

Statistics

The science of analyzing data

Yordan Darakchiev

Technical Trainer

iordan93@gmail.com





sli.do

#MathForDevs

Table of Contents

- Descriptive and inferential statistics
- Population and sample
- Properties of statistical distributions
- Visualizing data
- Covariance and correlation
- Common misconceptions



Basic Concepts

How to deal with many numbers

Descriptive Statistics

- Numbers which are used to **summarize** and **describe** data
 - We work with **all items of interest** – **statistical population**
 - We don't try to make predictions, just describe what we're seeing
- Not very useful on their own
 - But an important part of other methods
- Example: pet shop sales
 - 100 pets in one month: 40 dogs, 30 cats, 30 other
- What percent of all pets are dogs?
- What's the mean number of cats sold per month?
- We can also represent the information graphically
 - What does the distribution of dog sales per day look like?
 - What does the cumulative distribution of sales look like?
 - How do sales compare?

Inferential Statistics

- In many cases the **population** is too large (or even infinite)
 - We represent the population by a subset – **sample**
 - **The population characteristics can be estimated by using the sample**
 - We have to be extremely careful how to choose the sample
 - In most cases we need **random sampling** of the population
- Examples
 - Voting predictions
 - We ask a small number of people and we draw **inferences** about the entire country
 - Mean salary by age
 - We divide people into age groups (e.g. < 20 , $20 - 25$, $25 - 30$, $30 - 35$, ...) and ask several people within each age group
 - This also makes the continuous variable "age" easier to work with

Sampling

- The process of selecting a sample from the population
- Steps in the sampling process
 - Define the population
 - Specify the **sampling frame** – a set of items from the population
 - Specify the **sampling method** – how to select items from the frame
 - Determine the sample size
 - Implement the sampling and collect data
- A badly done sampling can induce **biases** and **errors**
 - **Selection bias** – selecting a non-random sample
 - E.g., asking only CEOs of companies when sampling data for salaries by age
 - **Random sampling error** – random variations in the results

Sampling Methods

- Non-random sampling
 - Can be biased
 - **Not representative** of the population
- Random sampling
 - Every member of the population has equal chance of being chosen
 - Example: insect population in trees
 - Trees are numbered 1-200, 10 trees are chosen at random
 - All insects are counted on the 10 random trees
- Stratified sampling
 - Divide the population into categories (subpopulations)
 - For each category, sample at random
 - Example: foot measurement study → male / female; age groups
 - Select samples for each combination { gender; age }

Properties of Distributions

Mean, standard deviation,
skewness, kurtosis

Summarizing Distributions

- A **histogram** is a **complete description** of the sample distribution
- We often summarize it using a few descriptive statistics
 - Central tendency
 - Do the values tend to cluster around a center?
 - Modes
 - How many clusters are there? Where are they?
 - Variance
 - How much variability is there (how "spread out" is the distribution)?
 - Tails
 - How quickly do probabilities drop off as we move away from the center(s)?
 - Outliers
 - Are there extreme values, far from the center(s)?
- These are also called **summary statistics**

Measures of Central Tendency

- **Average** – a number which describes a typical data point
 - Can be calculated in many ways

- **Arithmetic mean**

- The sum of all measurements divided by the number of observations

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

- **Median**

- The middle value of the distribution
 - To calculate it, the numbers must be sorted in ascending order
 - Examples:
 - $\text{Me}(\{1, 2, 2, 3, 4\}) = 2$; $\text{Me}(\{1, 2, 2, 3, 4, 10\}) = 2,5$

- **Mode**

- The most frequent item
 - $\text{Mo}(\{1, 3, 2, 3, 4, 3\}) = 3$
 - Many "most frequent items" \Rightarrow multimodal distribution

Variance

- Describes how far away a sample is from the sample mean
 - All differences from the mean $x_i - \bar{x}$ can be positive or negative
 - They all sum up to 0 (that's the definition of the mean)
 - So we square them to make them positive
$$S^2(x) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$
 - Standard deviation: $S(x) = \sqrt{S^2(x)}$
- In the sample variance formula, there is $n - 1$ in the denominator
 - It refers to "degrees of freedom" – how many items we can remove
 - The number of parameters that can vary
 - Because all distances sum up to 0, if we know $n - 1$ of them, we can find the last one
 - Gives us an unbiased estimator (more on that [here](#))

Variance (2)

- Why bother to take the standard deviation?
 - Instead of using variance directly
- It's all about units
- Example:
 - Let's say we're measuring length in m
 - By definition, the variance will have units of m^2
 - We want to see how far is a certain point from the center and the units don't match
 - Compare $d = 2m$, $S^2 = 0,25m^2$ to $d = (2 \pm 0,5) m$
 - In order to make units match, we take the square root
 - Example
 - If we measure $2,75m$, we can say "This measurement is located at **1,5 standard deviations above the mean**"
 - Comparisons like these are very useful in statistics

Population vs. Sample: Measures

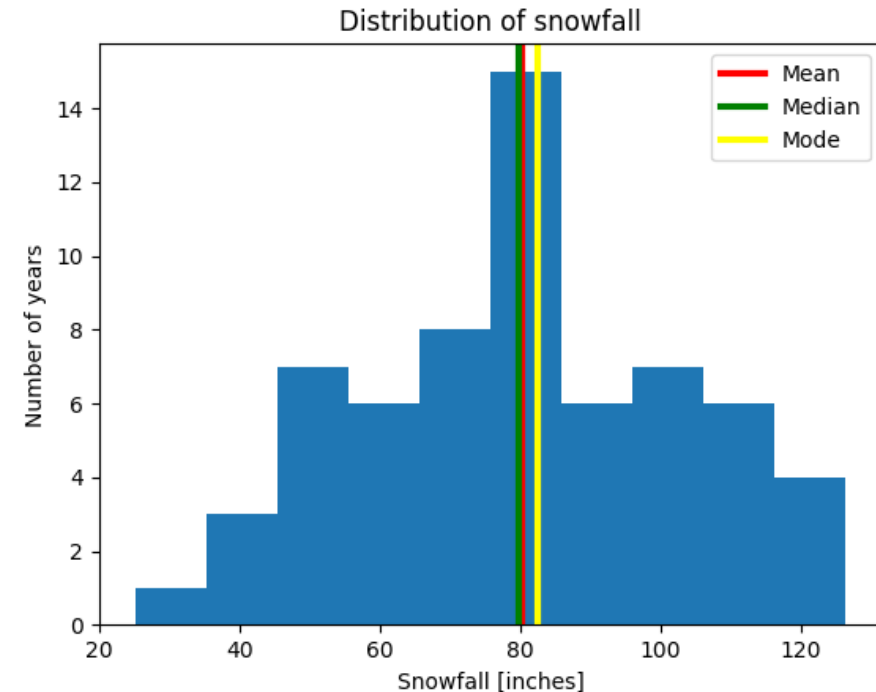
- There are differences between a population and samples from that population \Rightarrow we have different statistics
 - Notation
 - Sample statistics – sample mean, sample variance, etc. – Latin letters
 - Population statistics – Greek letters
- Population mean μ
 - Also called **expected value**
 - N – population size
- Population variance σ^2
 - Note how since we know the entire population, there is **no estimation** going on
 - So, there is N in the denominator
- Population standard error
 - $\sigma(x) = \sqrt{\sigma^2(x)}$

$$\mu(x) = E[x] = \frac{1}{N} \sum_{i=1}^N x_i$$

$$\begin{aligned}\sigma^2(x) &= E[(x_i - \mu)^2] = \\ &= \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2\end{aligned}$$

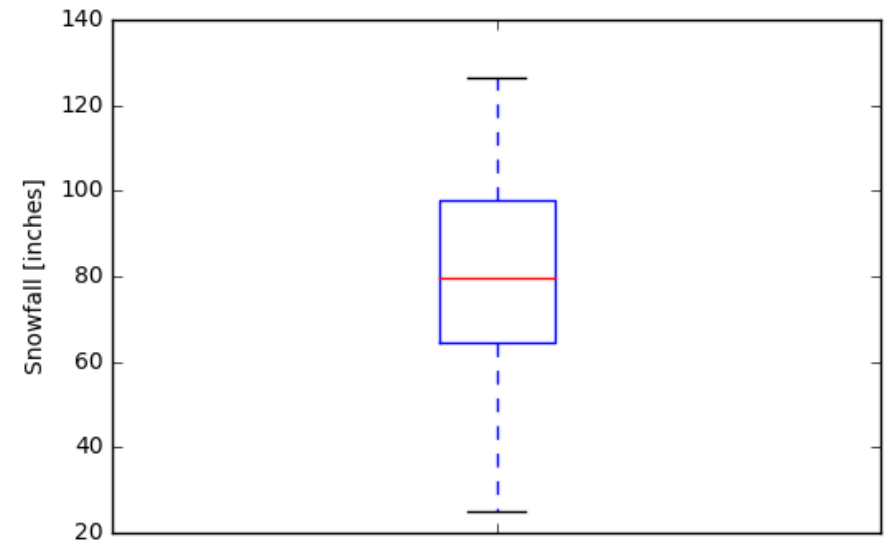
Example: Snowfall Data

- You are given data of snowfall in Buffalo, NY in inches for years 1910 – 1972 (**snowfall.csv**)
- Plot a histogram
- Print the mean, standard deviation and modes
- Print the standard deviation
 - Note: If you're using **numpy**, it returns the biased estimator of standard deviation. Pass a parameter **ddof = 1** (difference in degrees of freedom) to calculate the unbiased estimator
- Overlay the mean, median and first mode on the histogram



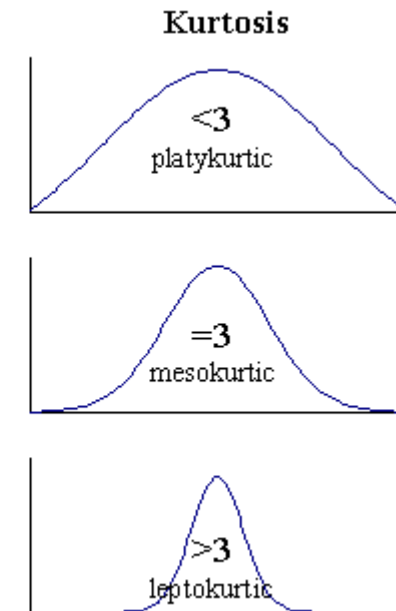
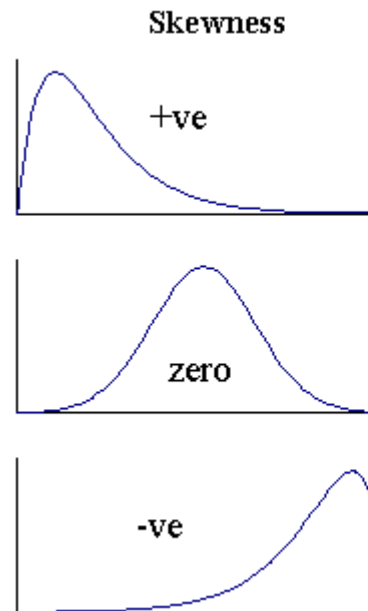
Five-Number Summary

- Conveys similar information to a histogram
 - How many percent of the data are less than or equal to a specified number
 - Minimum (0%); first quartile (25%); median (50%); third quartile (75%); maximum (100%)
 - Generalization: **quantiles** – divide the frequency distribution into equal groups
 - 100 groups = **percentiles**
- Visualization: boxplot
 - Middle line – median
 - Box – quartiles
 - Whiskers – largest "non-outliers" – 1.5 times the interquartile range
 - Points – outliers



Moments of Distributions

- r^{th} central moment: $\mu_r(x) = \frac{\sum (x - \mu)^r}{N}$
 - Defined for discrete and continuous variables
 - Measure the shape of the probability distribution
- Zeroth moment: 1 (**total probability**)
- First moment: **arithmetic mean** μ
- Second moment: **variance** σ^2
- Third moment: **skewness** γ
 - Asymmetry in the distribution
- Fourth moment: **kurtosis** β
 - Heaviness of the "tails"
 - "Normal": $\beta = 3$
 - Excess kurtosis: $\beta - 3$



Moments of the Gaussian Distribution

- Generalization of the binomial distribution
- Probability density function

$$N(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Mean: μ
- Median: μ
- Mode: μ
- Variance: σ^2
- Skewness: 0
- Excess kurtosis: 0

- *"And that, kids, is why I love the Gaussian distribution."*

Standard Score

- In order to compare different Gaussian distributions, we can "normalize" them
 - Change their parameters to get a "standard" Gaussian distribution with $\mu = 0$ and $\sigma = 1$
 - We need to "shift" the distribution left or right and "squish" or "stretch" to achieve the required standard deviation
 - The shift is denoted by the standard score (or z-score): $Z(x) = \frac{x - \mu}{\sigma}$
- Example: 50 student scores
 - Normal distribution, mean 60 (out of 100) and standard deviation 15
 - How well did a student perform if they had 70 / 100?
 - Top 25% of the class
 - What marks do the top 10% of the class have?
 - 79 and up



Many Variables

Extending what we know

Covariance

- Up to now, we've been looking at variables on their own
 - But in many cases, they interact with each other
- **Covariance** is a measure of the joint variability of two variables

$$\text{cov}(x, y) = \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})$$

- Positive: as one variable increases, the other also increases
 - Negative: as one variable increases, the other decreases
 - Zero: the two variables don't vary together at all
- We can see that $\text{cov}(X, X) = \sigma^2(X)$
- In higher dimensions, we calculate a **covariance matrix**
 - The same idea: element (i, j) is equal to the covariance of the i^{th} and j^{th} dimensions: $A_{ij} = \text{cov}(x_i, x_j)$

Correlation

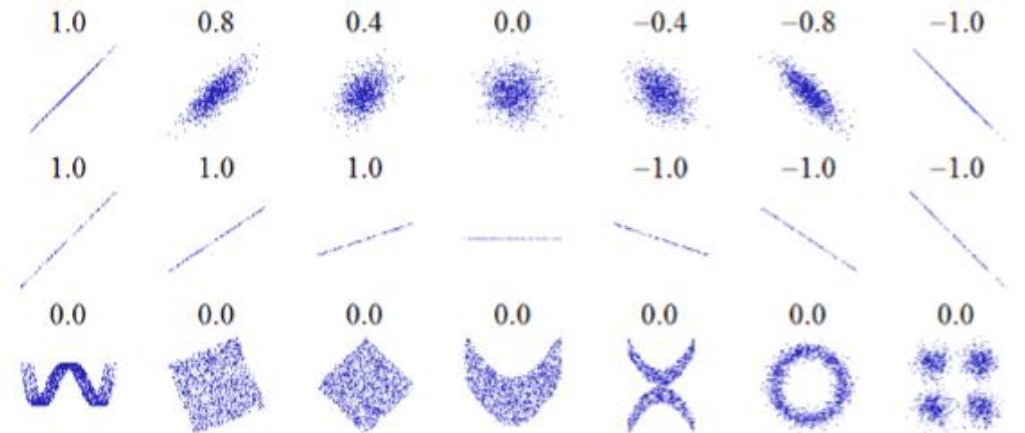
- Like the variance, covariance is in "weird" units
 - We divide by the standard deviations to normalize them
⇒ standard scores (similar to z-scores)

$$p_i = \frac{(x_i - \bar{x})}{s_x} \frac{(y_i - \bar{y})}{s_y}$$

- The mean value can be calculated as
 - This is called **Pearson's correlation coefficient**

$$\rho = \frac{1}{n} \sum p_i = \frac{\text{cov}(x, y)}{s_x s_y}$$

- The correlation coefficient can be in $[-1; 1]$
 - High absolute value
⇒ strong correlation
 - Measures the linearity of a relationship between two variables
 - Cannot express other, more complex relationships

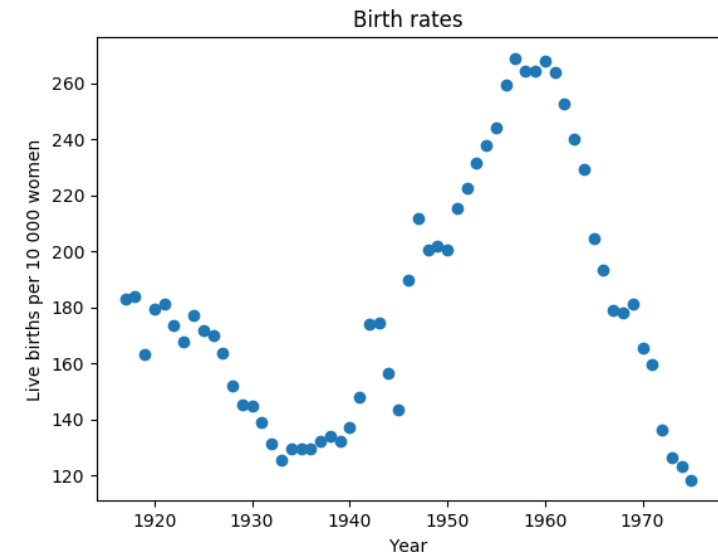


Scatter Plots

- The easiest way to see how two variables are correlated
- Two versions:
 - "Independent" variable – x -axis, "dependent" variable – y -axis
 - Two correlated variables (we can't say which is "independent")
- Besides, outliers usually become easily visible
- Best practices
 - **Label your axes**
 - If needed, include a legend
 - Scale / transform the variables if needed
 - Simplifies the relationship
 - Add trendlines if needed
 - You can also plot line charts if that's what your data suggests

Example: Birth Rates

- You are given the number of live births per 10 000 23-year-old women in the US between 1917 and 1975
 - File: **birth_rates.csv**
- Plot a scatter plot of the birth rates per year
 - What conclusions can you make?
 - This is called "time series analysis" – we are analyzing a process as it evolves with time
- Additionally, you can still inspect the variables one by one
 - Plot a histogram of the birth rates, disregarding the years
 - Are there any "typical" birth rates?
 - Are they distributed normally?



Example: Brain and Body Weights

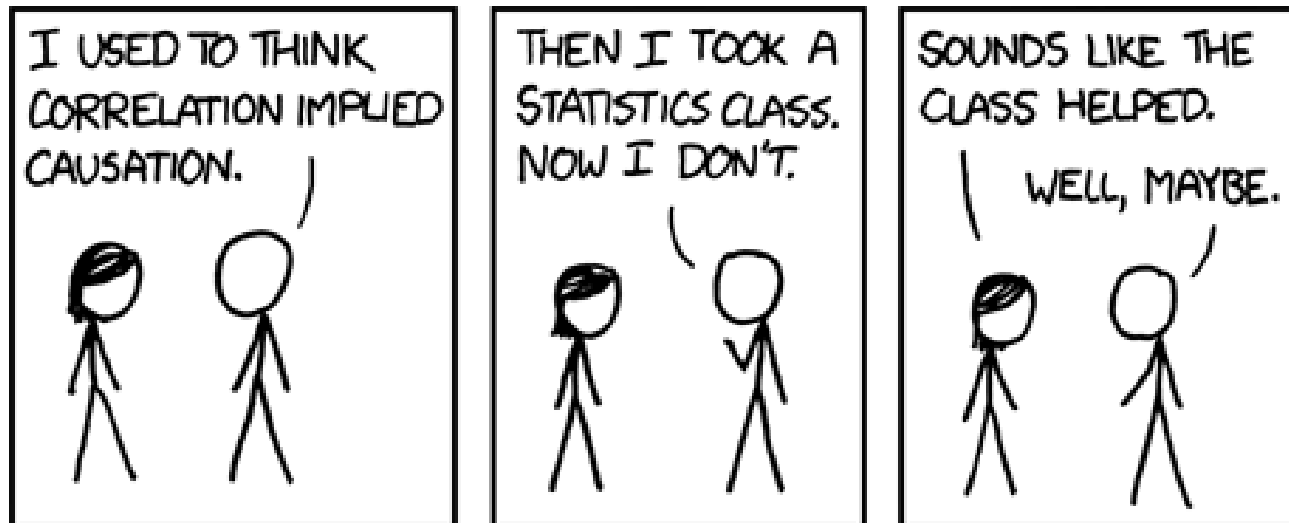
- File: `brain_weight.csv`
- Inspect the two variables: body weight [kg], brain weight [g]
 - Plot histograms, even boxplots if needed
- Create a scatterplot
 - The distribution is highly skewed, almost nothing is visible
- Transform the data
 - Take logarithms of both the body weight and the brain weight
 - Plot histograms of the logarithms
 - Create another (log-log) scatterplot
 - Is there any significant relationship?
 - If so, what is the **real** relationship (between the untransformed variables)?
 - To find it, you have to "reverse" the transformation

Common Pitfalls

Statistics can be dangerous
(and wrong)

Correlation Does Not Imply Causation!

- If two variables are correlated, this does not mean that necessarily the first causes the second
- Example: height and weight
 - Does a greater weight cause a greater height?
- We can still describe them
- **We can predict** height from weight and vice versa
- But that still does not say anything about one causing the other

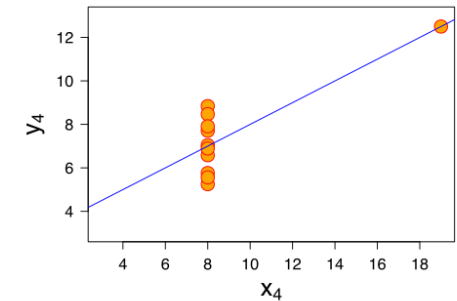
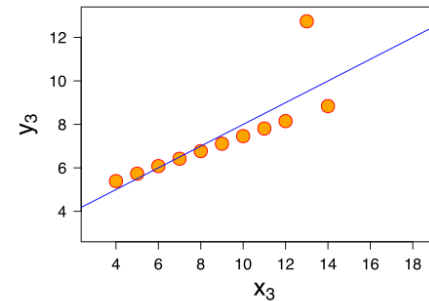
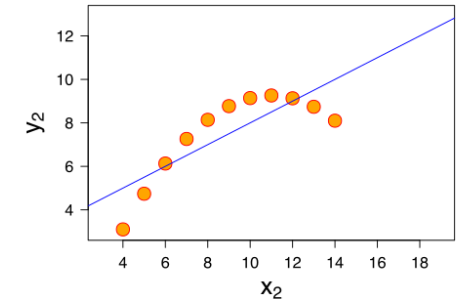
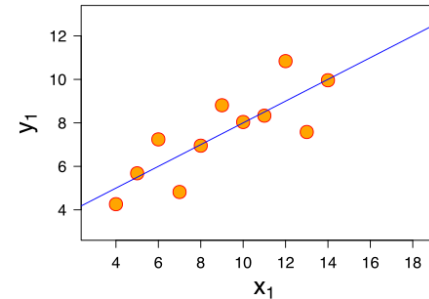


Correlation vs. Causation

- Reverse causation
 - The faster the windmills rotate, the more wind there is
⇒ Windmills cause wind
- Lurking variable
 - The more firefighters there are to put out a fire, the greater the damage caused
⇒ Firefighters being present at fires, cause more damage
- Bidirectional relationship
 - Predator numbers affect prey numbers, but prey numbers (amount of food) also affect predator numbers
- Coincidence
 - <http://tylervigen.com/spurious-correlations>
- More information about causal relationships: [minutephysics \(YouTube\)](#)

Anscombe's Quartet

- Four datasets with similar descriptive statistics which look completely different when plotted
 - More information: [Wikipedia](https://en.wikipedia.org/wiki/Anscombe%27s_quartet)
- Takeaways
 - Plot the data
 - In general, it's important to get to know your data
 - List as many assumptions and simplifications as possible
 - **Do not rely** simply on a bunch of numbers
 - Even worse, a single number



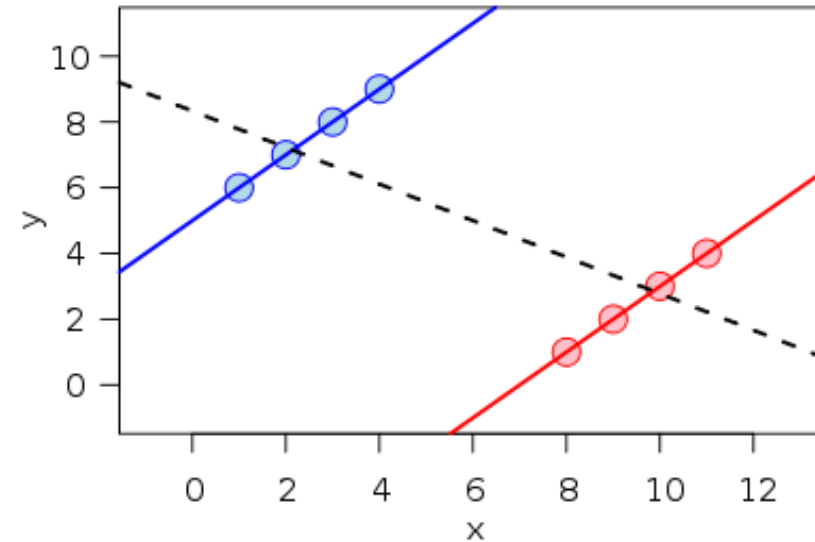
Simpson's Paradox

- 1973, University of California – Berkeley was sued for sex discrimination
 - Accepted 44% male applicants but 35% female applicants
 - When researches dug in, they found it was not so
 - *"If the data are properly pooled...there is a small but statistically significant bias in favor of women."*
- Simpson's paradox
 - A case of **omitted variable** bias
 - Observed explanatory variable → explained variable
 - **Lurking variable**
 - **Uneven sample sizes (in most cases)**
 - The effect of the observed explanatory variable reverses when we take the lurking variable into account



Simpson's Paradox (2)

- When we consider both samples together, it appears that x has a negative effect on y
 - When we take color into account, the relationship reverses
- Other example: kidney stone treatment
 - One treatment is better for large stones, and better for small stones; but the other one is better overall
 - Confounders – the severity of the illness + different sample sizes
- An article with [more info](#) on the topic



UCB Admissions – Explanation

- The research paper concluded that 6 departments were significantly biased towards men and 4 – towards women
 - The other 75 weren't (significantly) biased at all
- Actually, the overall bias was (slightly) in favor of women
- Women tended to apply to competitive departments with low admission rates
- Men tended to apply to less competitive departments with high admission rates
 - We cannot observe that directly from our dataset
 - **Lurking variable** – competitiveness
 - Students didn't have the same motivations to apply

Summary

- Descriptive and inferential statistics
- Population and sample
- Properties of statistical distributions
- Visualizing data
- Covariance and correlation
- Common misconceptions

The image features a white background with two blue decorative bars. The top bar is a solid blue strip. Below it is a thinner, darker blue wavy line. The bottom of the image is framed by a similar pattern of a thin dark blue wavy line above a solid blue strip.

Questions?