Economics 103 – Statistics for Economists

Francis J. DiTraglia

University of Pennsylvania

Lecture # 2

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.
- Participation is *strictly voluntary*: you can still earn full clicker credit for today's lecture without participating in the survey.

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

Numeric Entry – How long did you sleep?

For how many hours did you sleep last night?

Please round to the nearest half hour, e.g. 7 hours 23 minutes becomes 7.5.

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:

4ft $=$ 48in

5ft $=$ 60in

6ft $=$ 72in

7ft $=$ 84in

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

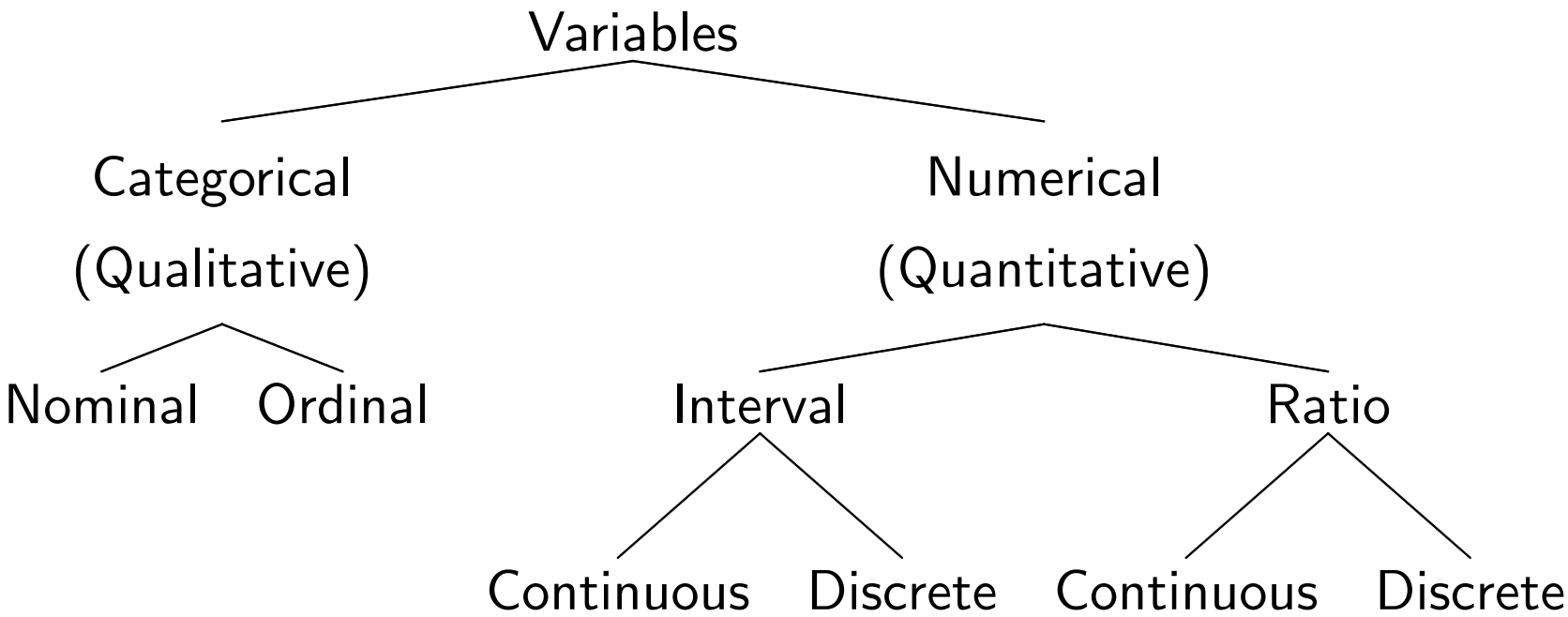
Please answer using your remote.

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

A Taxonomy of Variables



From Weakest to Strongest

Categorical

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

From Weakest to Strongest

Categorical

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

- **Nominal** no order to the categories

From Weakest to Strongest

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

From Weakest to Strongest

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

From Weakest to Strongest

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

Interval only differences meaningful, no natural zero

From Weakest to Strongest

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

Interval only differences meaningful, no natural zero

Ratio differences and ratios meaningful, natural zero

And For Numerical Variables (interval or ratio)...

Discrete

Takes value from discrete set of numbers, typically count data

And For Numerical Variables (interval or ratio)...

Discrete

Takes value from discrete set of numbers, typically count data

Continuous

Value could be any real number within some range (even though *measurements* are made with finite precision)

Note that in R, categorical variables are called *factors*

What kind of variable is...

...Handspan?

(a) Nominal

(b) Ordinal

(c) Interval

(d) Ratio

What kind of variable is...

...Temperature?

(a) Nominal

(b) Ordinal

(c) Interval

(d) Ratio

What kind of variable is...

...Eye Color?

(a) Nominal

(b) Ordinal

(c) Interval

(d) Ratio

What kind of variable?

On course evaluations you can rate your professor as follows:

0 = Poor, 1 = Fair, 2 = Good, 3 = Very Good, 4 = Excellent.

What kind of data is your rating?

(a) Nominal

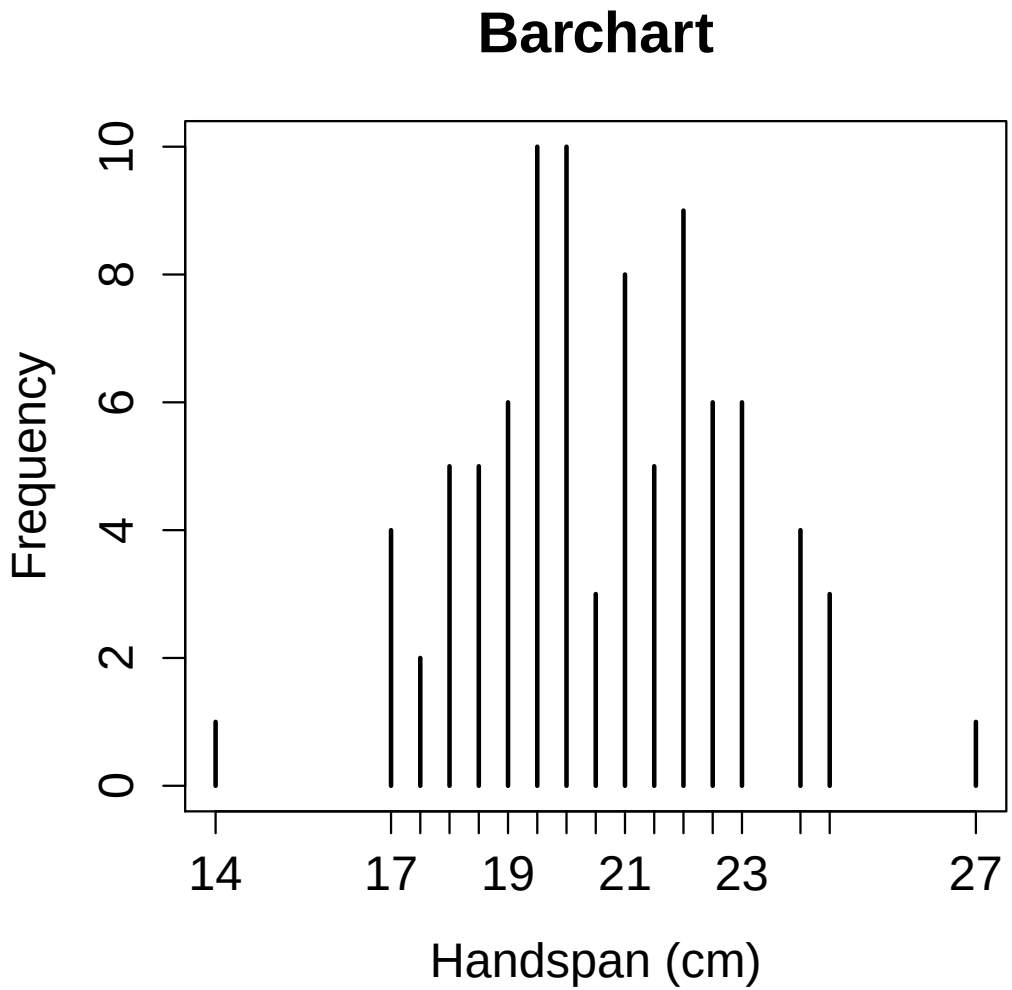
(b) Ordinal

(c) Interval

(d) Ratio

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"

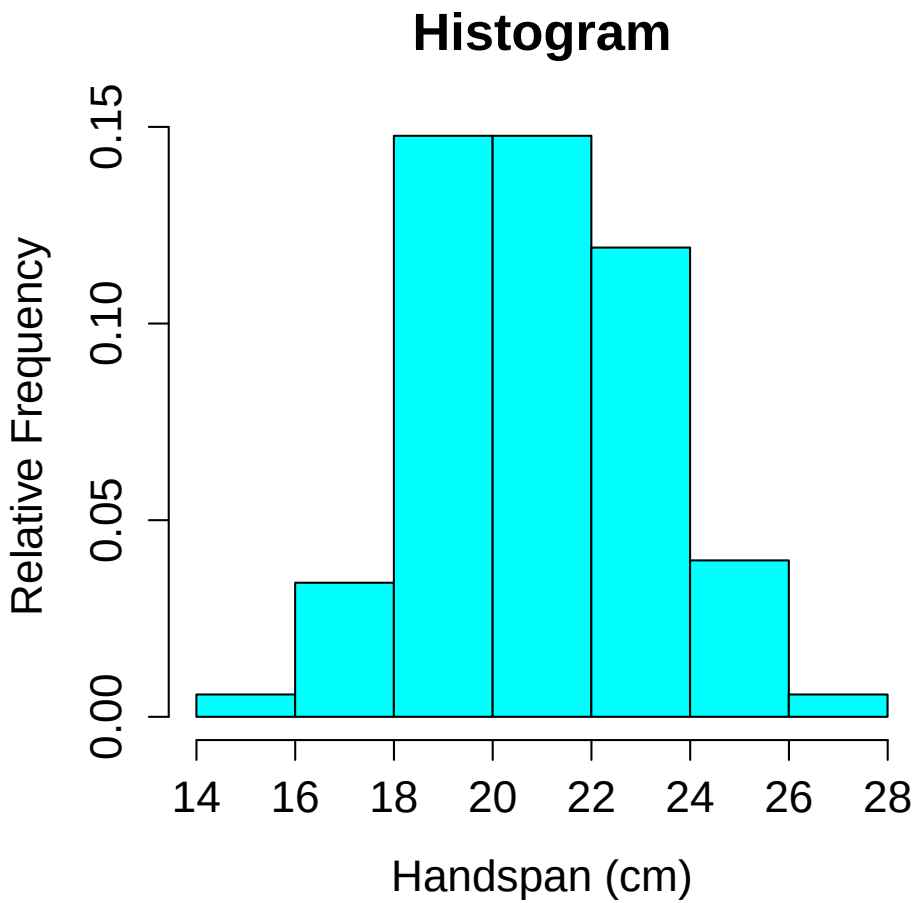
cm	Freq.	Rel. Freq.
14.0	1	0.01
17.0	4	0.05
17.5	2	0.02
18.0	5	0.06
18.5	5	0.06
19.0	6	0.07
19.5	10	0.11
20.0	10	0.11
20.5	3	0.03
21.0	8	0.09
21.5	5	0.06
22.0	9	0.10
22.5	6	0.07
23.0	6	0.07
24.0	4	0.05
24.5	3	0.03
27.0	1	0.01
	$n = 88$	1.00

Group data into non-overlapping bins of equal width:

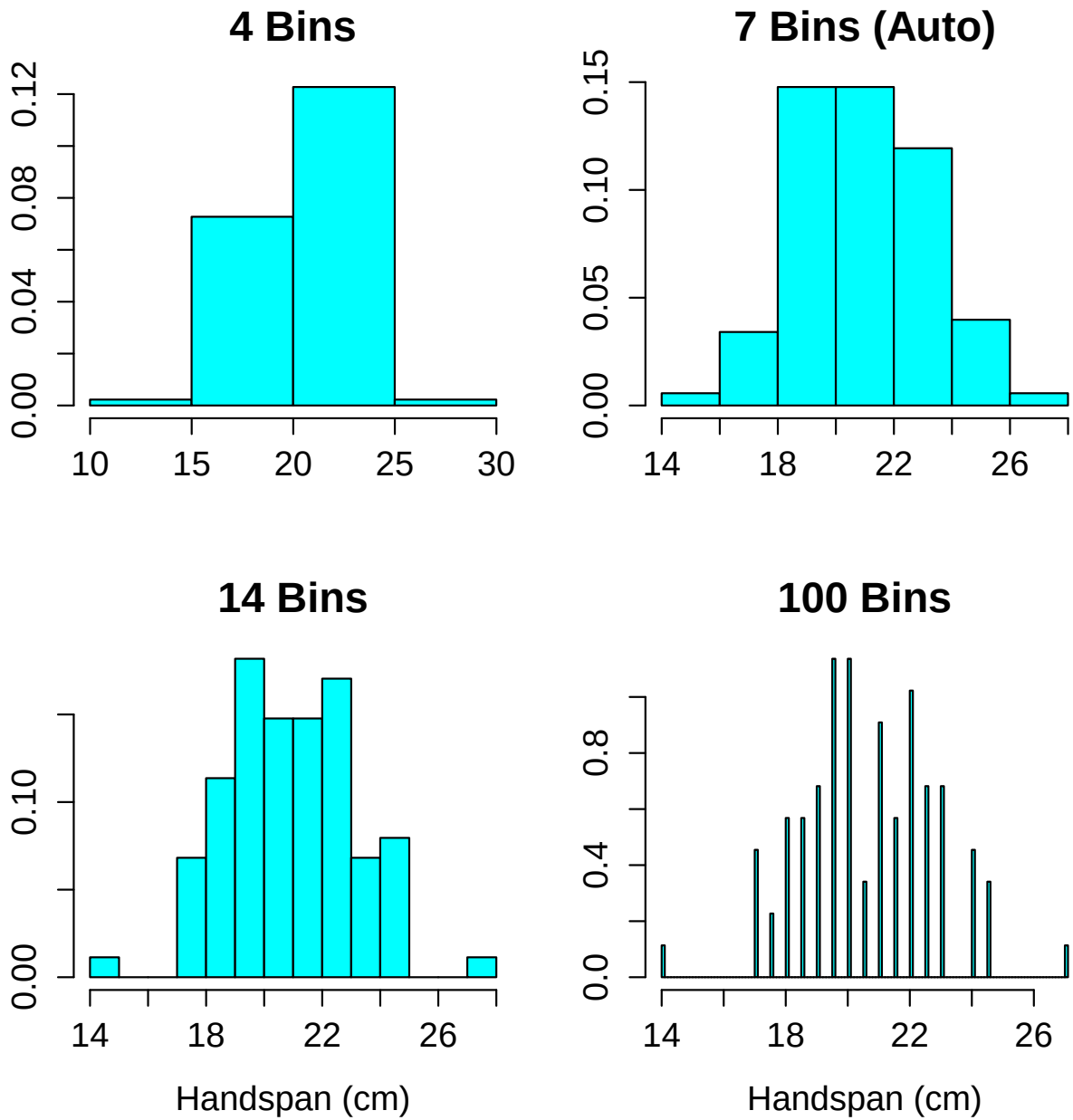
Bins	Freq.	Rel. Freq.
$[14, 16)$	1	0.01
$[16, 18)$	6	0.07
$[18, 20)$	26	0.30
$[20, 22)$	26	0.30
$[22, 24)$	21	0.24
$[24, 26)$	7	0.08
$[26, 28)$	1	0.01
	$n = 88$	1.00

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)

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)

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

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

Summary Statistic: Numerical Summary of Sample

1. Measures of Central Tendency
   - Mean
   - Median
2. Measures of Spread
   - Variance
   - Standard Deviation
   - Range
   - Interquartile Range (IQR)

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

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.

Measures of Central Tendency

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

Sample Mean

- $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
- Only for numeric data
- Works best when data are symmetric and there are no outliers

Measures of Central Tendency

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

Sample Mean

- $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
- Only for numeric data
- Works best when data are symmetric and there are no 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
- Robust to outliers and skewness

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)

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

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.)

What is an Outlier?

Outlier

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

Mean is Sensitive to Outliers, Median Isn't

First Dataset: 1 2 3 4 5

Mean = 3, Median = 3

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

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.

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

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

Q_1 = value in the $0.25(n+1)^{th}$ ordered position

An Example: n = 12

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

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

An Example: n = 12

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

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

An Example: n = 12

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

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

Student Debt

Guess the 90th percentile of student loan debt in the U.S. That is, guess the amount of money such that 10% college students graduate with *more* than this amount of debt and 90% graduate with less than or equal to this amount of debt.

Student Debt

Would you guess that the median amount of student loan debt in the U.S. is above, below, or equal to the mean amount?

(a) Median $>$ Mean

(b) Median $=$ Mean

(c) Median $<$ Mean

Source: Avery & Turner (2012)

Table 4

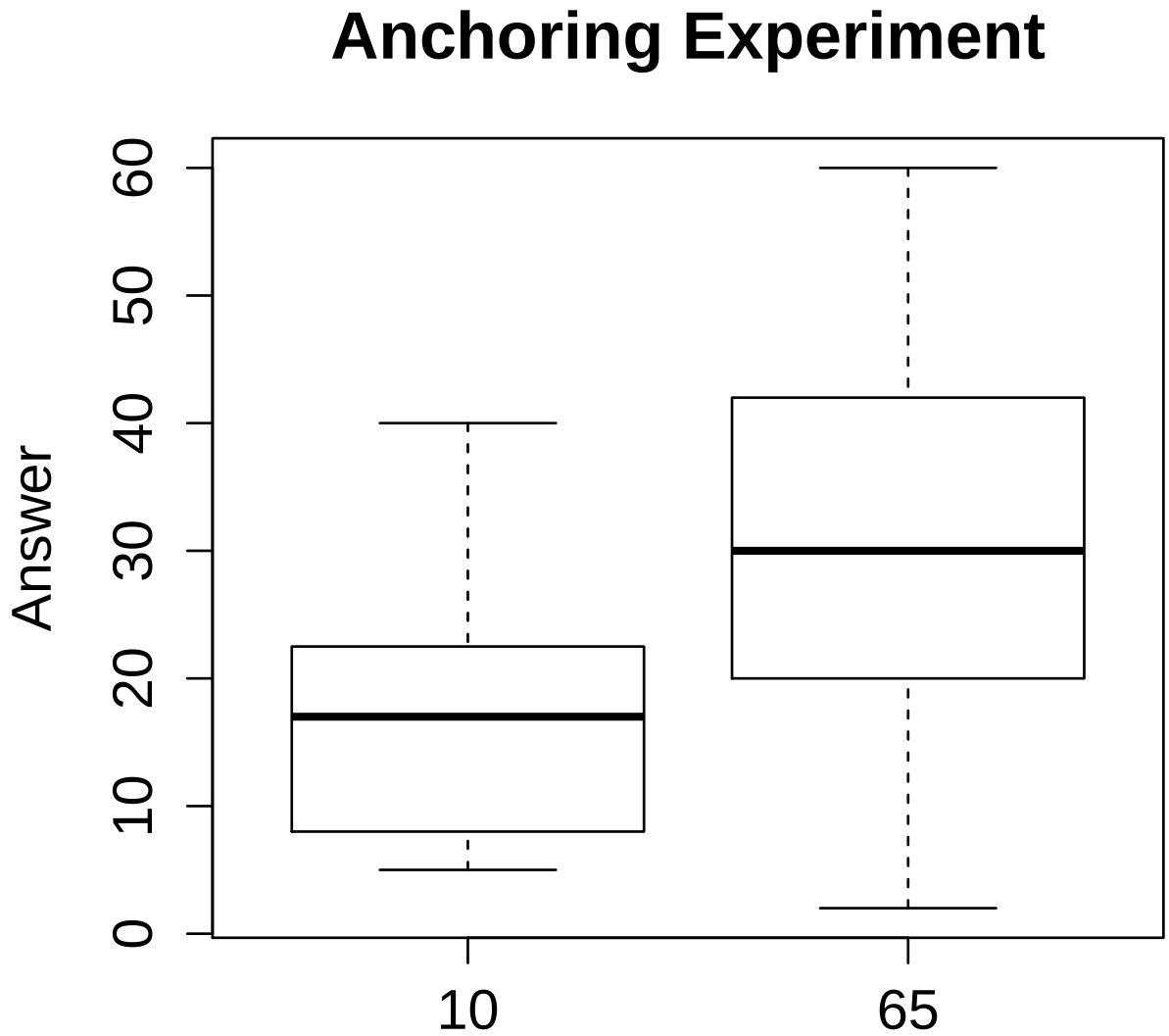
Borrowing Distribution after Six Years, by Degree Type and First Institution

	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				
10th	$0	$0	$0	$0
25th	$0	$0	$6,376	$0
50th	$6,000	$11,500	$13,961	$0
75th	$19,000	$24,750	$28,863	$6,625
90th	$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				
10th	$0	$0	$12,000	$0
25th	$0	$0	$30,000	$0
50th	$7,500	$15,500	$45,000	$11,971
75th	$20,000	$27,000	$50,000	$23,265
90th	$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

Range

Maximum Observation - Minimum Observation

Measures of Variability/Spread

Range

Maximum Observation - Minimum Observation

Interquartile Range (IQR)

IQR$= Q_3 - Q_1$

Measures of Variability/Spread

Range

Maximum Observation - Minimum Observation

Interquartile Range (IQR)

IQR$= Q_3 - Q_1$

Variance

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

Measures of Variability/Spread

Range

Maximum Observation - Minimum Observation

Interquartile Range (IQR)

IQR$= Q_3 - Q_1$

Variance

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

Standard Deviation

$$s = \sqrt{s^2}$$

Variance

Essentially the average squared distance from the mean. Sensitive to both skewness and outliers.

Variance

Essentially the average squared distance from the mean. Sensitive to both skewness and outliers.

Standard Deviation

$\sqrt{\text{Variance}}$, but more convenient since same units as data

Variance

Essentially the average squared distance from the mean. Sensitive to both skewness and outliers.

Standard Deviation

$\sqrt{\text{Variance}}$, but more convenient since same units as data

Range

Difference between larges and smallest observations. *Very* sensitive to outliers. Displayed in boxplot.

Variance

Essentially the average squared distance from the mean. Sensitive to both skewness and outliers.

Standard Deviation

$\sqrt{\text{Variance}}$, but more convenient since same units as data

Range

Difference between larges and smallest observations. *Very* sensitive to outliers. Displayed in boxplot.

Interquartile Range

Range of middle 50% of the data. Insensitive to outliers, skewness. Displayed in boxplot.