Week1

1. Data Basics

- Observations, variables, and data metrics

In data matix,

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

- types of variables

numerical (quantitative)	- take on numerical values - sensible to add, substract, take averages, etc.with these values	continuous	- take on any of an infinite number of values within a given range - ex : height
		discrete	- take on one of a specific set of numeric values where we're able to count or enumerate all of the possibilities - ex : count numbers
categorical (qualitative)	- take on a limited number of distinct categories - categories can be identified with numbers, but not sensible to do arithmetic operations	regular categorical	
		ordinal	- levels have an inherent ordering - ex : satisfaction levels

- 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	- observational studies collect data in a way that does not directly interfere with how the data or rise.

		- only establish an association (correlation between the explanatory and the response variables) - retrospective : uses past data - prospective : data are collected throughout the study
	experiment	- Randomly assign subjects to treatments - Establish causal connections

- 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

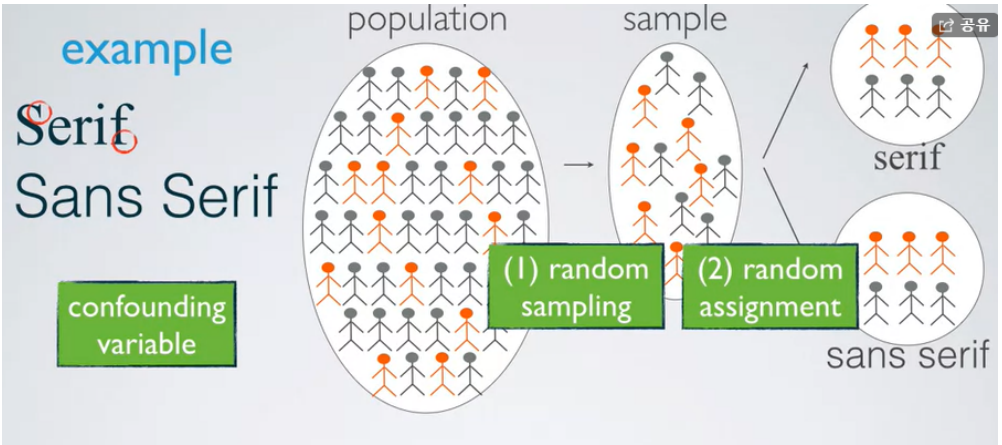
- Principles of experimental design
    1. control : compare treatment of interest to a control group
    2. randomize : randomly assign subjects to treatments
    3. replicate : Collect a sufficiently large sample or replicate the entire study
    4. block : block for variables known or suspected to affect the outcome

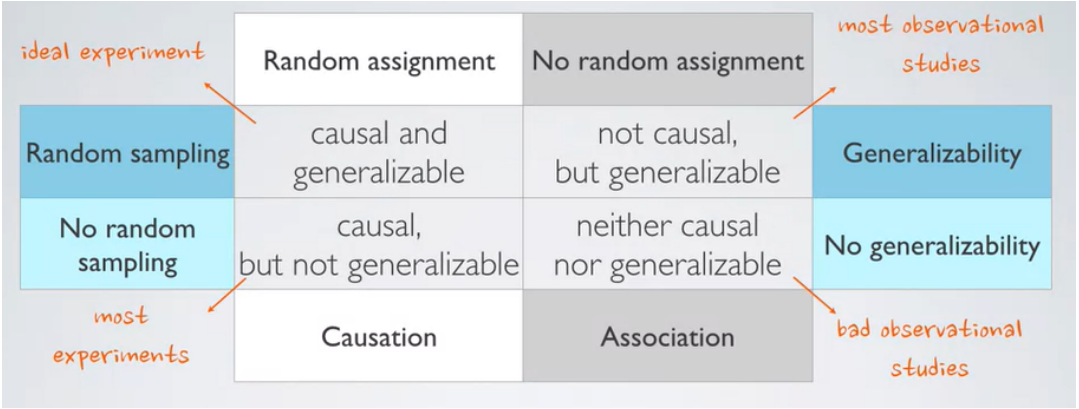
- blocking vs. explanatory variables
    - explanatory variables(factors) : conditions we can impose on experimental units
    - blocking variables : characteristics that the experimental units come with, that we would like to control for
    - blocking is like stratifying (계층화)
        - blocking during 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

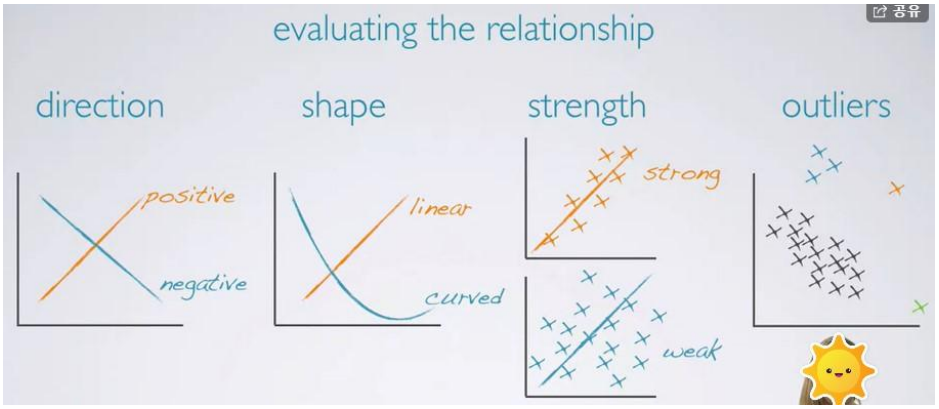
Week2

[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

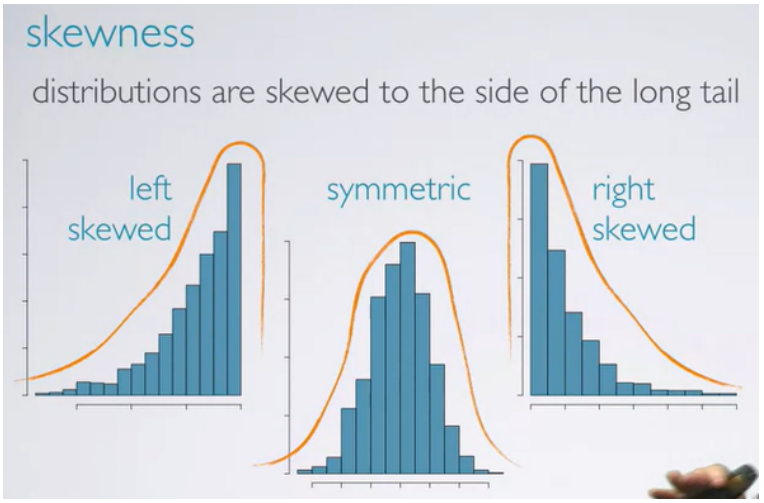


● Histogram

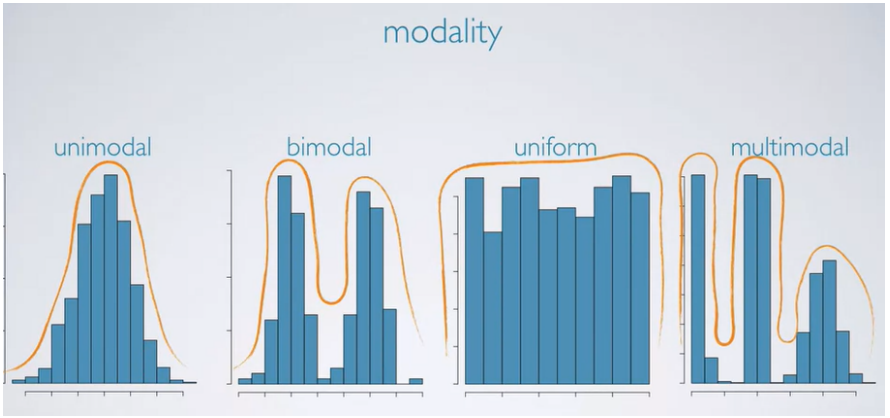
- One good way of visualizing the distribution of a numerical variable is a histogram.
- A histogram provides a view of the data density.
- especially useful for describing the shape of the distribution

● Skewness

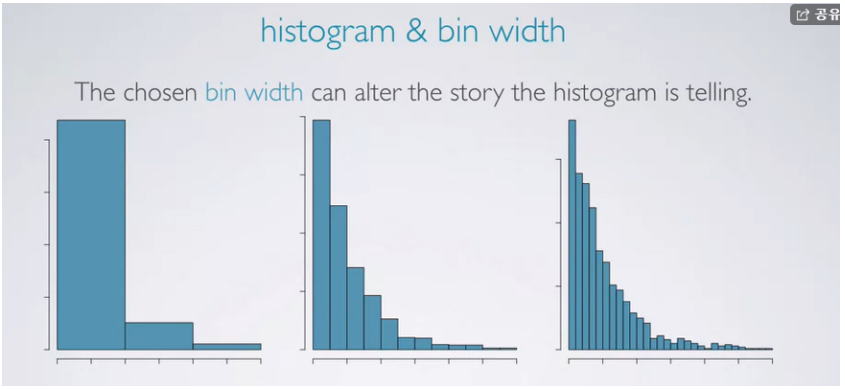
- Distributions are set to be skewed to the left or right side of the long tail.



- Modality



- histogram & bin width



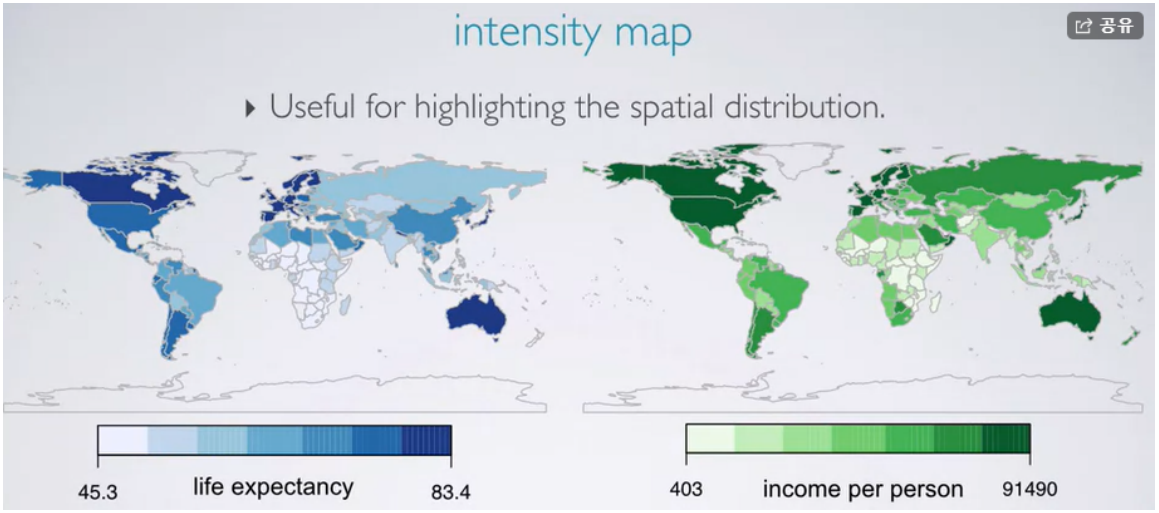
- Dotplot

- 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

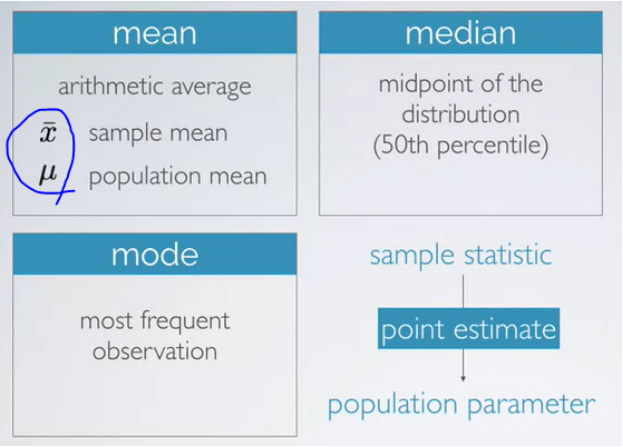
- Intensity map

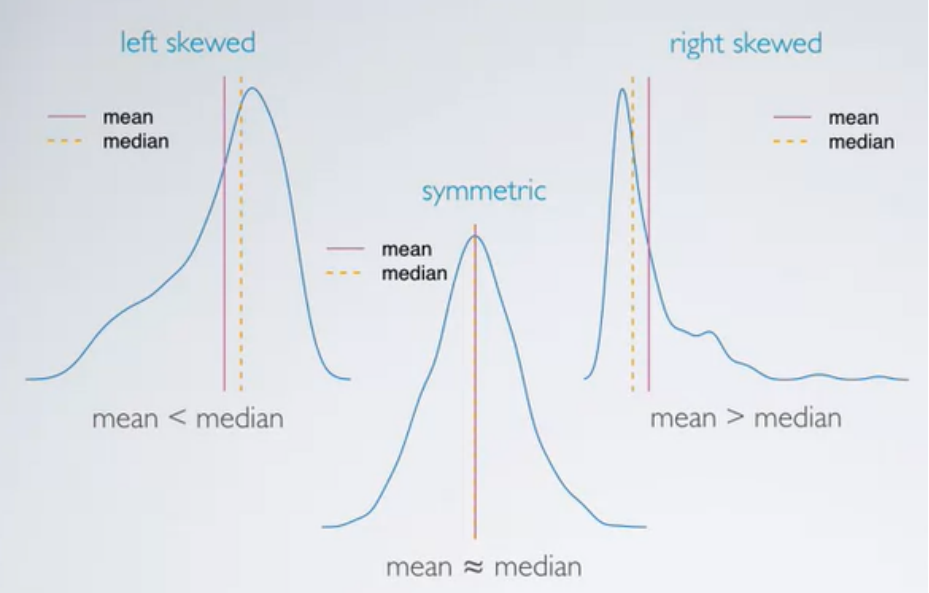


2. Measures of center

- mean : arithmetic average
- median : midpoint of the distribution
- mode : most frequent observation

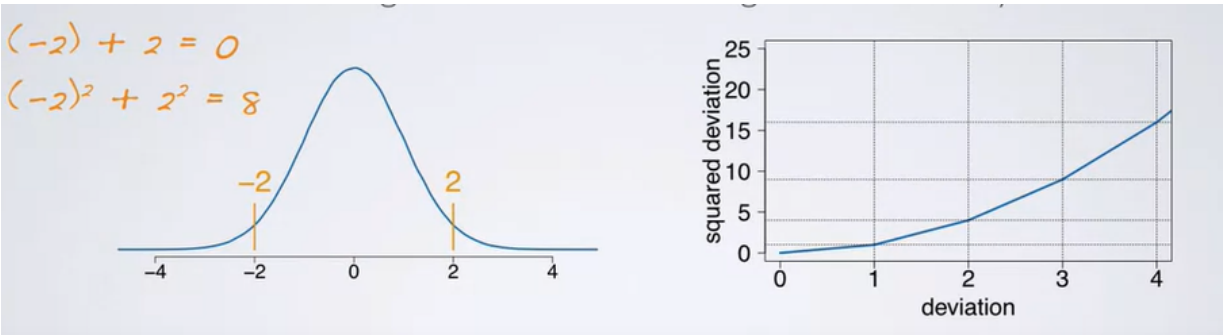
- If these measurements are calculated from a sample they are called sample statistics.
- Sample statistics are points estimates for the unknown population parameters.





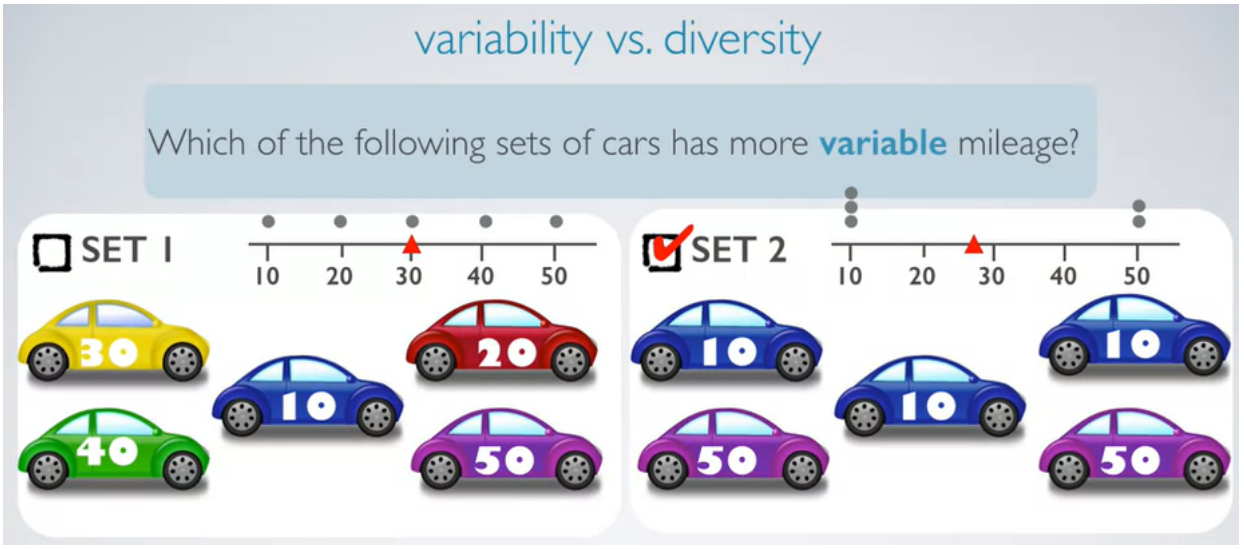
3. Measures of Spread

- 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 σ^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 σ

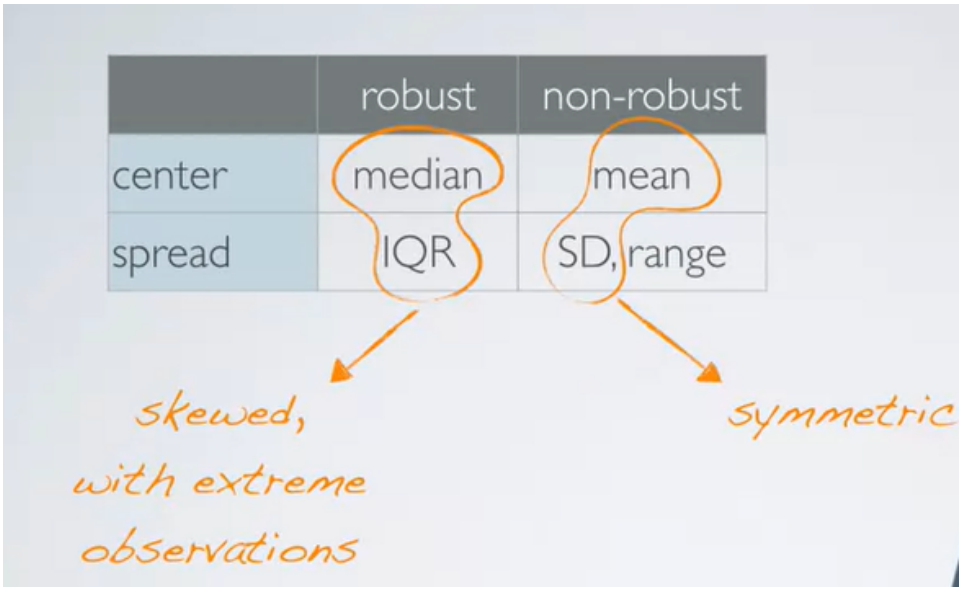
- variability vs. diversity
    - The set with more data at the ends of the distribution (away from the center) is more variable.



- 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

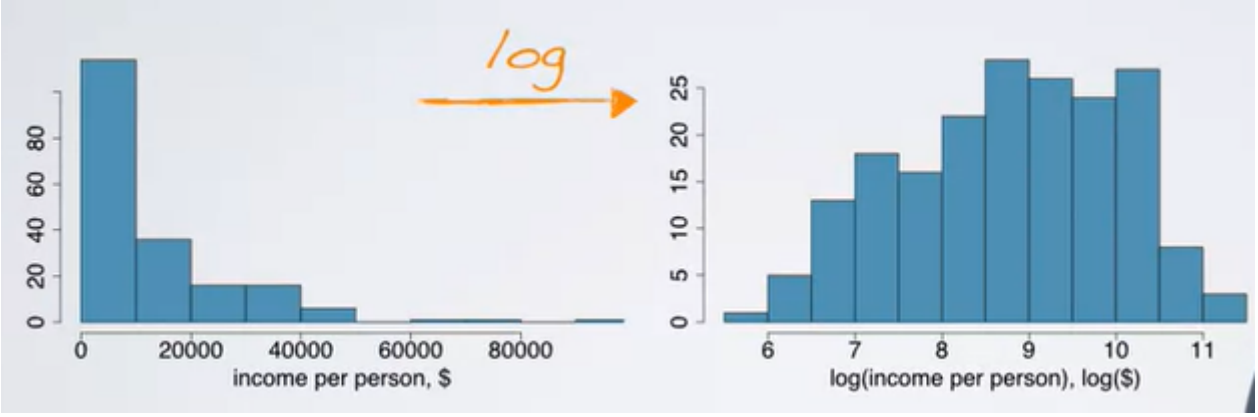
- In statistics, the term robust or robustness refers to the strength of a statistical model, tests, and procedures according to the specific conditions of the statistical analysis a study hopes to achieve. (https://www.thoughtco.com/what-is-robustness-in-statistics-3126323)
- We define robust statistics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



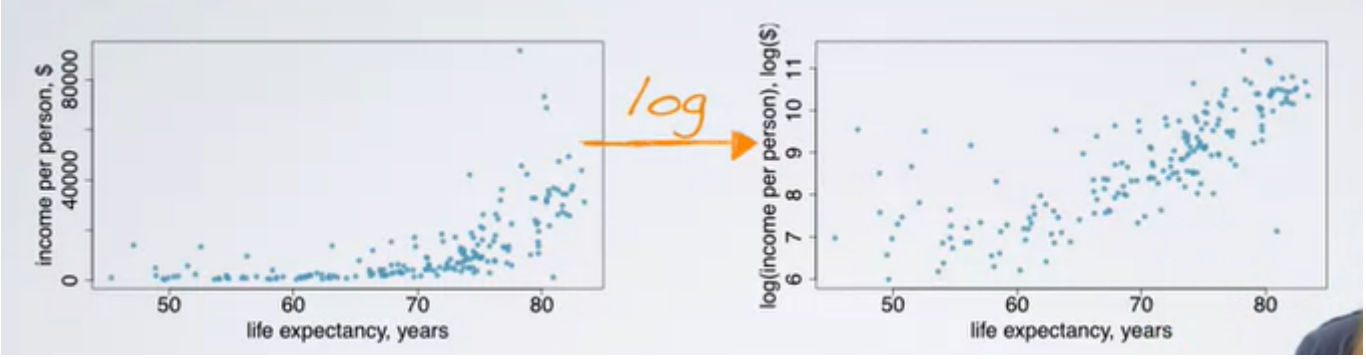
5. Transforming Data

● 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

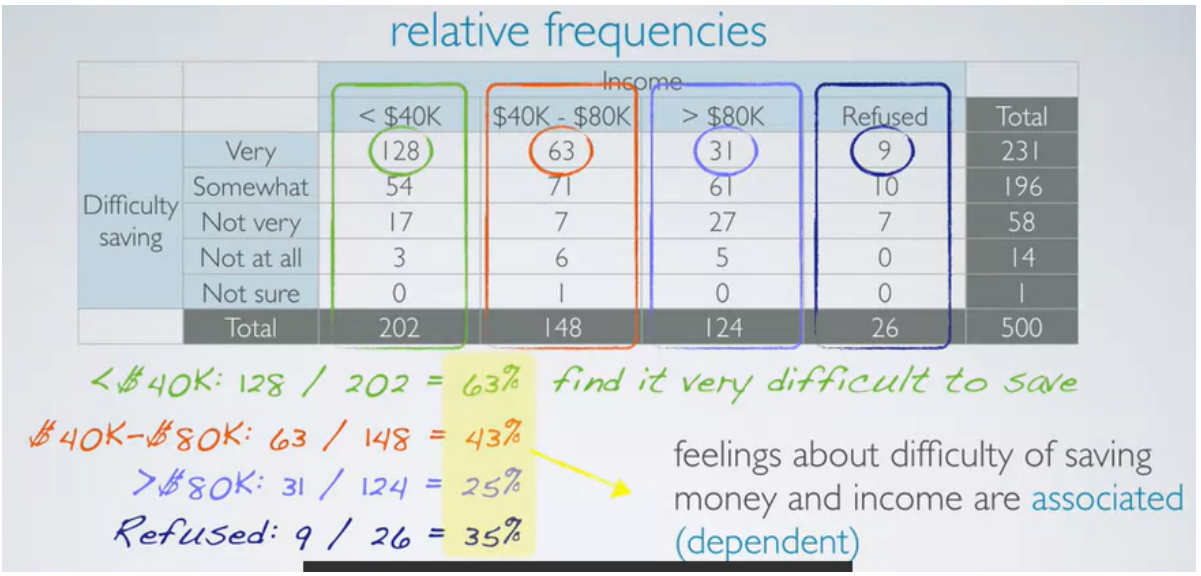
[Exploring Categorical Data and Introduction to Inference]

1. Exploring Categorical Variables

- Frequency table & bar plot
- How are bar plots different than histograms?

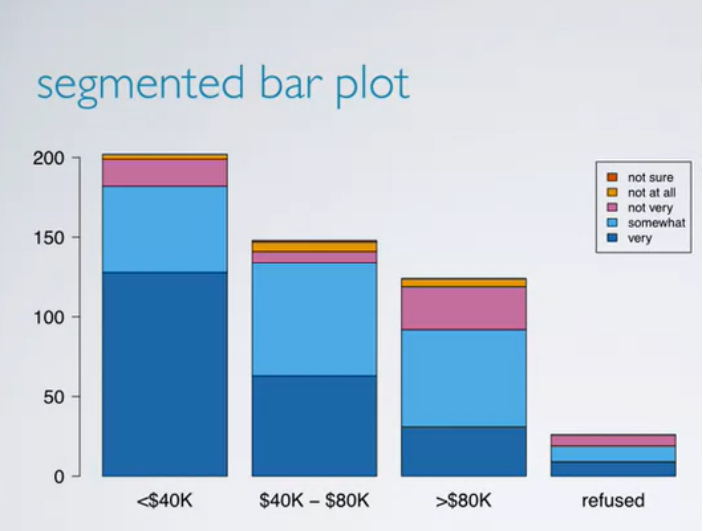
- Bar plots for displaying distributions of categorical variables, histograms for numerical variables
- x-axis on a histogram is a number line, and the ordering of the bars are not interchangeable
- In a bar plot, the categories can be listed in any order

- contingency table(교차표)



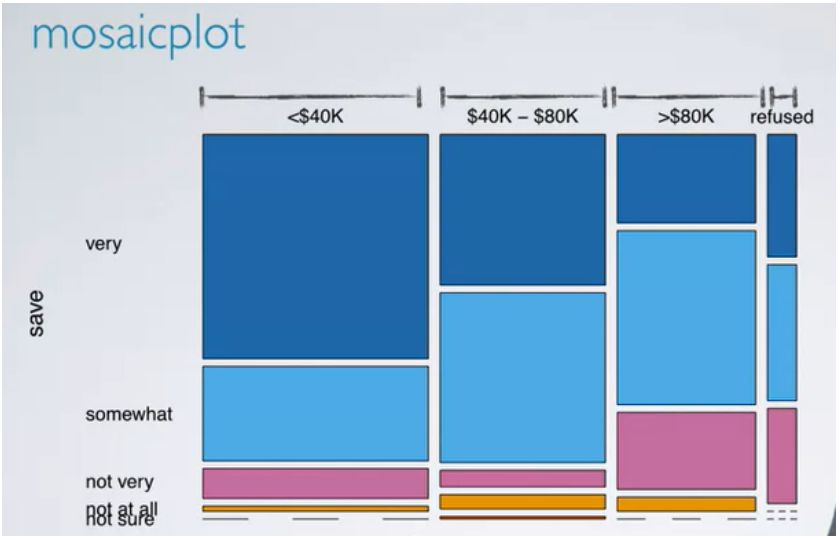
-

- segmented bar plot

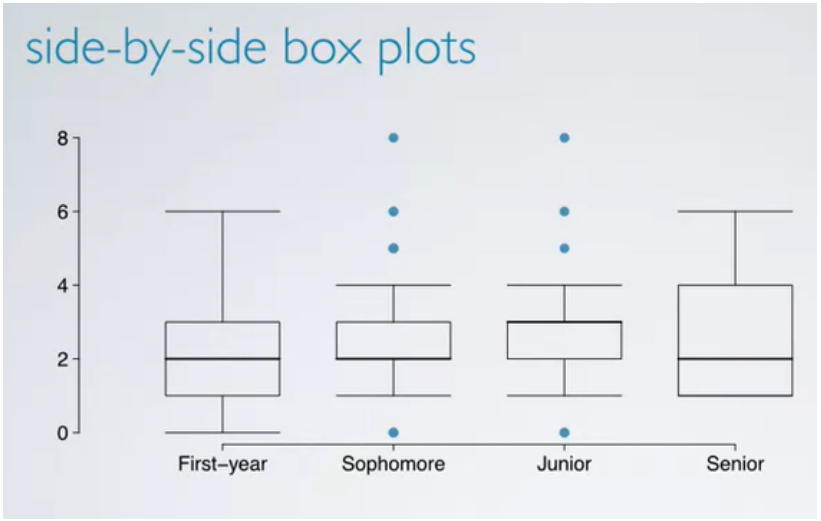


- useful for visualizing conditional frequency distributions
- compare relative frequencies to explore relationship between the variables

- 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

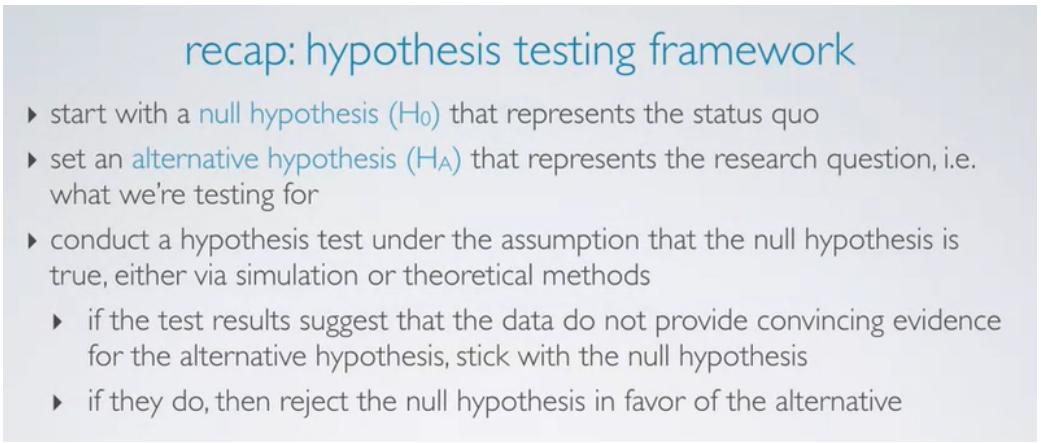
		promotion		
		promoted	not promoted	total
gender	male	21	3	24
	female	14	10	24
	total	35	13	48

% of males promoted = 21/24 ≈ 88%

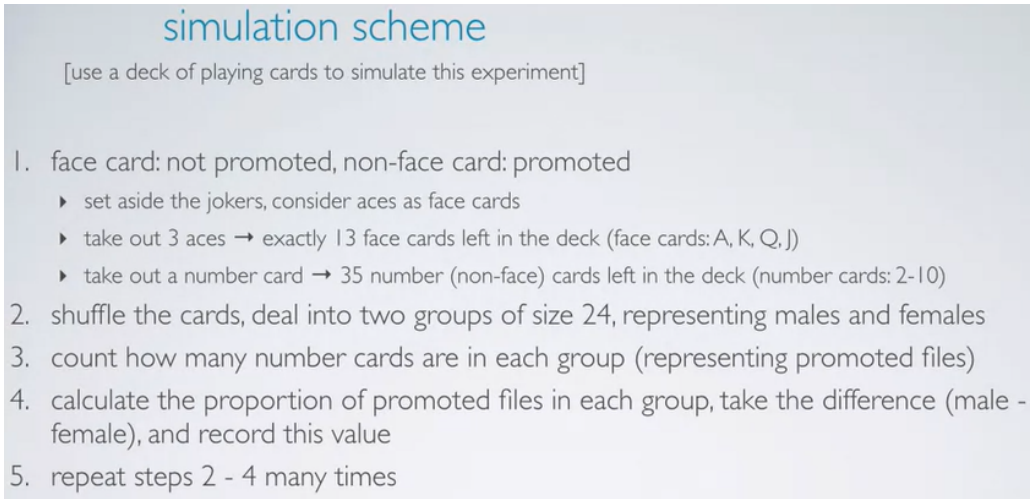
% of females promoted = 14/24 ≈ 58%

- two competing claims

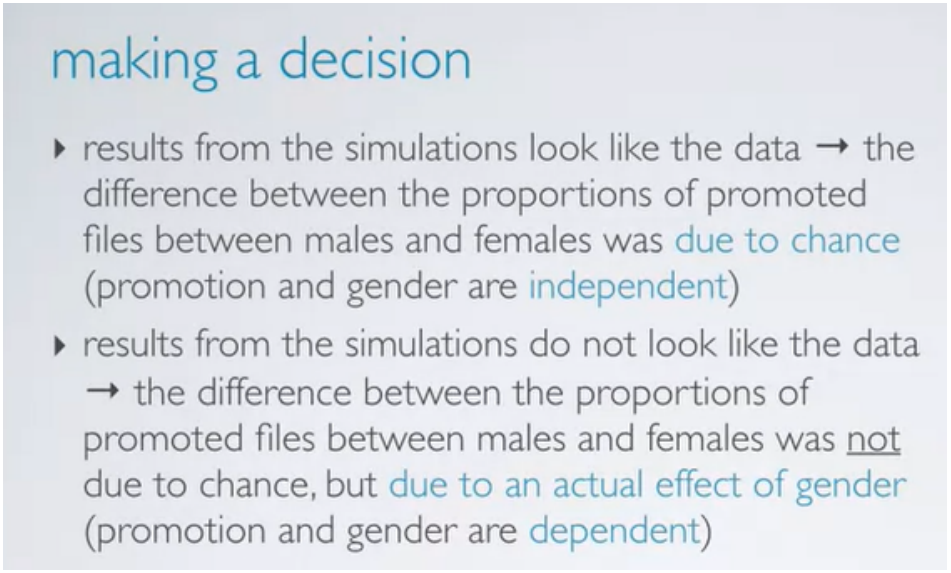
- **Null hypothesis(**귀무 가설**)**
  - Promotions and gender are independent
  - it's simply due to change
- **Alternative hypothesis (**대립 가설**)**
  - it is dependent, there is gender discrimination



-



-

- 

- 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 an 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

Week3

[Defining Probability]

1. Probability and Distributions

- Random Process
- In a random process, we know what outcomes could happen but we don't know which particular outcome will happen.
- ex : coin tosses, die rolls, the shuffle mode on music player, stock market

- probability (P(A) = Probability of event A)
- Frequentist interpretation : The probability of an outcome is the proportion of times the outcome would occur if we observed the random process an infinite number of times.
- Bayesian interpretation : A Bayesian interprets probability as a subjective degree of belief.

- law of large numbers : As more observations are collected, the proportion of occurrences with a particular outcome converges to the probability of that outcome .
- Common misunderstanding of it is the gambler's fallacy(law of averages).

2. Disjoint Events + General Addition Rule

- Disjoint (mutually exclusive)
- disjoint events cannot happen at the same time.
- ex : A student can’t both fail and pass a class.

- non-disjoint events can happen at the same time.

- union of disjoint events : What is the probability of disjoint event A or disjoint event B?
  -> P(A or B) = P(A) + P(B)
- union of non-disjoint events : What is the probability of non-disjoint event A or non-disjoint event B?
  -> P(A or B) = P(A) + P(B) - P(A and B)

- Sample space
- A collection of all possible outcomes of a trial

- Probability distributions
- It lists all possible outcomes in the sample space, and the probabilities with which they occur.
- rule

The events listed must be disjoint.
Each probability must be between 0 and 1
The probability distribution must total 1.

- complementary events
- Complementary events are two mutually exclusive events whose probabilities add up to 1.

-

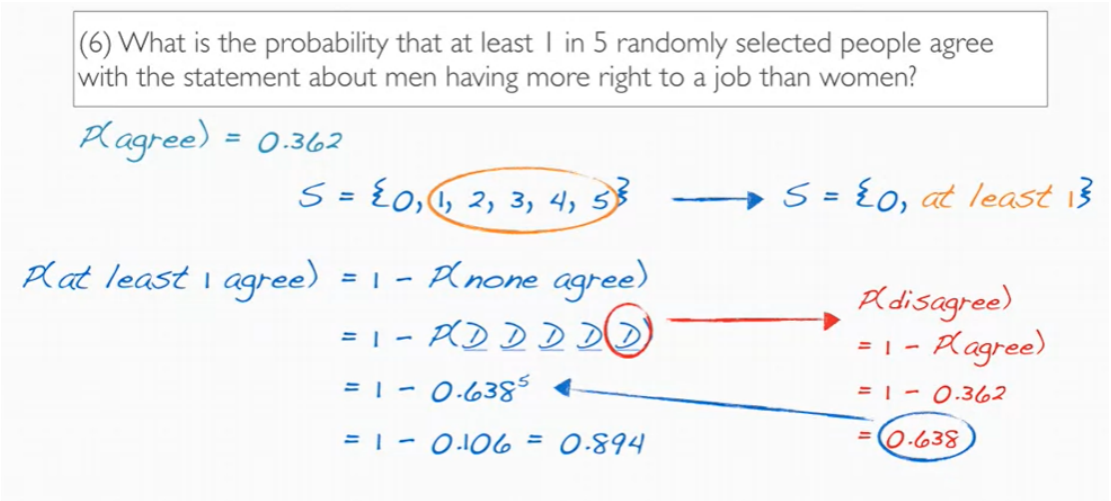
- disjoint vs. complementary
- Do the sum of probabilities of two disjoint outcomes always add up to 1 ?
  -> **Not necessarily**, there may be more than two outcomes in the sample space.
- Do the sum of probabilities of two complementary outcomes always add up to 1?
  -> **Yes**, that's the definition of complementary.

3. Independence

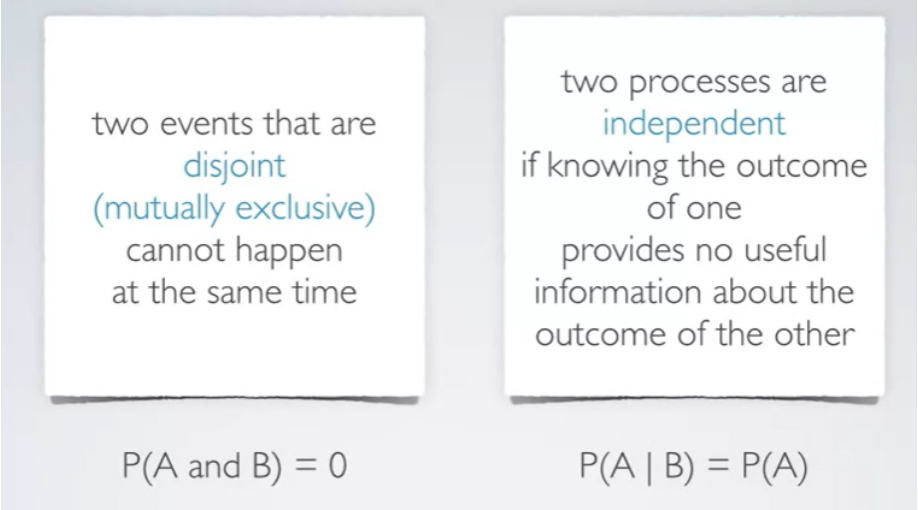
- Independence
- Two processes are independent if knowing the outcome of one provides no useful information about the outcome of the other.
- Checking for independence : P(A|B) = P(A), then A and B are independent.

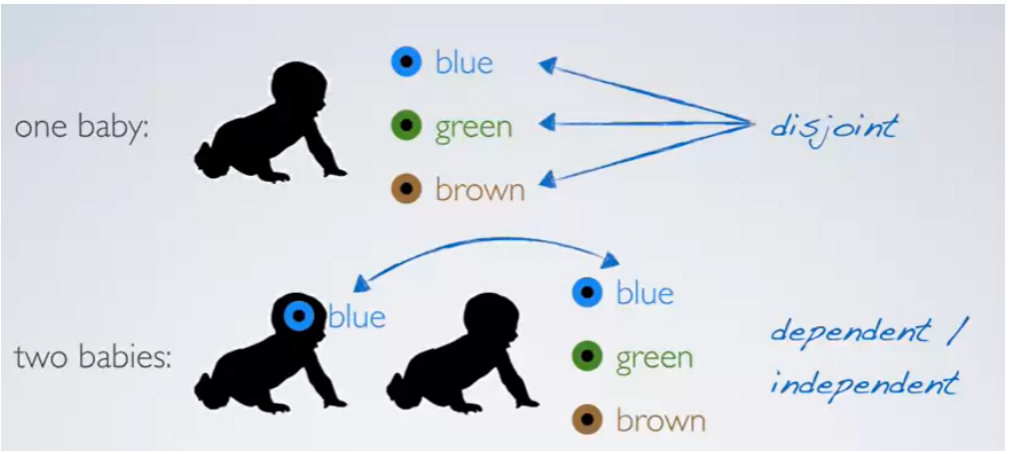
- determining dependence based on sample data
  <u>observed difference between conditional probabilities</u> -> dependence -> hypothesis test
- If the difference is large, there is stronger evidence that the difference is real .
- If sample size is large, even the small difference can provide strong evidence of a real difference .
- Product rule for an independent events : If A and B are independent, **P(A and B) = P(A) X P(B)**

4. Probability Examples

- 

5. Disjoint vs. Independent



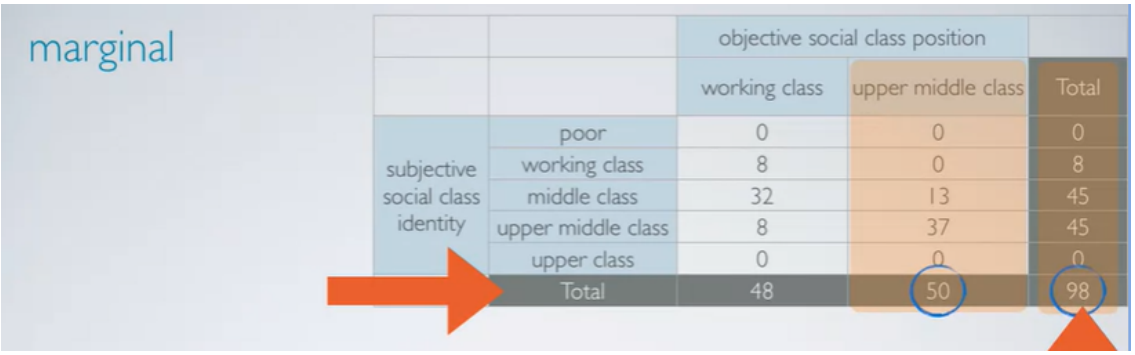


- We can generalize this to say that disjoint offense with non-zero probability are always dependent on each other because if we know that one happened we know that the other one cannot happen.

[Conditional Probability]

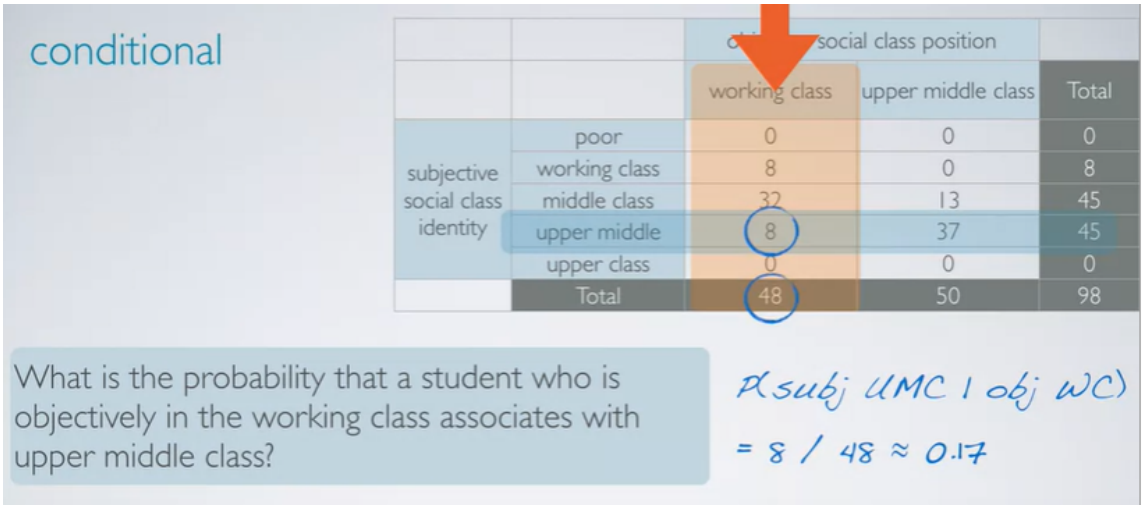
1. Conditional Probability

- marinal



- The term ‘marginal probability’ comes from the fact that the counts we use to calculate this probability came from the margins of the contingency table.
- joint
- the intersection of the two interested groups

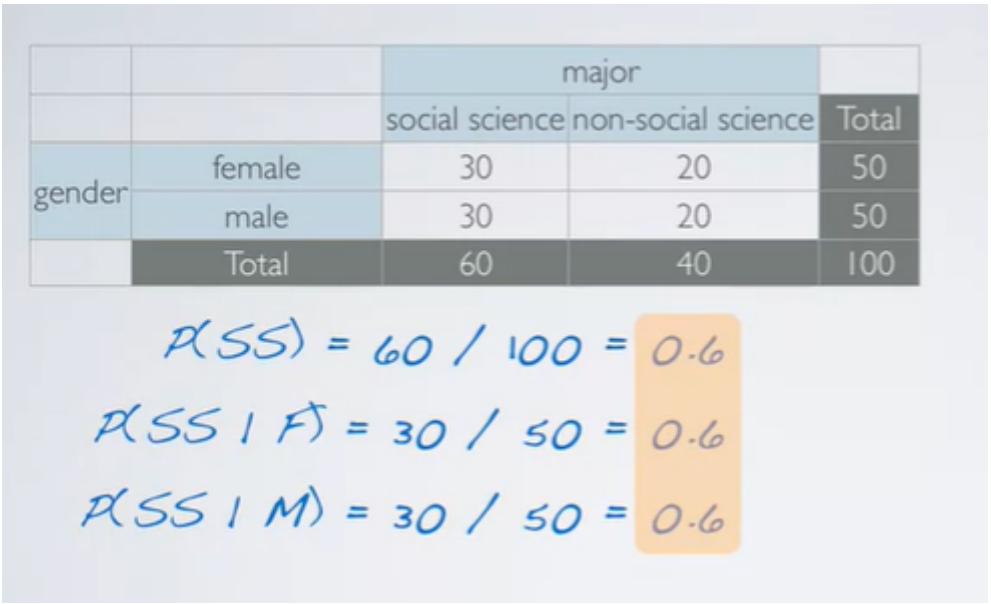
- conditional



-

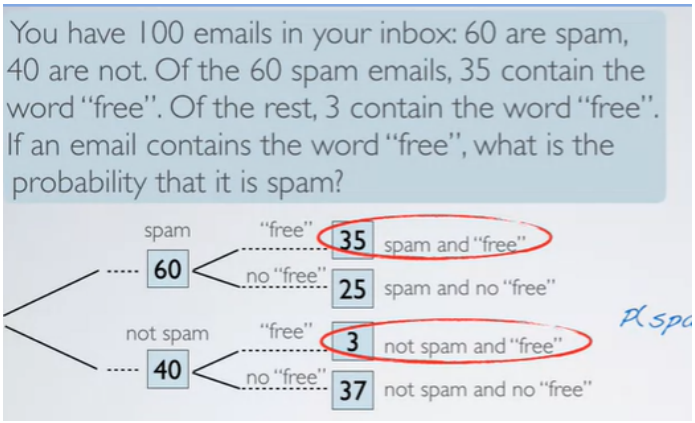
- Bayes' theorem

- P(A|B) = P(A and B) / P(B)
- General product rule -> **P(A and B) = P(A|B) X P(B)**
- If P(A|B) = P(A) then the events A and B are said to be independent.

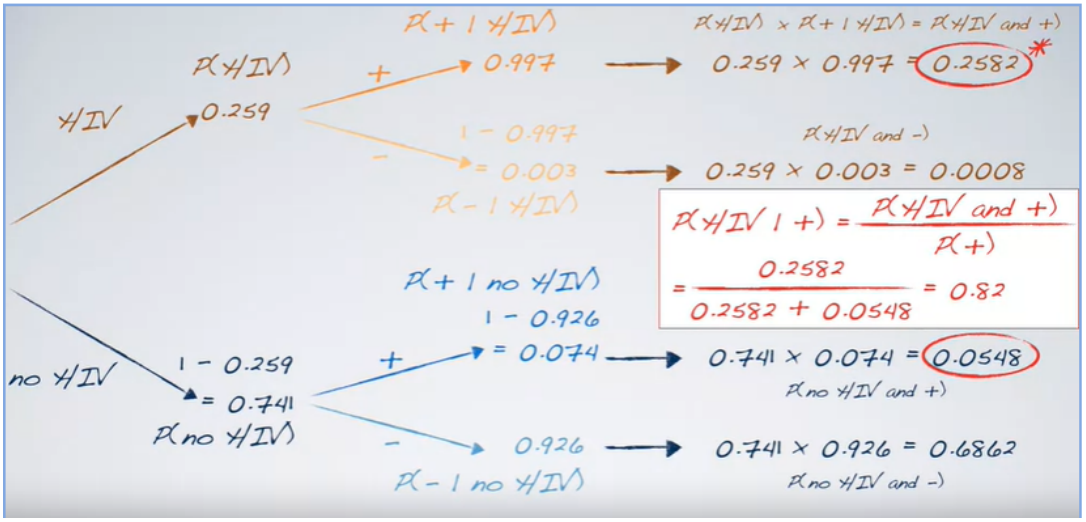


2. Probability Trees

- example 1.

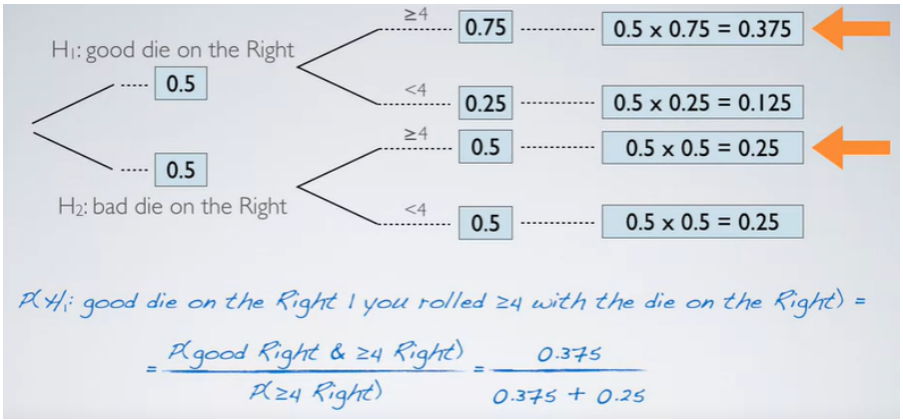


- example 2.



3. Bayesian Inference

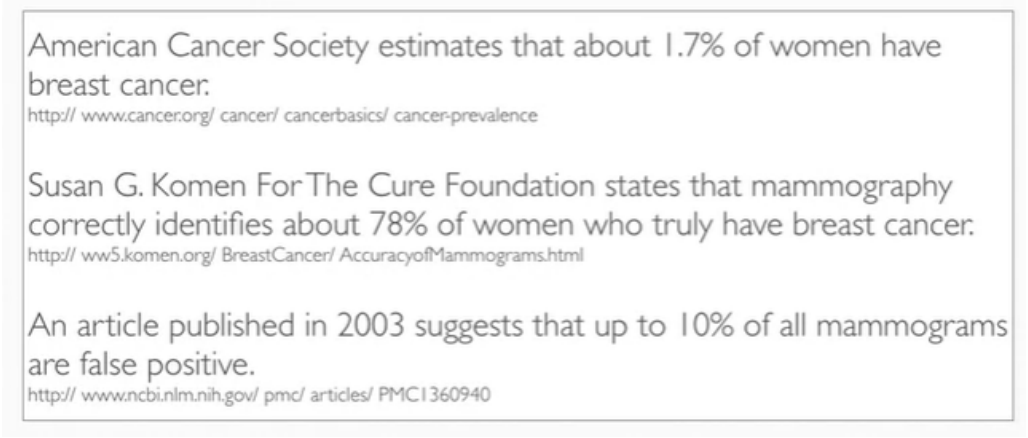
- set-up
- die rolling
    1. What is the probability of rolling >= 4 with a 6-sided die?
       S= {1,2,3,4,5,6} P(>=4) = 0.5
    2. What is the probability of rolling >= 12-sided die?
       S= {1,2,3,...,11,12} P(>=4) = 0.75



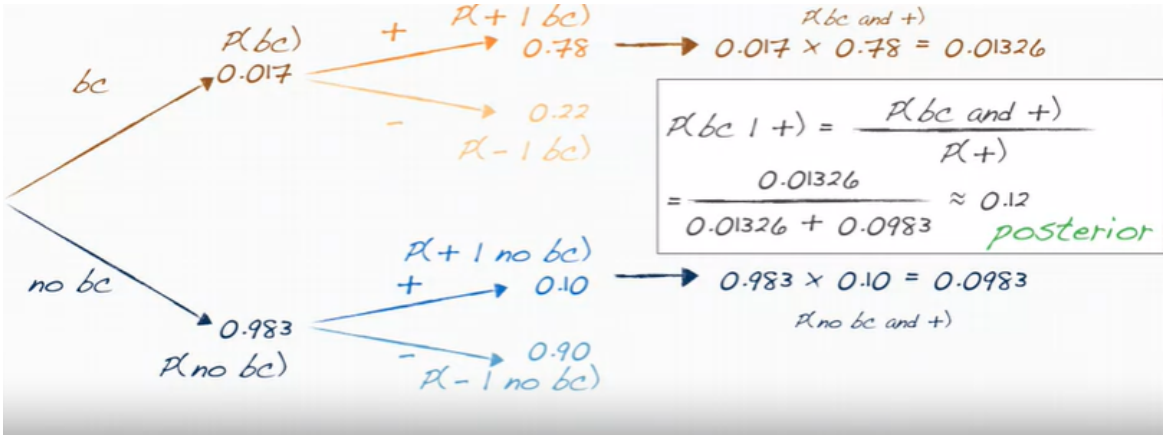
- posterior probability

- The probability we just calculated is also called the **posterior probability.**
- Posterior probability is generally defined as P(hypothesis|data).
- It tells us the probability of a hypothesis we set forth, given the data we just observed.
- It depends on both the prior probability we set and The observed data.
- This is different than what we calculated at the end of the randomization test on gender discrimination - the probability of observed or more extreme data given the null hypothesis being true.
  (확률변수에 대한 관측이나 증거에 대한 조건부 확률을 말한다. 즉 어떤 특정사건이 이미 발생하였는데 이 특정사건이 나온 연원이 무엇인지 불확실한 상황을 식으로 나타낸 것이며 P(A|B)로 표현될 수 있다.)
- Naturally integrate data as you collect it, and update our priors.
- Avoid the counter-intuitive definition of a p-value: P(observed of more extreme outcome | H0 is true)
- Instead base decisions on the posterior probability : P(hypothesis is true|observed data)
- A good prior helps, a bad prior hurts, but the prior matters less the more data you have.

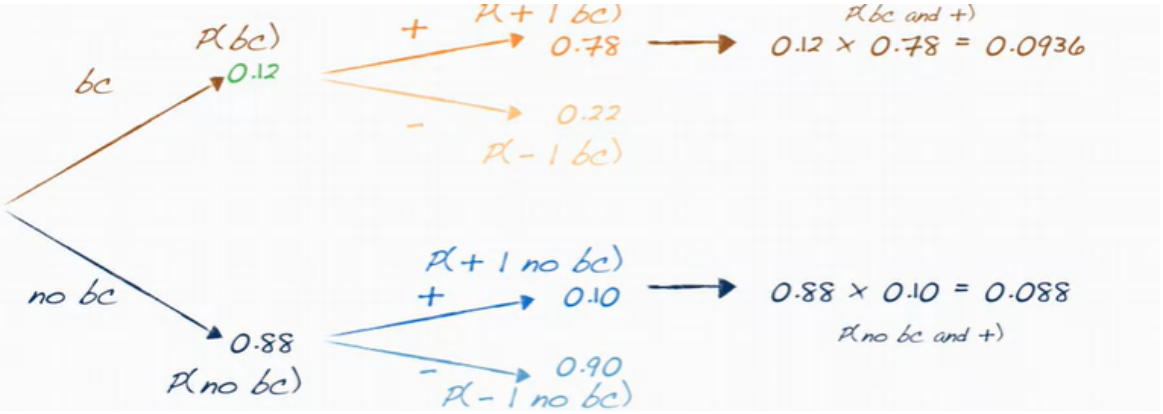
4. Examples of Bayesian Inference



- Prior to any testing and any information exchange between the patient and the doctor, what probability should the doctor assign to a female patient having breast cancer?
- P(bc) = 0.017 -> This is prior
- When a patient goes through breast cancer screening, there are two competing claims : patient has cancer and patient doesn't have cancer. Iif a mammogram is a positive result, what is the probability that the patient has cancer ?
- P(bc | +) = ?



- Since a positive mammogram doesn't necessarily mean that the patient actually has breast cancer, the doctor might decide to re-test the patient. What is the probability of having breast cancer if the second mammogram also yields a positive result ?



- 0.0936 / (0.0936+0.088) = 0.52