

Economics 103 – Statistics for Economists

Francis J. DiTraglia

University of Pennsylvania

Lecture # 3

Recall From Last Lecture:

Range

Maximum Observation - Minimum Observation

Interquartile Range (IQR)

$$\text{IQR} = Q_3 - Q_1$$

Variance

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

Standard Deviation

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

Recall From Last Lecture:

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 largest 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.

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

Why Squares?

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

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 right-skewed, negative left-skewed.

Why cubed?

To get the desired sign.

Why divide by s^3 ?

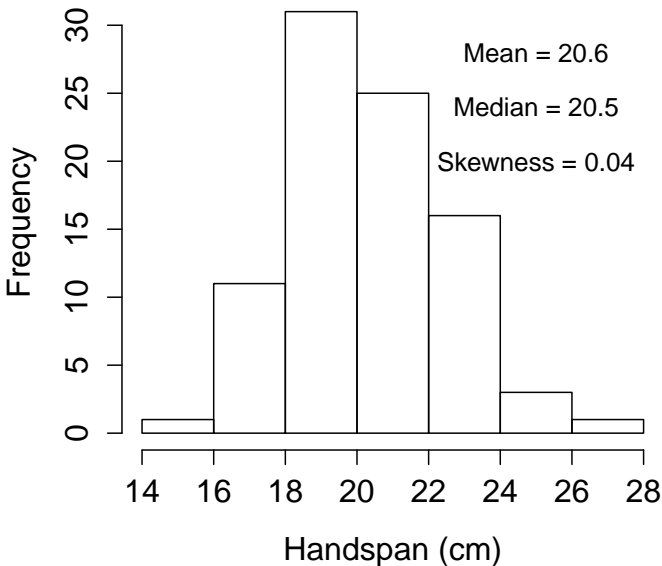
So that skewness is unitless

Rule of Thumb

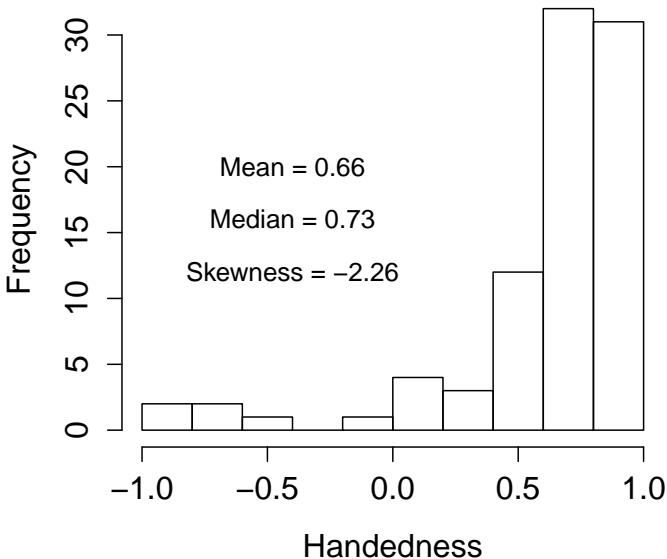
Typically (but not always), right-skewed \Rightarrow mean $>$ median

left-skewed \Rightarrow mean $<$ median

Histogram of Handspan



Histogram of 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, \dots, x_N

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

Sample: x_1, x_2, \dots, 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, \dots, 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)

<i>Eye Color</i>	<i>Sex</i>		Total
	Male	Female	
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, \dots, x_n and y_1, \dots, 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

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

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

$$s_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}$$

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

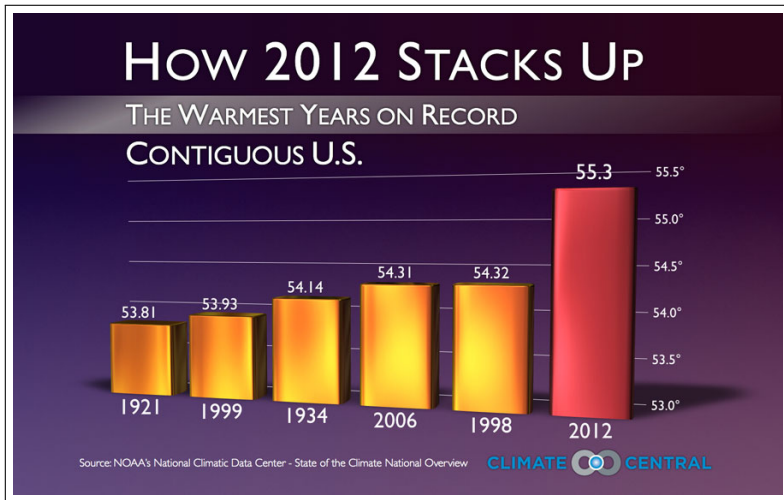
N individuals in the Population, n individuals in the Sample:

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

Some Thoughts on Statistical Graphics

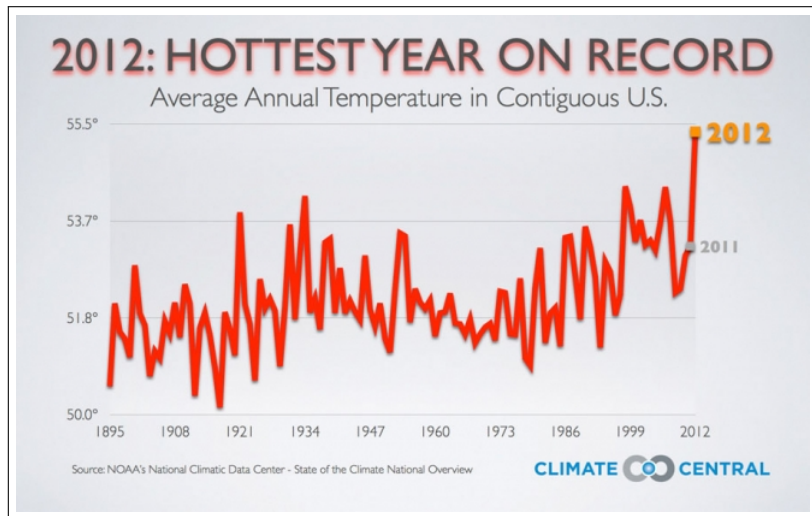
What's Wrong with This Graph?

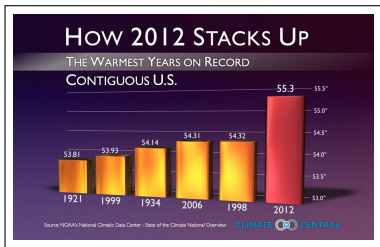
Source: [Climate Central](#)



Why is this one better?

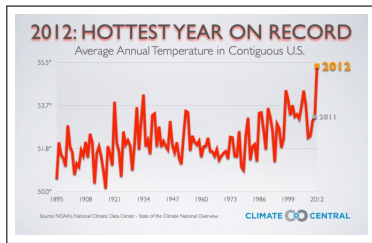
Source: [Climate Central](#)





Bad Graph:

- ▶ Unnatural ordering of years
- ▶ 3-D effect (perspective) and y-axis exaggerate 2012
- ▶ Why *six* warmest years?
- ▶ Lack of context: what about other 111 years?
- ▶ Gives no information on overall climate variability



Good Graph:

- ▶ Shows all years in order (context)
- ▶ Shows variability and trend
- ▶ y-axis chosen to comfortably fit all observations
- ▶ 2-D rather than 3-D

Why Graphics?

Humans naturally skilled at interpreting spatial information

Some Guidelines

- ▶ Use **distance** rather than area or perspective
- ▶ Avoid clutter (chartjunk)
- ▶ Use meaningful order where possible
- ▶ Make a **visual argument**, not a work of art

Best Statistical Graphic Ever

Charles Joseph Minard, 1861

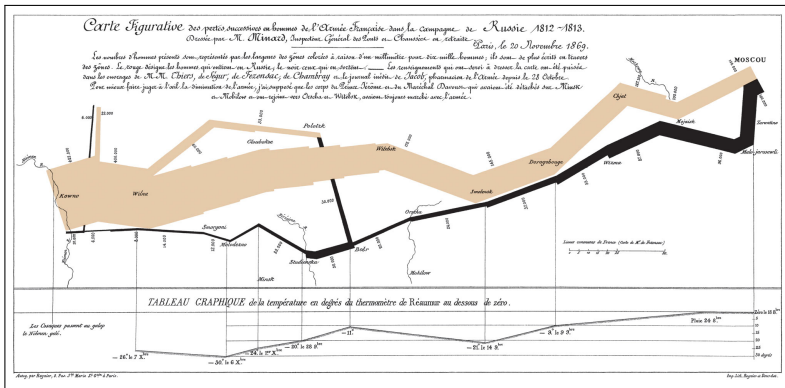
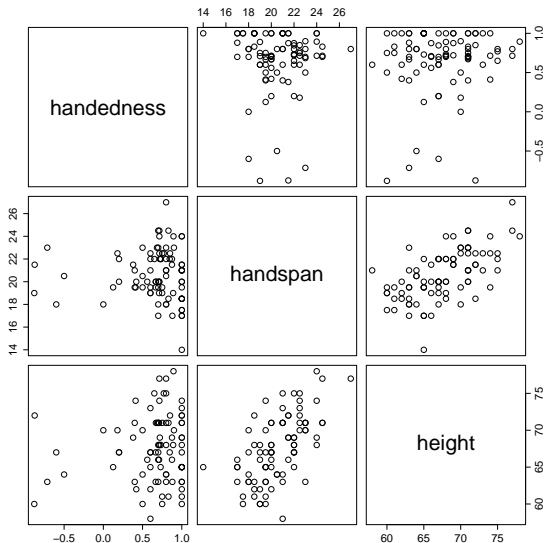


Figure : Napoleon's disastrous Russian Campaign of 1812. Depicts six variables: temperature, location (two dimensions), direction, number of troops, and date.

Pairs Plot – Handedness, Handspan and Height



Bubblechart – 2005 Crime Rates by State

Source: [Flowing Data](#)

**Burglaries per
100,000 population**

