

Week1

1. Data Basics

- Observations, variables, and data metrics

In data matrix,

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

- types of variables

numerical (quantitative)	<ul style="list-style-type: none"> - take on numerical values - sensible to add, subtract, take averages, etc. with these values 	continuous	<ul style="list-style-type: none"> - take on any of an infinite number of values within a given range - ex : height
		discrete	<ul style="list-style-type: none"> - 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)	<ul style="list-style-type: none"> - take on a limited number of distinct categories - categories can be identified with numbers, but not sensible to do arithmetic operations 	regular categorical	
		ordinal	<ul style="list-style-type: none"> - 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	<ul style="list-style-type: none"> - observational studies collect data in a way that does not directly interfere with how the data arise.
---------	---------------	---

		<ul style="list-style-type: none"> - only establish an association (correlation between the explanatory and the response variables) - retrospective : uses past data - prospective : data are collected throughout the study
	experiment	<ul style="list-style-type: none"> - 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 are 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

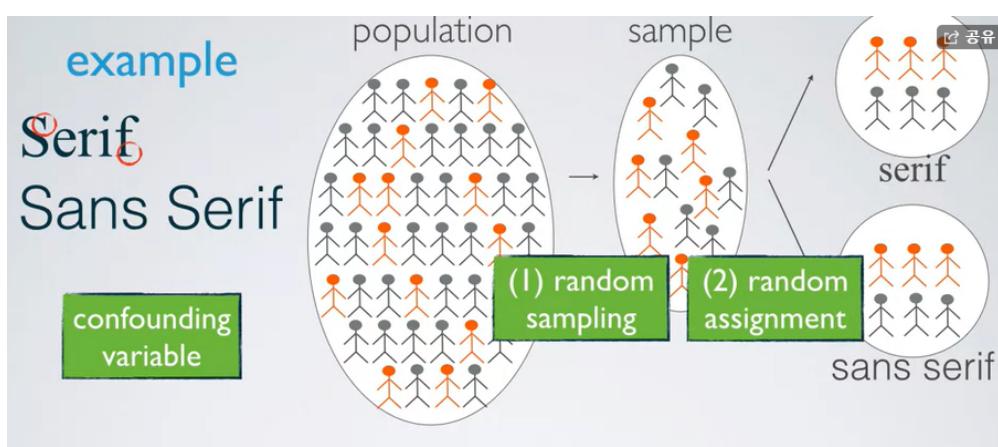
1. Simple Random Sample(SRS) : Each case is equally likely to be selected.
2. stratified sample : Firstly divide the population into homogeneous groups called Strata and then randomly sample from within each stratum.
3. cluster sample : we divide the population into clusters randomly sample a few clusters 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 :



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

ideal experiment

Random sampling

No random sampling

most experiments

most observational studies

Generalizability

No generalizability

bad observational studies

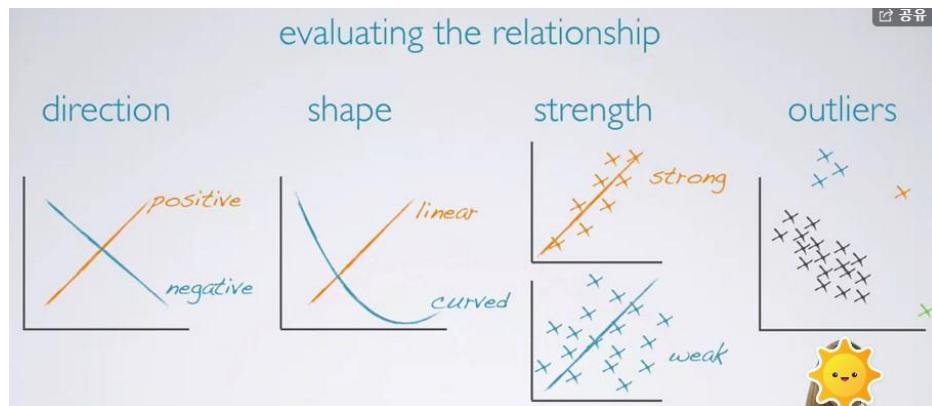
Week2

[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

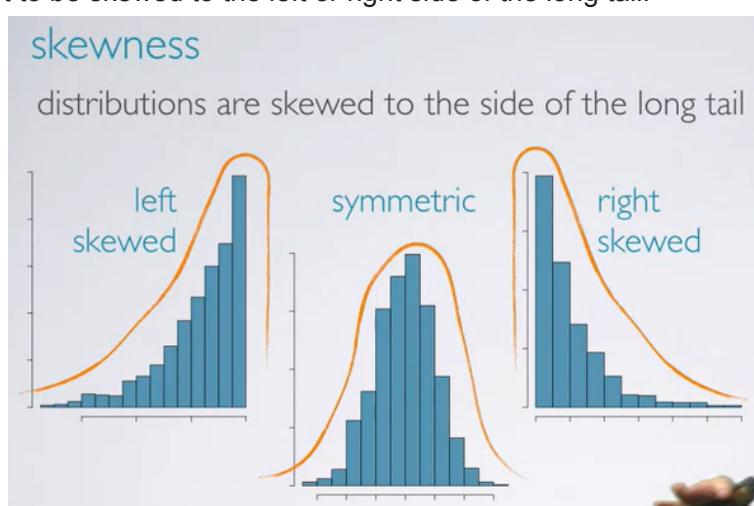


- Histogram

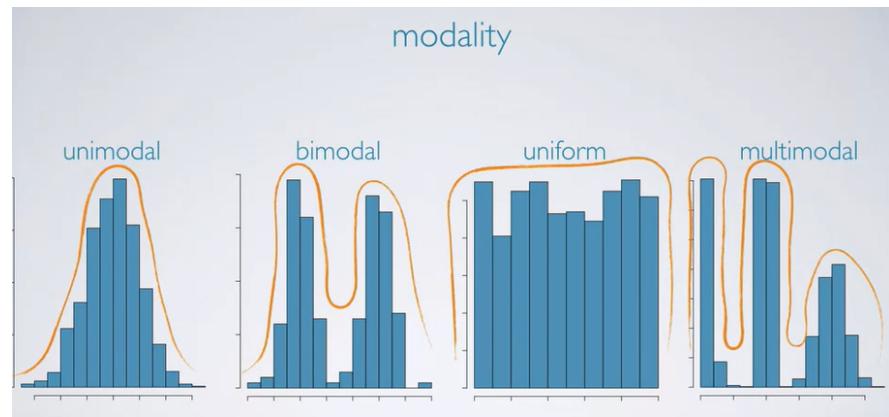
- One good way of visualizing the distribution of a numerical variable is a histogram.
- A histogram provides a view of the data density.
- especially useful for describing the shape of the distribution

- Skewness

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



- Modality



- histogram & bin width



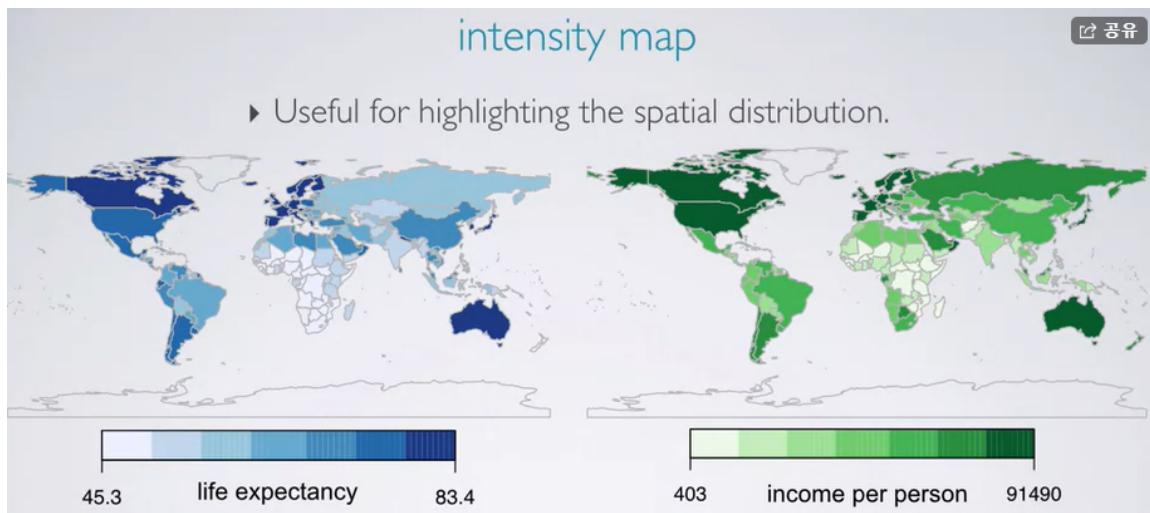
- Dotplot

- 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

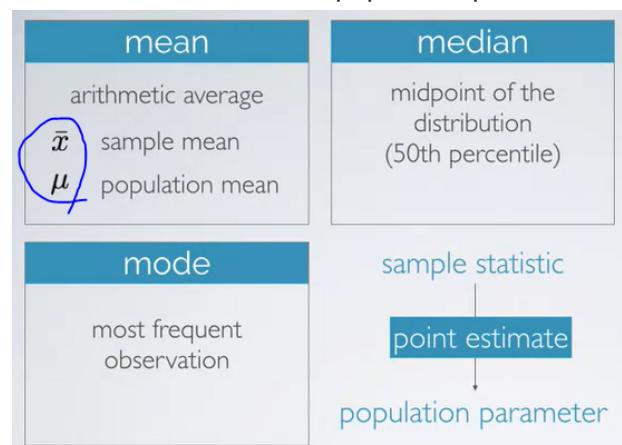
- Useful for highlighting outliers, median, IQR

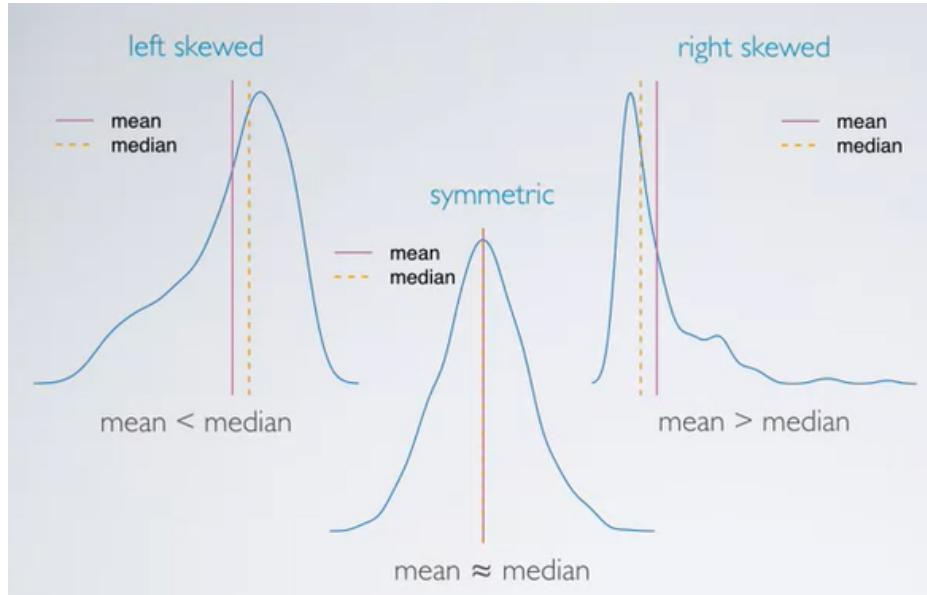
- Intensity map



2. Measures of center

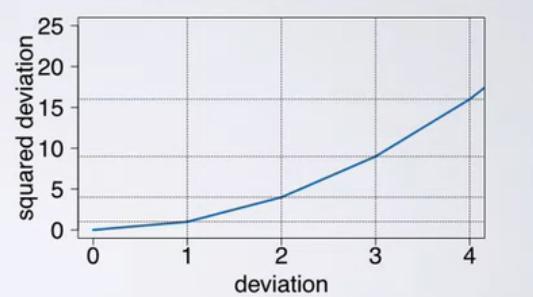
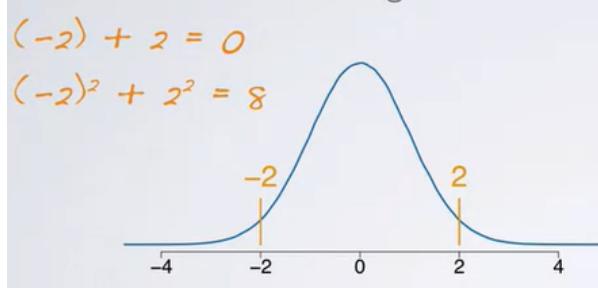
- mean : arithmetic average
- median : midpoint of the distribution
- mode : most frequent observation
- If these measurements are calculated from a sample they are called sample statistics.
- Sample statistics are points estimates for the unknown population parameters.





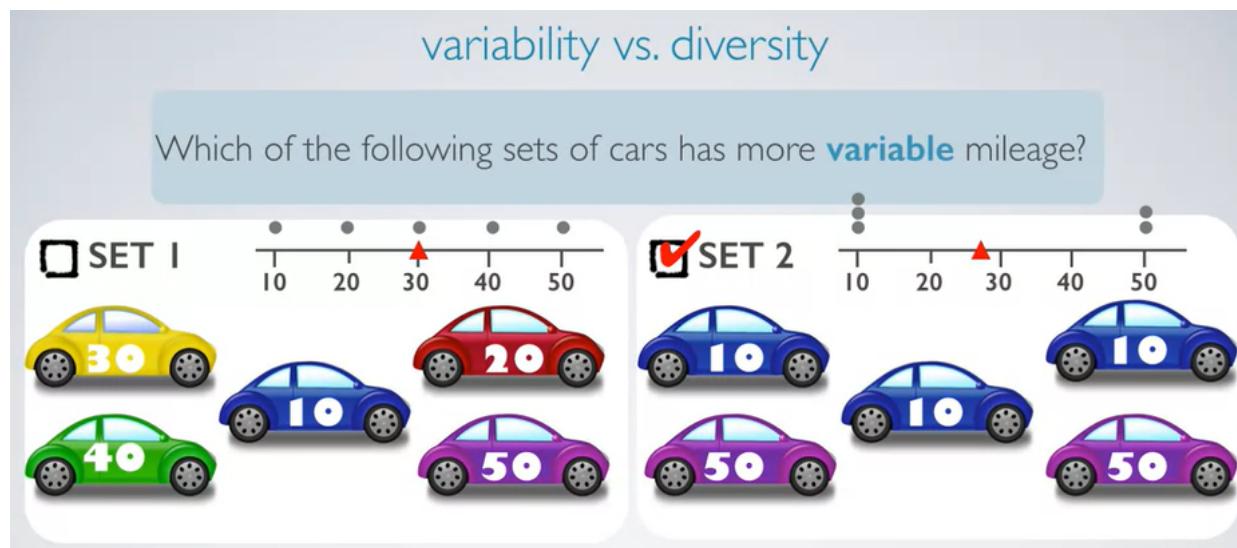
3. Measures of Spread

- 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.
 -

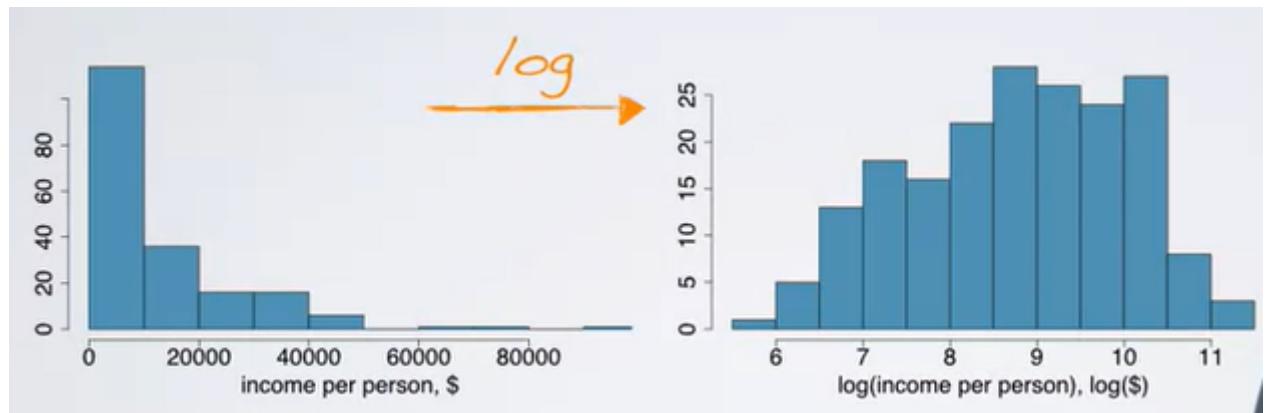
4. Robust Statistics

- 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. (이상치, 예러값으로부터 크게 영향을 받지 않는 값)

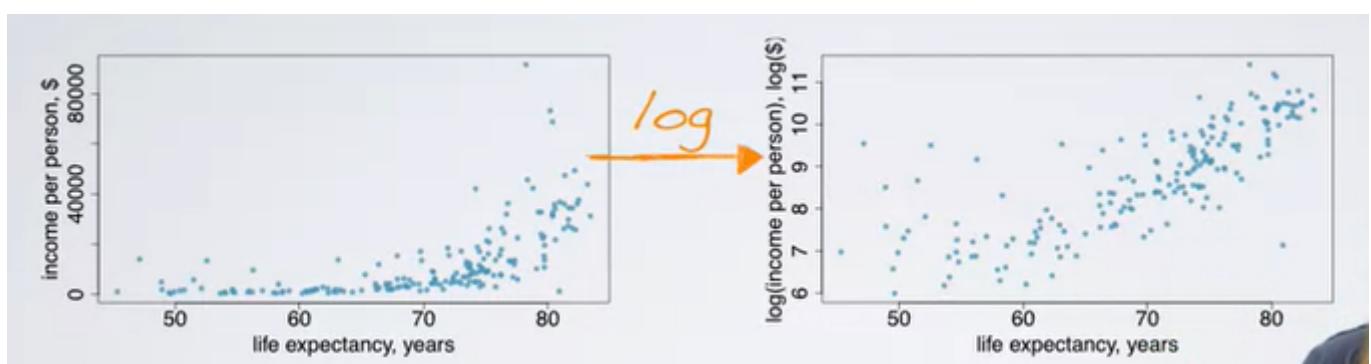
	robust	non-robust
center	median	mean
spread	IQR	SD, range
	<i>skewed, with extreme observations</i>	<i>symmetric</i>

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

[Exploring Categorical Data and Introduction to Inference]

1. Exploring Categorical Variables

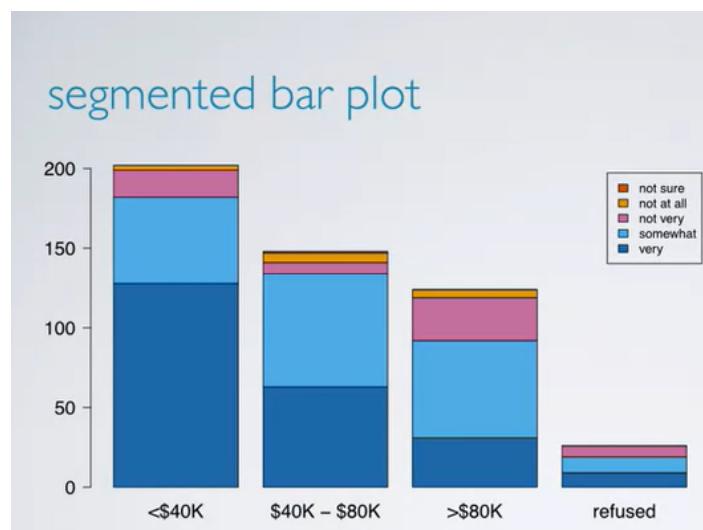
- Frequency table & bar plot
- How are bar plots different than histograms?
- Bar plots for displaying distributions of categorical variables, histograms for numerical variables
- x-axis on a histogram is a number line, and the ordering of the bars are not interchangeable
- In a bar plot, the categories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

		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: $9 / 26 = 35\%$

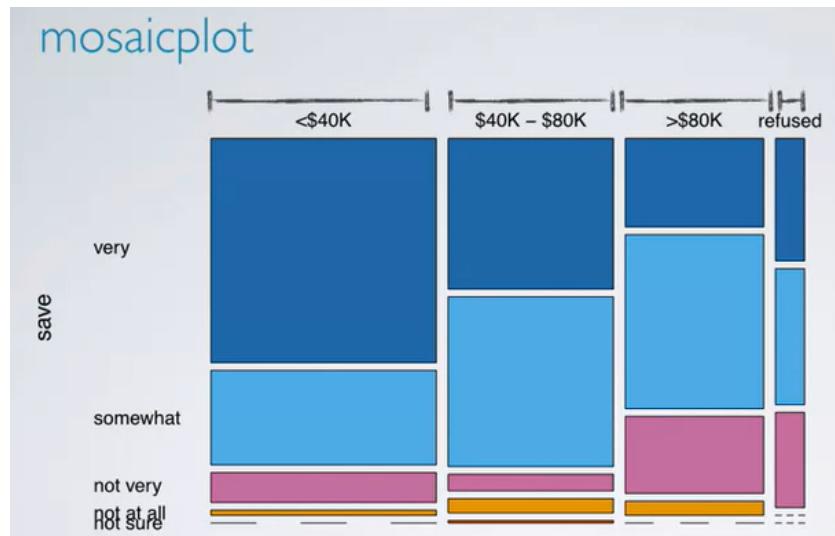
feelings about difficulty of saving money and income are associated (dependent)

- segmented bar plot



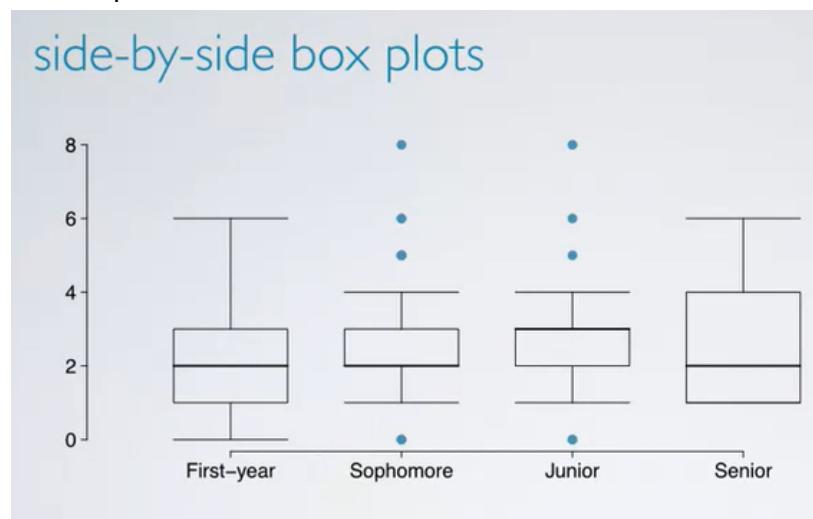
- useful for visualizing conditional frequency distributions
- compare relative frequencies to explore relationship between the variables

- 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
- 본인이 맞다고 생각하는 주장과 정반대되는 주장

- **Alternative hypothesis (대립 가설)**

- it is dependent, there is gender discrimination
- 실제로 맞다고 주장하고 싶은 내용

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
 - 귀무가설이 맞다고 가정할 때 실제로 해당 통계치가 관측될 확률
 - p-value가 0.05 미만이면 귀무가설에 해당되는 관측이 나올 확률이 5% 미만이라는 의미
 - 작을수록 귀무가설 기각
- if this probability is low, reject the null hypothesis in favor of the alternative
- <http://www.incodom.kr/%EA%B7%80%EB%AC%B4%EA%B0%80%EC%84%A4>

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?
-> $P(A \text{ or } B) = P(A) + P(B)$
- union of non-disjoint events : What is the probability of non-disjoint event A or non-disjoint event B?
-> $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
- 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.

complementary		
one toss	head	tail
probability	0.5	0.5

complementary				
two tosses	head - head	tail - tail	head - tail	tail - head
probability	0.25	0.25	0.25	0.25

- 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
 - observed difference between conditional probabilities -> 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, $P(A \text{ and } B) = P(A) \times P(B)$

4. Probability Examples

(6) What is the probability that at least 1 in 5 randomly selected people agree with the statement about men having more right to a job than women?

$$P(\text{agree}) = 0.362$$

$$S = \{\emptyset, 1, 2, 3, 4, 5\} \rightarrow S = \{\emptyset, \text{at least 1}\}$$

$$P(\text{at least 1 agree}) = 1 - P(\text{none agree})$$

$$= 1 - P(\underline{D} \underline{D} \underline{D} \underline{D} \underline{D})$$

$$= 1 - 0.638^5$$

$$= 1 - 0.106 = 0.894$$

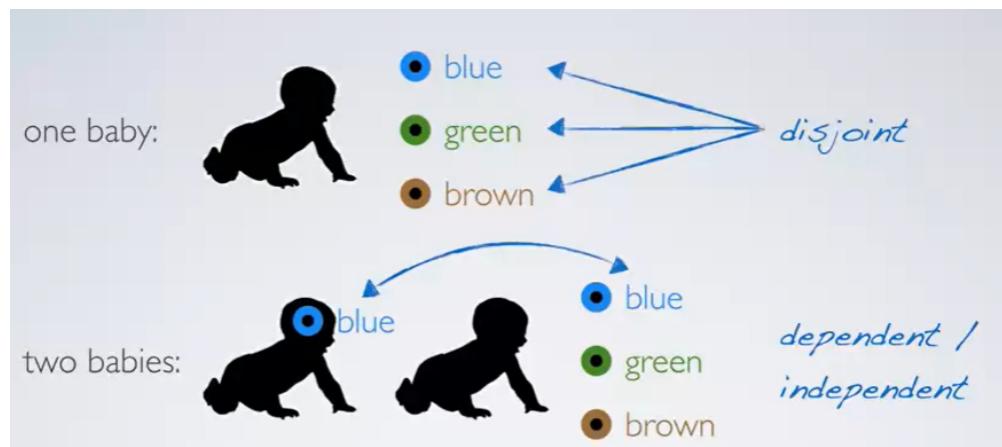
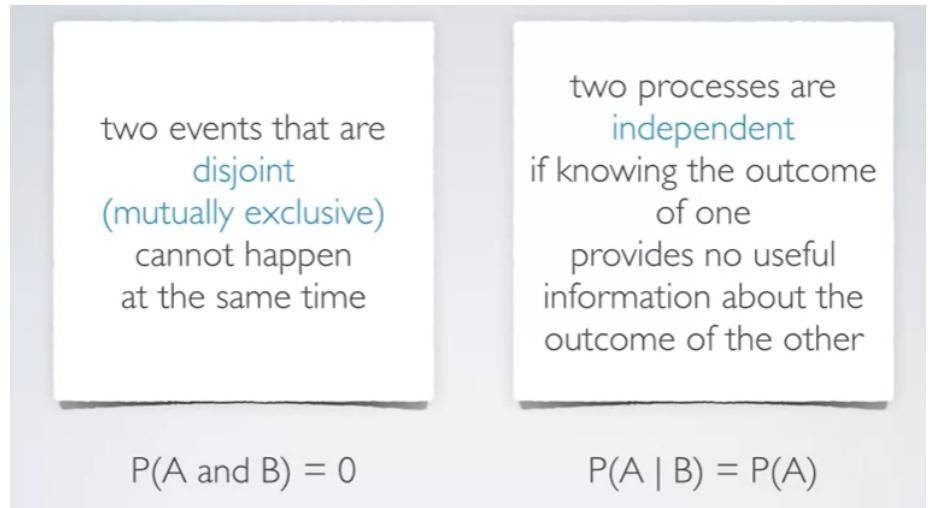
$$P(\text{disagree})$$

$$= 1 - P(\text{agree})$$

$$= 1 - 0.362$$

$$= 0.638$$

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

marginal

		objective social class position		Total
		working class	upper middle class	
subjective social class identity	poor	0	0	0
	working class	8	0	8
	middle class	32	13	45
	upper middle class	8	37	45
	upper class	0	0	0
	Total	48	50	98

- 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

conditional

		social class position		
		working class	upper middle class	Total
subjective social class identity	poor	0	0	0
	working class	8	0	8
	middle class	32	13	45
	upper middle	8	37	45
	upper class	0	0	0
	Total	48	50	98

What is the probability that a student who is objectively in the working class associates with upper middle class?

$P(\text{subj UMC} \mid \text{obj WC}) = 8 / 48 \approx 0.17$

- Bayes' theorem
- $P(A|B) = P(A \text{ and } B) / P(B)$
- General product rule $\rightarrow P(A \text{ and } B) = P(A|B) \times P(B)$
- If $P(A|B) = P(A)$ then the events A and B are said to be independent.

		major		
		social science	non-social science	Total
gender	female	30	20	50
	male	30	20	50
	Total	60	40	100

$P(SS) = 60 / 100 = 0.6$

$P(SS \mid F) = 30 / 50 = 0.6$

$P(SS \mid M) = 30 / 50 = 0.6$

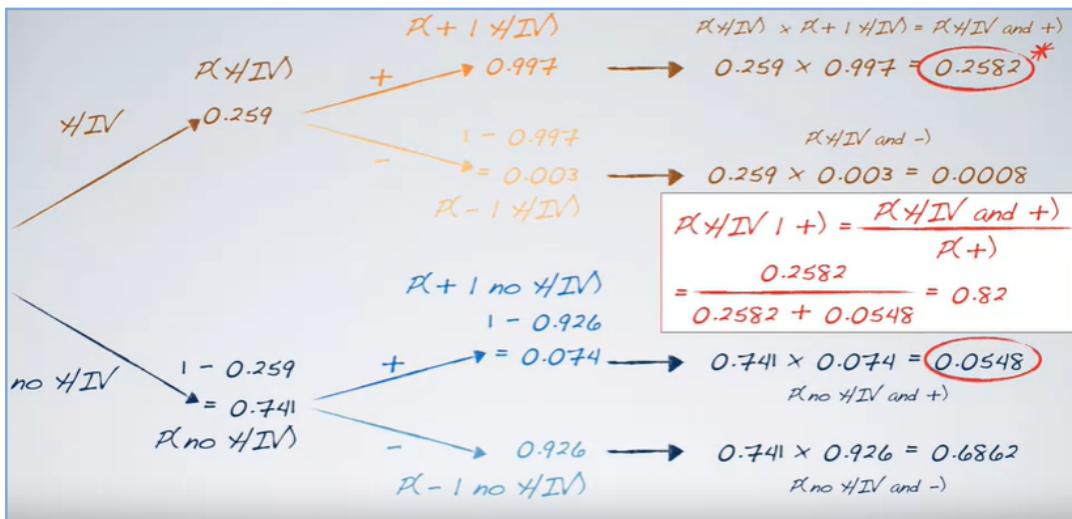
2. Probability Trees

- example 1.

You have 100 emails in your inbox: 60 are spam, 40 are not. Of the 60 spam emails, 35 contain the word "free". Of the rest, 3 contain the word "free". If an email contains the word "free", what is the probability that it is spam?

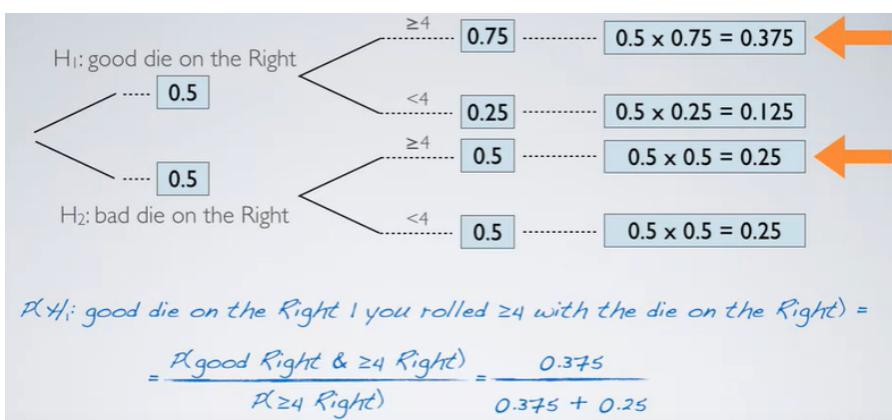


- example 2.



3. Bayesian Inference

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



- posterior probability

- The probability we just calculated is also called the **posterior probability**.
- Posterior probability is generally defined as $P(\text{hypothesis}|\text{data})$.
- 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: $P(\text{observed or more extreme outcome} | H_0 \text{ is true})$
- Instead base decisions on the posterior probability : $P(\text{hypothesis is true}|\text{observed data})$
- A good prior helps, a bad prior hurts, but the prior matters less the more data you have.

4. Examples of Bayesian Inference

American Cancer Society estimates that about 1.7% of women have breast cancer.

<http://www.cancer.org/cancer/cancerbasics/cancer-prevalence>

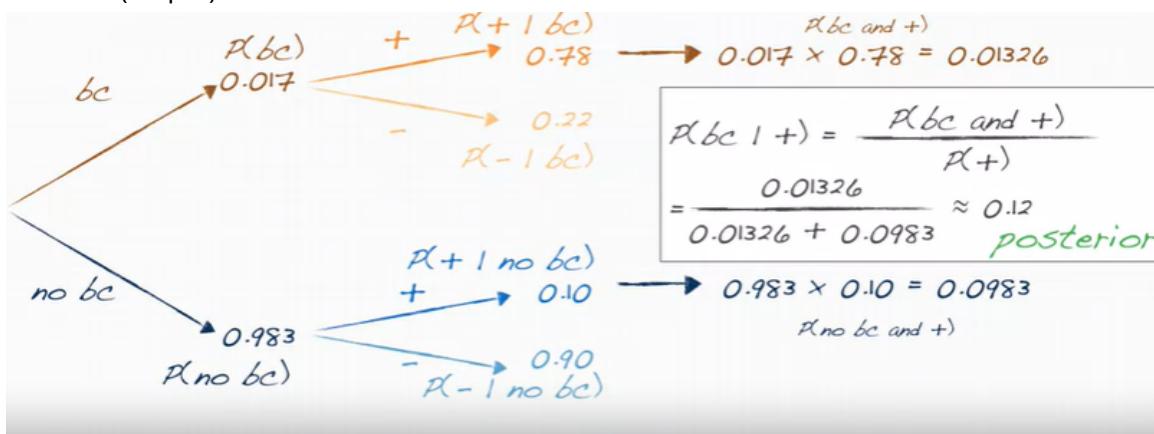
Susan G. Komen For The Cure Foundation states that mammography correctly identifies about 78% of women who truly have breast cancer.

<http://ww5.komen.org/BreastCancer/AccuracyofMammograms.html>

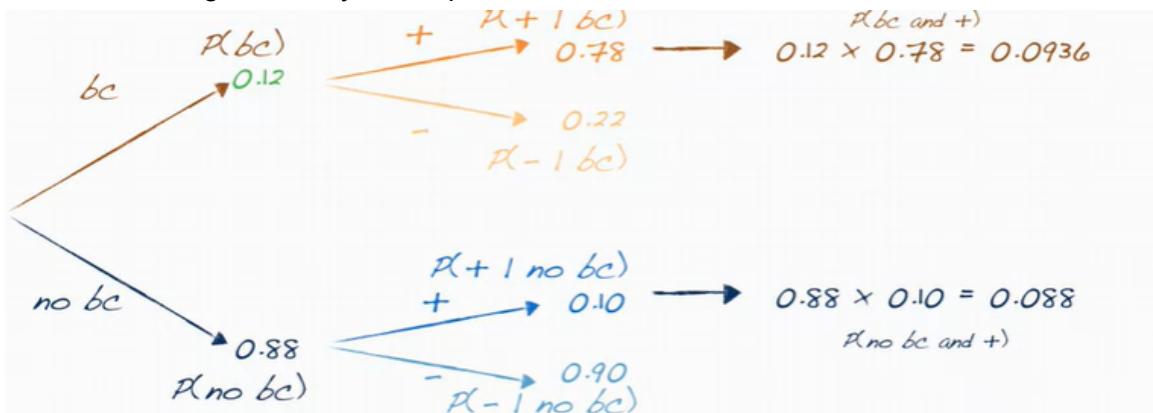
An article published in 2003 suggests that up to 10% of all mammograms are false positive.

<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1360940>

- 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?
- $P(bc) = 0.017 \rightarrow$ 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. If 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 ?



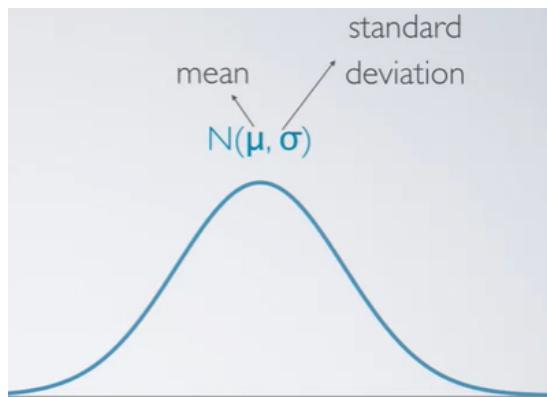
- $0.0936 / (0.0936+0.088) = 0.52$

Week4

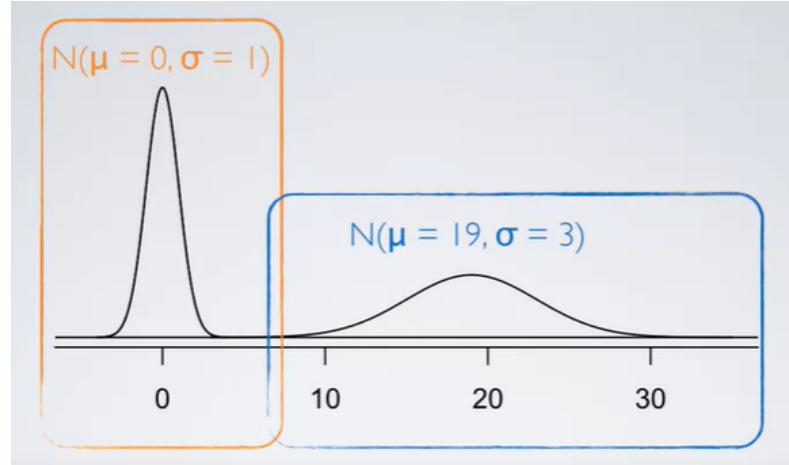
[The Normal Distribution]

1. Normal Distribution

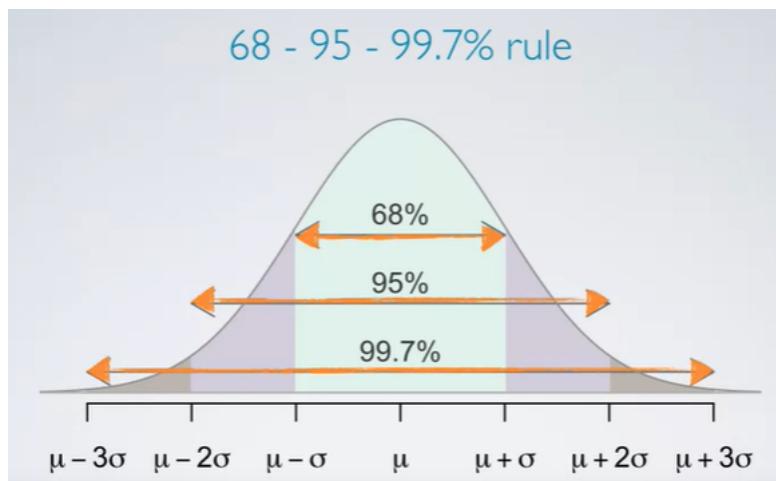
- Normal distribution
- unimodal and symmetric
- follows very strict guidelines about how variably the data are distributed around the mean
- many variables are nearly normal, but none are exactly normal



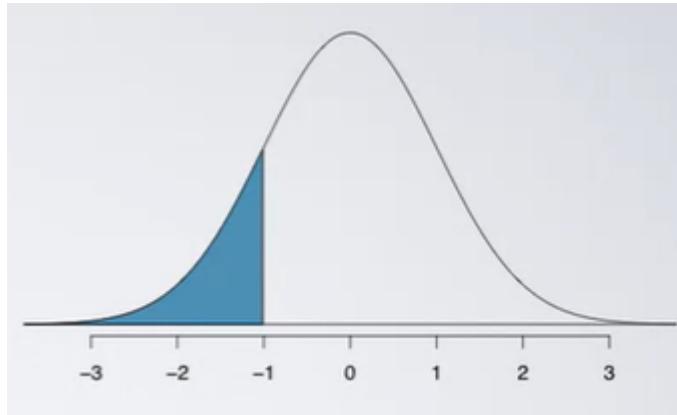
- example :



- 68-95-99.7% rule

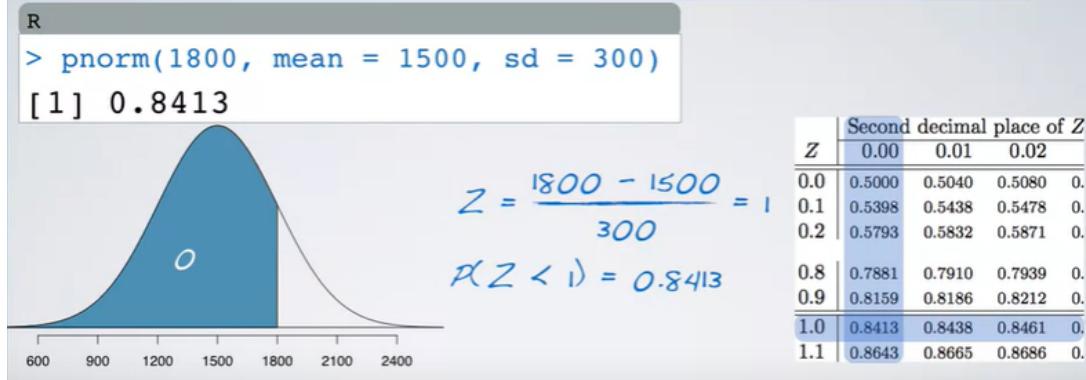


- Standardizing with Z scores
 - standardized (Z) score of an observation is the number of standard deviations it falls above or below the mean.
 - $Z = \frac{\text{observation} - \text{mean}}{SD}$
 - Z score of mean is 0.
 - unusual observation : $|Z| > 2$
 - defined for distributions of any shape.
- percentiles
 - When the distribution is normal, Z scores can be used to calculate percentiles.
 - percentile is the percentage of observations that fall below a given data point.
 - graphically, percentile is the area below the probability distribution curve to the LEFT of that observation.



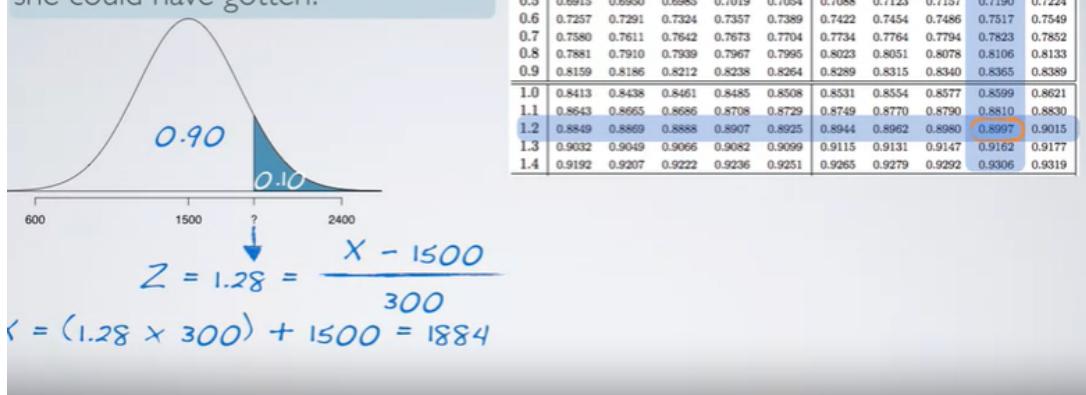
- https://gallery.shinyapps.io/dist_calc/
- practice1.

SAT scores are distributed normally with mean 1500 and SD 300. Pam earned an 1800 on her SAT. What is Pam's percentile score?



- practice2.

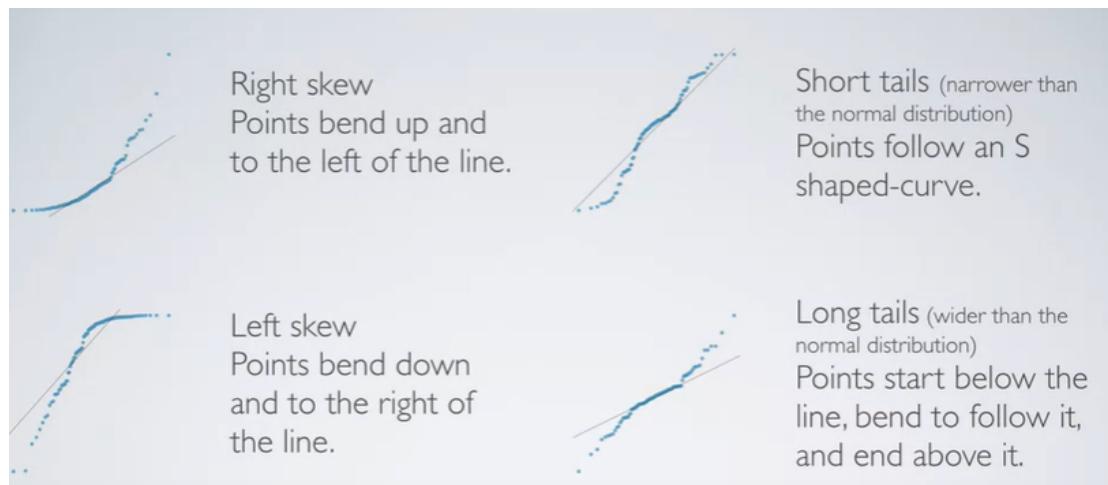
A friend of yours tells you that she scored in the top 10% on the SAT. What is the lowest possible score she could have gotten?



- Normal probability table : <https://drive.google.com/file/d/1CmcRQ1foAtvBVCqzjhcvGe9mfpGVQ1zl/edit>

2. Evaluating the Normal Distribution

- Anatomy of a normal probability plot
- Data are plotted on the y-axis of a normal probability plot, and theoretical quantiles (following a normal distribution) on the x-axis.
- If there is a one-to-one relationship between the data and the theoretical quantiles, then data follow a nearly normal distribution.
- Since a one-to-one relationship would appear as a straight line on a scatter plot, the closer the points are to a perfect straight line, the more confident we can be that the data follow the normal model.
- Constructing a normal probability plot requires calculating percentiles and corresponding z-scores for each observation, which is tedious. Therefore we generally rely on software when making these plots.
- shape of the data points



3. Working with the Normal Distribution

- example1

Suppose weights of checked baggage of airline passengers follow a nearly normal distribution with mean 45 pounds and standard deviation 3.2 pounds. Most airlines charge a fee for baggage that weigh in excess of 50 pounds.

$$\text{baggage} \sim N(\text{mean} = 45, SD = 3.2)$$

What percent of airline passengers are expected to incur this fee?

- method 1 :

Distribution Calculator

Distribution: Normal

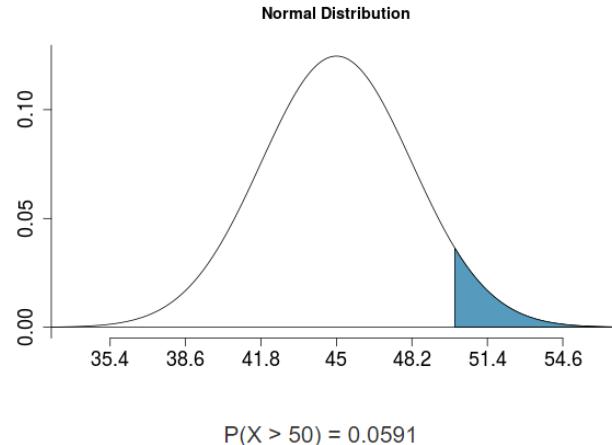
Mean: -50 to 50, set to 45

Standard deviation: 0.1 to 30, set to 3.2

Model: $P(X > a)$

Find Area: Upper Tail

a: 32.2 to 57.8, set to 50

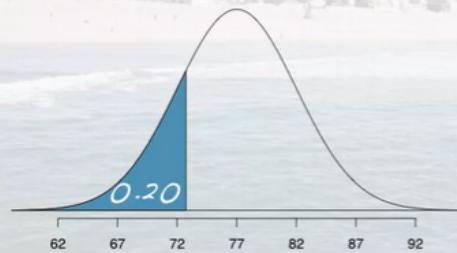


- method 2 : using R
`pnorm(50, mean = 45, sd = 3.2)`
- method 3 : by hand using Z score table
- example 2

The average daily high temperature in June in LA is 77 °F with a standard deviation of 5 °F. Suppose that the temperatures in June closely follow a normal distribution.

How cold are the coldest 20% of the days during June in LA?

temperatures $\sim N(\text{mean} = 77, \text{SD} = 5)$



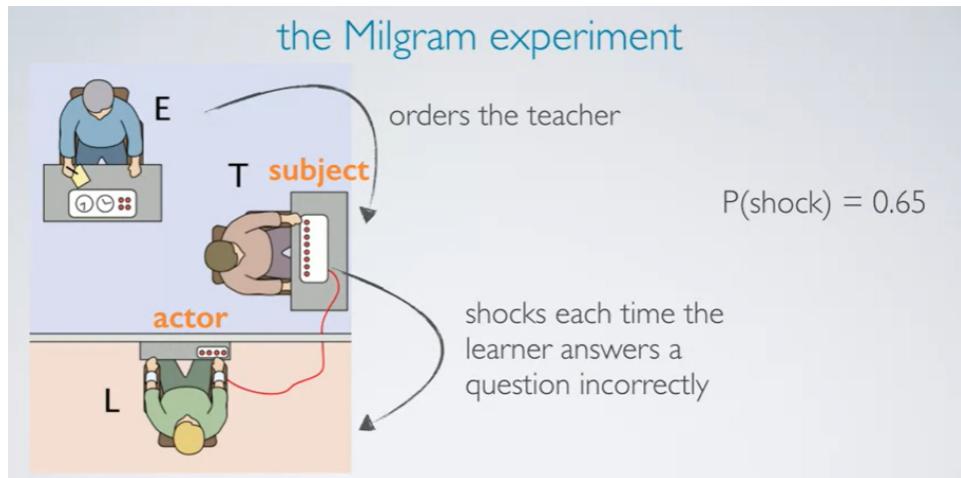
```
R
> qnorm(0.20, mean = 77, sd = 5)
[1] 72.79
```

- method 1. using R
- method 2. using Z score table

[Binomial Distribution]

1. Binomial Distribution(이항분포)

- The Milgram experiment



- Learner is an actor.
- Subjects(teachers) believe they actually give a shock to a learner.
- P(shock) is the probability that teachers obey the order.
- Bernoulli random variable
- Each person in Milgram's experiment can be thought of as a trial.
- A person is labeled a success if she refuses to administers a severe shock and failure if she administers such shock.
- Since only 35% of people refuse to administer a shock probability of success is $p = 0.35$.
- When the individual dryer has only two possible outcomes it is called Bernoulli random variable.

Suppose we randomly select four individuals to participate in this experiment.
What is the probability that exactly 1 of them will refuse to administer the shock?

- ▶ Four individuals:
(A) Anthony
(B) Brittany
(C) Clara
(D) Dorian

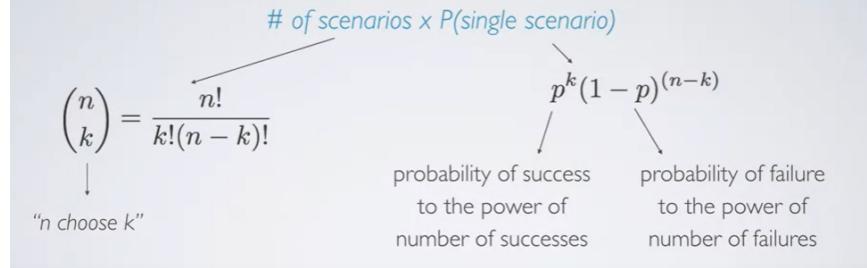
- ▶ Multiple scenarios where "exactly 1 refuses"

$$\begin{aligned}
 \text{Scenario 1: } & \frac{0.35}{\text{(A) refuse}} \times \frac{0.65}{\text{(B) shock}} \times \frac{0.65}{\text{(C) shock}} \times \frac{0.65}{\text{(D) shock}} = 0.0961 \\
 \text{OR} \\
 \text{Scenario 2: } & \frac{0.65}{\text{(A) shock}} \times \frac{0.35}{\text{(B) refuse}} \times \frac{0.65}{\text{(C) shock}} \times \frac{0.65}{\text{(D) shock}} = 0.0961 \\
 \text{OR} \\
 \text{Scenario 3: } & \frac{0.65}{\text{(A) shock}} \times \frac{0.65}{\text{(B) shock}} \times \frac{0.35}{\text{(C) refuse}} \times \frac{0.65}{\text{(D) shock}} = 0.0961 \\
 \text{OR} \\
 \text{Scenario 4: } & \frac{0.65}{\text{(A) shock}} \times \frac{0.65}{\text{(B) shock}} \times \frac{0.65}{\text{(C) shock}} \times \frac{0.35}{\text{(D) refuse}} = 0.0961 \\
 & + \\
 & 4 \times 0.0961 = 0.3844
 \end{aligned}$$

- Binomial distribution

binomial distribution

the **binomial distribution** describes the probability of having exactly k successes in n independent Bernoulli trials with probability of success p



How many scenarios yield 1 success in 4 trials?

$$n = 4 \quad k = 1$$

$$\binom{4}{1} = \frac{4!}{1! \times (4-1)!} = \frac{4 \times \cancel{3} \times \cancel{2} \times 1}{1 \times \cancel{3} \times \cancel{2} \times 1} = 4$$

How many scenarios yield 2 successes in 9 trials?

$$ssFFFFFFFFFF \quad n = 9 \quad k = 2$$

$$\begin{array}{c} SFSFFFFFFFFFF \\ SFSSFFFFFFFFFF \\ \dots \\ \binom{9}{2} = \frac{9!}{2! \times 7!} = \frac{9 \times 8 \times \cancel{7}!}{2 \times 1 \times \cancel{7}!} = 36 \end{array}$$

```
R
> choose(9,2)
[1] 36
```

- If p presents probability of success ($1-p$) represents probability of failure, n presents number of independent trials and k represents number of successes

$$P(k \text{ successes in } n \text{ trials}) = \binom{n}{k} p^k (1-p)^{(n-k)}$$

$$\text{where } \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- **binomial conditions**

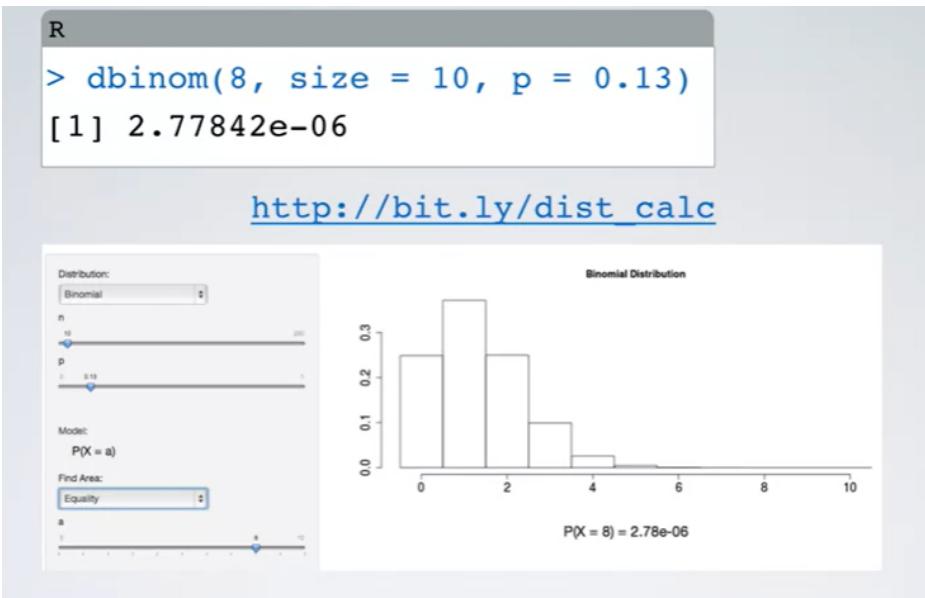
1. The trials must be independent
2. The number of trials, n , must be fixed
3. each trial outcome must be classified as a success or a failure
4. The probability of success, p , must be the same for each trial

- ex:

According to a 2013 Gallup poll, worldwide only 13% of employees are engaged at work (psychologically committed to their jobs and likely to be making positive contributions to their organizations). Among a random sample of 10 employees, what is the probability that 8 of them are engaged at work?

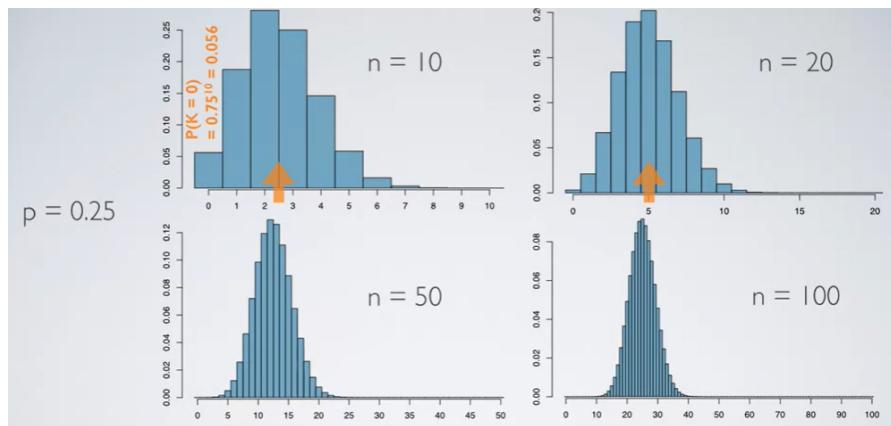
practice

$$\begin{aligned} n &= 10 & P(K=8) &= \binom{10}{8} 0.13^8 \times 0.87^2 \\ p &= 0.13 & &= \frac{10!}{8! \times 2!} \times 0.13^8 \times 0.87^2 \\ 1-p &= 0.87 & &= \frac{10 \times 9 \times \cancel{8}!}{\cancel{8}! \times 2 \times 1} \times 0.13^8 \times 0.87^2 \\ k &= 8 & &= 45 \times 0.13^8 \times 0.87^2 \\ & & &= 0.00000278 \end{aligned}$$



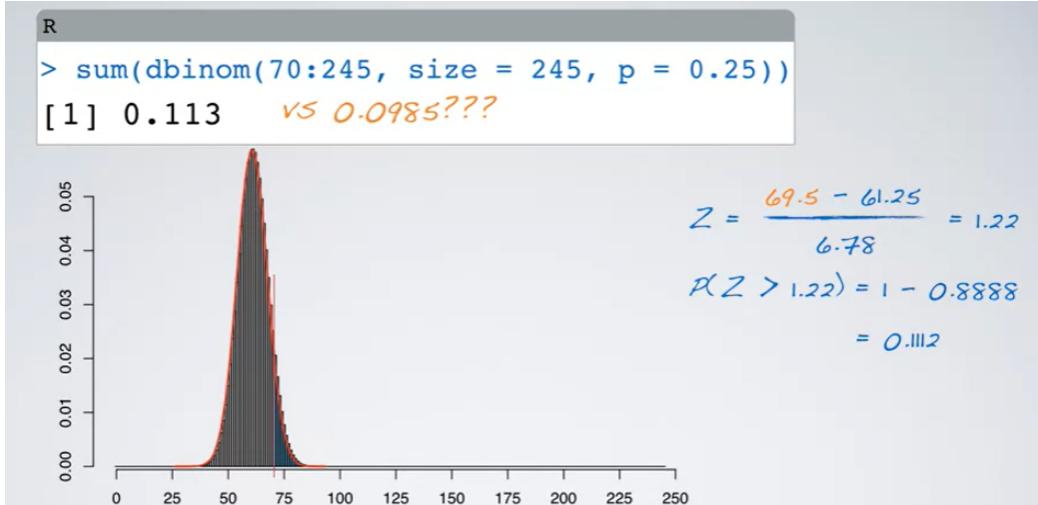
- Expected value (mean) of binomial distribution : $\mu = np$
- Standard deviation of binomial distribution : $\sigma = \sqrt{np(1 - p)}$

2. Normal Approximation to Binomial



- Recent study : Facebook users get more than they give
 - friend request : 40% made, 63% received at least one
 - likes : liked 14 times, had their content “liked” 20 times on average
 - messages : sent 9 messages, received 12 on average
 - tags : 12% tagged a friend in a photo, but 35% tagged
 - other findings : 25% considered power users, average FB user has 245 friends.
 - $P(70 \text{ or more power user friends}) = ?$
- $p = 0.24$, $n = 245$, $P(K \geq 70) ?$
- mean = $np = 61.25$
- $sd = 6.78$
- $Z = (70 - 61.25) / 6.78 = 1.29$
- $P(Z > 1.29) = 1 - 0.9015 = 0.0985$

- So there is a 9.85% chance that an average Facebook user with 245 friends has at least 70 friends who are considered power users.
- In R, $\text{sum(dbinom(70:245, size= 245, p = 0.25))} = 0.113$
- 1. Calculating the probabilities for each outcome $70 \sim 245$
- 2. Wrapping it around with the sum function
- Why are the results slightly different?!



1. We called the approach the normal approximation to the binomial after all, so it's just an approximation and not an exact result.
 2. If we need the exact probability, that difference may be frustrating .
 3. We can see that the red normal curve is slightly different than the bars representing the exact binomial probabilities. It falls a little bit short.
 4. So under the continuous normal distribution the probability of exactly 70 successes is undefined. So the Shaded area above 70 doesn't exactly include the probability of 70 successes..
- A common fix to this problem is a 0.5 adjustment to the observation of Interest.

Success-failure rule: A binomial distribution with at least 10 expected successes and 10 expected failures closely follows a normal distribution.

$$\begin{aligned} np &\geq 10 \\ n(1-p) &\geq 10 \end{aligned}$$

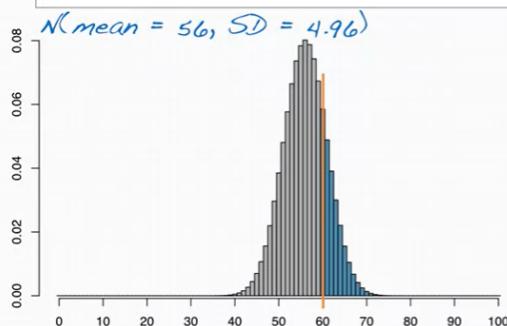
Normal approximation to the binomial: If the success-failure condition holds,

$$\text{Binomial}(n,p) \sim \text{Normal}(\mu, \sigma)$$

$$\text{where } \mu = np \text{ and } \sigma = \sqrt{np(1-p)}$$

- The binomial distribution with sufficient sample size start to look nearly normal.
- example

What is the probability that at least 60 out of a random sample of 100 uninsured Americans plan to get health insurance through a government health insurance exchange?



$$Z = \frac{60 - 56}{4.96} \approx 0.81$$

$$P(Z > 0.81) = 0.209$$

$$Z = \frac{59.5 - 56}{4.96} \approx 0.71$$

$$P(Z > 0.71) = 0.239$$

What is the probability that in a random sample of 1000 people exactly 600 plan to get health insurance through a government health insurance exchange?

- (a) 0.243, same as $P(K = 6)$
- (b) less than 0.243
- (c) more than 0.243

$$P = 0.56$$

$$n_1 = 10 \quad \mu_1 = 10 \times 0.56 = 5.6$$

$$\Delta = 6 - 5.6 = 0.4$$

$$n_2 = 1000 \quad \mu_2 = 1000 \times 0.56 = 560$$

$$\Delta = 600 - 560 = 40$$

5. Which of the following is **false**? Hint: It might be useful to sketch the distributions.

- The Z score for the median of a left skewed distribution is most likely negative.
- Calculating percentiles based on the Z table is only appropriate for observations that come from (nearly) normal distributions.
- Z scores are defined for observations from distributions of any shape and skew.
- The Z score for the median of a symmetric distribution is approximately 0.

틀립니다

This question refers to the following learning objective: Depending on the shape of the distribution determine whether the median would have a negative, positive, or 0 Z score keeping in mind that the mean always has a Z score of 0.

Z scores are defined for all distributions, but calculation of percentiles based on a Z score using the normal distribution (e.g. the table) require normality.

7. One strange phenomenon that sometimes occurs at U.S. airport security gates is that an otherwise law-abiding passenger is caught with a gun in his/her carry-on bag. Usually the passenger claims he/she forgot to remove the handgun from a rarely-used bag before packing it for airline travel. It's estimated that every day 3,000,000 gun owners fly on domestic U.S. flights. Suppose the probability a gun owner will mistakenly take a gun to the airport is 0.00001. What is the probability that tomorrow more than 35 domestic passengers will accidentally get caught with a gun at the airport? Choose the closest answer.

- 0.82
- 0.18
- 0.91
- 0.28
- 0.02

✖ 틀립니다

This question refers to the following learning objective: When number of trials is sufficiently large, use normal approximation to calculate binomial probabilities, and explain why this approach works.

This calculation would involve the sum of many binomial probabilities, so after checking conditions for the normal approximation to the binomial, ($\mu = np = 3,000,000 \times 0.00001 = 30 > 10$ and $n(1 - p) = 2,999,970 > 10$) and calculating $\sigma = \sqrt{np(1 - p)} = 5.48$, we let G denote the number of domestic passengers who get caught with a gun on a particular day. Then calculate

$$P(G > 35) = P((G - 30)/5.48)$$

$$= P((35 - 30)/5.48)$$

$$\approx P(Z > .91)$$

$$\approx 0.18$$

5. Which of the following is **false**? Hint: It might be useful to sketch the distributions.

- If you calculated the Z score for the median of a right skewed distribution, you'd most likely get a positive number.
- If observations don't come from a distribution that's nearly normal, we can't calculate percentiles based on the Z table.
- Regardless of the shape and skew of a distribution, Z scores are still defined and can be calculated for observations from that distribution.
- The Z score for the median of a symmetric distribution is approximately 0.

✖ 틀립니다

This question refers to the following learning objective: Depending on the shape of the distribution determine whether the median would have a negative, positive, or 0 Z score keeping in mind that the mean always has a Z score of 0.

Calculating percentiles using the Z score and the normal probability distribution (e.g. the table) require normality, while calculation of Z scores do not.

