# Summary Statistics

**Nicholas Mattei,** *Tulane University*

*CMPS3660 – Introduction to Data Science – Fall 2019*

**https://rebrand.ly/TUDataScience**

Thanks to William Green (NYU Stern IOMS)
and Hector Corrada Bravo (UMD)
and Zico Kolter [CMU}

# Announcements

- Rest of Semester Schedule

- Lab 8 Review

- More EDA/Stats!

# The Data LifeCycle

Data Collection → Data Processing → Exploratory Analysis & Data Visualization → Analysis, Hypothesis Testing, & ML → Insight & Policy Decision
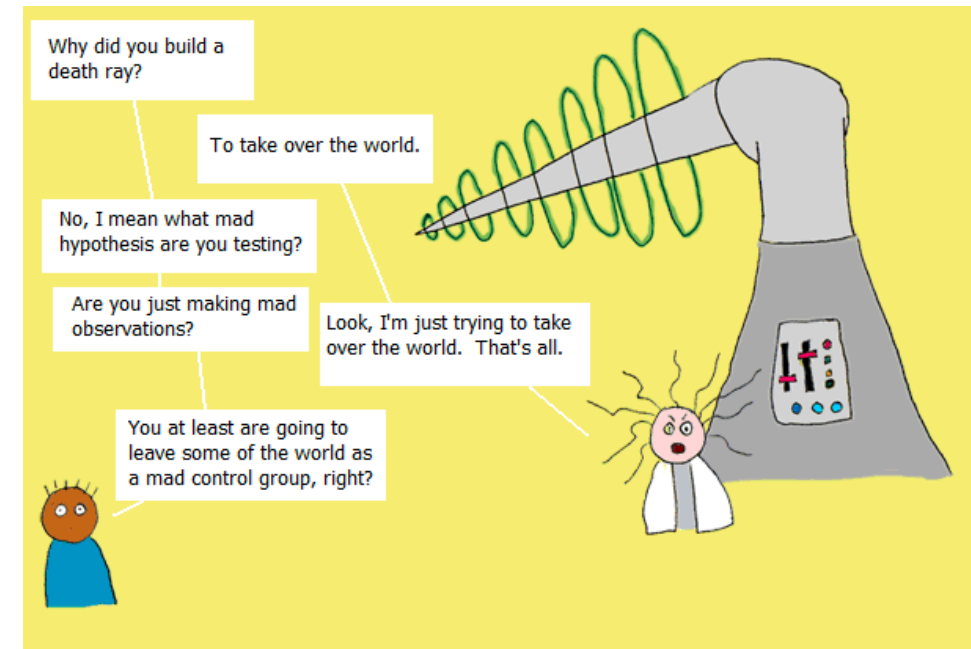
Today
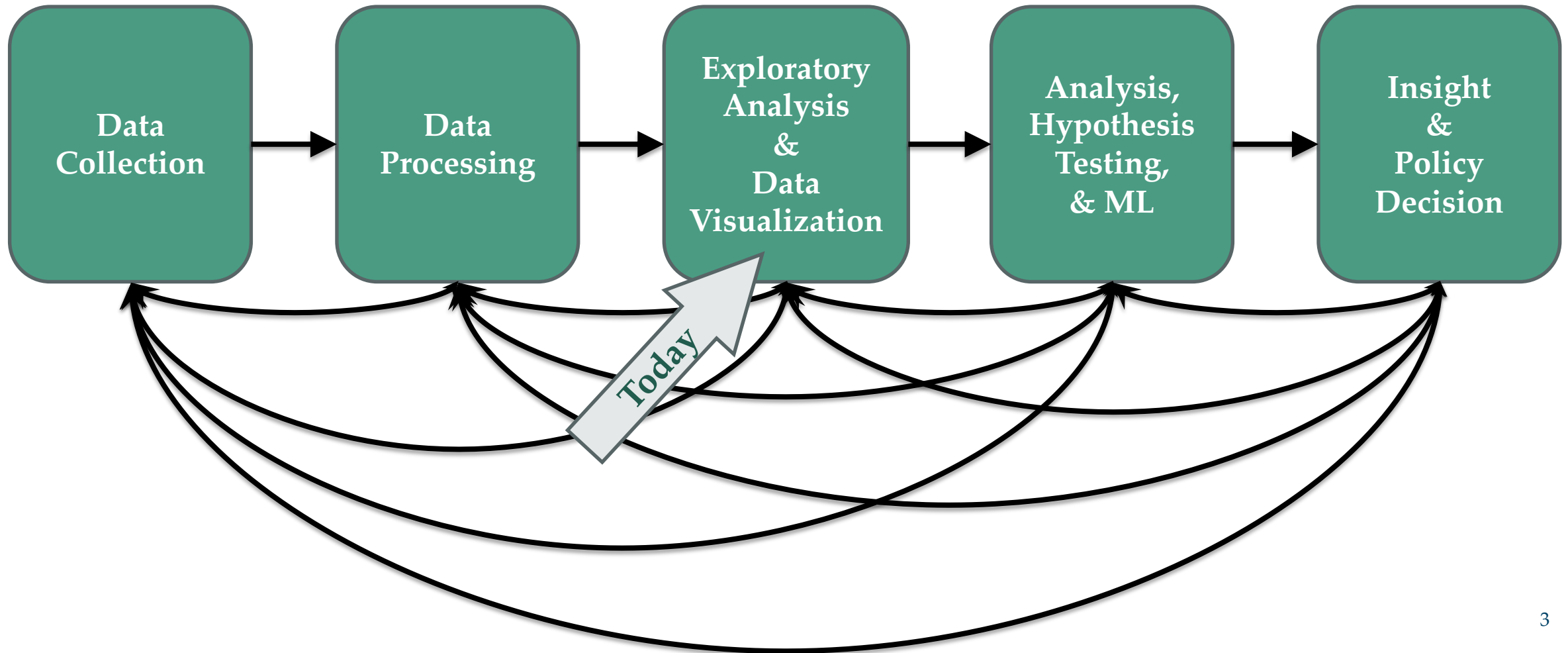
# Exploratory Data Analysis

- Seen so far:
  - Manipulations that prepare datasets into tidy form.
  - Join tables and compute summaries.
  - Form relationships between different entities.

- **EDA** is the last step before Big Time Statistics and ML™:
  - Want to quickly "get a feel" for the data through summary statistics, visualization…
  - Spot nuances like skew, how distributed the data is, trends, how pairs of variables interact, problems.
  - Suggests which Stats/ML assumptions to make and approaches to take.

# But Why?

- We've only done a bit of "Science" so far…
  - Typically, "Science" is "determining some truth about the world…."

- Suppose you work for a company that is considering a redesign of their website; does their new design (design B) offer any statistical advantage to their current design (design A)?

Why did you build a death ray?

To take over the world.

No, I mean what mad hypothesis are you testing?

Are you just making mad observations?

Look, I'm just trying to take over the world. That's all.

You at least are going to leave some of the world as a mad control group, right?

Sad truth: Most "mad scientists" are actually just mad engineers

http://cowbirdsinlove.com/46

- In a linear regression, does a certain variable impact the response?
  - Does energy consumption depend on whether or not a day is a weekday or weekend?

- Both: concerned with making actual statements about the nature of the world.

[ZK]

# Population v. Sample Statistics

**Population**

**Random Sample**
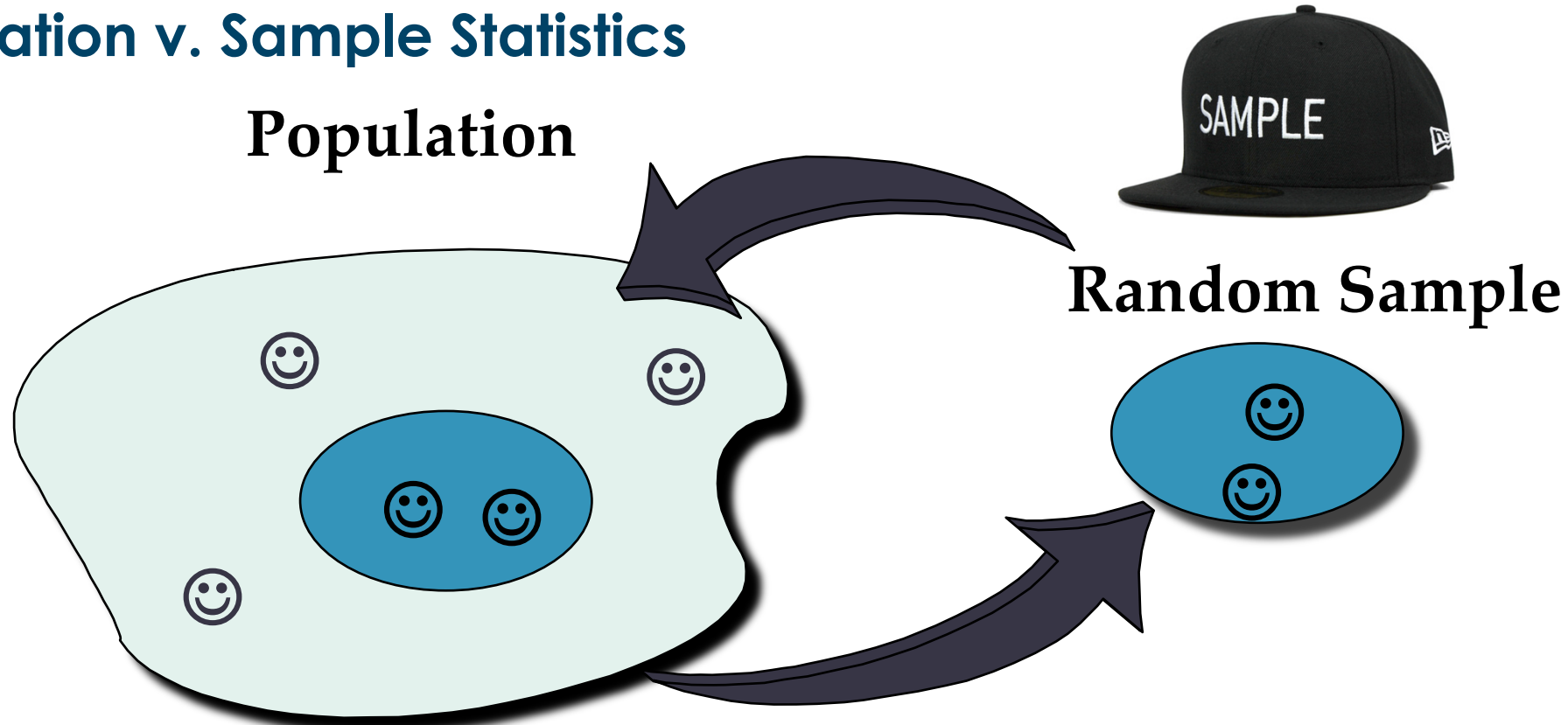
SAMPLE

**Population**

| | |
|---|---|
| **Mean** | $\mu = \mathbf{E}[X]$ |
| **Variance** | $\sigma^2 = \mathbf{E}[(X - \mu)^2]$ |

**Sample**

| | |
|---|---|
| **Mean** | $\bar{x} = \dfrac{1}{m} \sum\limits_{i=1}^{m} x^{(i)}$ |
| **Variance** | $s^2 = \dfrac{1}{m-1} \sum\limits_{i=1}^{m} (x^{(i)} - \bar{x})^2$ |

[ZK]

# Sample mean as random variable

The same mean is an empirical average over $m$ independent samples from the distribution; it can also be considered as a random variable

This new random variable has the mean and variance

$$\mathbf{E}[\bar{x}] = \mathbf{E}\left[\frac{1}{m}\sum_{i=1}^{m}x^{(i)}\right] = \frac{1}{m}\sum_{i=1}^{m}\mathbf{E}[X] = \mathbf{E}[X] = \mu$$

$$\mathbf{Var}[\bar{x}] = \mathbf{Var}\left[\frac{1}{m}\sum_{i=1}^{m}x^{(i)}\right] = \frac{1}{m^2}\sum_{i=1}^{m}\mathbf{Var}[X] = \frac{\sigma^2}{m}$$

where we used the fact that for *independent* random variables $X_1, X_2$
$$\mathbf{Var}[X_1 + X_2] = \mathbf{Var}[X_1] + \mathbf{Var}[X_2]$$

When estimating variance of sample, we use $s^2/m$ (the square root of this term is called the **standard error**)

8

[ZK]

# Central limit theorem

Central limit theorem states further that $\bar{x}$ (for "reasonably sized" samples, in practice $m \geq 30$) actually has a Gaussian distribution *regardless of the distribution of $X$*

$$\bar{x} \to \mathcal{N}\left(\mu, \frac{\sigma^2}{m}\right) \text{ (or equivalently) } \frac{\bar{x} - \mu}{\sigma/m^{1/2}} \to \mathcal{N}(0,1)$$

In practice, for $m < 30$ and for estimating $\sigma^2$ using sample variance, we use a Student's t-distribution with $m - 1$ degrees of freedom

$$\frac{\bar{x} - \mu}{s/m^{1/2}} \to T_{m-1}, \qquad p(x; \nu) \propto \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

[ZK]

# Last Week's Lesson

**Having a really big sample does not assure you of an accurate result.**

**It may assure you of a really solid, really bad (inaccurate) result.**

- Not all randomness is create equal when it comes to random sampling of a population:
  - Ask **why** data are missing!  MCAR, MAR, MNAR.
  - Ask how the data were collected.

# Today's Lesson (and Labs): Summary Statistics

- Part of **descriptive statistics**, used to summarize data:
  - Convey lots of information with extreme simplicity

- Descriptive statistics for a variable:
  - Measures of location: mean, median, mode
  - Measure of dispersion: variance, standard deviation

- Measuring correlation of two variables:
  - Understanding correlation
  - Measuring correlation
  - Scatter plots and regression

# Measures of Location

> These are 30 hours of average defect data on sets of circuit boards. Roughly what is the typical value?
>
> ```
> 1.45  1.65  1.50  2.25  1.65  1.60  2.30  2.20  2.70  1.70
> 2.35  1.70  1.90  1.45  1.40  2.60  2.05  1.70  1.05  2.35
> 1.90  1.55  1.95  1.60  2.05  2.05  1.70  2.30  1.30  2.35
> ```

- Location and central tendency
  - There exists a distribution of values
  - We are interested in the "center" of the distribution

- Two measures are the **sample mean** and the **sample median**
- They look similar, and measure the same thing

- They differ systematically (and predictably) when the data are not **symmetric.**

[WG]

# The Mean of Aggregate Data

| State | Listing | IncomePC | State | Listing | IncomePC | State | Listing | IncomePC |
|---|---|---|---|---|---|---|---|---|
| Hawaii | 896800 | 24057 | Rhode Island | 432534 | 22251 | Texas | 266388 | 19857 |
| California | 713864 | 22493 | Delaware | 420845 | 22828 | Mississippi | 255774 | 15838 |
| New York | 668578 | 25999 | Oregon | 417551 | 20419 | Tennessee | 255064 | 19482 |
| Connecticut | 654859 | 29402 | Idaho | 415885 | 18231 | Wisconsin | 243006 | 21019 |
| Dist. Columbia | 577921 | 31136 | Illinois | 377683 | 23784 | Michigan | 241107 | 22333 |
| Nevada | 549187 | 24023 | New Hampshire | 361691 | 23434 | Missouri | 221773 | 20717 |
| New Jersey | 529201 | 23038 | New Mexico | 358369 | 17106 | South Dakota | 220708 | 19577 |
| Massachusetts | 521769 | 25616 | Vermont | 346469 | 20224 | West Virginia | 219275 | 17208 |
| Wyoming | 499674 | 20436 | South Carolina | 340066 | 17695 | Arkansas | 217659 | 16898 |
| Maryland | 480578 | 24933 | North Carolina | 330432 | 19669 | Ohio | 209189 | 20928 |
| Utah | 475060 | 17043 | Georgia | 326699 | 20251 | Kentucky | 208391 | 17807 |
| Colorado | 467979 | 22333 | Alaska | 324774 | 23788 | Oklahoma | 203926 | 17744 |
| Arizona | 448791 | 19001 | Minnesota | 306009 | 22453 | Kansas | 201389 | 20896 |
| Florida | 447698 | 21677 | Maine | 299796 | 19663 | Indiana | 200683 | 20378 |
| Montana | 446584 | 17865 | Pennsylvania | 295133 | 22324 | Iowa | 184999 | 20265 |
| Virginia | 443618 | 22594 | Louisiana | 280631 | 17651 | North Dakota | 173977 | 18546 |
| Washington | 440542 | 22610 | Alabama | 269135 | 18010 | Nebraska | 164326 | 20488 |

**Average list price:**
**1/51 ($898,800 + $713,864 + … + $164,326) = $369,687**

[WG]

# Averaging Averages?

- Hawaii's average listing = $896,800
- Hawaii's population = 1,275,194
- Illinois' average listing = $377,683
- Illinois' population = 12,763,371

- Illinois and Hawaii each get an equal weight of 1/51 = .019607 when the mean is computed.
- Looks like Hawaii is getting too much influence ...

[WG]

# Weighted Average

Simple average $= \overline{\text{Listing}} = \sum_{\text{States}} \text{Weight}_{\text{State}} \; \text{Listing}_{\text{State}}$

$\text{Weight} = \dfrac{1}{51} = .019607$

Illinois is 10 times as big as Hawaii.  Suppose we use weights that are in proportion to the state's population. (The weights sum to 1.0.)

$\text{Weight}_{\text{State}}$ varies from .001717 for Wyoming to .121899 for California

New average is \$409,234 compared to \$369,687 without weights, an error of 11%!

**Sometimes an unequal weighting of the observations is necessary**

State population data: http://www.factmonster.com/ipka/A0004986.html

# Averages & Time Series

- Averaging **trending** time series is usually not helpful
- Mean changes completely depending on time interval
- What about **periodic** time series data ?????????

**Ask yourself:**
- **Does the mean over the entire observation period mean anything?**
- **Does it estimate anything meaningful?**



Time Series Plot of US Car Thefts
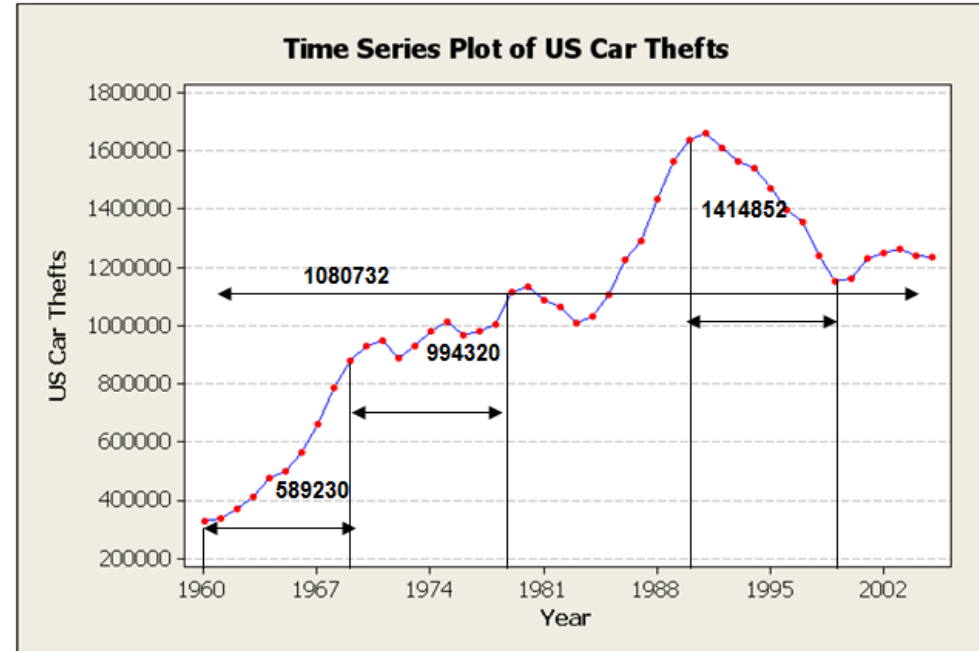
# The Sample Median

- Median:
  - Sort the data
  - Take the middle point*

- Odd number:
  - Central observation: Med[1,2,4,6,8,9,17]

- Even number:
  - Midpoint between the two central observations Med[1,2,4,6,8,9,14,17] = (6+8)/2=7

# What is the center?

- The mean and median measure the **central tendency** of data
- Generally, the **center** of of a dataset is a point in its range that is close to the data.
- Close?  Need a **distance metric** between two points $x_1$ and $x_2$.

- We've talked about some already!
  - Absolute deviation: $| x_1 - x_2 |$
  - Squared deviation: $(x_1 - x_2)^2$

- We'll define the center based on these metrics

# Dataset for this part

- 53,940 measurements of diamonds
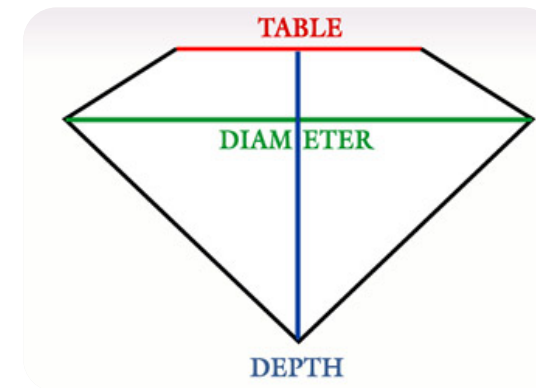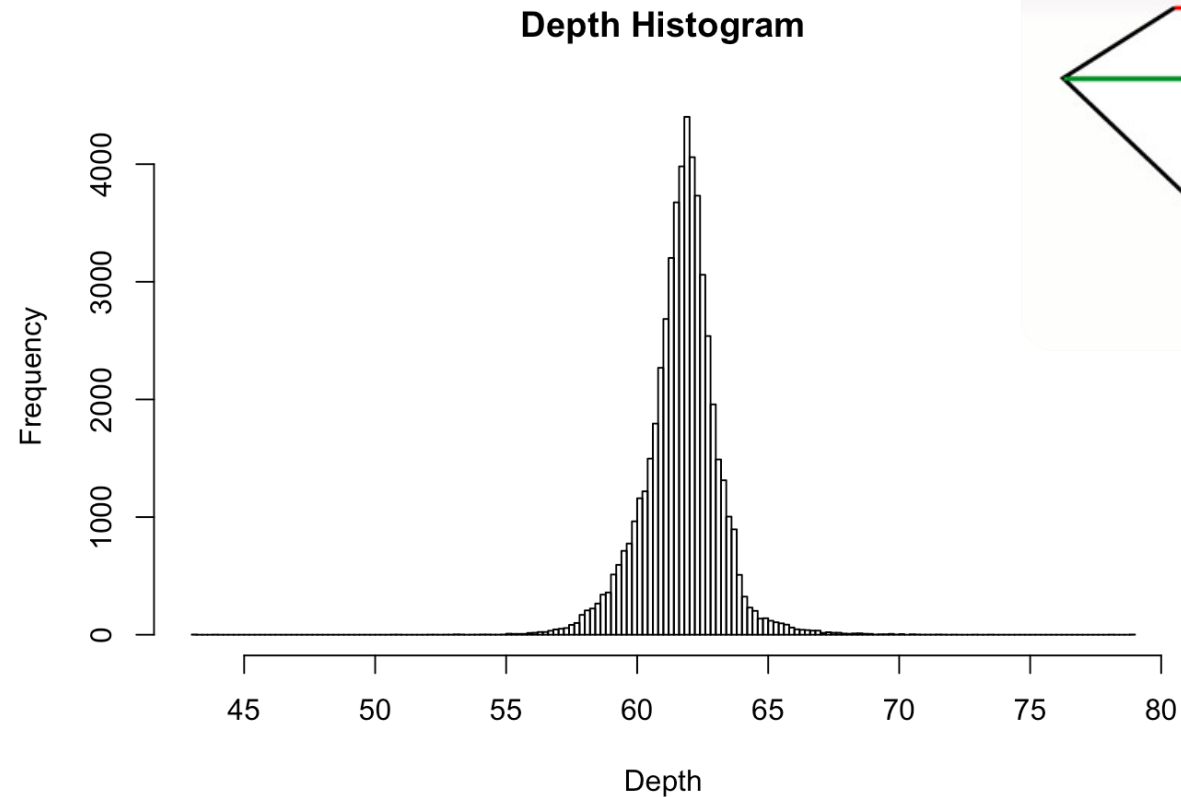
# The Mean Revisited

- Define a center point μ based on some function of the distance from each data point to that center point

  – Residual sum of squares (RSS) for a point μ:

$$RSS(\mu) = \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2$$



Residual Sum of Squares
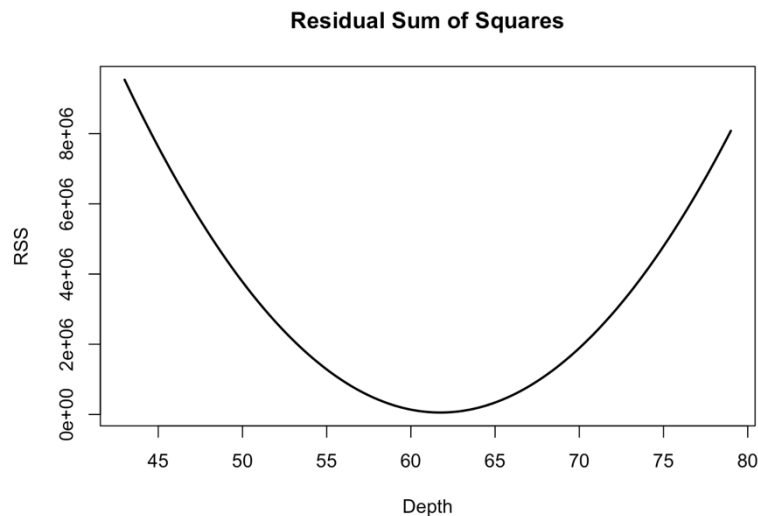
So what should our estimate of the "center" of this dataset be, based on the RSS metric?
?????????????

# The mean Revisited

- Want the point μ that minimizes the RSS   ??????????
- Find the derivative of RSS and set it to zero, solve for μ!

$$\frac{\partial}{\partial \mu} \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2 = \frac{1}{2} \sum_{i=1}^{n} \frac{\partial}{\partial \mu} (x_i - \mu)^2$$

$$= \frac{1}{2} \sum_{i=1}^{n} 2(x_i - \mu) \times (-1)$$

# The mean revisited

$$= \frac{1}{2} \sum_{i=1}^{n} 2(x_i - \mu) \times (-1)$$
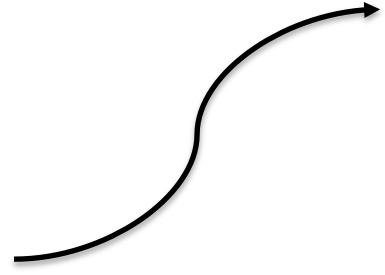
$$= \frac{1}{2} 2 \sum_{i=1}^{n} (\mu - x_i)$$

$$= \sum_{i=1}^{n} \mu - \sum_{i=1}^{n} x_i$$

$$= n\mu - \sum_{i=1}^{n} x_i$$
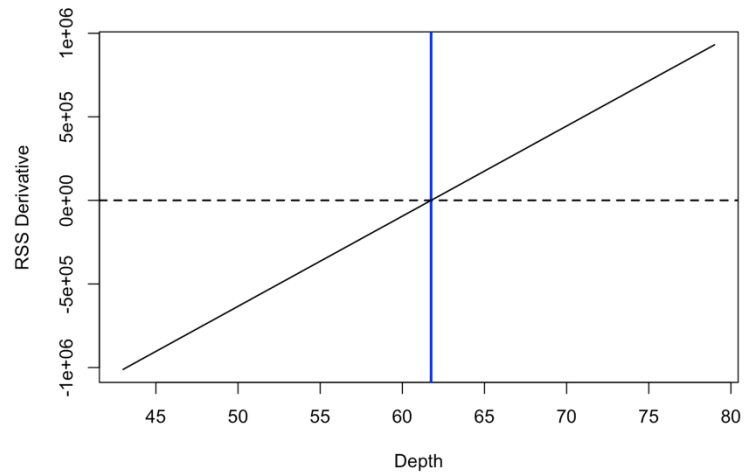
**Derivative of RSS**
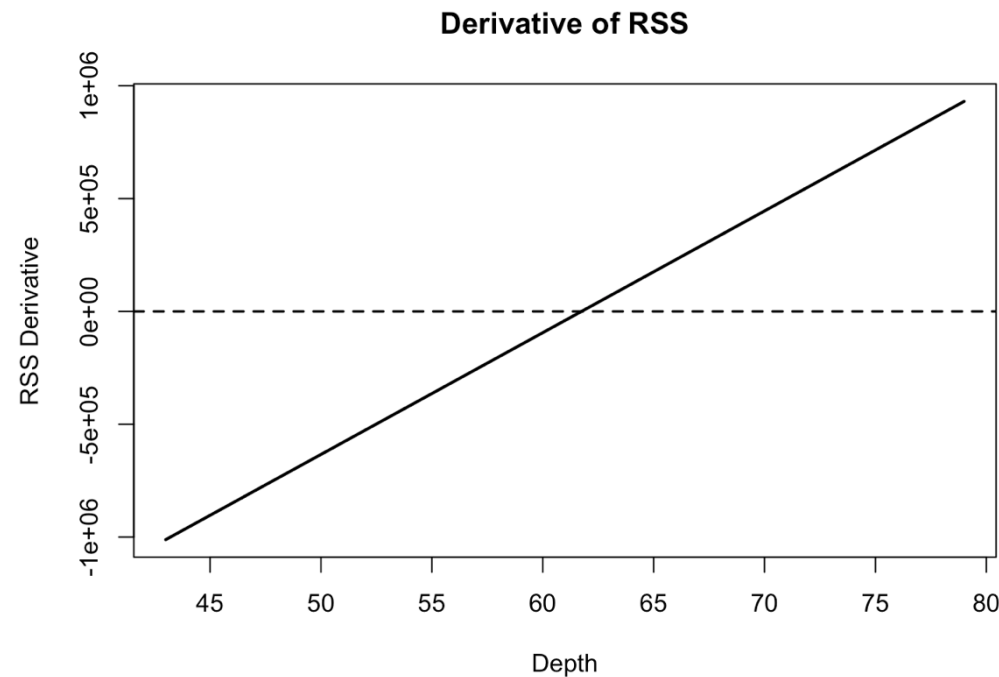
# The Mean Revisited

- Set the derivative to zero and solve for μ:

$$\frac{\partial}{\partial \mu} = 0$$

$$n\mu - \sum_{i=1}^{n} x_i = 0$$

$$n\mu = \sum_{i=1}^{n} x_i$$

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

The mean is the point μ that minimizes the RSS for a dataset.
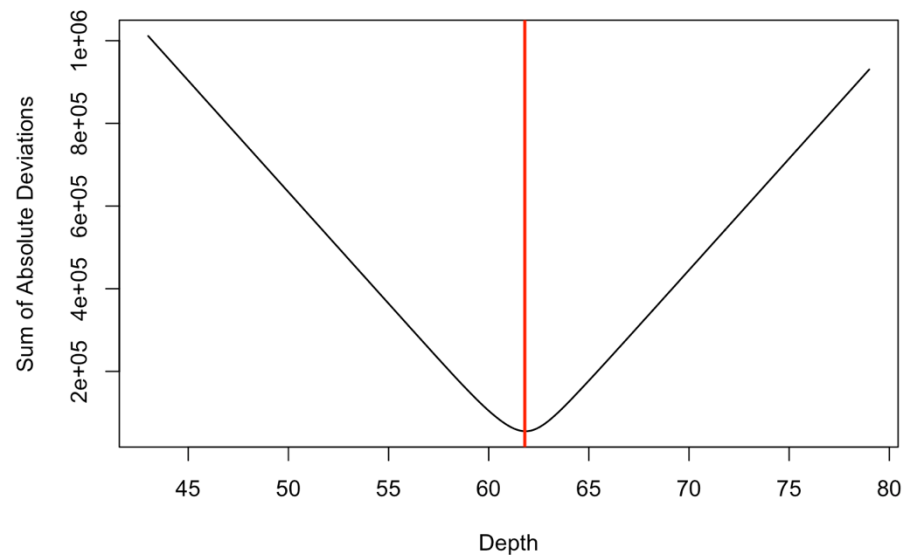
# The Mean Revisited

What about a weighted average ??????
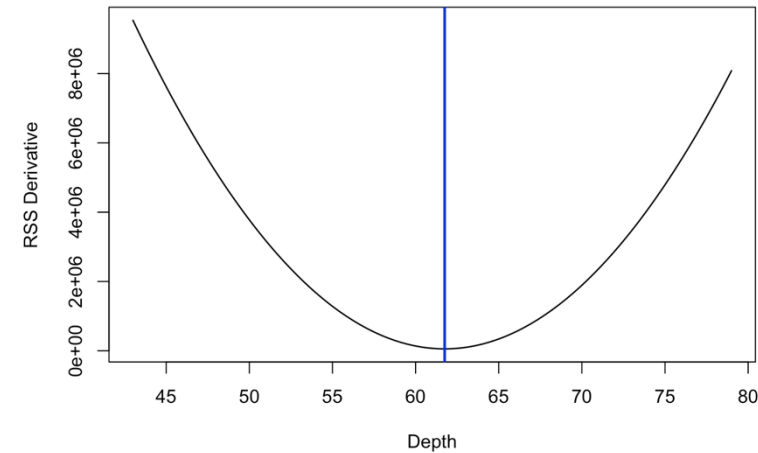


The mean is the point μ that minimizes the RSS for a dataset.
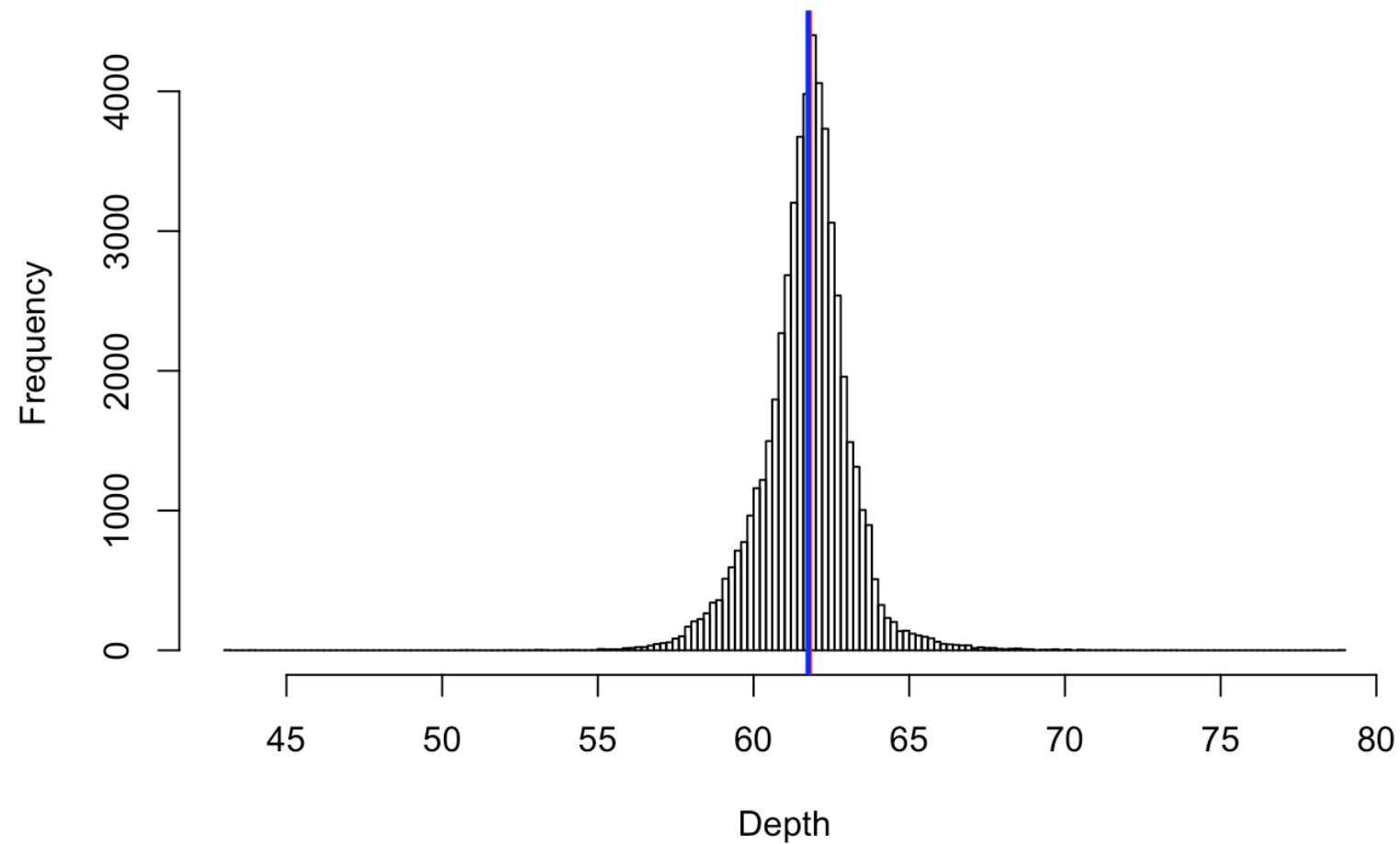
# The Median Revisited

- Define a center point $m$ based on some function of the distance from each data point to that center point.
  - The median $m$ minimizes the sum of absolute differences:
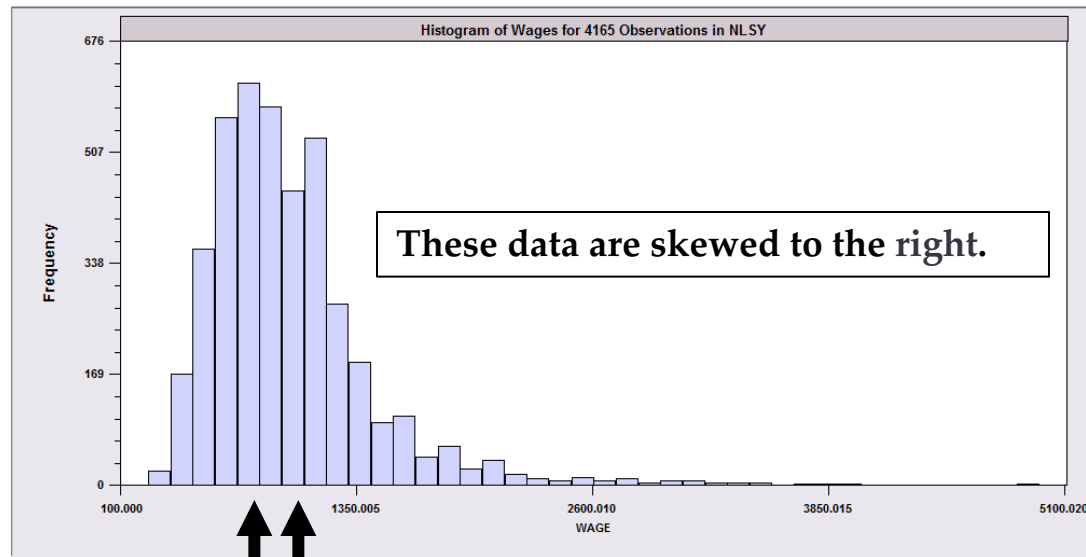
$$\sum_{i=1}^{n} |x_i - m|$$

# Mean != Median

**Depth Histogram**

# Skewed Data



Histogram of Wages for 4165 Observations in NLSY

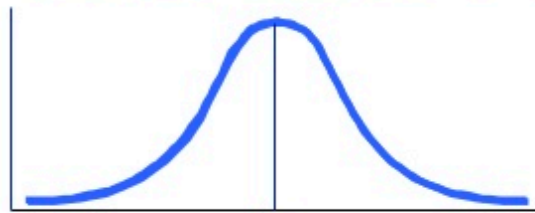**These data are skewed to the right.**

Median Mean

Monthly Earnings
N = 595,
Median = 800
Mean   = 883

**The mean will exceed the median when the distribution is skewed to the right.**

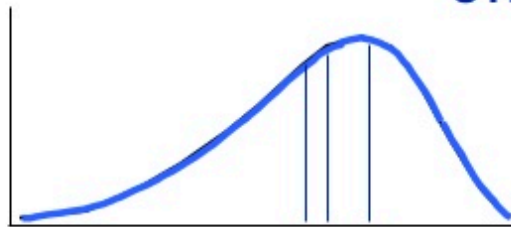**Skewness is in the direction of the long tail**
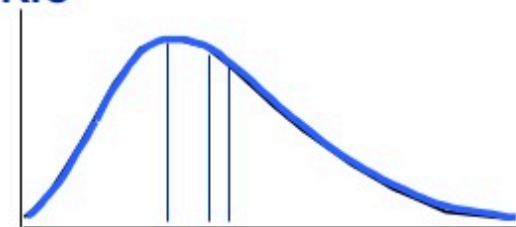
# A Visual Guide to Skew



- We skew in the direction of the long tail.

- Image: https://guides.douglascollege.ca/c.php?g=408742&p=2970198

41

# Skewness

- Extreme observations distort means but not medians.
- Outlying observations distort the mean:
  - Med  [1,2,4,6,8,9,17] = 6
  - Mean[1,2,4,6,8,9,17] = 6.714
  - Med  [1,2,4,6,8,9,17000] = 6 (still)
  - Mean[1,2,4,6,8,9,17000] = 2432.8 (!)

- Typically occurs when there are some outlying observations, such as in cross sections of income or wealth and/or when the sample is not very large.

**The New York Times**

## Business Day

DATAPOINTS

# Income Gap Grows Wider (and Faster)

By ANNA BERNASEK
Published: August 31, 2013

INCOME inequality in the United States has been growing for decades, but the trend appears to have accelerated during the Obama administration. One measure of this is the relationship between median and average wages.

**1.7%**

Increase in **median** annual wage

**3.9%**

Increase in **average** annual wage

**2009 through 2011**

The median wage is straightforward: it's the midpoint of everyone's wages. Interpreting the average, though, can be tricky. If the income of a handful of people soars while everyone else's remains the same, the entire group's average may still rise substantially. So when average wages grow faster than the median, as happened from 2009 through 2011, it means that lower earners are falling further behind those at the top.

One way to see the acceleration in inequality is to look at the ratio of average to median annual wages. From 2001 through 2008, during the George W. Bush administration, that ratio grew at 0.28 percentage point per year. From 2009 through 2011, the latest year for which the data is available, the ratio increased 1.14 percentage points annually, or roughly four times faster.
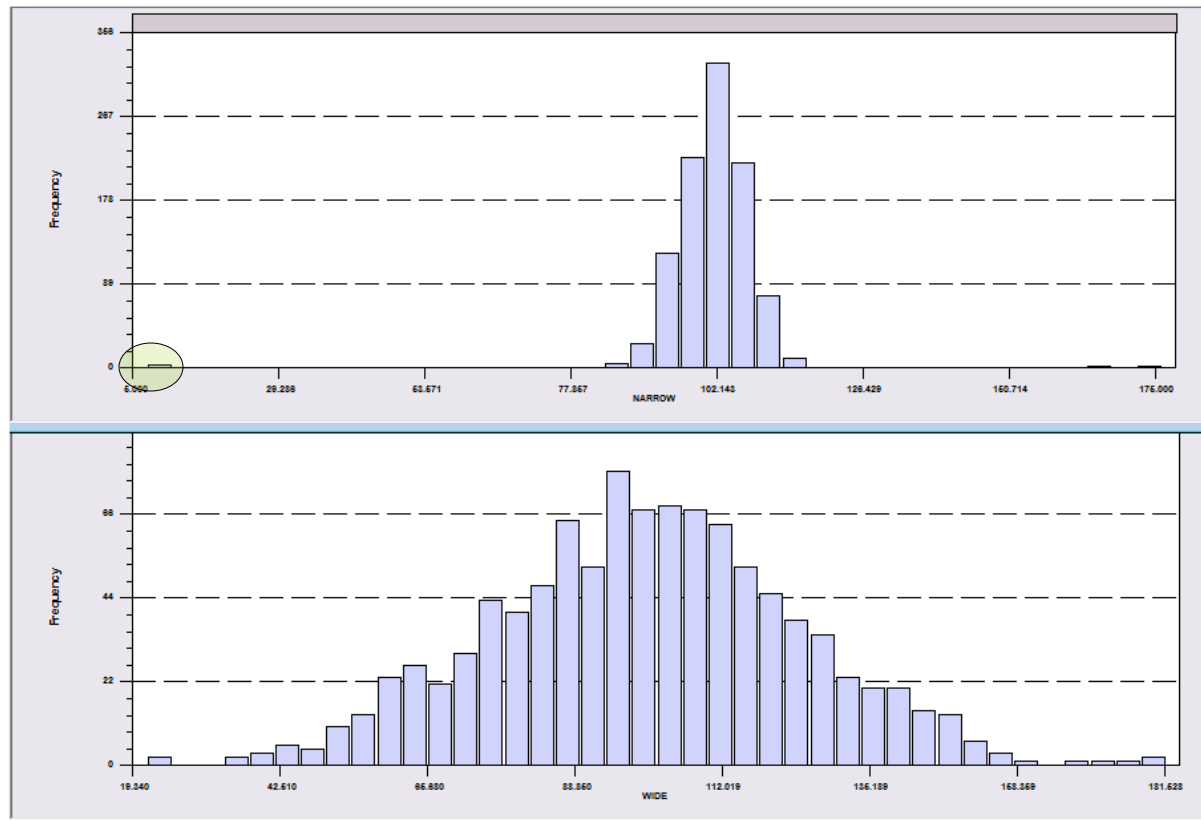
# Announcements

- Lab 9 Review

- Project 2 Due Tonight

- Project 3 Posted

- Grad Project 4 Posted

- Stats and Maybe ML!



https://xkcd.com/2225/
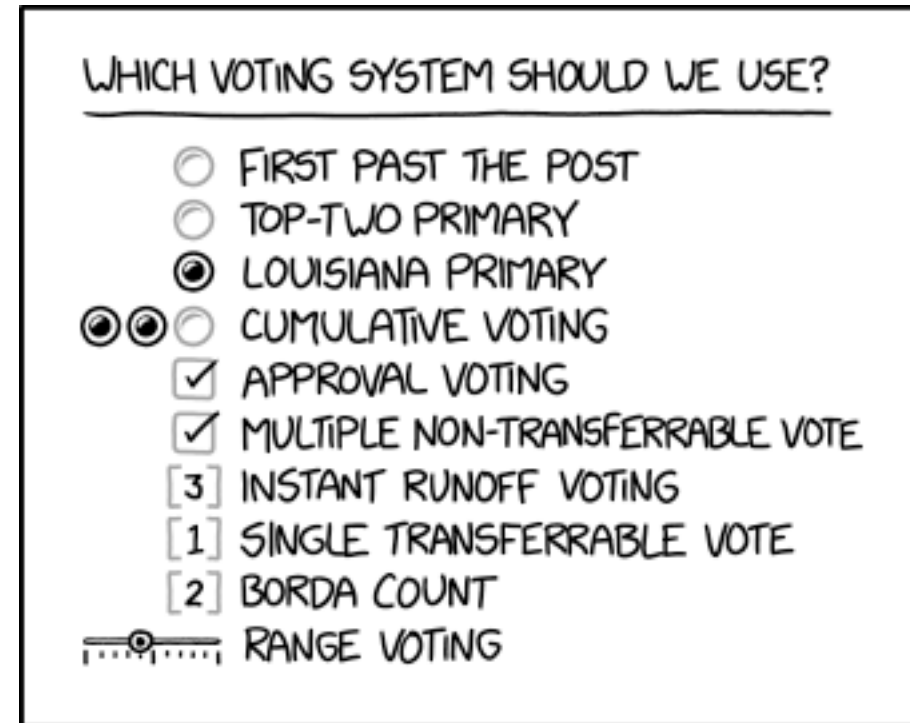
# More Information Needed!



Both data sets have a mean of about 100.

# Dispersion of the Observations

```
30 hours of average defect data on sets of circuit boards.

1.45   1.65   1.50   2.25   1.65   1.60   2.30   2.20   2.70   1.70
2.35   1.70   1.90   1.45   1.40   2.60   2.05   1.70   1.05   2.35
1.90   1.55   1.95   1.60   2.05   2.05   1.70   2.30   1.30   2.35
```



Histogram of Defects
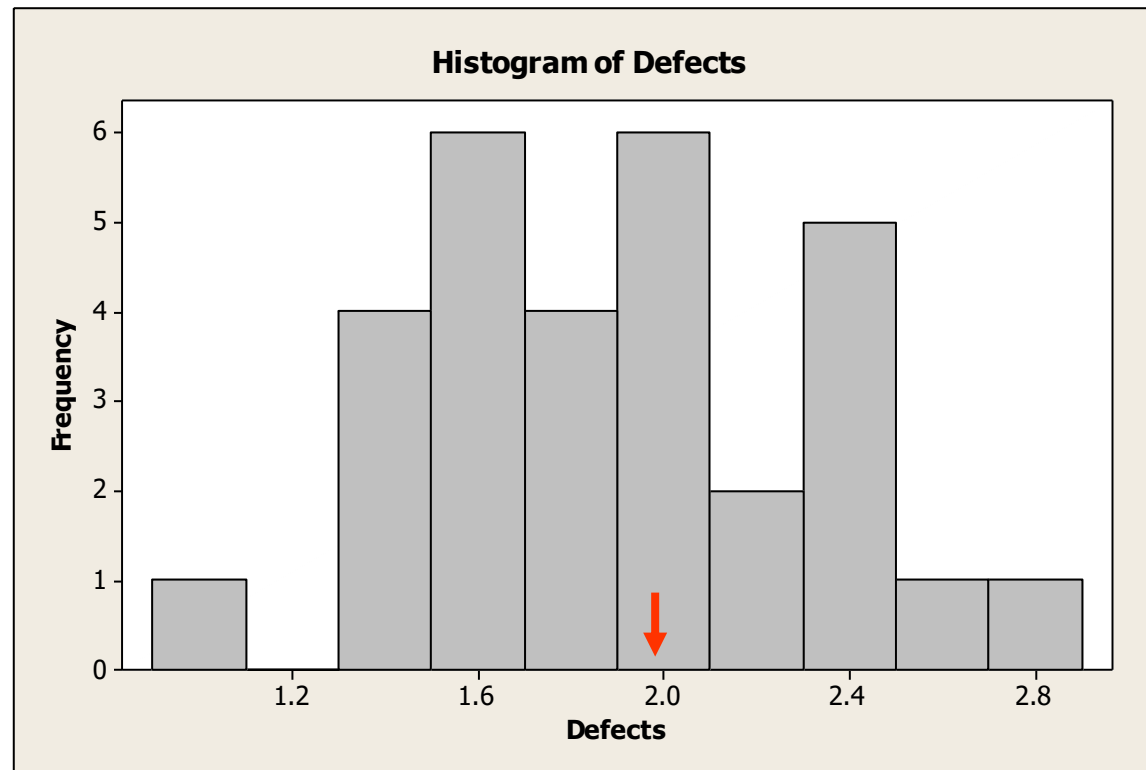
We quantify the variation of the values around the mean.

Note the range is from 1.05 to 2.70. This gives an idea where the data lie.

The mean plus a measure of the variation do the same job.

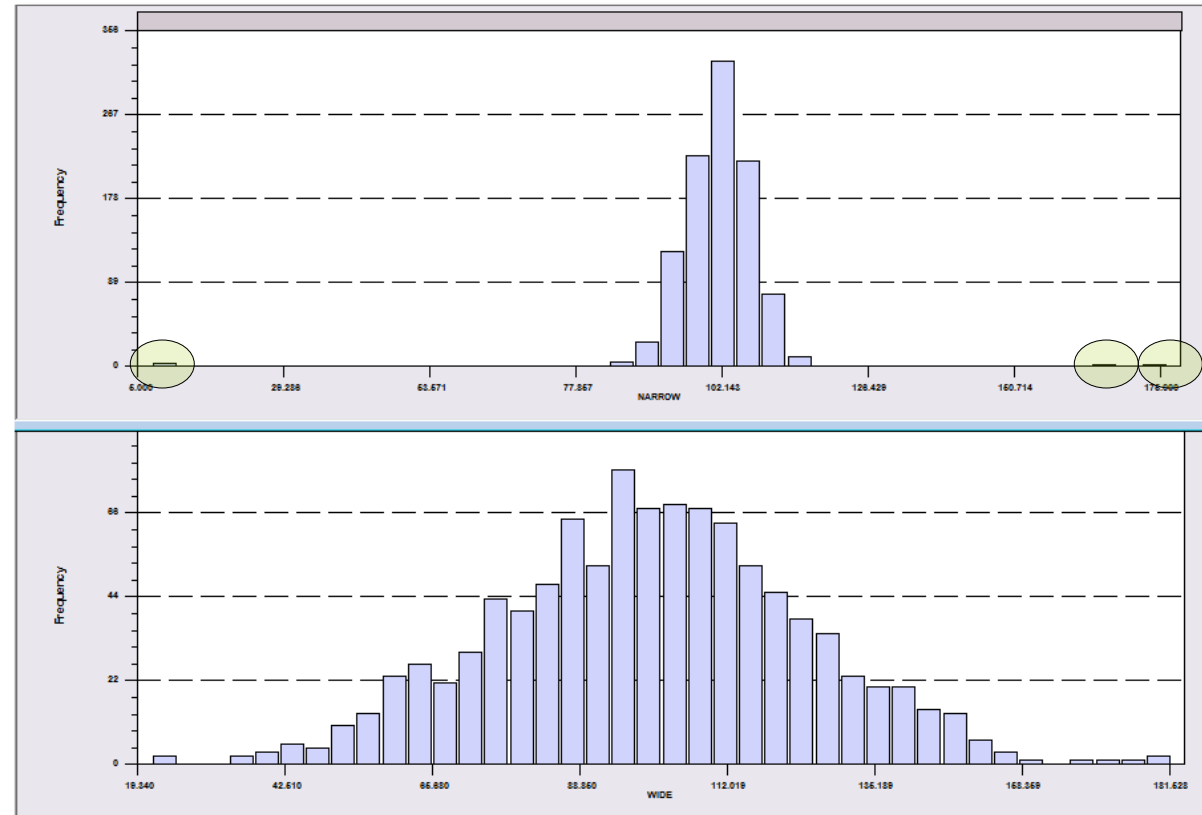# Range as a Measure of Dispersion

- Problems ????????



These two data sets both have 1,000 observations that range from about 10 to about 180.

# Variance & STDEV: Univariate Measures of dispersion

- Variance = $s_X^2$ = $\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^2$   or   $\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^2$

- Standard deviation = $s_X$ = $\sqrt{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^2}$

- The variance is commonly used statistic for spread
  - What are the units of the variance ??????????

- Standard deviation "fixes this," can be used as an **interpretable unit of measurement.**

Why n-1? http://nebula.deanza.edu/~bloom/math10/m10divideby_nminus1.pdf

# Variance, Aside: Why Divide by n-1?

- Remember: we are typically calculating the mean / median / variance / etc of a **sample** of a population.
    - Want that {mean, median, variance, …} to be an "unbiased" estimate of the true population's {mean, median, variance, …}

- **Unbiased? Consider variance …**
    1. Look at every possible sample of the population
    2. Compute sample variance of each population
    3. Is the average of those variances equal to the population variance? If so, then this is an "unbiased" estimator.

- Extra Reading: https://en.wikipedia.org/wiki/Bias_of_an_estimator

# Variance, Aside: Why Divide by n-1?

- Dividing by n-1 in the sample variance computation leads to an unbiased estimate of the population variance

- Intuition. Fix a sample …

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

  – Variance measures distribution around a mean

  – Sampled values are, on average, closer to sample mean than to true population mean

  – So, we will underestimate the true variance slightly

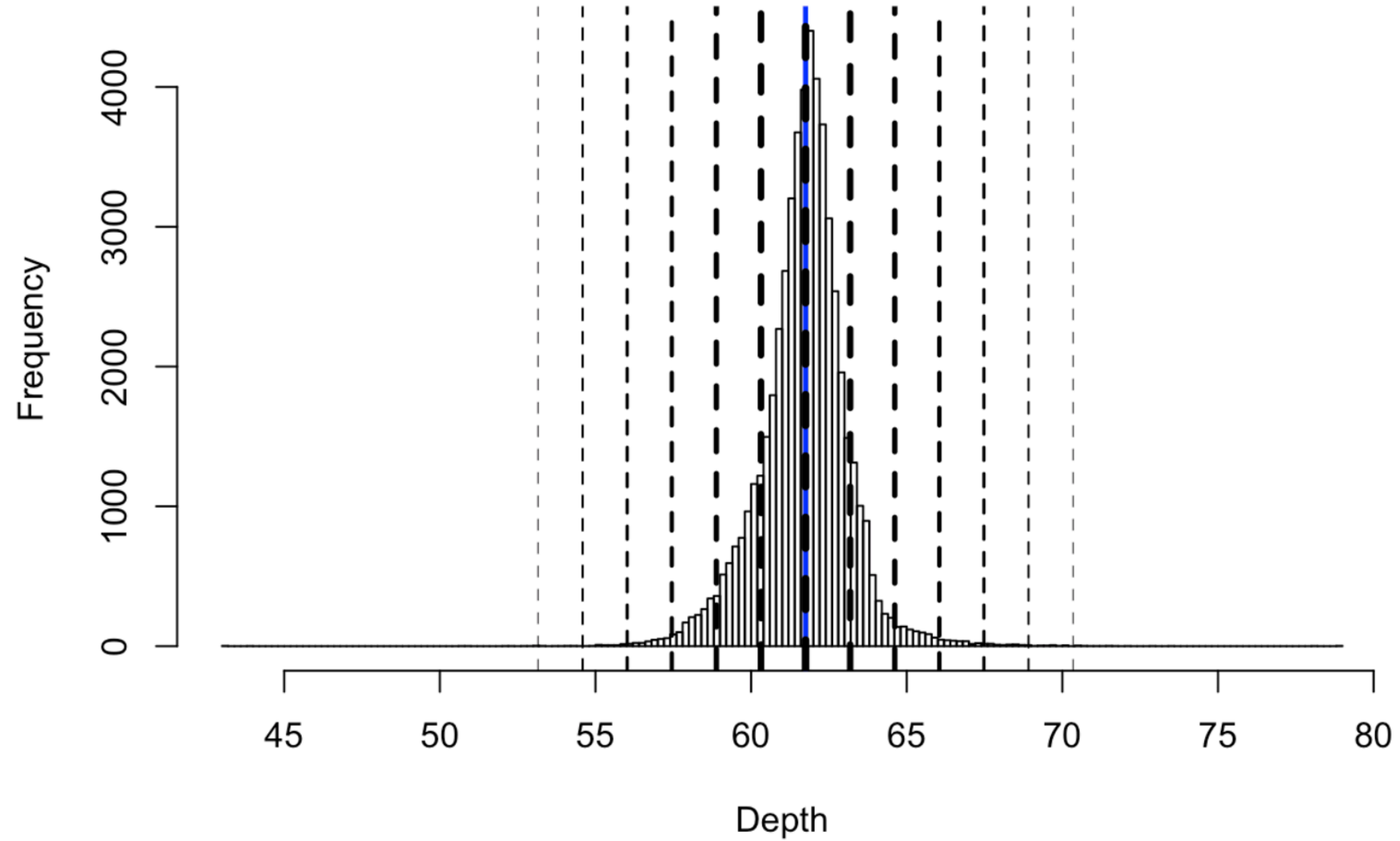  – Using n-1 instead of n makes our variance calculation bigger

- This "embiggening" impacts smaller $n$ more than larger $n$

  – Larger samples are better estimates of population

  – If sample is the population, just divide by $n$ …

Depth Histogram

# Using "standard deviations from the mean" as a unit

| SDs | Proportion | Interpretation |
|-----|-----------|----------------|
| 1 | 0.68 | 68% of the data is within ± 1 sds |
| 2 | 0.95 | 95% of the data is within ± 2 sds |
| 3 | 0.9973 | 99.73% of the data is within ± 3 sds |
| 4 | 0.999937 | 99.9937% of the data is within ± 4 sds |
| 5 | 0.9999994 | 99.999943% of the data is within ± 5 sds |
| 6 | 1 | 99.9999998% of the data is within ± 6 sds |
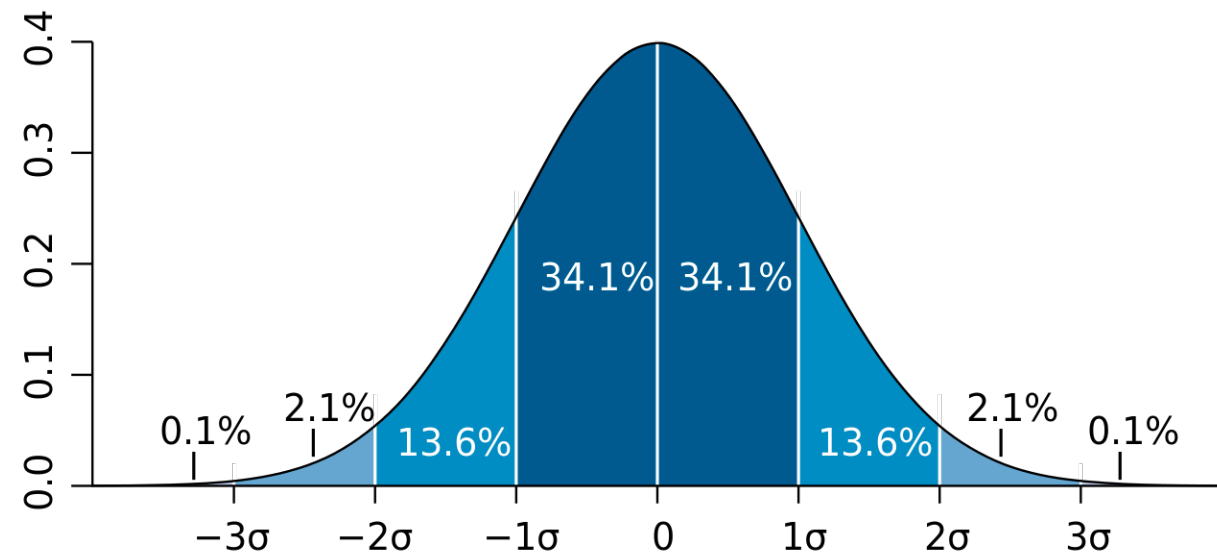
# Pairs of data Points?
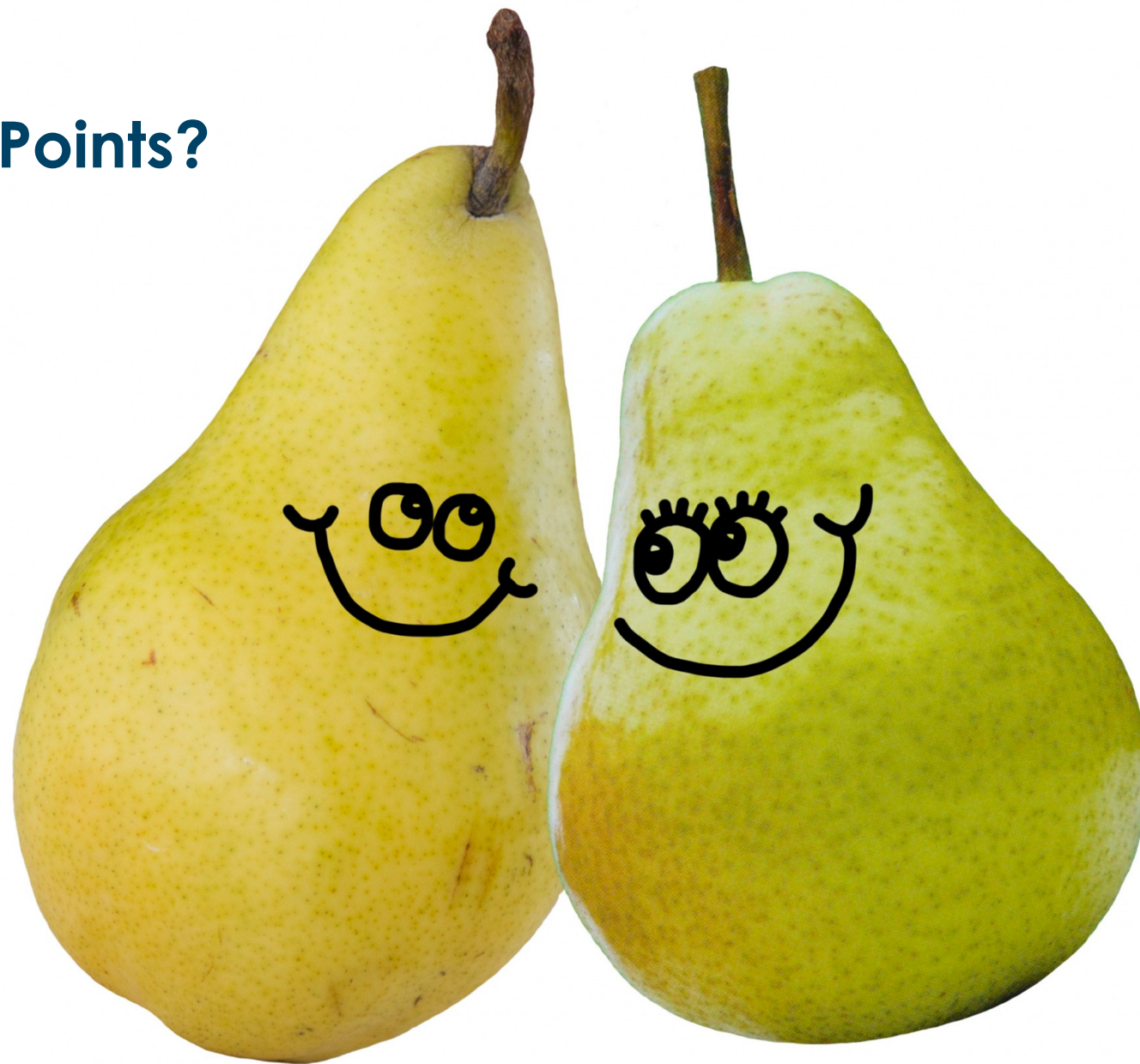
# Correlation

- Variables Y and X vary together
- Causality vs. Correlation: Does movement in X "cause" movement in Y in some metaphysical sense?
- Correlation
  - Simultaneous movement through a statistical relationship
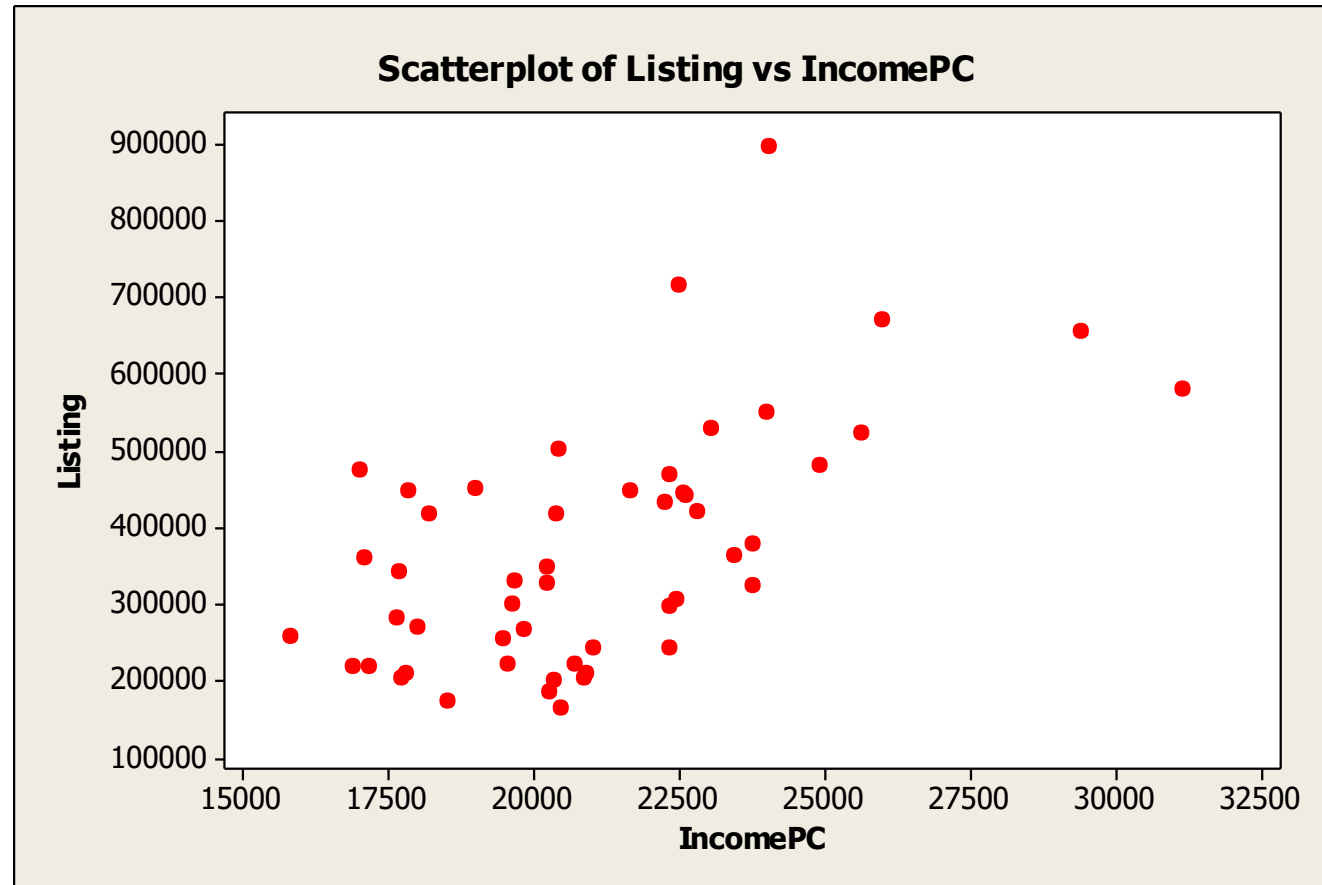  - Simultaneous variation "induced" by the variation of a common third effect
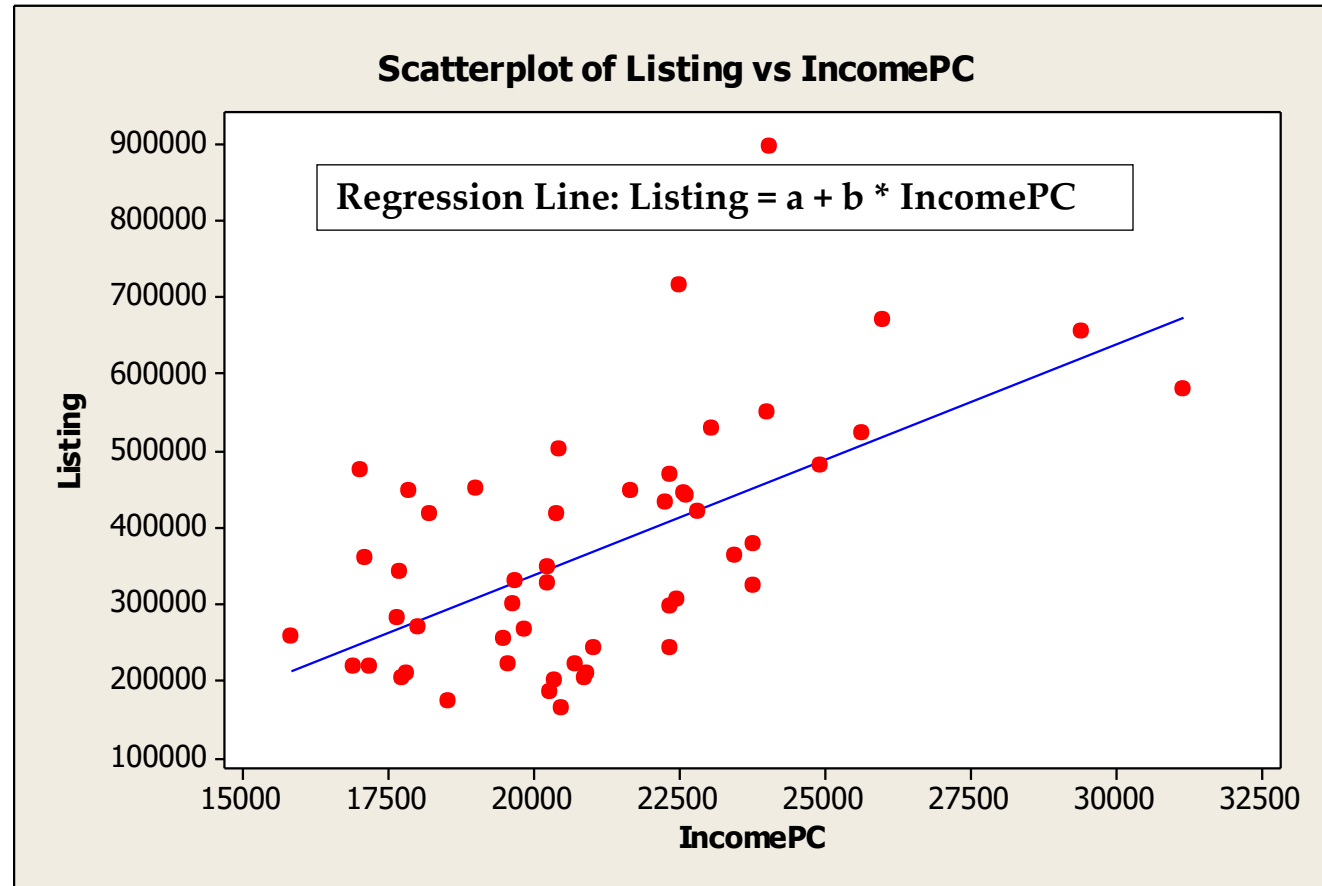
# House Prices & Per Capita Income

| State | Listing | IncomePC |
|---|---|---|
| Hawaii | 896800 | 24057 |
| California | 713864 | 22493 |
| New York | 668578 | 25999 |
| Connecticut | 654859 | 29402 |
| Dist.Columbia | 577921 | 31136 |
| Nevada | 549187 | 24023 |
| New Jersey | 529201 | 23038 |
| Massachusetts | 521769 | 25616 |
| Wyoming | 499674 | 20436 |
| Maryland | 480578 | 24933 |
| Utah | 475060 | 17043 |
| Colorado | 467979 | 22333 |
| Arizona | 448791 | 19001 |
| Florida | 447698 | 21677 |
| Montana | 446584 | 17865 |
| Virginia | 443618 | 22594 |
| Washington | 440542 | 22610 |

| State | Listing | IncomePC |
|---|---|---|
| Rhode Island | 432534 | 22251 |
| Delaware | 420845 | 22828 |
| Oregon | 417551 | 20419 |
| Idaho | 415885 | 18231 |
| Illinois | 377683 | 23784 |
| New Hampshire | 361691 | 23434 |
| New Mexico | 358369 | 17106 |
| Vermont | 346469 | 20224 |
| South Carolina | 340066 | 17695 |
| North Carolina | 330432 | 19669 |
| Georgia | 326699 | 20251 |
| Alaska | 324774 | 23788 |
| Minnesota | 306009 | 22453 |
| Maine | 299796 | 19663 |
| Pennsylvania | 295133 | 22324 |
| Louisiana | 280631 | 17651 |
| Alabama | 269135 | 18010 |

| State | Listing | IncomePC |
|---|---|---|
| Texas | 266388 | 19857 |
| Mississippi | 255774 | 15838 |
| Tennessee | 255064 | 19482 |
| Wisconsin | 243006 | 21019 |
| Michigan | 241107 | 22333 |
| Missouri | 221773 | 20717 |
| South Dakota | 220708 | 19577 |
| West Virginia | 219275 | 17208 |
| Arkansas | 217659 | 16898 |
| Ohio | 209189 | 20928 |
| Kentucky | 208391 | 17807 |
| Oklahoma | 203926 | 17744 |
| Kansas | 201389 | 20896 |
| Indiana | 200683 | 20378 |
| Iowa | 184999 | 20265 |
| North Dakota | 173977 | 18546 |
| Nebraska | 164326 | 20488 |

# Scatter Plot Suggests Positive Correlation



Scatterplot of Listing vs IncomePC

# Linear Regression Measures Correlation



Scatterplot of Listing vs IncomePC

Regression Line: Listing = a + b * IncomePC

# Correlation Is Not Causation

- Price and income seem to be **positively** correlated.



**Scatterplot of Income vs GasPrice**

Does a rise in income cause a rise in gas prices ?????????????

US gasoline prices, 1953-2004, plotted against per-capita US income

# A Hidden Relationship

**Not positively "related" to each other;
both positively related to "time."**

# "Related" ...?

- Want to capture: some variable X varies in the same direction and at the same scale as some other variable Y

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

- What happens if:
  - X varies in the opposite direction as Y ????????
  - X varies in the same direction as Y ????????

- What are the units of the covariance ????????

- Pearson's correlation coefficient is **unitless** in [-1,+1]:

$$cor(x, y) = \frac{cov(x, y)}{sd(x)sd(y)}$$

# Correlation


Scatterplot of Listing vs IncomePC

$$r_{\text{Income,Listing}} = +0.591$$

# Correlations



Scatterplot of cost vs Defects

r = +1.0



Scatterplot of Noise vs Defects

r = 0.0



Scatterplot of Noise vs MoreNoise

r = +0.5

# Correlation is not causation!!!



Correlation: 99%   Sources: US Census & USDA   tylervigen.com
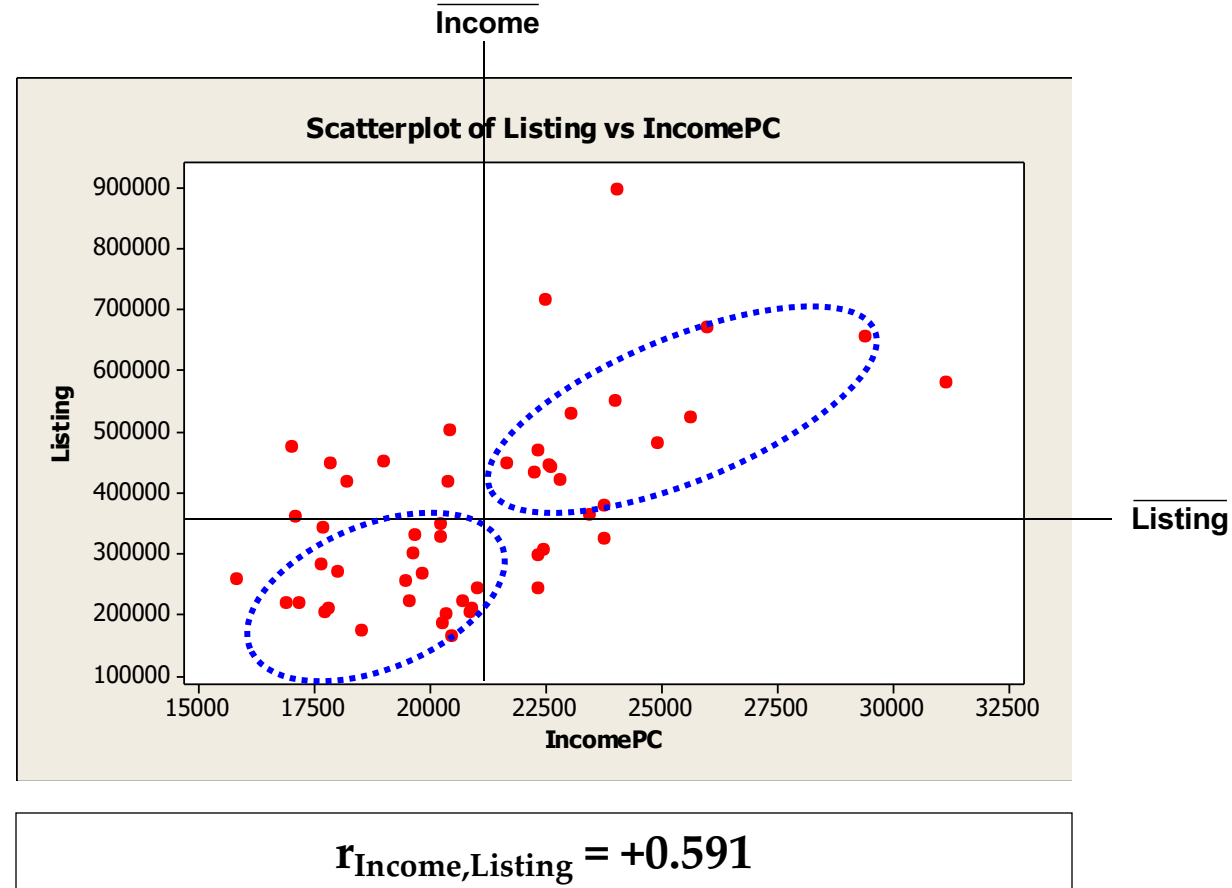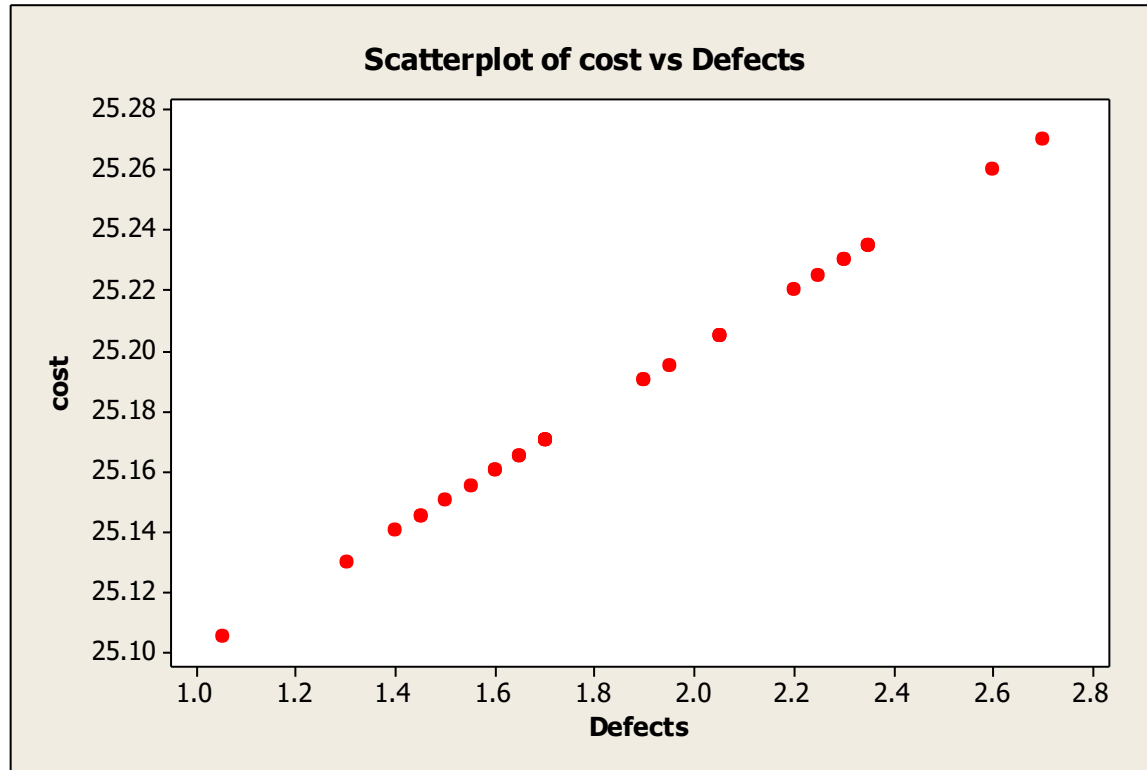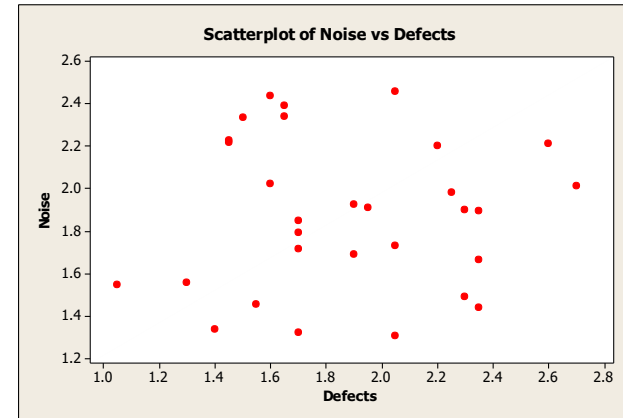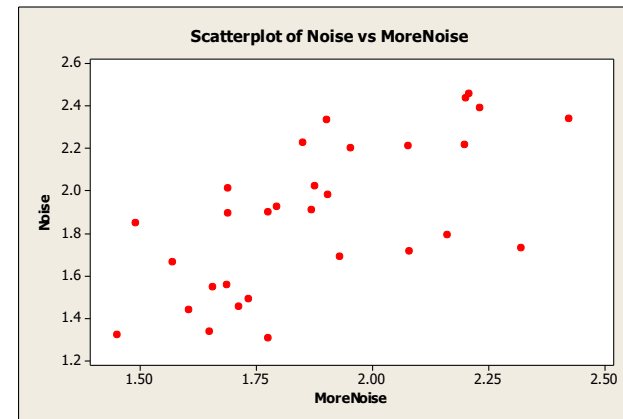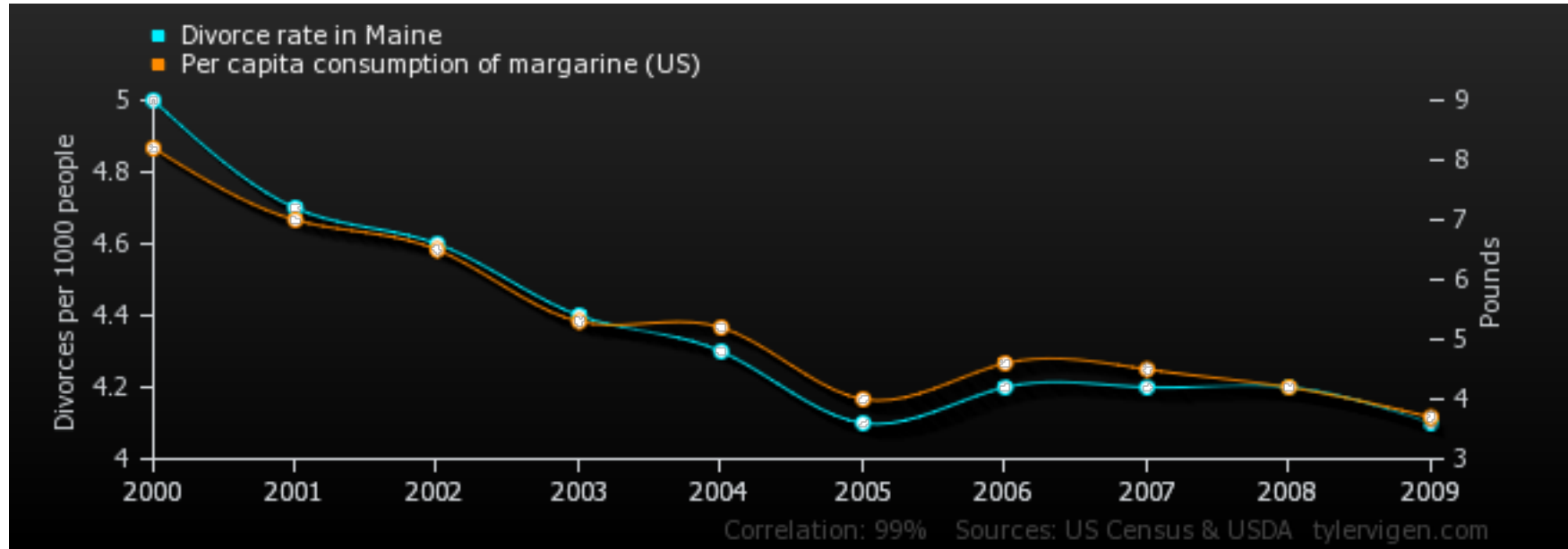
|  | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
|---|---|---|---|---|---|---|---|---|---|---|
| Divorce rate in Maine Divorces per 1000 people (US Census) | 5 | 4.7 | 4.6 | 4.4 | 4.3 | 4.1 | 4.2 | 4.2 | 4.2 | 4.1 |
| Per capita consumption of margarine (US) Pounds (USDA) | 8.2 | 7 | 6.5 | 5.3 | 5.2 | 4 | 4.6 | 4.5 | 4.2 | 3.7 |

r=0.993

?????????

http://tylervigen.com/spurious-correlations

# Just to drive the point home …



Per capita cheese consumption correlates with Number of people who died by becoming tangled in their bedsheets. Correlation: 94.71% (r=0.947091). Data sources: U.S. Department of Agriculture and Centers for Disease Control & Prevention. tylervigen.com
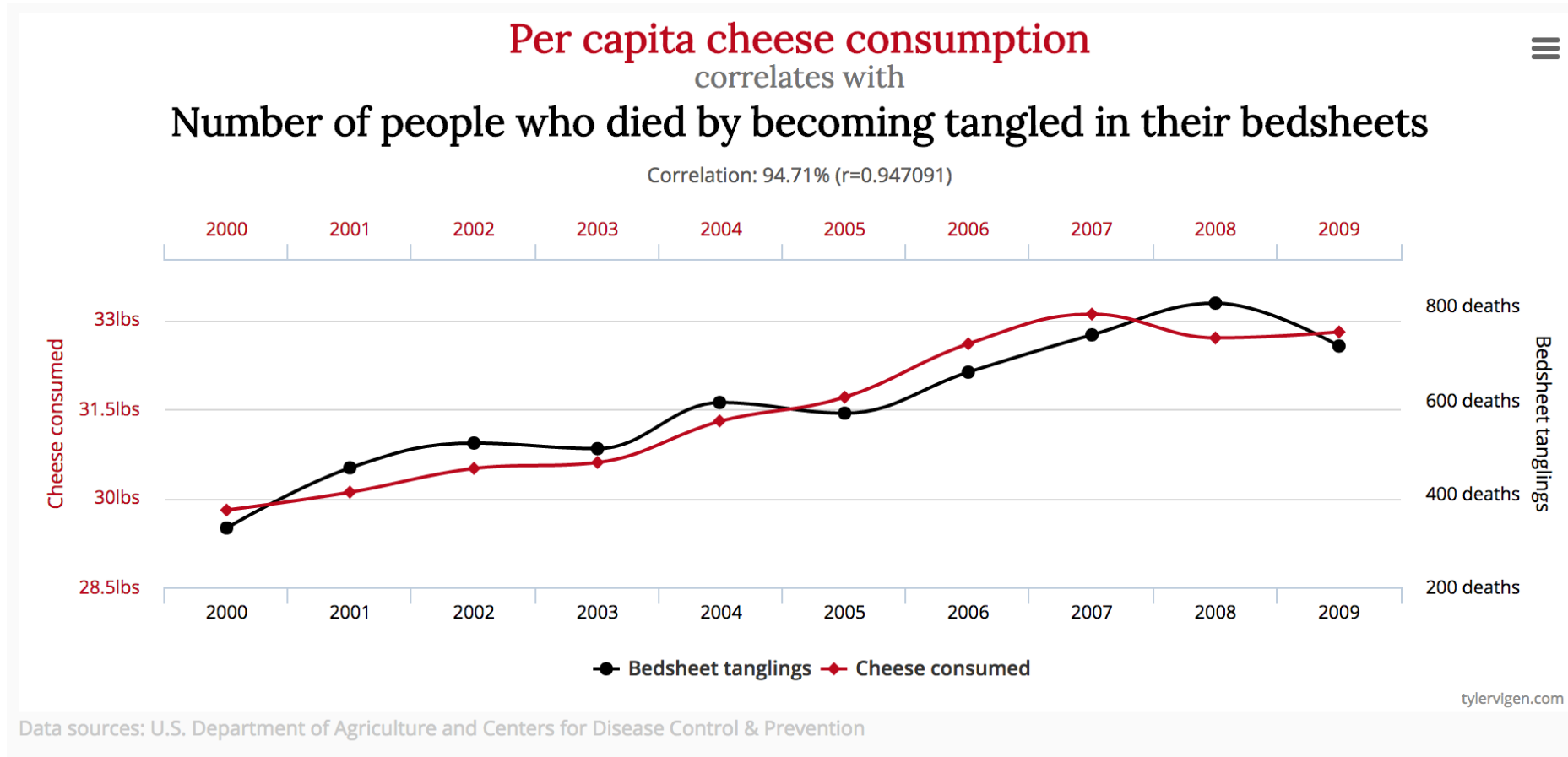
**Transformations**

# Transformations

- So, you've figured out that your data are:
  - Skewed
  - Have vastly different ranges across datasets and/or different units

- What do you do?

Transform the variables to:
- ease the validity and interpretation of data analyses
- change or ease the type of Stat/ML models you can use

# Standardization

- Transforming the variable to a comparable metric
  - known unit
  - known mean
  - known standard deviation
  - known range
- Three ways of standardizing:
  - P-standardization (percentile scores)
  - Z-standardization (z-scores)
  - D-standardization (dichotomize a variable)

# When you should always standardize

- When averaging multiple variables, e.g. when creating a socioeconomic status variable out of income and education.
- When comparing the effects of variables with unequal units, e.g. does age or education have a larger effect on income?

*Slides adapted from Maarten Buis*

# *P*-Standardization

- Every observation is assigned a number between 0 and 100, indicating the percentage of observation **beneath** it.

- Can be read from the cumulative distribution

- In case of knots: assign midpoints

- The median, quartiles, quintiles, and deciles are special cases of P-scores.

| | rent | cum % | percentile |
|---|---|---|---|
| room 1 | 175 | 5,3% | 5,3% |
| room 2 | 180 | 10,5% | 10,5% |
| room 3 | 185 | 15,8% | 15,8% |
| room 4 | 190 | 21,1% | 21,1% |
| room 5 | 200 | 26,3% | 26,3% |
| room 6 | **210** | 31,6% | **36,8%** |
| room 7 | **210** | 36,8% | **36,8%** |
| room 8 | **210** | 42,1% | **36,8%** |
| room 9 | 230 | 47,4% | 47,4% |
| room 10 | **240** | 52,6% | **55,3%** |
| room 11 | **240** | 57,9% | **55,3%** |
| room 12 | **250** | 63,2% | **65,8%** |
| room 13 | **250** | 68,4% | **65,8%** |
| room 14 | 280 | 73,7% | 73,7% |
| room 15 | **300** | 78,9% | **81,6%** |
| room 16 | **300** | 84,2% | **81,6%** |
| room 17 | 310 | 89,5% | 89,5% |
| room 18 | 325 | 94,7% | 94,7% |
| room 19 | 620 | 100,0% | 100,0% |

# P-standardization

- Turns the variable into a ranking, i.e. it turns the variable into a ordinal variable.

- It is a **non-linear** transformation: relative distances change

- Results in a fixed mean, range, and standard deviation; M=50, SD=28.6.
  – This can change slightly due to knots

- A histogram of a P-standardized variable approximates a uniform distribution.

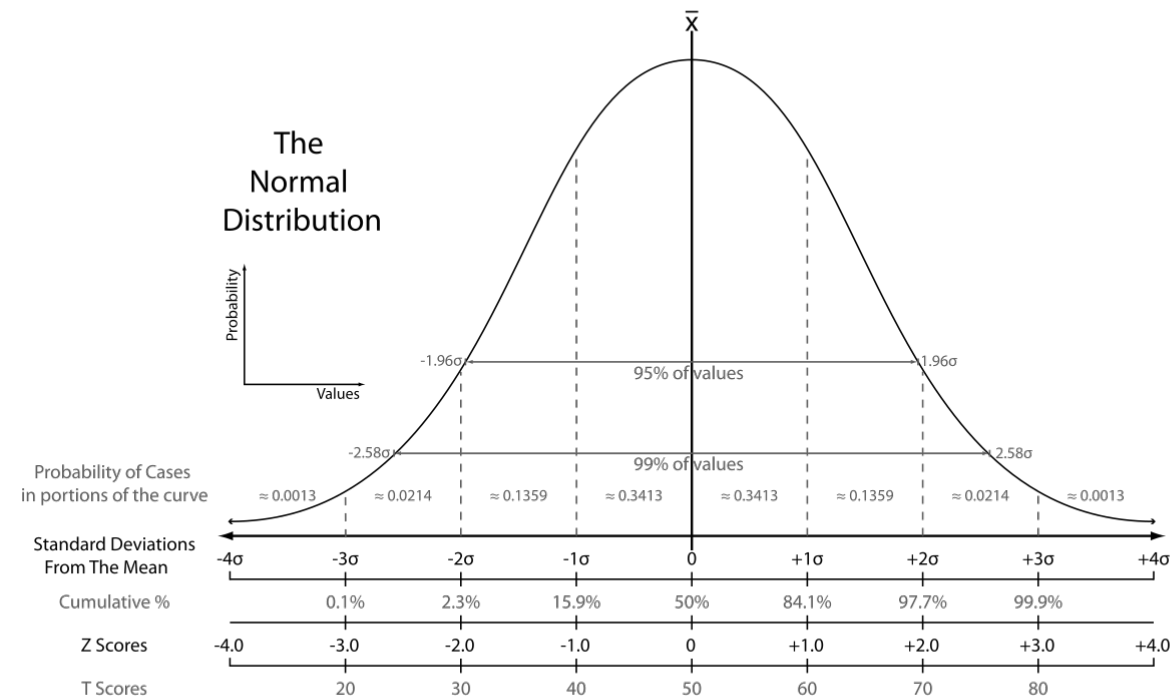|  | rent | cum % | percentile |
|---|---|---|---|
| room 1 | 175 | 5,3% | 5,3% |
| room 2 | 180 | 10,5% | 10,5% |
| room 3 | 185 | 15,8% | 15,8% |
| room 4 | 190 | 21,1% | 21,1% |
| room 5 | 200 | 26,3% | 26,3% |
| room 6 | **210** | 31,6% | **36,8%** |
| room 7 | **210** | 36,8% | **36,8%** |
| room 8 | **210** | 42,1% | **36,8%** |
| room 9 | 230 | 47,4% | 47,4% |
| room 10 | **240** | 52,6% | **55,3%** |
| room 11 | **240** | 57,9% | **55,3%** |
| room 12 | **250** | 63,2% | **65,8%** |
| room 13 | **250** | 68,4% | **65,8%** |
| room 14 | 280 | 73,7% | 73,7% |
| room 15 | **300** | 78,9% | **81,6%** |
| room 16 | **300** | 84,2% | **81,6%** |
| room 17 | 310 | 89,5% | 89,5% |
| room 18 | 325 | 94,7% | 94,7% |
| room 19 | 620 | 100,0% | 100,0% |

# Centering And Scaling – Z-Standardization.

- Transform your data into a **unitless** scale
  - Put data into "standard deviations from the mean" units
  - This is called standardizing a variable, into standard units

- Given data points $x = x_1, x_2, ..., x_n$:

$$z_i = \frac{(x_i - \bar{x})}{\text{sd}(x)}$$

- Translates $x$ into a scaled and centered variable $z$
- What is the mean of $z$ ?????????
- What is the standard deviation of z ?????????



The Normal Distribution

| | -4σ | -3σ | -2σ | -1σ | 0 | +1σ | +2σ | +3σ | +4σ |
|---|---|---|---|---|---|---|---|---|---|
| Probability of Cases in portions of the curve | | ≈ 0.0013 | ≈ 0.0214 | ≈ 0.1359 | ≈ 0.3413 | ≈ 0.3413 | ≈ 0.1359 | ≈ 0.0214 | ≈ 0.0013 |
| Standard Deviations From The Mean | -4σ | -3σ | -2σ | -1σ | 0 | +1σ | +2σ | +3σ | +4σ |
| Cumulative % | | 0.1% | 2.3% | 15.9% | 50% | 84.1% | 97.7% | 99.9% | |
| Z Scores | -4.0 | -3.0 | -2.0 | -1.0 | 0 | +1.0 | +2.0 | +3.0 | +4.0 |
| T Scores | | 20 | 30 | 40 | 50 | 60 | 70 | 80 | |

# Centering or Scaling

- Maybe you just want to center the data:

$$z_i = (x_i - \bar{x})$$

- What is the mean of $z$ ?????????
- What is the standard deviation of $z$ ?????????

- Maybe you just want to scale the data:

$$z_i = \frac{x_i}{\text{sd}(x_i)}$$

- What is the mean of $z$ ?????????
- What is the standard deviation of $z$ ?????????

# Discrete to Continuous Variables

- Some models only work on continuous numeric data
- Convert a binary variable to a number ??????????
  - health_insurance = {"yes", "no"} → {1, 0}
- Why not {-1, +1} or {-10, +14}?
  - 0/1 encoding lets us say things like "if a person has healthcare then their income increases by $X."
  - Might need {-1,+1} for certain ML algorithms (e.g., SVM)

# Discrete to Continuous Variables

- What about non-binary variables?
- My main transportation is a {BMW, Bicycle, Hovercraft}
- One option: { BMW → 1, Bicycle → 2, Hovercraft → 3 }
  - Problems ??????????

- **One-Hot Encoding:** convert a categorical variable with N values into a N-bit vector:
  - BMW → [1, 0, 0]; Bicycle → [0, 1, 0]; Hovercraft → [0, 0, 1]

```
# Converts dtype=category to one-hot-encoded cols
cols = ['my_transportation']
df = df.get_dummies( columns = cols )
```

# Continuous to discrete variables

- Do doctors prescribe a certain medication to older kids more often?  Is there a difference in wage based on age?
- Pick a discrete set of bins, then put values into the bins
- Equal-length bins:
- Bins have an equal-length range and skewed membership
- Good/Bad ????????
- Equal-sized bins:
- Bins have variable-length ranges but equal membership
- Good/Bad ????????

# Skewed Data

- Skewed data often arises in multiplicative processes:
- Some points float around 1, but one unlucky draw → 0
- Logarithmic transforms reduce skew:
- If values are all positive, apply $\log_2$ transform
- If some values are negative:
  - Shift all values so they are positive, apply $\log_2$
  - Signed log: $\text{sign}(x) * \log_2(\ |x|\ + 1)$
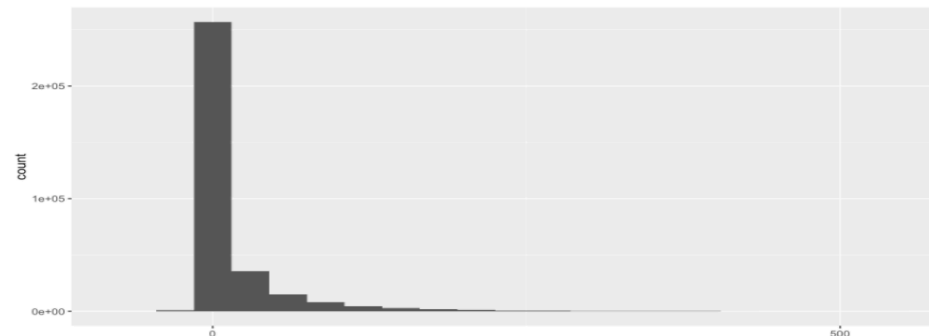
# Skewed Data
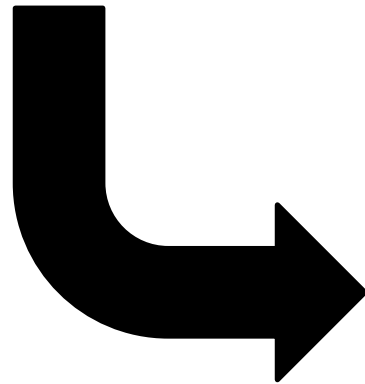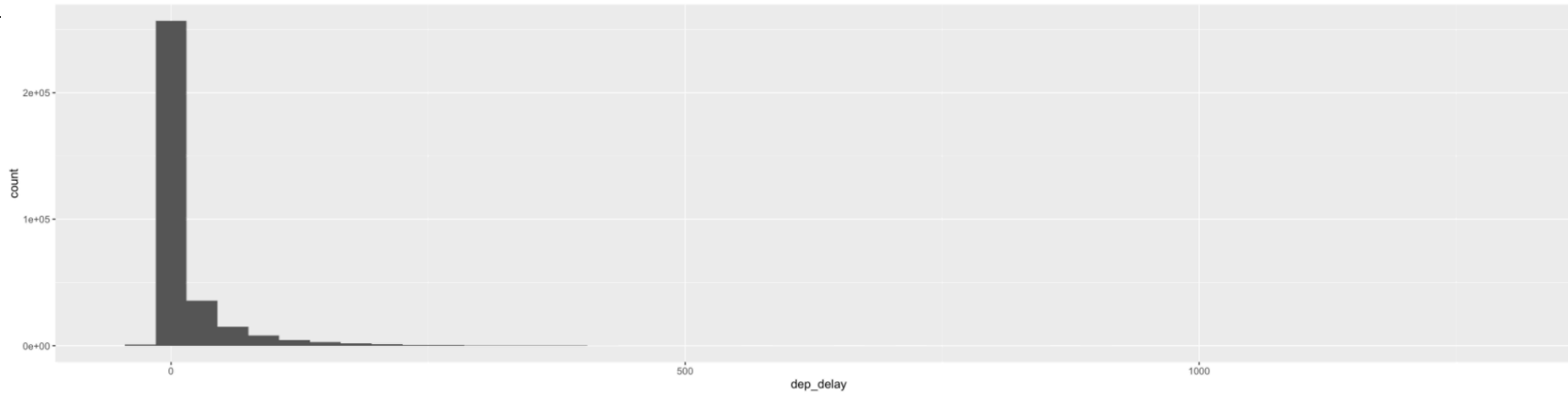


log₂ transform on airline takeoff delays