

# Week1

## 1. Data Basics

- Observations, variables, and data metrics

In data matrix,

- each row = observation(case)
- each column = variable

- types of variables

<b>numerical (quantitative)</b>	<ul style="list-style-type: none"> <li>- take on numerical values</li> <li>- sensible to add, subtract, take averages, etc. with these values</li> </ul>	continuous	<ul style="list-style-type: none"> <li>- take on any of an infinite number of values within a given range</li> <li>- ex : height</li> </ul>
		discrete	<ul style="list-style-type: none"> <li>- take on one of a specific set of numeric values where we're able to count or enumerate all of the possibilities</li> <li>- ex : count numbers</li> </ul>
<b>categorical (qualitative)</b>	<ul style="list-style-type: none"> <li>- take on a limited number of distinct categories</li> <li>- categories can be identified with numbers, but not sensible to do arithmetic operations</li> </ul>	regular categorical	
		ordinal	<ul style="list-style-type: none"> <li>- levels have an inherent ordering</li> <li>- ex : satisfaction levels</li> </ul>

- relationships between variables

- When two variables that show some connection with one another are called associated(dependent)
- Association can be further described as positive or negative.
- If two variables are not associated, they are said to be independent.

## 2. Observational studies and experiments

- Observational studies and experiments

studies	observational	<ul style="list-style-type: none"> <li>- observational studies collect data in a way that does not directly interfere with how the data or rise.</li> </ul>
---------	---------------	---

		<ul style="list-style-type: none"> <li>- only establish an association (correlation between the explanatory and the response variables)</li> <li>- retrospective : uses past data</li> <li>- prospective : data are collected throughout the study</li> </ul>
	experiment	<ul style="list-style-type: none"> <li>- Randomly assign subjects to treatments</li> <li>- Establish causal connections</li> </ul>

- confounding variables : Extraneous variables that affect both the explanatory and the response variable, and that make it seem like there is a relationship between them (교란변수)

- Correlation and causation
  - Correlation does not imply causation.
  - What determines whether we can infer causation or just correlation is the type of study that we are basing our conclusions on.

### 3. Sampling and sources of bias

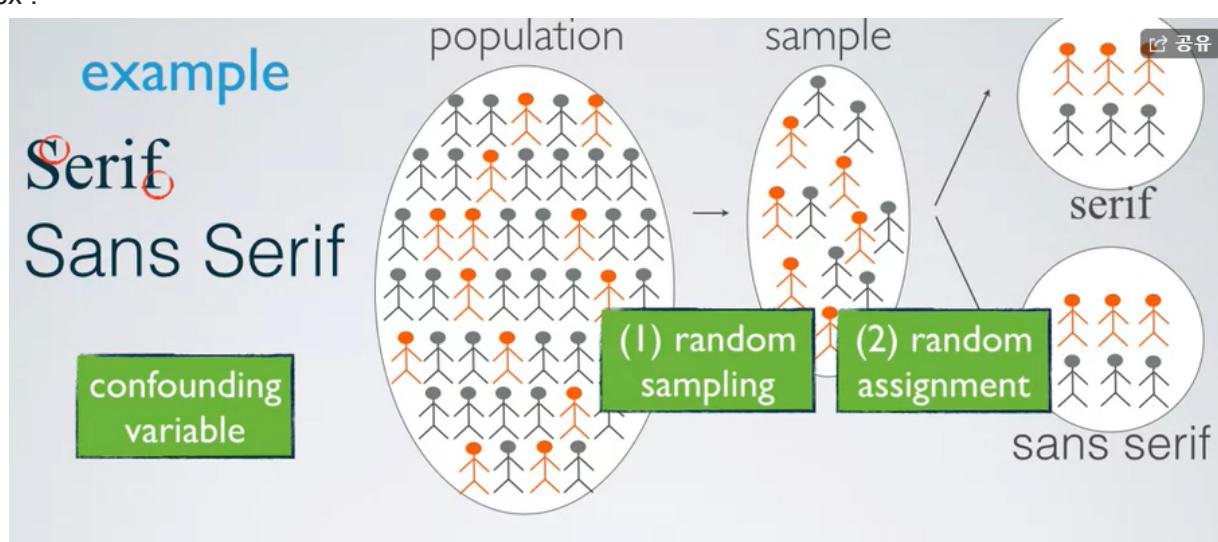
- Census
  - the procedure of systematically calculating, acquiring and recording **information** about the members of a given **population**.
  - There are reasons why conducting a census is not a good idea.
    1. Some individuals are hard to locate or measure, and these people be different from the rest of the population.
    2. Populations rarely stand still.
  - ex : Cooking
    1. Explanatory analysis : When you taste a spoonful of soup and decide that Spoonful you're tasted isn't salty enough
    2. Inference : You then generalize and conclude that your entire needs salt, that's making an inference.
    3. Representative sample : If you first stir the soup thoroughly before you taste, your spoonful will be more likely to be representative of the whole pot.
- A few sources of sampling bias
  1. Convenience sample : Individuals who are easily accessible are more likely to be included in the sample
  2. Non-response : If only a (non-random) fraction of the randomly sampled people respond to a survey such that the sample is no longer representative of the population.
  3. Voluntary response : Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue.
- Sampling methods
  1. Simple Random Sample(SRS) : Each case is equally likely to be selected.
  2. stratified sample : Firstly divide the population into homogeneous groups called Strada and then randomly sample from within each stratum.
  3. cluster sample : we divide the population into clusters randomly sample a few clusters and then sample all observations within these clusters.
  4. multistage sample : It adds another step to Cluster sampling. Just like in cluster sampling we divide the population into clusters, randomly sample a few clusters, and then we randomly sample observations from within these clusters. Usually we use cluster sampling and multi-stage sampling for economical reasons.

## ● Experimental Design

- Principles of experimental design
  1. control : compare treatment of interest to a control group
  2. randomize : randomly assign subjects to treatments
  3. replicate : Collect a sufficiently large sample or replicate the entire study
  4. block : block for variables known or suspected to affect the outcome
- blocking vs. explanatory variables
  - explanatory variables(factors) : conditions we can impose on experimental units
  - blocking variables : characteristics that the experimental units come with, that we would like to control for
  - blocking is like stratifying (계층화)
    - blocking during random assignment
    - stratifying during random sampling
- Experimental terminology
  - placebo : fake treatment, often used as the control group for medical studies
  - placebo effect : showing change despite being on the placebo
  - blinding : experimental units don't know which group they're in
  - double-blind : both the experimental units and the researchers don't know the group assignment
  -

## ● Random Sample Assignment

- Random Sampling
  - occurs when subjects are being selected for a study
  - the resulting sample is likely representative of the population
  - the study's results are **generalizable** to the population at large
- Random Assignment
  - occurs only in experimental settings, where subjects are being assigned to various treatments
  - we usually see that the subjects exhibit slightly different characteristics from one another
  - These different characteristics are represented equally in the treatment and control groups.
  - In other words, random assignment allows us to make **causal conclusions** based on the study.
- ex :



ideal experiment	Random assignment	No random assignment	most observational studies
Random sampling	causal and generalizable	not causal, but generalizable	Generalizability
No random sampling	causal, but not generalizable	neither causal nor generalizable	No generalizability
most experiments	Causation	Association	bad observational studies

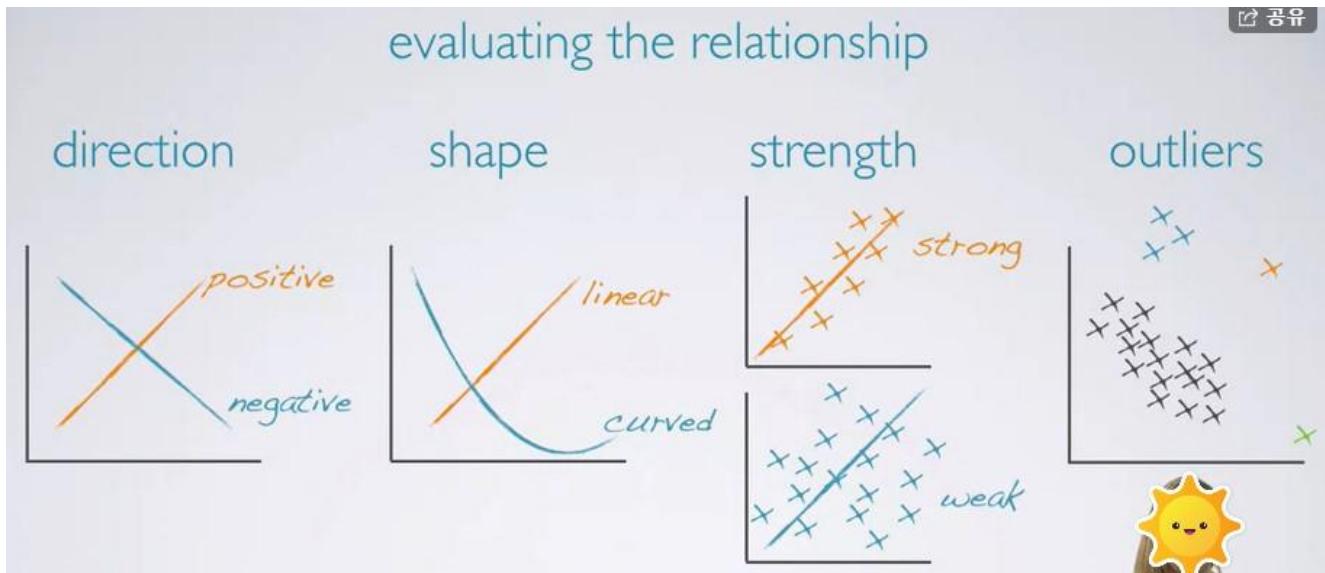
# Week2

[Exploring Numerical Data]

## 1. Visualizing Numerical Data

- Scatter Plot

- A common tool for visualizing the relationship between two numerical variables is a **scatter plot**.
- It's very important to note that labeling variables as explanatory and response does not guarantee that the relationship between the two is actually causal.
- evaluating the relationship



- Histogram

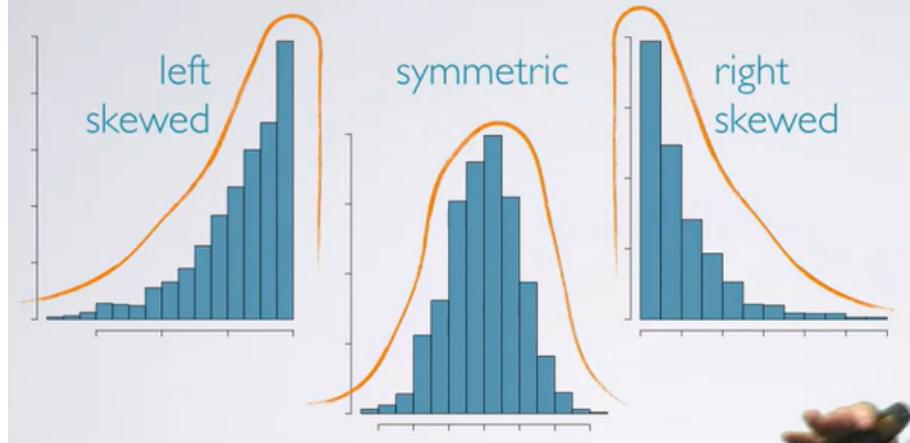
- One good way of visualizing the distribution of a numerical variable is a histogram.
- A histogram provides a view of the data density.
- especially useful for describing the shape of the distribution

- Skewness

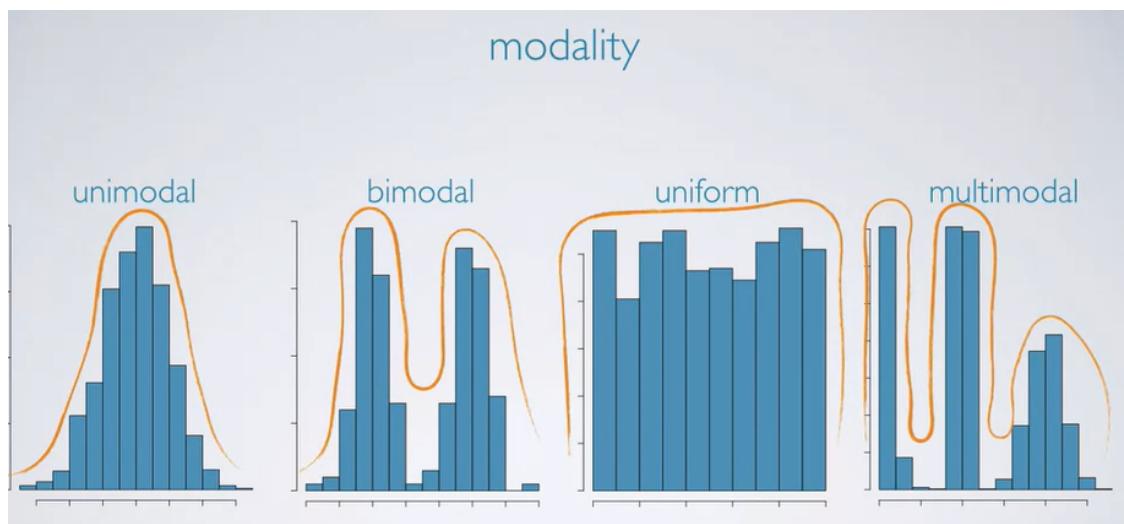
- Distributions are set to be skewed to the left or right side of the long tail.

## skewness

distributions are skewed to the side of the long tail



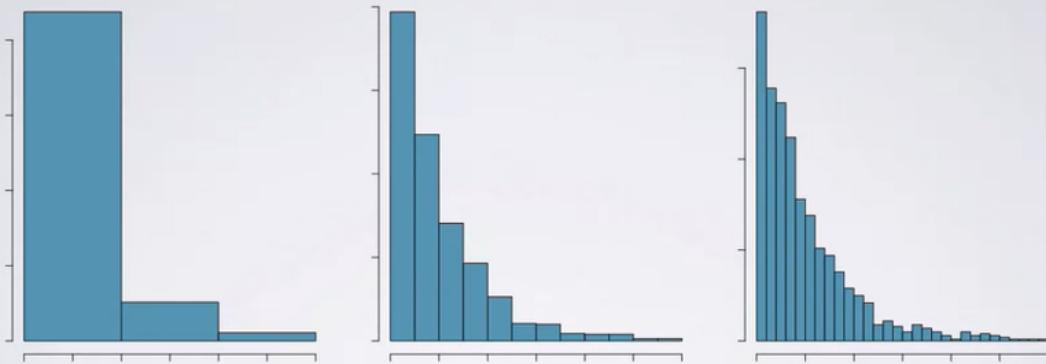
- Modality



- histogram & bin width

## histogram & bin width

The chosen **bin width** can alter the story the histogram is telling.



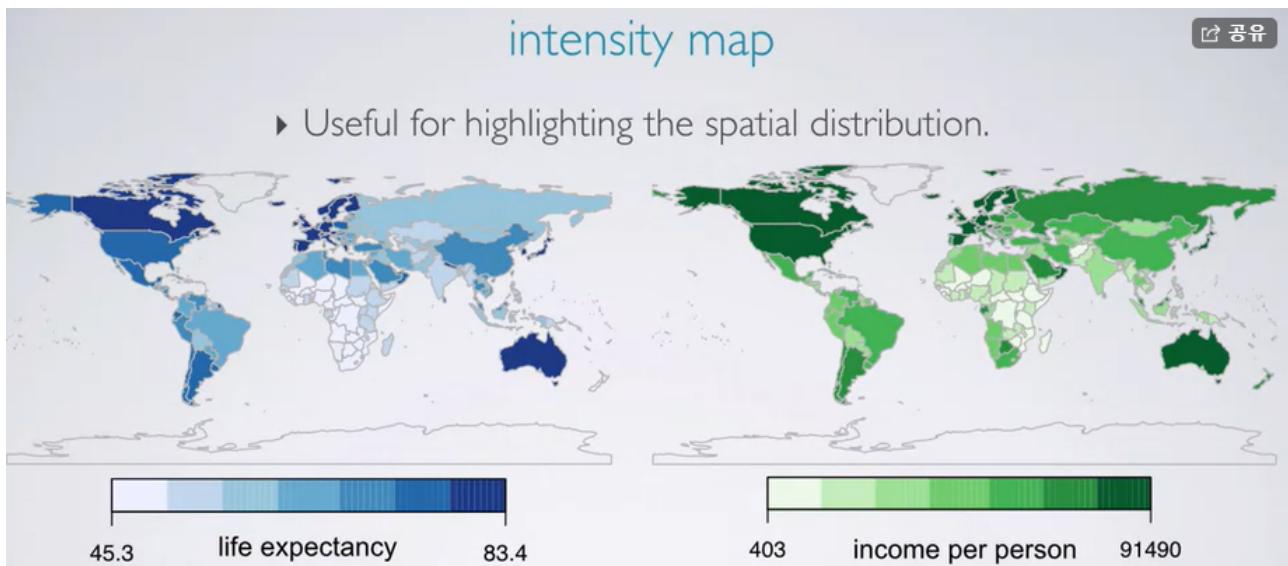
- Dotplot

- Another technique for visualizing such data is a Dot Plot.
- A dot plot is especially useful when individual values are of interest.
- Can get busy as the sample size increases.

- Box plot

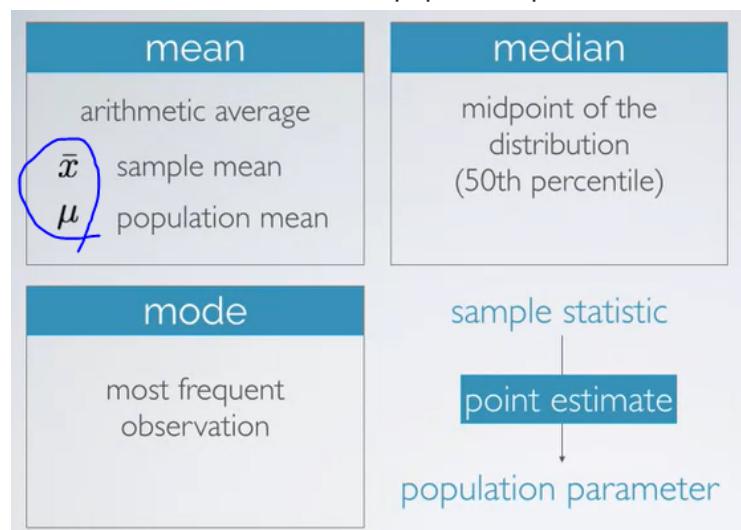
- Useful for highlighting outliers, median, IQR

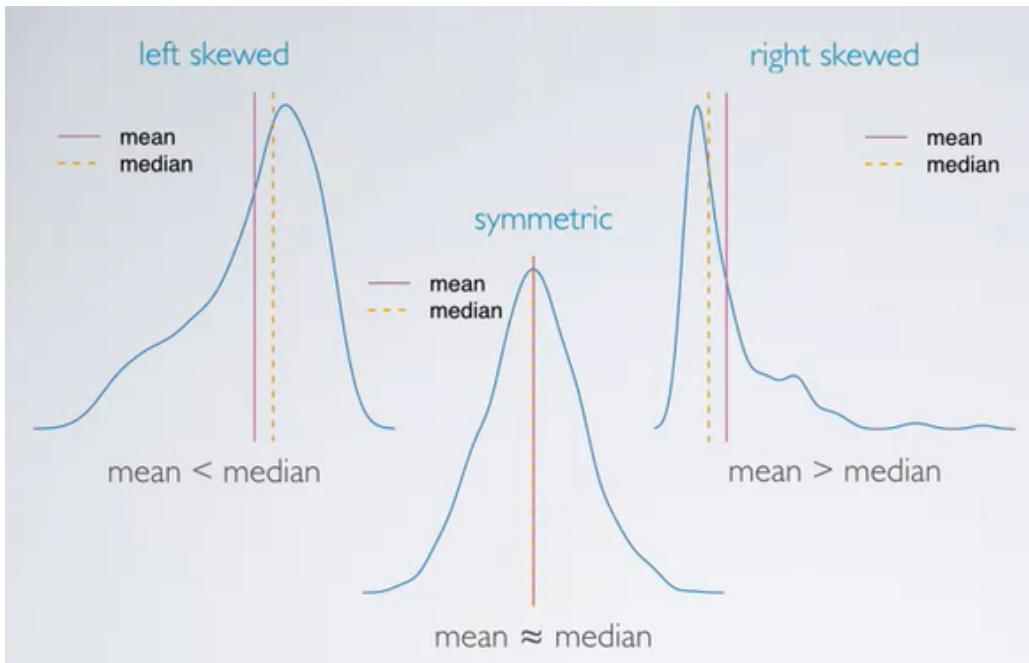
- Intensity map



## 2. Measures of center

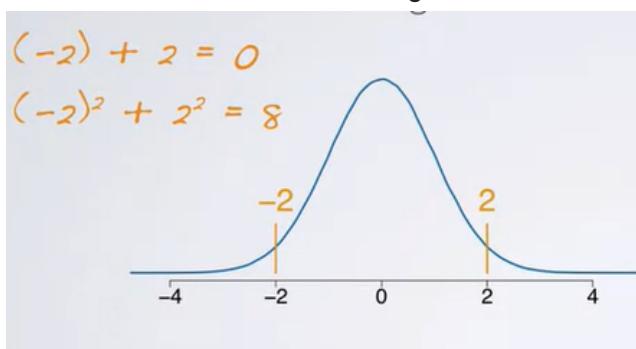
- mean : arithmetic average
- median : midpoint of the distribution
- mode : most frequent observation
- If these measurements are calculated from a sample they are called sample statistics.
- Sample statistics are points estimates for the unknown population parameters.





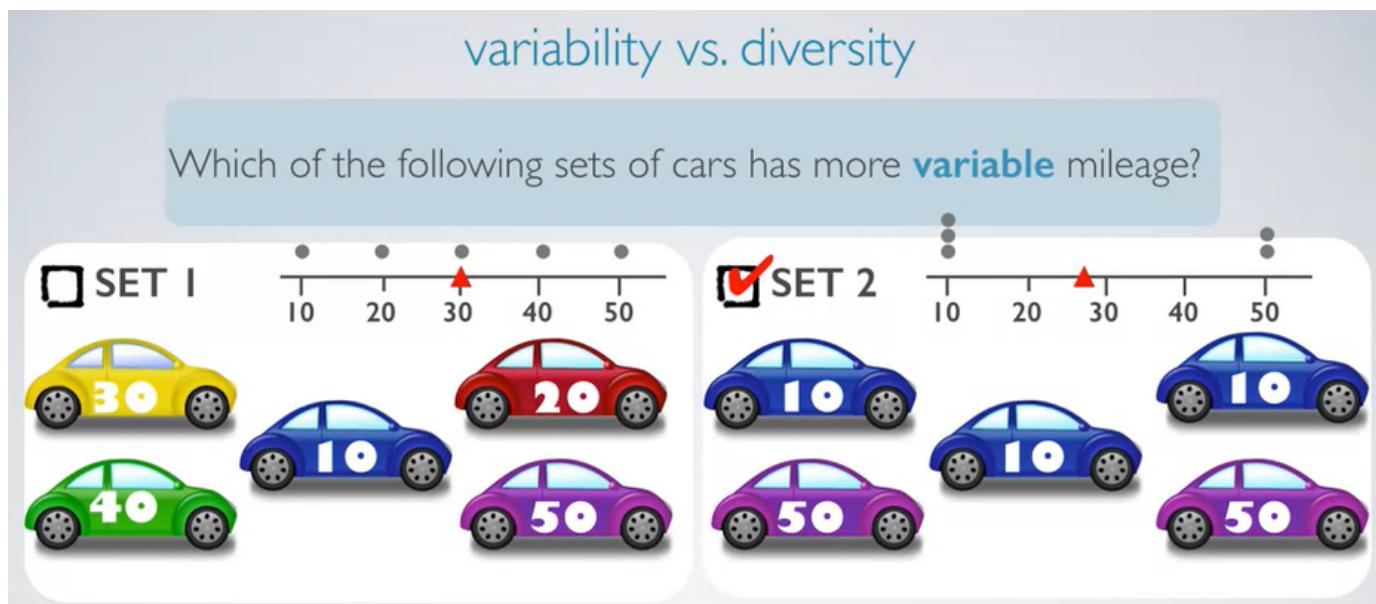
### 3. Measures of Spread

- range : (max-min)
  - not so reliable as it relies on two extreme values
- variance
  - roughly the average squared deviation from the mean
  - sample variance  $s^2$
  - population variance  $\sigma^2$
  - $$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$
  - Why do we square the differences?
    - to get rid of negatives so that negatives and positives don't cancel each other when added together
    - To increase larger deviations more than smaller ones so that they are weighed more heavily



- standard deviation
  - Roughly the average deviation around the mean and has the same units as the data
  - sample sd  $s$
  - population sd  $\sigma$

- variability vs. diversity
  - The set with more data at the ends of the distribution (away from the center) is more variable.



- inter-quartile range
  - Range of the middle 50% of the data, distance between the first quartile (25th percentile) and 3rd quartile (75th percentile)
  - $IQR = Q3 - Q1$
  - This measure is most readily available in a box plot.
  -

## 4. Robust Statistics

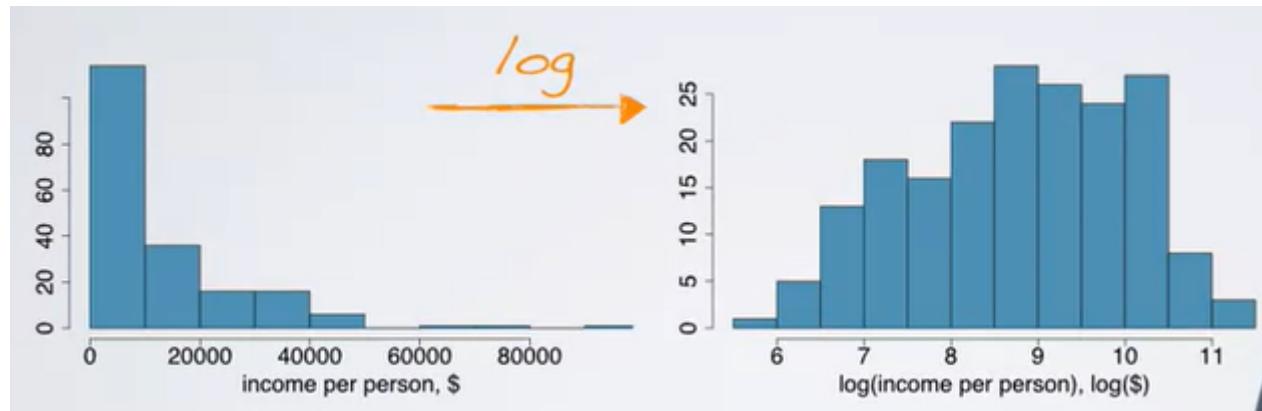
- In statistics, the term robust or robustness refers to the strength of a statistical model, tests, and procedures according to the specific conditions of the statistical analysis a study hopes to achieve.  
(<https://www.thoughtco.com/what-is-robustness-in-statistics-3126323>)
- We define robust statistics as measures on which extreme observations have little effect.

	robust	non-robust
center	median	mean
spread	IQR	SD, range
<i>skewed, with extreme observations</i>		<i>symmetric</i>

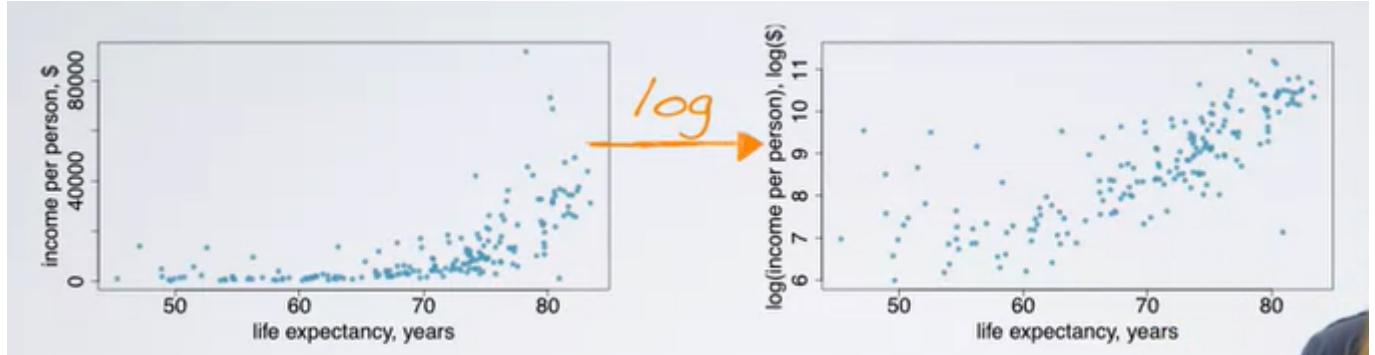
## 5. Transforming Data

- Transformations

- A transformation is a **rescaling** of the data using a function
- when data are very strongly skewed, we sometimes transform them so they are easier to model
- **(natural) transformation** is often applied when much of the data cluster is near zero (relative to the larger values in the data set) and all observations are positive.



- **log transformation** : To make the relationship between the variables more **linear** and hence easier to model with simple methods



- goals of transformations
  - To see the data structure differently
  - To reduce skew assist in modeling
  - To straighten a nonlinear relationship in a scatter plot

## 1. Exploring Categorical Variables

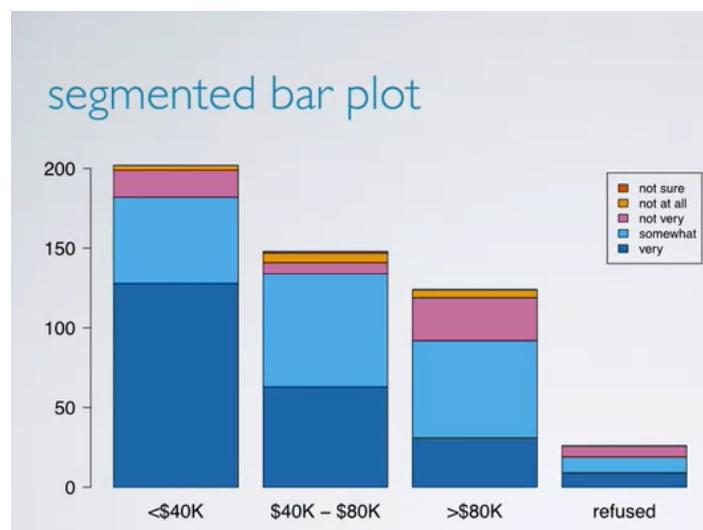
- Frequency table & bar plot
- How are bar plots different than histograms?
- Bar plots for displaying distributions of categorical variables, histograms for numerical variables
- x-axis on a histogram is a number line, and the ordering of the bars are not interchangeable
- In a bar plot, the categories can be listed in any order
- contingency table(교차표)

		relative frequencies				
		< \$40K	\$40K - \$80K	> \$80K	Refused	Total
Difficulty saving	Very	128	63	31	9	231
	Somewhat	54	71	61	10	196
	Not very	17	7	27	7	58
	Not at all	3	6	5	0	14
	Not sure	0	1	0	0	1
Total		202	148	124	26	500

$\text{<} \$40K: 128 / 202 = 63\%$  find it very difficult to save  
 $\$40K-\$80K: 63 / 148 = 43\%$   
 $\text{>} \$80K: 31 / 124 = 25\%$   
 Refused:  $9 / 26 = 35\%$

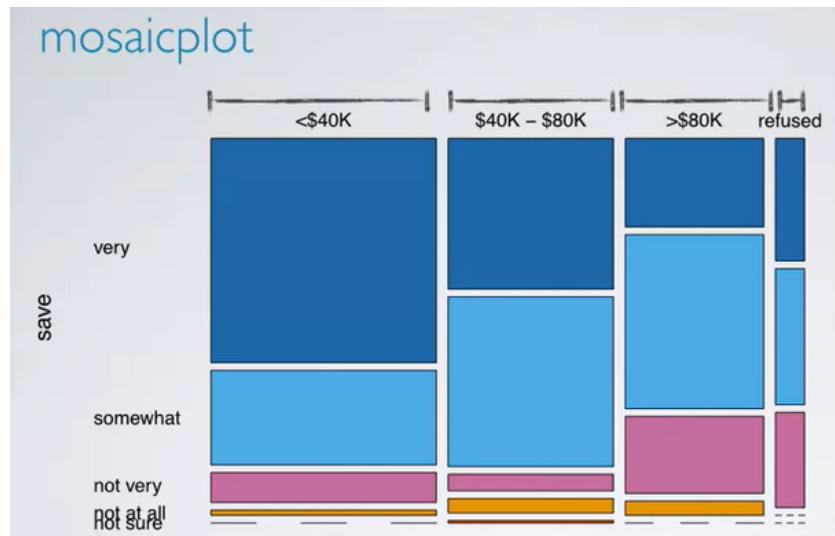
feelings about difficulty of saving money and income are associated (dependent)

- segmented bar plot



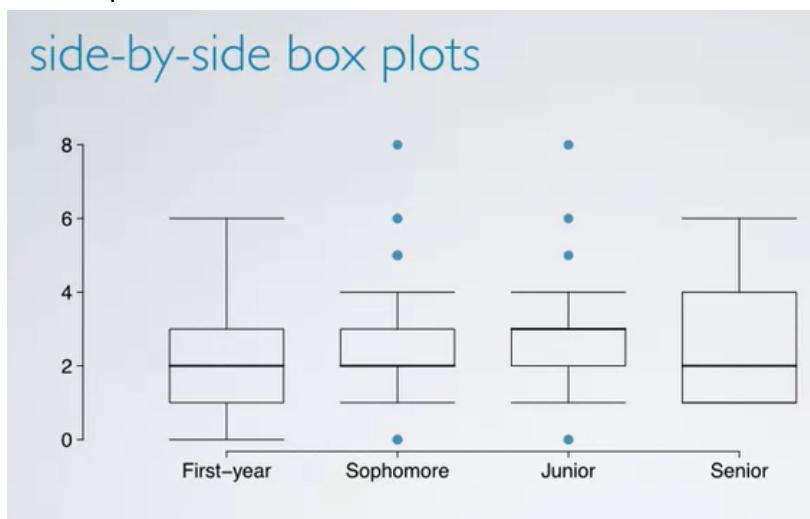
- useful for visualizing conditional frequency distributions
- compare relative frequencies to explore relationship between the variables

- mosaic plot



- relationship between numerical and categorical variables

- side-by-side box plots



## 2. Introduction to Inference

- Gender discrimination

- 48 male bank supervisors given the same personnel file, asked to judge whether the person should be promoted
- files were identical, except for gender of applicant
- random assignment (experiment)
- 35 / 48 promoted

## data

		promotion		
		promoted	not promoted	total
gender	male	21	3	24
	female	14	10	24
	total	35	13	48

$\% \text{ of males promoted} = 21/24 \approx 88\%$

$\% \text{ of females promoted} = 14/24 \approx 58\%$

- two competing claims
- **Null hypothesis(귀무 가설)**
  - Promotions and gender are independent
  - it's simply due to chance
- **Alternative hypothesis (대립 가설)**
  - it is dependent, there is gender discrimination

### recap: hypothesis testing framework

- ▶ start with a **null hypothesis ( $H_0$ )** that represents the status quo
- ▶ set an **alternative hypothesis ( $H_A$ )** that represents the research question, i.e. what we're testing for
- ▶ conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation or theoretical methods
  - ▶ if the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, stick with the null hypothesis
  - ▶ if they do, then reject the null hypothesis in favor of the alternative

### simulation scheme

[use a deck of playing cards to simulate this experiment]

1. face card: not promoted, non-face card: promoted
  - ▶ set aside the jokers, consider aces as face cards
  - ▶ take out 3 aces → exactly 13 face cards left in the deck (face cards: A, K, Q, J)
  - ▶ take out a number card → 35 number (non-face) cards left in the deck (number cards: 2-10)
2. shuffle the cards, deal into two groups of size 24, representing males and females
3. count how many number cards are in each group (representing promoted files)
4. calculate the proportion of promoted files in each group, take the difference (male - female), and record this value
5. repeat steps 2 - 4 many times

## making a decision

- ▶ results from the simulations look like the data → the difference between the proportions of promoted files between males and females was **due to chance** (promotion and gender are **independent**)
- ▶ results from the simulations do not look like the data → the difference between the proportions of promoted files between males and females was **not** due to chance, but **due to an actual effect of gender** (promotion and gender are **dependent**)

- summary
- set a null and an alternative hypothesis
- simulate the experiment assuming that the null hypothesis is true
- **evaluated the probability of observing an outcome at least as extreme as the one observed in the original data -> p-value**  
(귀무가설이 맞다고 가정할 때 얻은 결과보다 극단적인 결과가 실제로 관측될 확률)
- if this probability is low, reject the null hypothesis in favor of the alternative