Economics 103 – Statistics for Economists

Rossa O'Keeffe-O'Donovan

(heavily indebted to Francis J. DiTraglia)

University of Pennsylvania

Lecture # 2

Today: What we will cover

Summary Statistics

Goals:

- Understand each of the main summary statistics:
  - What they measure
  - How they are constructed

Z-scores

'Standardize' data to see how extreme an observation is

Relationships between variables

Goals:

- Understand crosstabs, covariance, correlation

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 (tomorrow)

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

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

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

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

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

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

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.

Measures of Variability/Spread

Range

Maximum Observation - Minimum Observation

Interquartile Range (IQR)

IQR$= Q_3 - Q_1$

(Sample) Variance

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

(Sample) Standard Deviation

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

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.

Interquartile Range

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

Sample Variance - 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!

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

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

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

What do the values indicate?

- Zero $\Rightarrow$ symmetry
- Positive $\Rightarrow$ right-skewed
- Negative $\Rightarrow$ left-skewed

Where does the formula come from?

- Why cubed? To get the desired sign.
- Why divide by s^3? So that skewness is unitless

Skewness – A Measure of Symmetry

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

Rules of Thumb

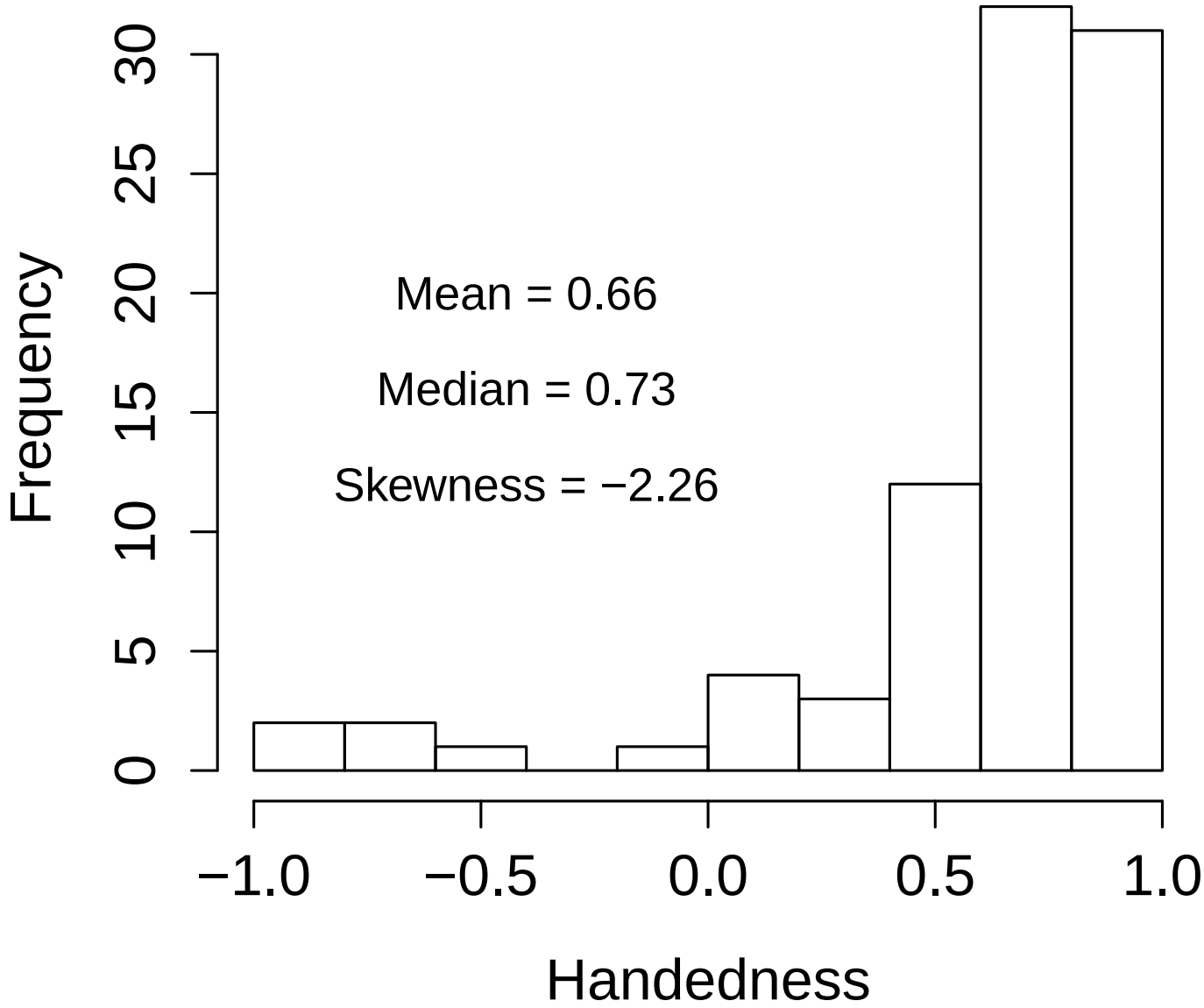
Typically (but not always):

- right-skewed $\Rightarrow$ mean $>$ median
- left-skewed $\Rightarrow$ mean $<$ median
- no skew $\Rightarrow$ mean $\approx$ median

Direction of the 'tail':

- right/positive skew $\Rightarrow$ long tail goes off in positive direction
- left/negative skew $\Rightarrow$ long tail goes off in negative direction


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
Frequency
0
5
10
15
20
25
30
Mean = 0.66
Median = 0.73
Skewness = −2.26
−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 (for later):

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 because 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 - we'll come back to this.

Z-scores: How "extreme" is an observation?

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 (99.7%) 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 10.7in (27cm)

(b) Height of 78in (198cm)

The means are:

- Mean handspan: 8.1in (20.6cm)
- Mean height: 67.6in (171.7cm)

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

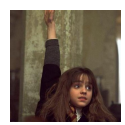
What is the sample mean of the z-scores?

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

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?



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? 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? (Units of x) $\times$ (Units of y)

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.	$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$	$r = \frac{s_{xy}}{s_x s_y}$

R demonstration

See links on last page of syllabus to set up R:

- Try out R at `http://tryr.codeschool.com/` (up to and including Level 6)
- Download and install R from `http://cran.r-project.org/`.
- Download and install RStudio by visiting `http://rstudio.org/download/desktop` and clicking the link listed under "Recommended for Your System."

Some handy R tips

- Write code into your script so you can easily edit and save it
- Highlight lines of code and click run (or press CTRL+Enter)
- Use the workspace
- Comments - you can use # to comment your code and describe what you are doing
- If you are having difficulty - ask on Piazza!