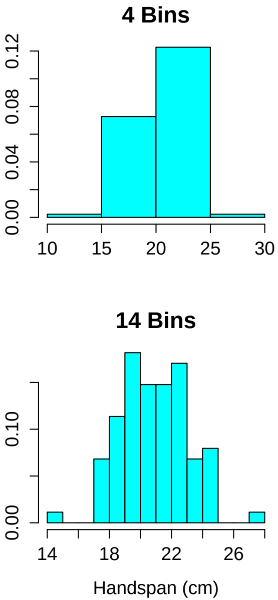
Bins	Freq.	Rel. Freq.
$[14, 16)$	1	0.01
$[16, 18)$	6	0.07
$[18, 20)$	26	0.30
$[20, 22)$	26	0.30
$[22, 24)$	21	0.24
$[24, 26)$	7	0.08
$[26, 28)$	1	0.01
	$n = 88$	1.00

Histogram – Density Estimate by Smoothing Barchart

Bins	Freq.	Rel. Freq.
$[14, 16)$	1	0.01
$[16, 18)$	6	0.07
$[18, 20)$	26	0.30
$[20, 22)$	26	0.30
$[22, 24)$	21	0.24
$[24, 26)$	7	0.08
$[26, 28)$	1	0.01
	$n = 88$	1.00



Number of Bins Controls Degree of Smoothing



Histograms are *Really* Important

Why Histogram?

Summarize numerical data, especially continuous (few repeats)

Too Many Bins – Undersmoothing

No longer a summary (lose the shape of distribution)

Too Few Bins – Oversmoothing

Miss important detail

Don't confuse with barchart!

Summary Statistic: Numerical Summary of Sample

1. Measures of Central Tendency
   - Mean
   - Median
2. Measures of Spread
   - Variance
   - Standard Deviation
   - Range
   - Interquartile Range (IQR)
3. Measures of Symmetry
   - Skewness
4. Measures of relationship between variables
   - Covariance
   - Correlation
   - Regression

Questions to Ask Yourself about Each Summary Statistic

1. What does it measure?
2. What are its units compared to those of the data?
3. (How) do its units change if those of the data change?
4. What are the benefits and drawbacks of this statistic?

Some of the information regarding items 2 and 3 is on the homework rather than in the slides because working it out for yourself is a good way to check your understanding.

What is an Outlier?

Outlier

A very unusual observation relative to the other observations in the dataset (i.e. very small or very big).

Measures of Central Tendency

Suppose we have a dataset with observations $x_1, x_2, \ldots, x_n$

Sample Mean

- $\bar{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$
- Only for numeric data
- Sensitive to asymmetry and outliers

Sample Median

- Middle observation if n is odd, otherwise the mean of the two observations closest to the middle.
- Applicable to numerical or ordinal data
- Insensitive to outliers and skewness

Mean is Sensitive to Outliers, Median Isn't

First Dataset: 1 2 3 4 5

Mean = 3, Median = 3

Second Dataset: 1 2 3 4 4990

Mean = 1000, Median = 3

When Does the Median Change?

Ranks would have to change so that 3 is no longer in the middle.

Percentage of UN Countries that are in Africa

You Were a Subject in a Randomized Experiment!

- There were only two numbers in the bag: 10 and 65
- Randomly assigned to Low group (10) or High group (65)

Anchoring Heuristic (Kahneman and Tversky, 1974)

Subjects' estimates of an unknown quantity are influenced by an irrelevant previously supplied starting point.

Are Penn students subject to to this cognitive bias?

Last Semester's Class

	Mean	Median
Low ($n = 43$)	17.1	17
High ($n = 46$)	30.7	30

Kahneman and Tversky (1974)

Low Group (shown 10) $\rightarrow$ median answer of 25

High Group (shown 65) $\rightarrow$ median answer of 45

(Kahneman shared 2002 Economics Nobel Prize with Vernon Smith.)

Percentiles (aka Quantiles) – Generalization of Median

Approx. $P\%$ of the data are at or below the P^{th} percentile.

Percentiles (aka Quantiles)

P^{th} Percentile = Value in $(P/100) \cdot (n+1)^{th}$ Ordered Position

Quartiles

Q1 = 25th Percentile

Q2 = Median (i.e. 50th Percentile)

Q3 = 75th Percentile

An Example: n = 12

60 63 65 67 70 72 75 75 80 82 84 85

$$
\begin{aligned}
Q_1 &= \text{value in the } 0.25(n+1)^{th} \text{ ordered position} \\
&= \text{value in the } 3.25^{th} \text{ ordered position} \\
&= 65 + 0.25 * (67 - 65) \\
&= 65.5
\end{aligned}
$$

Quantiles of Student Debt

Source: Avery & Turner (2012)

Table 4

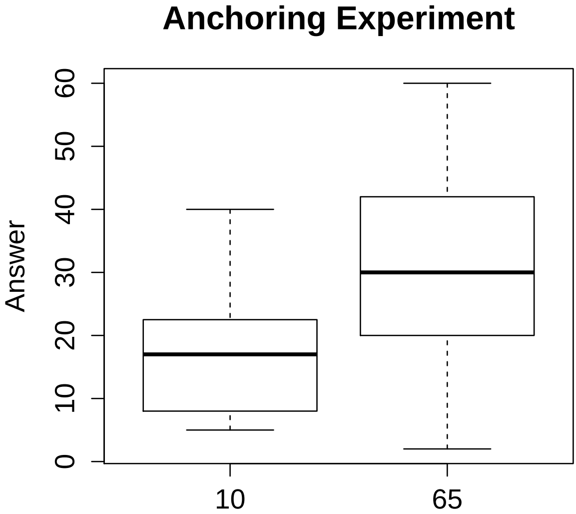
Borrowing Distribution after Six Years, by Degree Type and First Institution

	Type of institution of first enrollment			
	Public 4-year	*Private nonprofit 4-year*	*Private for-profit 4-year*	*Public 2-year*
All students beginning in 2004				
% Borrowing	61%	68%	89%	41%
Percentile of borrowers				
10^{th}	\$0	\$0	\$0	\$0
25^{th}	\$0	\$0	\$6,376	\$0
50^{th}	\$6,000	\$11,500	\$13,961	\$0
75^{th}	\$19,000	\$24,750	\$28,863	\$6,625
90^{th}	\$30,000	\$40,000	\$45,000	\$18,000
Mean	**\$11,706**	**\$16,606**	**\$19,726**	**\$5,586**
BA recipients				
BA completion	61.5%	70.7%	14.8%	13%
% Borrowing	59%	66%	92%	69%
Percentile of borrowers				
10^{th}	\$0	\$0	\$12,000	\$0
25^{th}	\$0	\$0	\$30,000	\$0
50^{th}	\$7,500	\$15,500	\$45,000	\$11,971
75^{th}	\$20,000	\$27,000	\$50,000	\$23,265
90^{th}	\$32,405	\$45,000	\$100,000	\$40,000
Mean	**\$12,922**	**\$18,700**	**\$45,042**	**\$15,960**

Source: Authors' tabulations based on the Beginning Postsecondary Survey 2004:2009.

Boxplots and the Five-Number Summary

Minimum $<$ Q1 $<$ Median $<$ Q3 $<$ Maximum



Measures of Variability/Spread – 1

Range

- Range = Maximum Observation - Minimum Observation
- *Very* sensitive to outliers.
- Displayed in boxplot.

Interquartile Range (IQR)

- IQR$= Q_3 - Q_1$
- IQR = Range of middle 50% of the data.
- Insensitive to outliers.
- Displayed in boxplot.

Measures of Variability/Spread – 2

Variance

- $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Essentially the average squared distance from the mean.
- (We'll talk about $n-1$ versus n later in the semester)
- Sensitive to both skewness and outliers.

Standard Deviation

- $s = \sqrt{s^2}$
- Same information as variance but more convenient since it has the same units as the data

Measures of Spread for Anchoring Experiment

Past Semester's Data

Treatment:	$X = 10$	$X = 65$
Range	35	58
IQR	14.5	21
S.D.	9.3	15.9
Var.	86.1	253.5

Lecture #3 – Summary Statistics Part II

Why squares in the definition of variance?

Outliers, Skewness, & Symmetry

Sample versus Population, Empirical Rule

Centering, Standardizing, & Z-Scores

Relating Two Variables: Cross-tabs, Covariance, & Correlation

Why Squares?

$$\boxed{s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

What's Wrong With This?

$$\begin{aligned}
\frac{1}{n-1}\sum_{i=1}^{N}(x_i - \bar{x}) &= \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\bar{x}\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i - n\bar{x}\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i - n \cdot \frac{1}{n}\sum_{i=1}^{n} x_i\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i\right] = 0
\end{aligned}$$

Variance is Sensitive to Skewness and Outliers

And so is Standard Deviation!

$$\boxed{s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Outliers

Differentiate with respect to $(x_i - \bar{x}) \Rightarrow$ the farther an observation is from the mean, the *larger* its effect on the variance.

Skewness

Variance measures average squared distance from center, taking mean as the center, but the mean is sensitive to skewness!

Skewness – A Measure of Symmetry

$$\boxed{\text{Skewness} = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}}$$

What do the values indicate?

Zero $\Rightarrow$ symmetry, positive right-skewed, negative left-skewed.

Why cubed?

To get the desired sign.

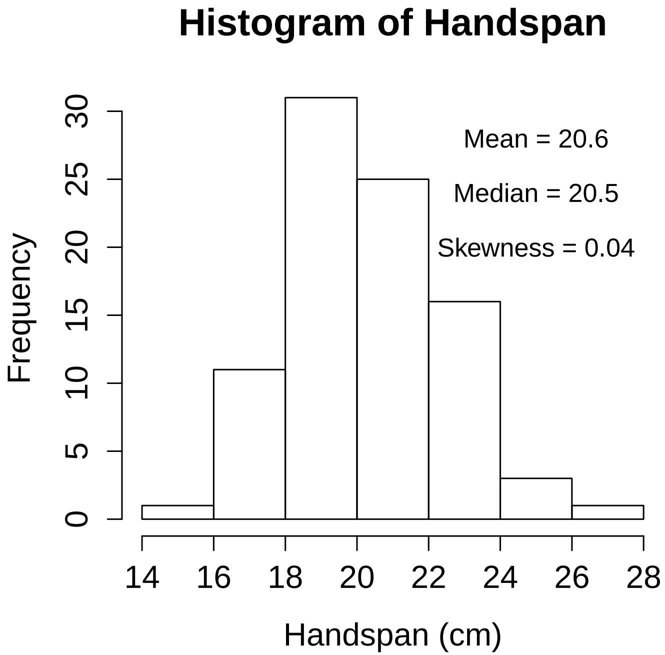
Why divide by s^3?

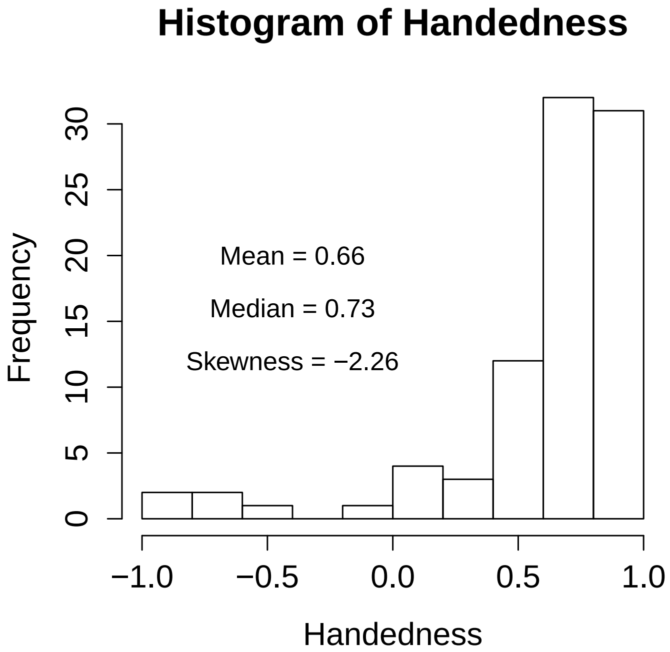
So that skewness is unitless

Rule of Thumb

Typically (but not always), right-skewed $\Rightarrow$ mean $>$ median

left-skewed $\Rightarrow$ mean $<$ median


Histogram of Handspan
Mean = 20.6
Median = 20.5
Skewness = 0.04
Frequency
0
5
10
15
20
25
30
14
16
18
20
22
24
26
28
Handspan (cm)


Histogram of Handedness
Mean = 0.66
Median = 0.73
Skewness = −2.26
Frequency
0
5
10
15
20
25
30
−1.0
−0.5
0.0
0.5
1.0
Handedness

Essential Distinction: Sample vs. Population

For now, you can think of the population as a list of N objects:

Population: $x_1, x_2, \ldots, x_N$

from which we draw a sample of size $n < N$ objects:

Sample: $x_1, x_2, \ldots, x_n$

Important Point:

Later in the course we'll be more formal by considering probability models that represent the *act of sampling* from a population rather than thinking of a population as a list of objects. Once we do this we will no longer use the notation N as the population will be *conceptually infinite*.

Essential Distinction: Parameter vs. Statistic

N individuals in the Population, n individuals in the Sample:

	Parameter (Population)	**Statistic** (Sample)
Mean	$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Var.	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
S.D.	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$

Key Point

We use a sample $x_1, \ldots, x_n$ to calculate statistics (e.g. $\bar{x}$, s^2, s) that serve as estimates of the corresponding population parameters (e.g. μ, σ^2, σ).

Why Do Sample Variance and Std. Dev. Divide by $n-1$?

Pop. Var.	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$	Sample Var.	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Pop. S.D.	$\sigma = \sqrt{\sigma^2}$	Sample S.D.	$s = \sqrt{s^2}$

There is an important reason for this, but explaining it requires some concepts we haven't learned yet.

Why Mean and Variance (and Std. Dev.)?

Empirical Rule

For large populations that are approximately bell-shaped, std. dev. tells where most observations will be relative to the mean:

- $\approx$ 68% of observations are in the interval $\mu \pm \sigma$
- $\approx$ 95% of observations are in the interval $\mu \pm 2\sigma$
- Almost all of observations are in the interval $\mu \pm 3\sigma$

Therefore

We will be interested in $\bar{x}$ as an estimate of μ and s as an estimate of σ since these population parameters are so informative.

Which is more "extreme?"

(a) Handspan of 27cm

(b) Height of 78in

Centering: Subtract the Mean

Handspan	Height
$27\text{cm} - 20.6\text{cm} = 6.4\text{cm}$	$78\text{in} - 67.6\text{in} = 10.4\text{in}$

Standardizing: Divide by S.D.

Handspan	Height
$27\text{cm} - 20.6\text{cm} = 6.4\text{cm}$	$78\text{in} - 67.6\text{in} = 10.4\text{in}$
$6.4\text{cm}/2.2\text{cm} \approx 2.9$	$10.4\text{in}/4.5\text{in} \approx 2.3$

The units have disappeared!

Z-scores: How many standard deviations from the mean?

Best for Symmetric Distribution, No Outliers (Why?)

$$z_i = \frac{x_i - \bar{x}}{s}$$

Unitless

Allows comparison of variables with different units.

Detecting Outliers

Measures how "extreme" one observation is relative to the others.

Linear Transformation

What is the sample mean of the z-scores?

$$
\begin{aligned}
\bar{z} &= \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i - \bar{x}}{s} = \frac{1}{n\cdot s}\left[\sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\bar{x}\right] \\
&= \frac{1}{n\cdot s}\left[\sum_{i=1}^{n} x_i - n\bar{x}\right] = \frac{1}{n\cdot s}\left[\sum_{i=1}^{n} x_i - n\cdot\frac{1}{n}\sum_{i=1}^{n} x_i\right] \\
&= \frac{1}{n\cdot s}\left[\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i\right] = 0
\end{aligned}
$$

What is the variance of the z-scores?

$$
\begin{aligned}
s_z^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(z_i-\bar{z})^2 = \frac{1}{n-1}\sum_{i=1}^{n} z_i^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)^2 \\
&= \frac{1}{s_x^2}\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right] = \frac{s_x^2}{s_x^2} = 1
\end{aligned}
$$

So what is the *standard deviation* of the z-scores?

Population Z-scores and the Empirical Rule: $\mu \pm 2\sigma$

If we knew the population mean μ and standard deviation σ we could create a *population version* of a z-score. This leads to an important way of rewriting the Empirical Rule:

Bell-shaped population $\Rightarrow$ approx. 95% of observations x_i satisfy

$$\mu - 2\sigma \leq x_i \leq \mu + 2\sigma$$

$$-2\sigma \leq x_i - \mu \leq 2\sigma$$

$$-2 \leq \frac{x_i - \mu}{\sigma} \leq 2$$

Relationships Between Variables

Crosstabs – Show Relationship between Categorical Vars.

(aka Contingency Tables)

Eye Color	*Sex*		
	Male	Female	Total
Black	5	2	7
Blue	6	4	10
Brown	26	31	57
Copper	1	0	1
Dark Brown	0	1	1
Green	4	1	5
Hazel	2	2	4
Maroon	1	0	1
Total	45	41	86

Example with Crosstab in *Percents*

Who Supported the Vietnam War?

In January 1971 the Gallup poll asked: "A proposal has been made in Congress to require the U.S. government to bring home all U.S. troops before the end of this year. Would you like to have your congressman vote for or against this proposal?"

Guess the results, for respondents in each education category, and fill out this table (the two numbers in each column should add up to 100%):

	Adults with:			
	Grade school education	High school education	College education	Total adults
% for withdrawal of U.S. troops (doves)				73%
% against withdrawal of U.S. troops (hawks)				27%
Total	100%	100%	100%	100%

Who Were the Doves?

Which group do you think was most strongly in favor of the withdrawal of US troops from Vietnam?

(a) Adults with only a Grade School Education

(b) Adults with a High School Education

(c) Adults with a College Education

Please respond with your remote.

Who Were the Hawks?

Which group do you think was most strongly opposed to the withdrawal of US troops from Vietnam?

(a) Adults with only a Grade School Education

(b) Adults with a High School Education

(c) Adults with a College Education

Please respond with your remote.

From The Economist –"Lexington," October 4th, 2001

"Back in the Vietnam days, the anti-war movement spread from the intelligentsia into the rest of the population, eventually paralyzing the country's will to fight."

Who *Really* Supported the Vietnam War

Gallup Poll, January 1971

	Adults with: Grade school education	High school education	College education	Total adults
% for withdrawal of U.S. troops (doves)	80%	75%	60%	73%
% against withdrawal of U.S. troops (hawks)	20%	25%	40%	27%
Total	100%	100%	100%	100%

What about numeric data?

Covariance and Correlation: Linear Dependence Measures

Two Samples of Numeric Data

$x_1, \ldots, x_n$ and $y_1, \ldots, y_n$

Dependence

Do x and y both tend to be large (or small) at the same time?

Key Point

Use the idea of centering and standardizing to decide what "big" or "small" means in this context.

Notation

$$
\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\
\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \\
s_x &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
s_y &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}
\end{aligned}
$$

Covariance

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- Centers each observation around its mean and multiplies.
- Zero $\Rightarrow$ no linear dependence
- Positive $\Rightarrow$ positive linear dependence
- Negative $\Rightarrow$ negative linear dependence
- Population parameter: σ_{xy}
- Units?

Correlation

$$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) = \frac{s_{xy}}{s_x s_y}$$

- Centers *and* standardizes each observation
- Bounded between -1 and 1
- Zero $\Rightarrow$ no linear dependence
- Positive $\Rightarrow$ positive linear dependence
- Negative $\Rightarrow$ negative linear dependence
- Population parameter: ρ_{xy}
- Unitless

We'll have more to say about correlation and covariance when we discuss linear regression.

Essential Distinction: Parameter vs. Statistic

And Population vs. Sample

N individuals in the Population, n individuals in the Sample:

	Parameter (Population)	**Statistic** (Sample)
Mean	$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Var.	$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$	$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
S.D.	$\sigma_x = \sqrt{\sigma_x^2}$	$s_x = \sqrt{s^2}$
Cov.	$\sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}$	$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$
Corr.	$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$	$r = \frac{s_{xy}}{s_x s_y}$

Lecture #4 – Linear Regression I

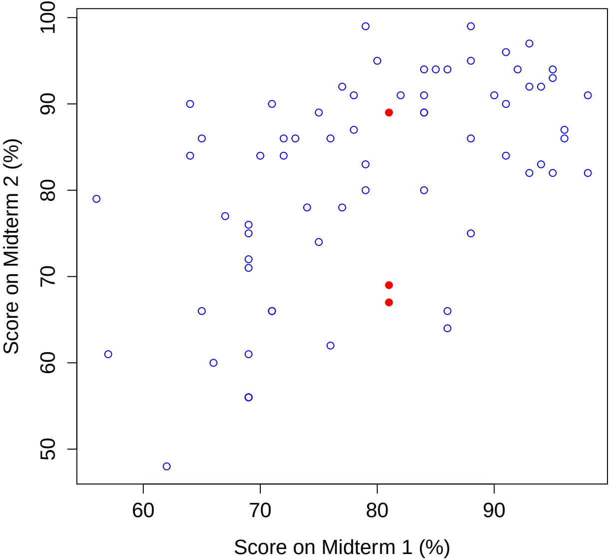
Overview / Intuition for Linear Regression

Deriving the Regression Equations

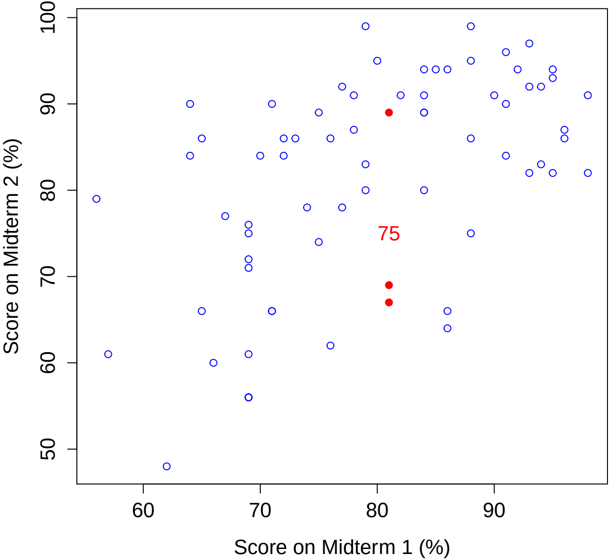
Relating Regression, Covariance and Correlation

Regression to the Mean

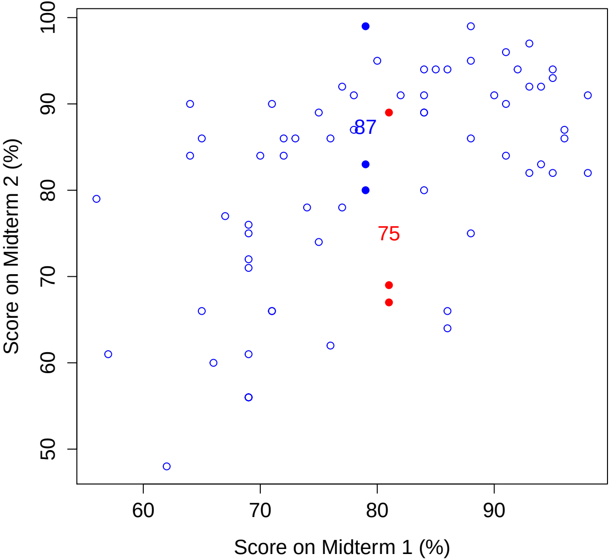
Predict Second Midterm given 81 on First



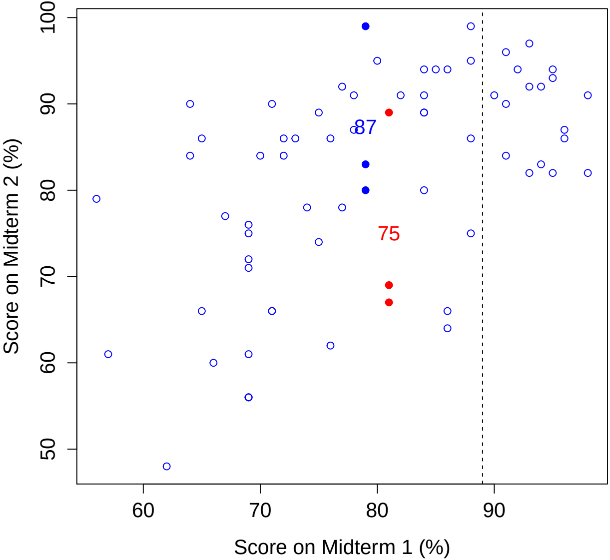
Predict Second Midterm given 81 on First



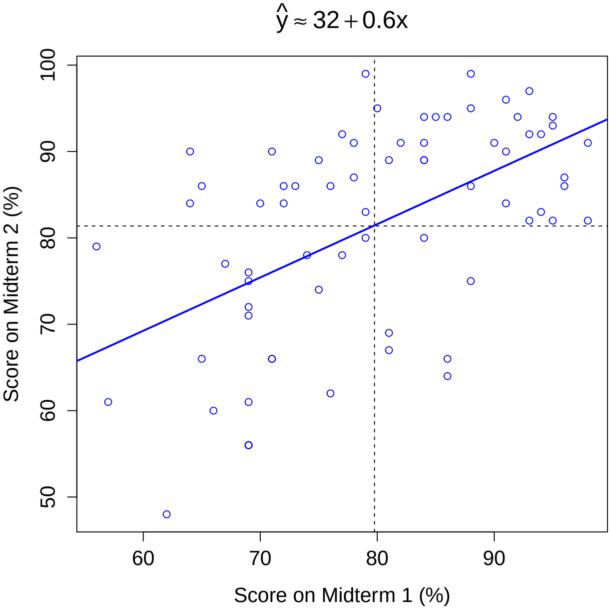
But if they'd only gotten 79 we'd predict higher?!



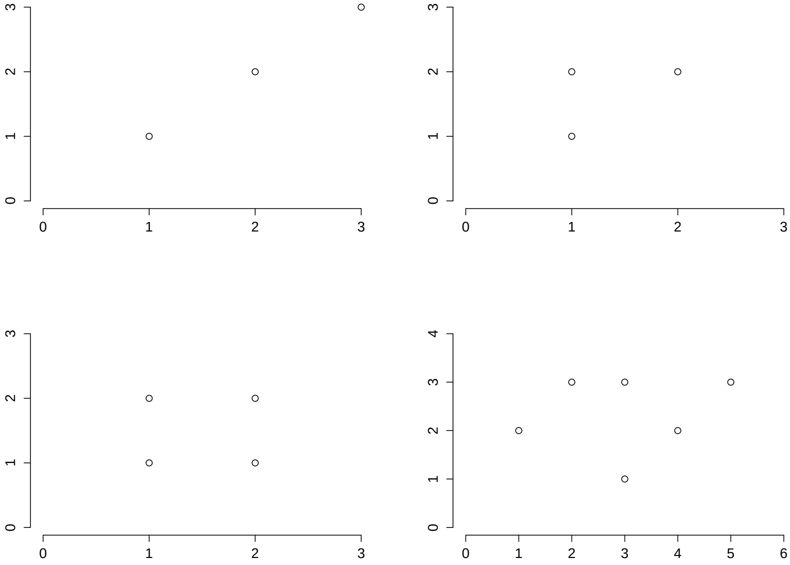
No one who took both exams got 89 on the first!

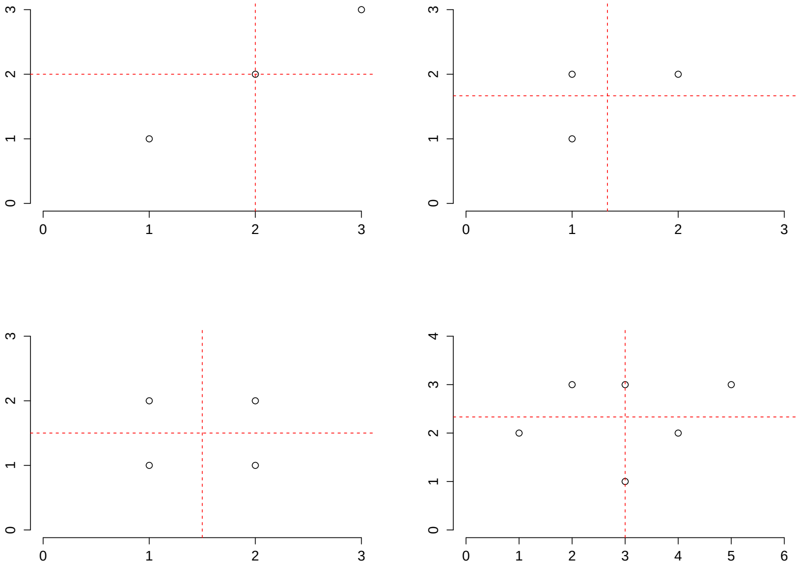


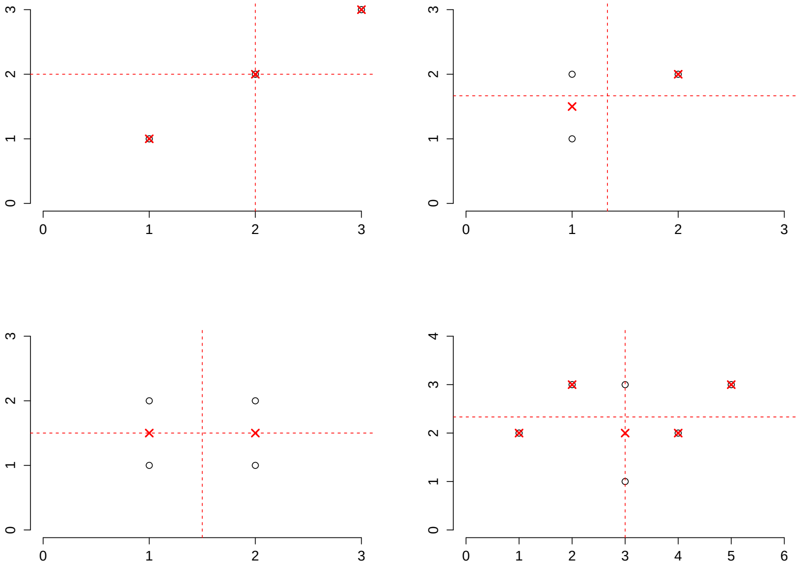
Regression: "Best Fitting" Line Through Cloud of Points

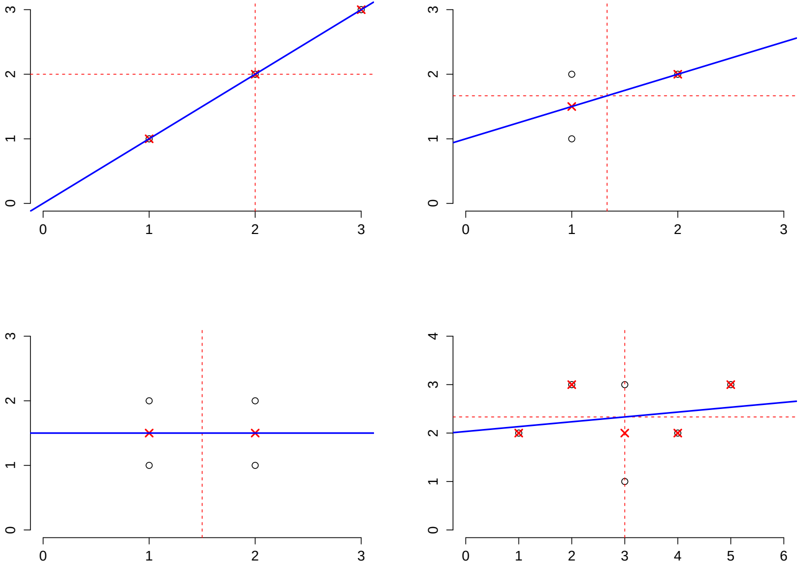


Fitting a Line by Eye


0 1 2 3
0 1 2 3
0 1 2 3
0 1 2 3
0 1 2 3
0 1 2 3
0 1 2 3 4
0 1 2 3 4 5 6


0
1
2
3
0
1
2
3
0
1
2
3
0
1
2
3
0
1
2
3
0
1
2
3
0
1
2
3
4
5
6
0
1
2
3
4


0 1 2 3
0 1 2 3
0 1 2 3
0 1 2 3
0 1 2 3
0 1 2 3
0 1 2 3 4 5 6
0 1 2 3 4


0
1
2
3
0
1
2
3
0
1
2
3
0
1
2
3
0
1
2
3
0
1
2
3
0
1
2
3
4
5
6
0
1
2
3
4

But How to Do this Formally?

Least Squares Regression – Predict Using a Line

The Prediction

Predict score $\hat{y} = a + bx$ on 2nd midterm if you scored x on 1st

How to choose (a, b)?

Linear regression chooses the slope (b) and intercept (a) that

minimize the sum of squared vertical deviations

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Why Squared Deviations?

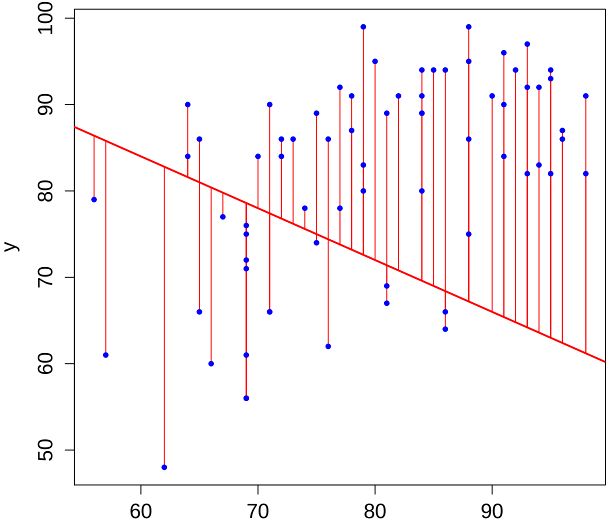
Important Point About Notation

$$\boxed{\underset{a,b}{\text{minimize}} \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2}$$

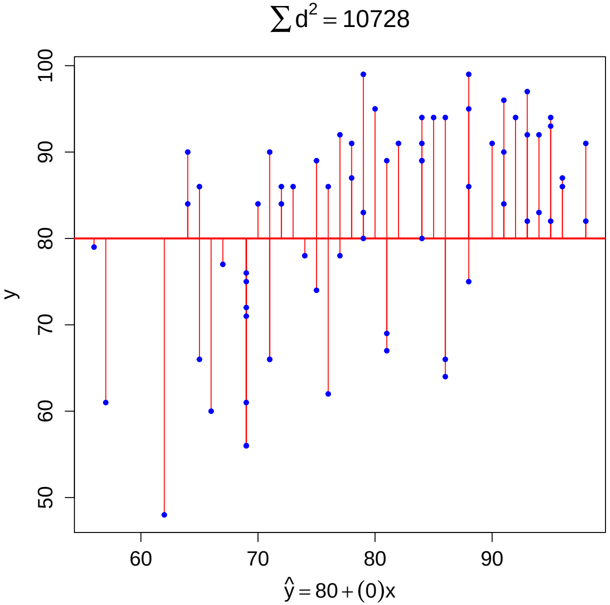
$$\boxed{\hat{y} = a + bx}$$

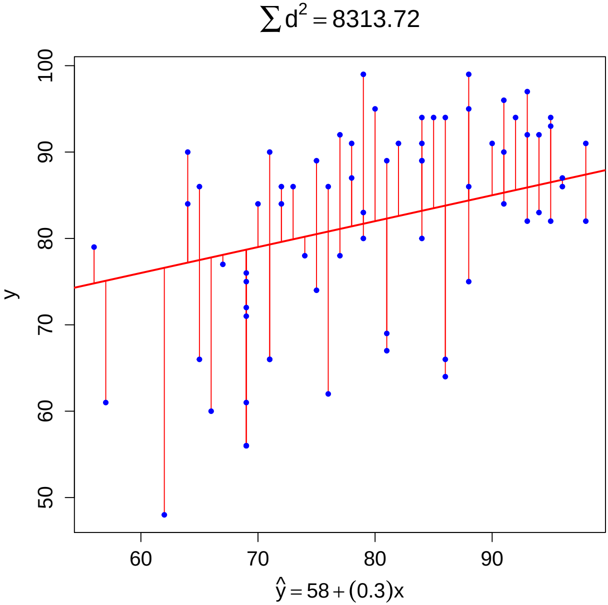
- $(x_i, y_i)_{i=1}^{n}$ are the observed data
- $\widehat{y}$ is our prediction for a given value of x
- Neither x nor $\widehat{y}$ needs to be in out dataset!

$$\sum d^2 = 25596.88$$

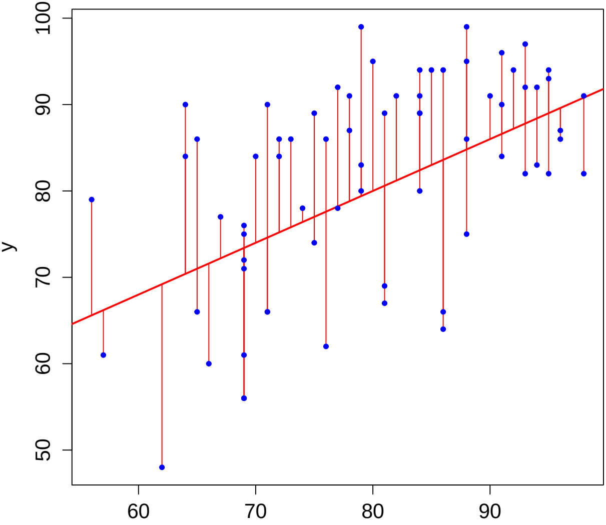


$$\hat{y} = 120 + (-0.6)x$$


$\sum d^2 = 10728$
y
100
90
80
70
60
50
60
70
80
90
$\hat{y} = 80 + (0)x$

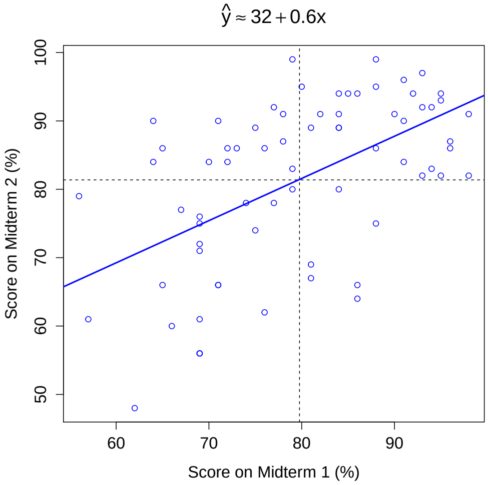

$\sum d^2 = 8313.72$
y
100
90
80
70
60
50
60
70
80
90
$\hat{y} = 58 + (0.3)x$

$\sum d^2 = 7650.48$



$\hat{y} = 32 + (0.6)x$

Prediction given 89 on Midterm 1?



$$32 + 0.6 \times 89 = 32 + 53.4 = 85.4$$

You Need to Know How To Derive This

Minimize the sum of squared vertical deviations from the line:

$$\min_{a,b} \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

How should we proceed?

(a) Differentiate with respect to x

(b) Differentiate with respect to y

(c) Differentiate with respect to x, y

(d) Differentiate with respect to a, b

(e) Can't solve this with calculus.

Objective Function

$$\min_{a,b} \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

FOC with respect to a

$$\begin{aligned}
-2\sum_{i=1}^{n} (y_i - a - bx_i) &= 0 \\
\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a - b\sum_{i=1}^{n} x_i &= 0 \\
\frac{1}{n}\sum_{i=1}^{n} y_i - \frac{na}{n} - \frac{b}{n}\sum_{i=1}^{n} x_i &= 0 \\
\bar{y} - a - b\bar{x} &= 0
\end{aligned}$$

Regression Line Goes Through the Means!

$$\boxed{\bar{y} = a + b\bar{x}}$$

Substitute $a = \bar{y} - b\bar{x}$

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - a - bx_i)^2 &= \sum_{i=1}^{n}(y_i - \bar{y} + b\bar{x} - bx_i)^2 \\
&= \sum_{i=1}^{n}\left[(y_i - \bar{y}) - b\left(x_i - \bar{x}\right)\right]^2
\end{aligned}$$

FOC wrt b

$$\begin{aligned}
-2\sum_{i=1}^{n}\left[(y_i - \bar{y}) - b\left(x_i - \bar{x}\right)\right]\left(x_i - \bar{x}\right) &= 0 \\
\sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right) - b\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 &= 0
\end{aligned}$$

$$b = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

Simple Linear Regression

Problem

$$\min_{a,b} \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Solution

$$b = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

Relating Regression to Covariance and Correlation

$$b = \frac{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right)}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} = \frac{\frac{1}{n-1} \sum_{i=1}^{n} \left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right)}{\frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} = \frac{s_{xy}}{s_x^2}$$

$$r = \frac{s_{xy}}{s_x s_y} = b\frac{s_x}{s_y}$$

Comparing Regression, Correlation and Covariance

Units

Correlation is unitless, covariance and regression coefficients (a, b) are not. (What are the units of these?)

Symmetry

Correlation and covariance are symmetric, regression isn't.

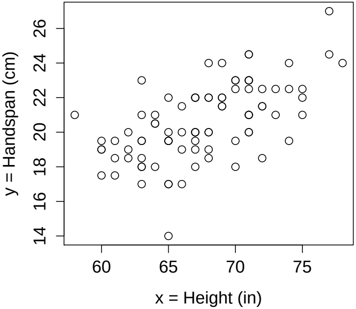
(Switching x and y axes changes the slope and intercept.)

On the Homework

Regression with z-scores rather than raw data gives $a = 0, b = r_{xy}$

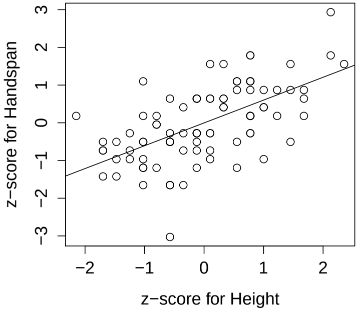
$$s_{xy} = 6, \quad s_x = 5, \quad s_y = 2, \quad \bar{x} = 68, \quad \bar{y} = 21$$

What is the sample correlation between height (x) and handspan (y)?



$$s_{xy} = 6, \quad s_x = 5, \quad s_y = 2, \quad \bar{x} = 68, \quad \bar{y} = 21$$

What is the sample correlation between height (x) and handspan (y)?



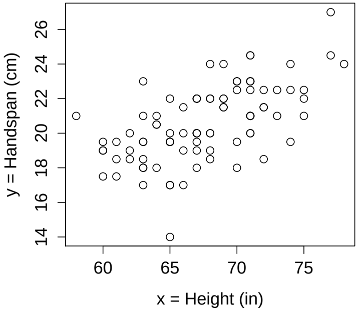
$$r = \frac{s_{xy}}{s_x s_y} = \frac{6}{5 \times 2} = 0.6$$

$$s_{xy} = 6, \quad s_x = 5, \quad s_y = 2, \quad \bar{x} = 68, \quad \bar{y} = 21$$

What is the value of b for the regression:

$$\hat{y} = a + bx$$

where x is height and y is handspan?

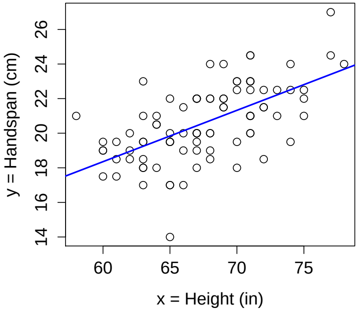


$$s_{xy} = 6, \quad s_x = 5, \quad s_y = 2, \quad \bar{x} = 68, \quad \bar{y} = 21$$

What is the value of b for the regression:

$$\hat{y} = a + bx$$

where x is height and y is handspan?



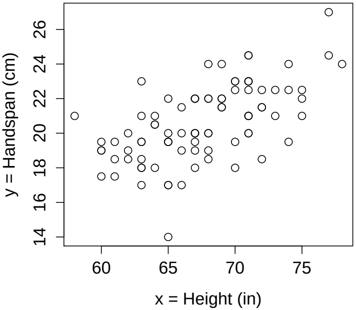
$$b = \frac{s_{xy}}{s_x^2} = \frac{6}{5^2} = 6/25 = 0.24$$

$$s_{xy} = 6, \quad s_x = 5, \quad s_y = 2, \quad \bar{x} = 68, \quad \bar{y} = 21$$

What is the value of a for the regression:

$$\hat{y} = a + bx$$

where x is height and y is handspan?
(prev. slide $b = 0.24$)

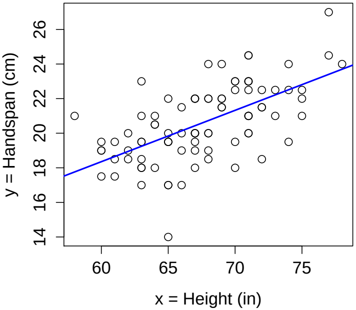


$$s_{xy} = 6, \quad s_x = 5, \quad s_y = 2, \quad \bar{x} = 68, \quad \bar{y} = 21$$

What is the value of a for the regression:

$$\hat{y} = a + bx$$

where x is height and y is handspan? (prev. slide $b = 0.24$)



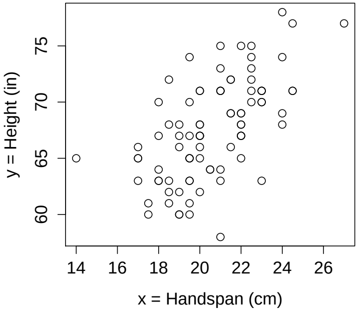
$$a = \bar{y} - b\bar{x} = 21 - 0.24 \times 68 = 4.68$$

$s_{xy} = 6, \quad s_y = 5, \quad s_x = 2, \quad \bar{y} = 68, \quad \bar{x} = 21$

What is the value of b for the regression:

$$\hat{y} = a + bx$$

where x is handspan and y is height?

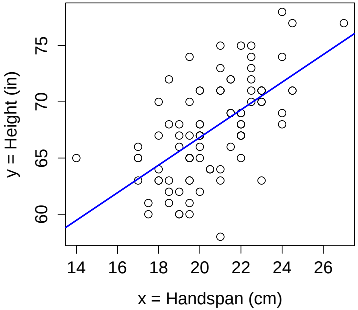


$$s_{xy} = 6, \quad s_y = 5, \quad s_x = 2, \quad \bar{y} = 68, \quad \bar{x} = 21$$

What is the value of b for the regression:

$$\hat{y} = a + bx$$

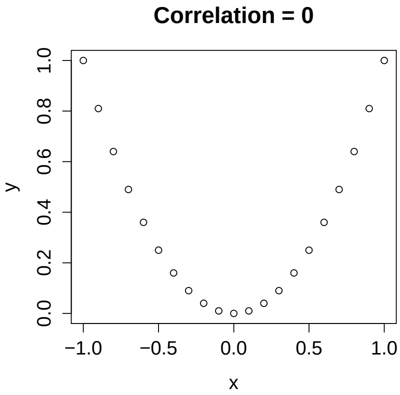
where x is handspan and y is height?



$$b = \frac{s_{xy}}{s_x^2} = 6/2^2 = 1.5$$

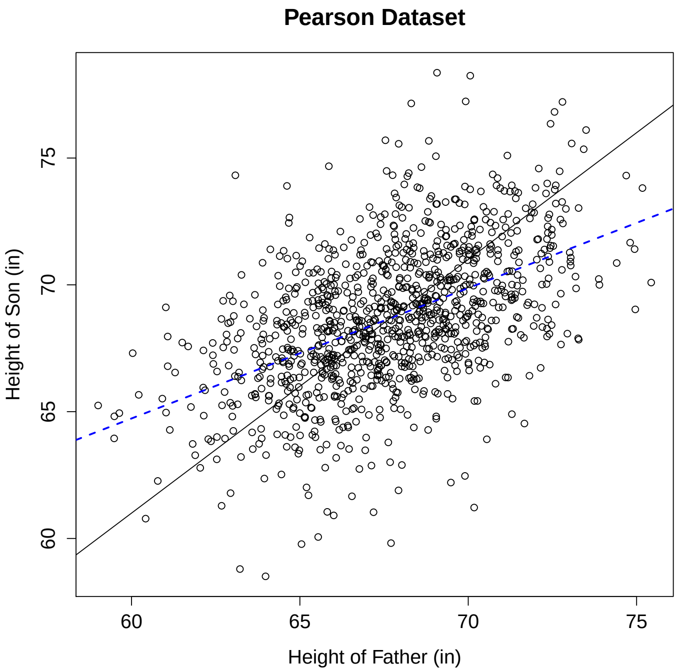
EXTREMELY IMPORTANT

- Regression, Covariance and Correlation: linear association.
- Linear association $\neq$ causation.
- Linear is not the only kind of association!



Regression to the Mean and the Regression Fallacy

Please read Chapter 17 of “Thinking Fast and Slow” by Daniel Kahnemann which I have posted on Piazza. This reading is fair game on an exam or quiz.


Pearson Dataset
Height of Son (in)
75
70
65
60
Height of Father (in)
60
65
70
75

Regression to the Mean

Skill and Luck / Genes and Random Environmental Factors

Unless $r_{xy} = 1$, There Is Regression to the Mean

$$\frac{\hat{y} - \bar{y}}{s_y} = r_{xy} \frac{x - \bar{x}}{s_x}$$

Least-squares Prediction $\hat{y}$ closer to $\bar{y}$ than x is to $\bar{x}$

You will derive the above formula in this week's homework.

Next Up: Basic Probability

Please do the following before our next class:

1. Complete the "Odd Questions" quiz posted on Piazza `OddQuestions.pdf` – we'll be discussing these in class.
2. If you're rusty on permutations, combinations, etc. from High School math, read this review `http://ditraglia.com/Econ103Public/ClassicalProbability.pdf`