Introduction to statistics

Population: all people or things that you are studying.

Parameter: a numerical description measuring the variable of interest in the POPULATION.

Sample: A representative subset of the population.

Statistic: a numerical description measuring the variable of interest in the SAMPLE.

Variable: A value or characteristic that changes for each member of the population.

Data: counts, measurements or observations recorded about a specific variable to study in a population.

Census: when data is studied from every member of the POPULATION.

Descriptive statistics: → is about sample. Data that is collected, organized, summarized and displayed.

Inferential statistics: → is about population. using data that is gathered to make predictions (or inferences) about the population parameters.

Data
- Qualitative or nominal or categorical
- Quantitative or numerical (mean, median, mode, ...)

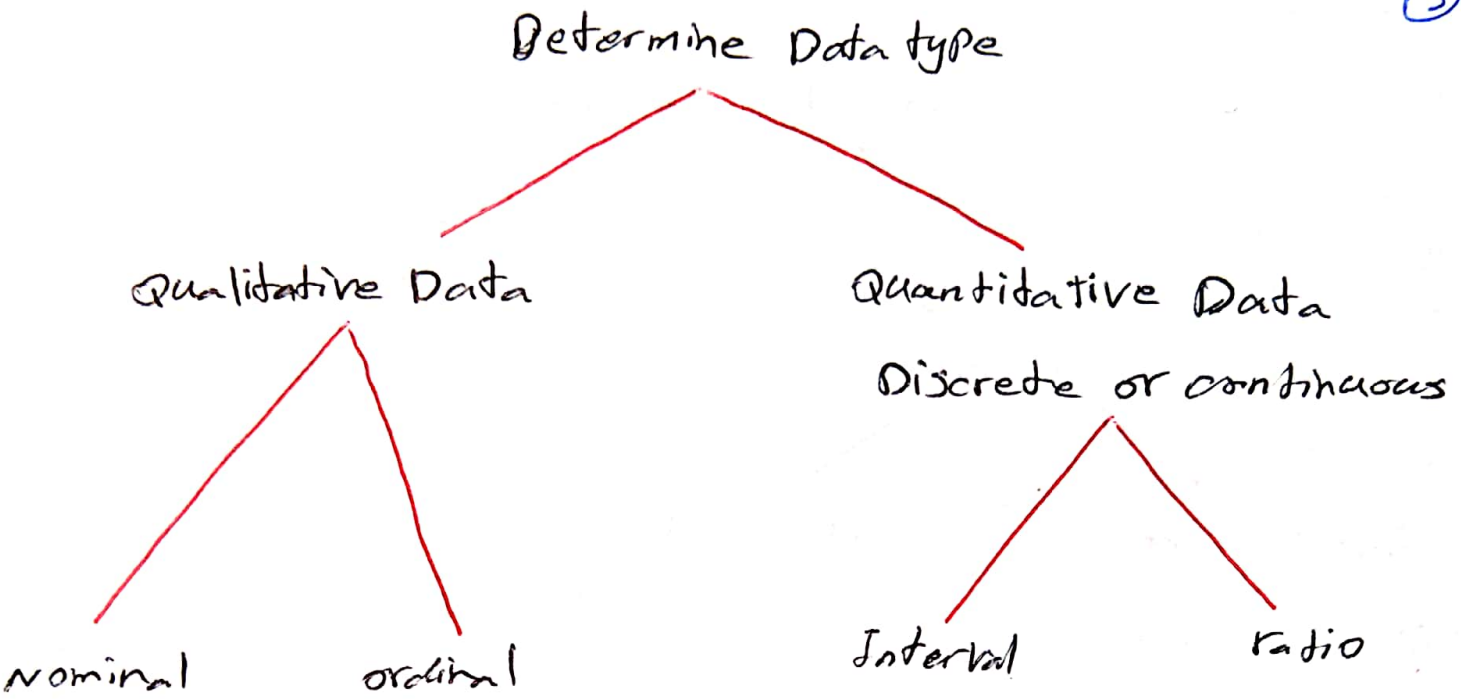
Quantitative
- continuous (measurment)
- Discrete (whole numbers) (counting)

Level of measurment
1) nominal (qualitative)
2) ordinal (qualitative)
3) Interval (quantitative)
4) Ratio (quantitative)

Exp: Birth years of your friends

Interval: can be arranged in a meaningful order and where differences between data entries are meaningful, but not Zero.

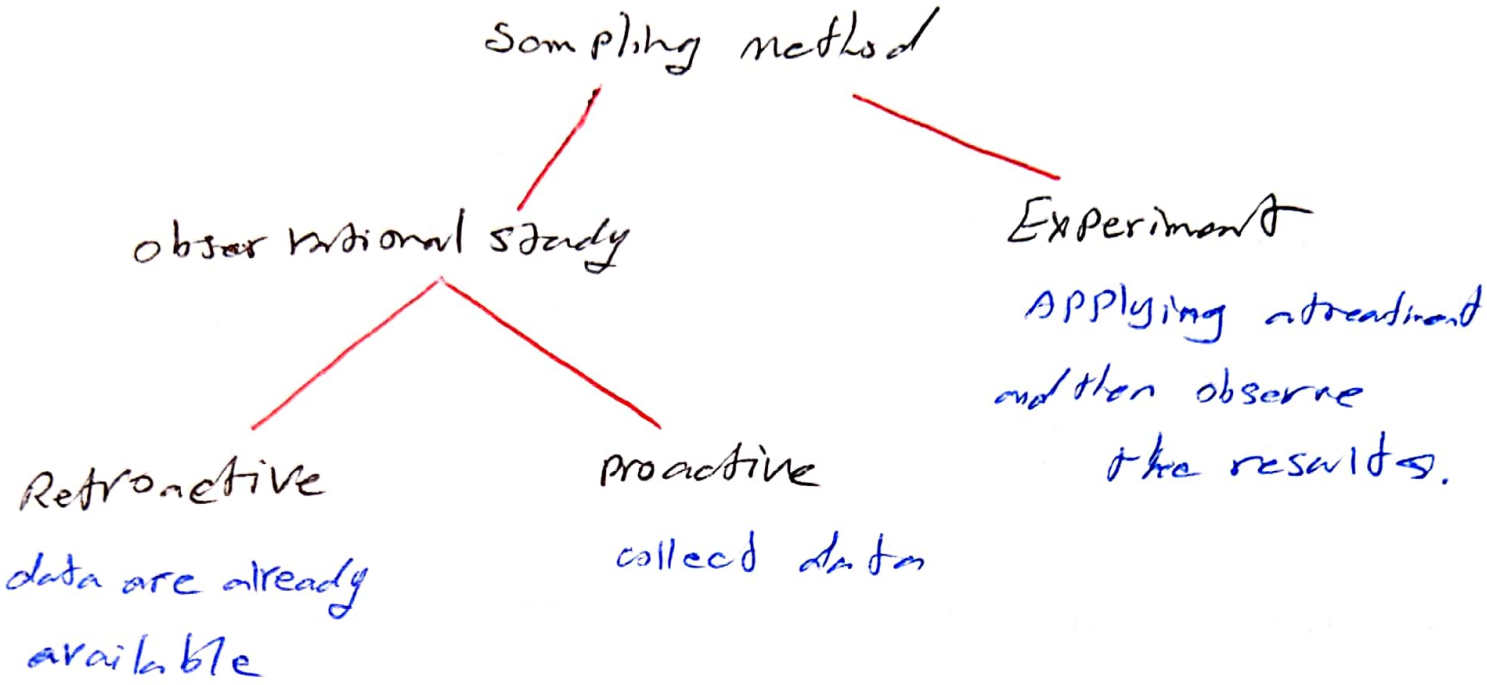
Ratio: can be arranged in a meaning ful order, when differences between data entries are meaningful and the Zero point indicates ~~of~~ the absence of something.



The Process of a statistical study

1) Determine the design of the study
2) collect the data
3) organize the data
4) Analyze the data to answer the question.

Determine the design of the study
- what question are we trying to answer?
- what is the population?
- what variables will we record?
- what is the sampling method?
- observational study
  - retroactive
  - proactive
- Experiment.



observational studies

- our goal is to collect data about our population.
- a census may be to expensive
- so we use Sample!

types of collect Sample's ~~types~~ or selected.

1) random sample: every member has same chance to be
2) simple random sample: every sample has same chance to be selected
3) stratified Sample: the population is divided in subgroups and a proportional random sample from each subgroups.
4) Cluster sample: same with stratified but without subgroups called clusters.
5) systematic sample.
6) convenience sample;

Types of observational studies

1) cross-sectional study: data are collected at a single point in time.
2) longitudinal study: data are gathered over a period of time.
3) meta-analysis: information from previous studies.
4) case study: look at multiple variables that affect a single event.

"Experiments"

- Experiments are the only way to show a cause-and-effect relationship.

- Treatment: some condition applied to a group of subjects in an experiment

- subjects: people (also known as participants) or things being studied in an experiment.

- Response variable: The variable in an experiment that responds to the treatment.

- Explanatory variable: The variable are alter in an experiment that caused the change in the response variable.

-control group: a group of subjects to which no treatment or a placebo is applied.

-Placebo effect: a response to the power of suggestion, rather than the treatment itself.

-Treatment groups: a group of subjects to which researchers apply a treatment.

~~single~~

-confounding variables: factors other than the treatment that cause an effect on the subjects of an experiment.

Experiment's steps:

1) Randomize the control and treatment groups.

2) control for outside effects on the response variable.

3) Replicate the experiment a significant number of times to see meaningful patterns.

"Critiquing a published study"

A) consider the study
- who paid for the study?
- where were the data collected?
- when was the information collected?
- who published the study?

B) consider the variable

EXP:
Bottle of a shampoo" makes hair 60% smoother
- smoother than what?
- under what condition?
- How did they means "smoothness"?
- in what type of weather was it tested?
- under what condition?
- what types of hair were tested?

C) consider the set-up:

Bias: is favoring of a certain outcome in a study.

Sampling bias: occures when the sample does not accurately represent the population.

- Dropouts: are participants who begin a study but fail to complete it.
- processing errors: errors that occur from the data processing
- nonadherent: a participant who remains in the study but strays from the directions.
- Researcher bias: when a researcher influences the result of a study.
- Response bias: when a researcher's behavior causes a participant to alter their response or when a participant gives an inaccurate response.
- Participation bias: occures when there is a lack of participat

-nonresponse bias: when a person refuses to participate in a survey or when a respondent omits questions when answering a survey.

D) consider the conclusion:
- Do the data support the conclusion?
- Do the result represent the whole picture or just a part?
- could there be other conclusion drawn?
- could there be other reasons for the same conclusion drawn?
- Does the study have any practical applications?

"Measures of center"

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

median: ?

mode: ?

"measures of spread"

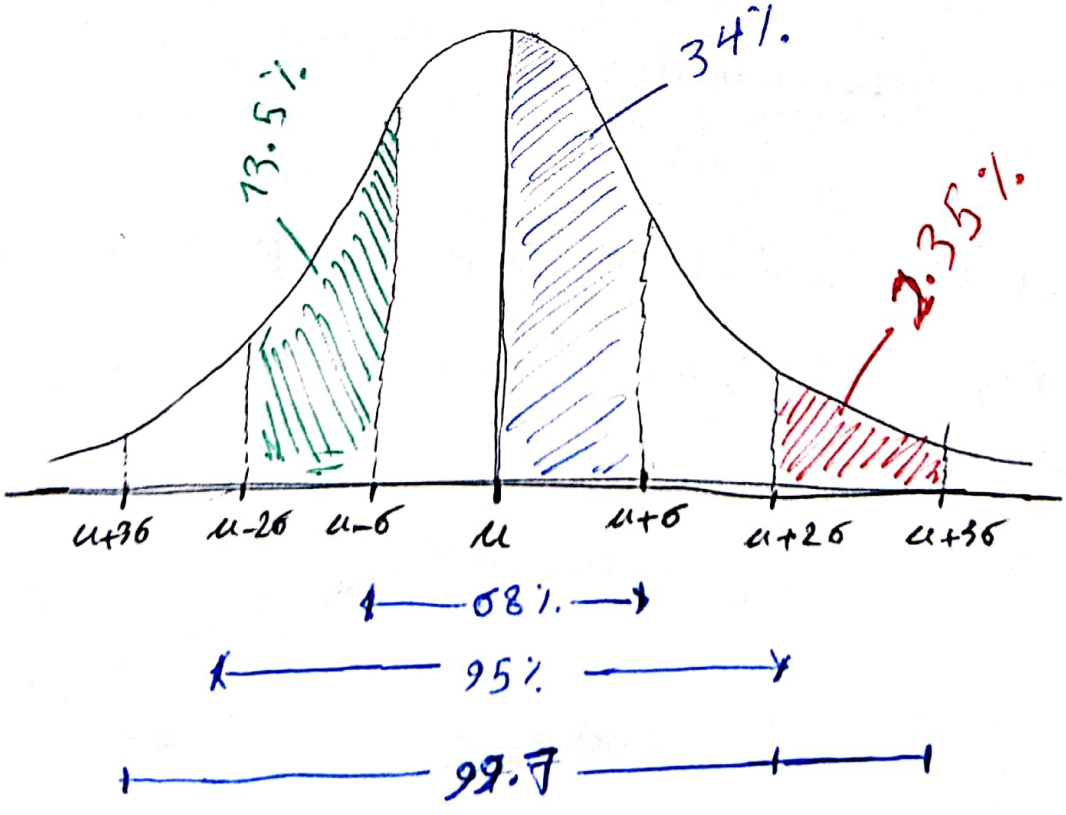
Range: max - min

Variance
- population: $\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$
- sample: $s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$

Standard deviation
- population: $\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$
- sample: $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$

"Empirical Rule" (~~for~~ used for normal destribution)

empirical rule for normal destributions: it gives us an estimate for how much data falls within 1, 2 or 3 standard deviations of our mean.



"Chebyshev's theorem" [for other distribution]

Chebyshev's theorem: estimate of spread of data based on standard deviation for non-normal data.

The proportion of data that lie within K standard deviations of the mean is at least $1-\frac{1}{K^2}$ for $K>1$.

Exp: $K=2 = 1-\frac{1}{2^2} = 1-\frac{1}{4} = \frac{3}{4} = 75\%$ falls within 2 std..

$K=3 = 1-\frac{1}{9} = \frac{8}{9} \simeq 88.9\%$

$K=4 = 1-\frac{1}{16} = \frac{15}{16} = 93.75\%$

"Percentiles"

- To find the data value for the P^{th} percentile, we use the location (L) of the data value (n) in the data set is given by,

$$L = n \cdot \frac{P}{100}$$

Exp: $\begin{cases} n=135 \\ P=10\% \end{cases}$ $L = n \cdot \frac{P}{100} = 135 \times \frac{10}{100} = 14$

that is the position of the value of our data set.

Find the value of the 10 the percentile.

Quartiles (Q_i)

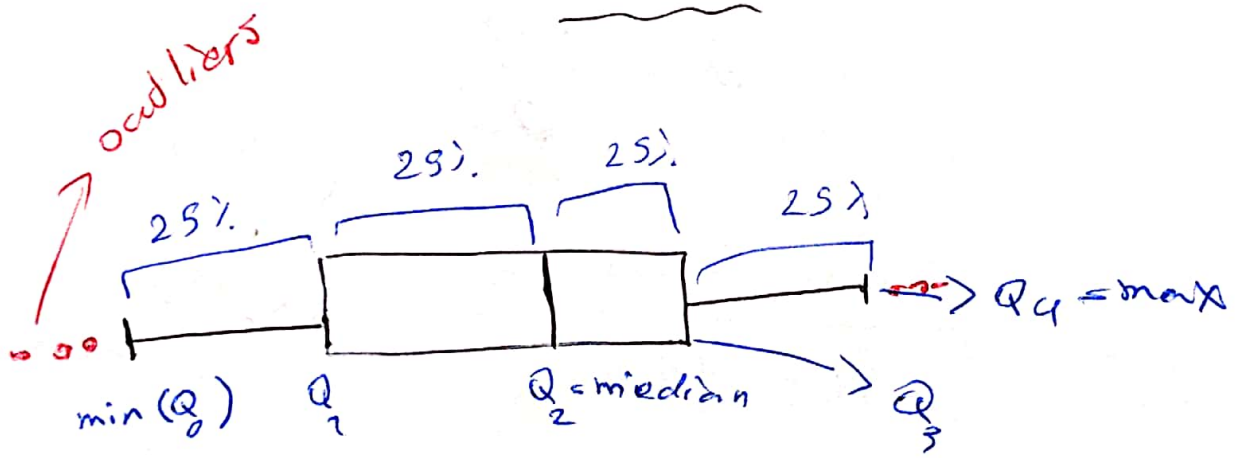
- a special type of percentile.

Q_1 = 25th percentile

Q_2 = 50th percentile or median

Q_3 = 75th percentile

"Box-Plot"



$$IQR = Q_3 - Q_1$$

- To determine if our data has outliers, we can move 1.5 IQR's outside of our quartiles.

$$\text{lower fence} = Q_1 - 1.5(IQR)$$
$$\text{upper fence} = Q_3 + 1.5(IQR)$$

every data outside these two numbers is outlier.

"Z-scores" or Standard scores

- it is a standard score, tells use how far a value is from the mean by computing how many std. it is away from the mean.

population Z-score $$Z = \frac{x - \mu}{\sigma}$$

sample Z-score $$Z = \frac{x - \bar{x}}{s}$$

$$Z = \frac{\text{observed} - \text{expected}}{\text{std. } (\sigma \text{ or } s)}$$

$\rightarrow (x - \bar{x})$ or $(x - \mu)$

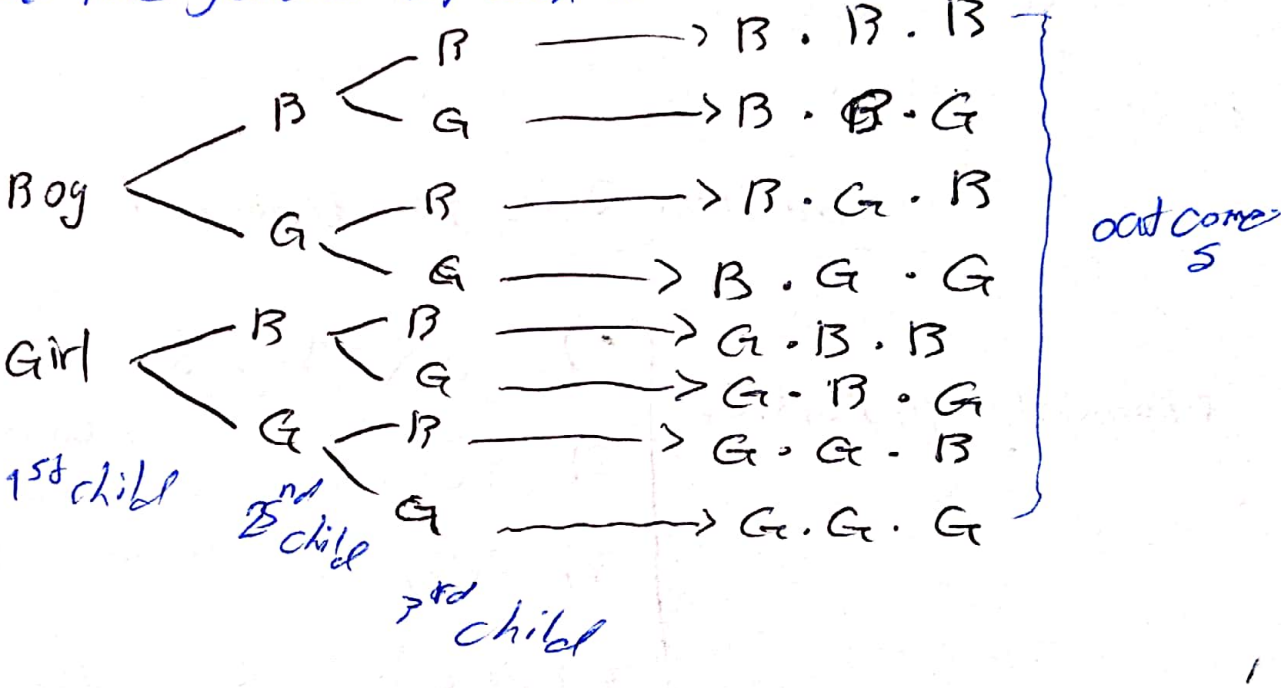
"probability"

- A probability experiment or trial is any process with a result determind by chance.
- An outcome is each individual result that is possible for a probability experiment.
- The sample space is the set of all possible outcomes for a given probability experiment.

- an Event is a subset of outcomes from the sample space.

A Tree diagram.

- a family with three children. use the tree diagram to find the gender of each child.



- There are three methods for calculating the probability of an outcome:

1) subjective probability: regarding of chance, the least precise type of probability. Exp: predicting the weather.

2) Experimental probability (aka Empirical): based on actual data. Exp: $P(E) = \frac{f}{n'} = \frac{93}{100} = 0.93$

f = frequency of event E

n' = the total number of time the experiment is performed

3) Classical probability (aka Theoretical): based on what should happen.

$P(E) = \frac{n(E)}{n(S)}$ ~~if~~ for example in a coin-flipping then $P(\text{Heads}) = \frac{1}{2} = 0.5$

- The law of large numbers: The more Time an experiment performed, The closer the probability will get to the classical probability (experimental probability)

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

"Addition Rule for Probability"

Probability properties:

1) $0 \leq P(E) \leq 1$
2) $P(S) = 1$ sum probability of the Sample space
3) $P(\phi) = 0$ event not in the sample space
4) $P(\bar{E}) = 1 - P(E)$ the complement rule.

- The addition rule for probability - "OR"

6-sided die

$P(3 \text{ or } 2) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$

$P(\text{less than } 4) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$ → P(1) or P(2) or P(3)

$P(\text{not less than } 4) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$ → $1 - P(\text{less than } 4)$

→ general addition rule

$P(\text{less than } 4 \text{ or even}) = ?$ $P(A \text{ or } B) = P(A) + P(B) - P(A \cap B)$

$P(\text{less than } 4) = P(1) + P(2) + P(3) = \frac{3}{6} = \frac{1}{2}$

$P(\text{even}) = P(2) + P(4) + P(6) = \frac{3}{6} = \frac{1}{2}$

$P(\text{less than } 4 \cap \text{even}) = P(2) = \frac{1}{6}$

so $\frac{1}{2} + \frac{1}{2} - \frac{1}{6} = \frac{5}{6}$

"multiplication Rule for probability"

multiplication Rule: $P(A \text{ and } B) = P(A) \times P(B)$ → with replacement

General Rule: $P(A \text{ and } B) = P(A) \times P(B|A)$

without replacement

$P(B|A)$ → P of given A

conditional probability - "Given"

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{P(A \cap B)}{P(A)}$$

- Fundamental counting Principle: if event A can happen in n ways and B can happen in m ways, then A and B can happen in $m \times n$ ways.

"Permutation (order matter) and combination (order doesn't matter)"

- factorials

$7! = 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$

$\frac{4!}{0!} = \frac{4 \times 3 \times 2 \times 1}{1}$

$\frac{95!}{93!} = \frac{95 \times 94 \times 93!}{93!} = 95 \times 94$

- Combination (order does not matter):

$$_nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

(n: number of items; r: number of choices)

- Permutation (order matters):

$$_nP_r = P(n,r) = \frac{n!}{(n-r)!}$$

- Permutation with repetition:

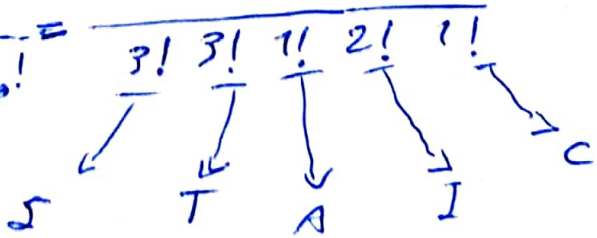
$$\frac{n!}{k_1!\, k_2! \ldots k_p!}$$

Exp: How many different ways can you arrange the letters in the word STATISTICS?

S = 3
T = 3
A = 1
I = 2
C = 1

$n = 10$

$$\frac{n!}{k_1!\, k_2! \ldots k_p!} = \frac{10!}{3!\,3!\,1!\,2!\,1!}$$



Discrete probability Distributions

expected value of

Expected value = weighted mean = $\sum_{i=1}^{n} (x_i \times P(X = x_i))$

Exp: rolling a six-sided dice:

~~$P(3) = \frac{1}{6}$~~ ~~50~~

x	1	2	3	4	5	6
P(x=x)	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

$$\sum_{i=1}^{n} (x_i \times P(X = x_i)) = \left(\frac{1}{6} \times 1\right) + \left(2 \times \frac{1}{6}\right) + \left(3 \times \frac{1}{6}\right) + \left(4 \times \frac{1}{6}\right)$$

$$+ \left(5 \times \frac{1}{6}\right) + \left(6 \times \frac{1}{6}\right) = \frac{1}{6} + \frac{2}{6} + \frac{3}{6} + \frac{4}{6} + \frac{5}{6} + \frac{6}{6}$$

Exp: let's play a game called Hearts. It costs $5 to play. from a deck of cards, you randomly draw 1 card. If you draw:

- Queen or King of Hearts you get $ 150
- Any other Queen or King you get $ 10
- any other Heart you get $5

Is this a good game to play?

Outcome	Q/K hearts	other Q/K	other Hearts	other cards
x	$ 145 $	$ 5	0	-5 $
P(x=x)	$\frac{2}{52}$	$\frac{6}{52}$	$\frac{11}{52}$	$\frac{33}{52}$

$$\sum_{i=1}^{n} \left(x_i \cdot P(X=x_i)\right) = \left(145 \times \frac{2}{52}\right) + \left(5 \times \frac{5}{52}\right) + \left(0 \times \frac{11}{52}\right) - \left(5 \times \frac{33}{52}\right) \simeq 2.98$$

we pay $5 to play the game but just can win ≃ $3 per game so that is not a good game to play.

- Variance and standard diviation of Discrete Prob.

$$\sigma^2 = \sum \left[(x_i - \mu)^2 \cdot P(X = x_i)\right]$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum \left[(x_i - \mu)^2 \cdot P(X = x_i)\right]}$$

"The Binomial Distribution"

- in binomial distribution, the experiment must have the following characteristics:

A) The experiment consists of a fixed number, n, of identical and independent trials.

B) There are only two possible outcomes - success or failure.

probability of success = p

probability of failure = $1 - p = q$

number of trials = n

The random variable (x) representing the number of successes in n trials.

combination $= \frac{n!}{x!(n-x)!}$

$$P(X=x) = {}_nC_x \cdot p^x (1-p)^{(n-x)}$$

EXP: what is the probability of getting 6 heads in 10 tosses of a coin?

$p = 0.5$

$q = 0.5$

$n = 10$

$X = 6$

$$P(X=6) = {}_{10}C_6 \cdot (0.5)^6 (1-0.5)^{(10-6)}$$

$$P(X=6) = {}_{10}C_6 (0.5)^6 (0.5)^4$$

$$P(X=6) = \frac{10!}{6!(10-6)!} \cdot (0.5)^6 \cdot (0.5)^4$$

$$P(X=6) = \frac{10 \times 9 \times 8 \times 7}{4 \times 3 \times 2 \times 1} \cdot (0.5)^6 \cdot (0.5)^4$$

$$P(X=6) \approx 0.2051$$

"Camulative Bihomial Distribution"

EXP: in an experiment of fliping a coin 10 times, what is the probability of fliping 7 or more heads?

$$P(X \geq 7) = P(X=7) + P(X=8) + P(X=9) + P(X=10)$$

or $1 - P(X<7)$

"The Poisson Distribution"

- in the Poisson distribution, the experiment must have the following characteristics:

A) Each success must be independent of other successes

B) The mean number of successes in a given interval must remain constant

- mean number of success in a given interval = λ
- The remain variable = X → it represent the number of success in the given interval.

$$P(X=x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

Exp: A barber usually finishes one haircut every 15 mins. calculate the probability that our barber is feeling extra speedy one day and finishes 6 haircuts in one hour.

$$P(X=x) = \frac{e^{-\lambda} \lambda^x}{x!} = \frac{e^{-4} 4^6}{6!} = ?$$

$\lambda = \frac{60}{15} = 4$

$x = \frac{60}{10} = 6$

"The Hypergeometric Distribution"

- in this case, the experiment must have the following characteristics:

A) each trial consists of an outcome of P or Q.

B) The population is of a known size.

C) The Trials are dependent (that is, selections are made without replacement).

- number of successes: X
- number of dependent trials: n
- number of successes in the population: K
- number of items in the population: N

$$P(X=x) = \frac{({}_kC_x)({}_{N-K}C_{n-x})}{({}_NC_n)}$$

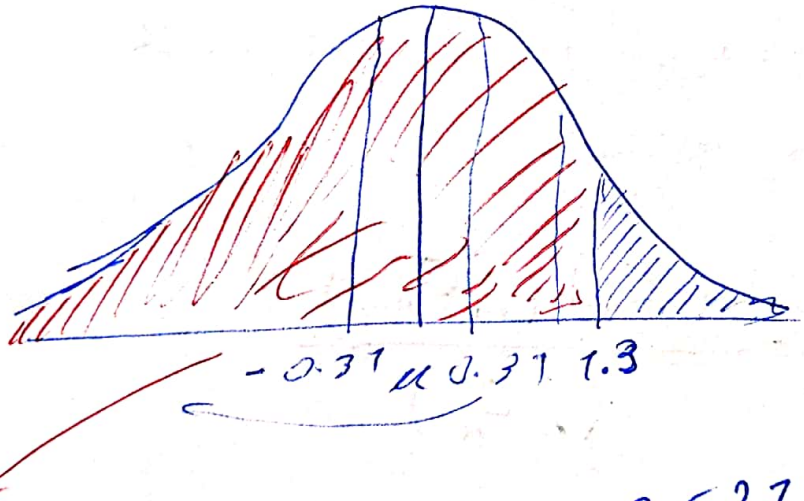
Exp: At a local grocery store there are twenty boxes of cereal on one shelf, half of which contain a prize, suppose that you buy three boxes of cereal. what is the probability that all three boxes contain a prize?

population	sample
$N = 20$	$n = 3$
$K = 10$	$x = 3$

$$P(x=3) = \frac{({}_{10}C_3)\,({}_{20-10}C_{3-3})}{{}_{20}C_3} = \frac{({}_{10}C_3)({}_{10}C_0)}{{}_{20}C_3}$$

$$= \frac{(120)(1)}{(1140)} = \frac{2}{19} \simeq 0.1053$$

"The normal Distribution and Z-Score"



(X)

$P(Z \leq 0.31)$ = using Z table $= 0.6217$

$P(Z < 1.37) = 0.9146$

$P(Z < -2.03) = 0.02117$

Exp: Body temperatures for adults are normally distributed with a mean of 98.6 °f and a standard deviation of 0.73 °F.

A) what is the probability of a healthy adult having a body temperature less than 97 °f?

$$Z = \frac{x-\mu}{\sigma} = \frac{97-98.6}{0.73} = -2.19178 \xrightarrow{\text{use } Z\text{-table}}$$

$$P(Z < 97°f) \Longrightarrow 0.014197$$

Probability of less than 97°f

B) what is the probability of a healthy adult having a body temperature more than 99.2 °f?

$$I = \frac{x-\mu}{\sigma} = \frac{99.2-98.6}{0.73} = 0.82$$

$$P(Z > 0.82) = 1 - P(Z < 0.82) = 0.2056$$

Exp: Compute the Z-scores for each situations

A) an area to the left of 0.239:

$$0.239 \xrightarrow{\text{use-Z-table}} \boxed{-0.7095} \rightarrow \text{Z-score}$$

if $\begin{cases} \text{mean} = 52 \\ \text{std} = 3 \end{cases} \Longrightarrow I = \frac{x-\mu}{\sigma} \Rightarrow x = Z\sigma + \mu$

$$\Rightarrow x = (-0.7095)(3) + 52$$

$$x = 49.871$$

B) The 80th percentile: $\mu = 52, \sigma = 3$

$0.8 \xrightarrow[\text{using } Z\text{-table in reverse}]{\text{refer}} Z = 0.8416$

$X = Z\sigma + \mu$

$= (0.8416)(3) + 52 = 54.525$

C) an area to the right of 0.963

$0.963 \longrightarrow Z = -1.79$

Approximating a Binomial Dist. with a normal Dist.

- If the conditions that $np \geq 5$ and $n(1-p) \geq 5$ are met for a given binomial distribution, and the distribution is very large, then a normal distribution can be used to approximate the binomial probability distribution with the mean and Stdv. given by.

$\mu = np \qquad \sigma = \sqrt{np(1-p)}$

one problem:

- we are trying to use the normal distribution, which is continuous, to approximate the binomial distribution, which is discrete. Therefore, a continuity correction must be used to convert the whole number value of the discrete binomial random variable to an interval range of continuous normal random variable.

- To make this correction, determine the value x of the binomial random variable and convert it to the interval from $(x-0.5)$ to $(x+0.5)$.

- Exp: use a normal distribution to estimate the probability of more than 55 girls being born in 100 births. Assume that the probability of a girl being born in an individual birth is 50%.

first check the conditions

$$np = 100(0.5) = 50 \geq 5$$

and

$$n(1-p) = 100(0.5) = 50 \geq 5$$

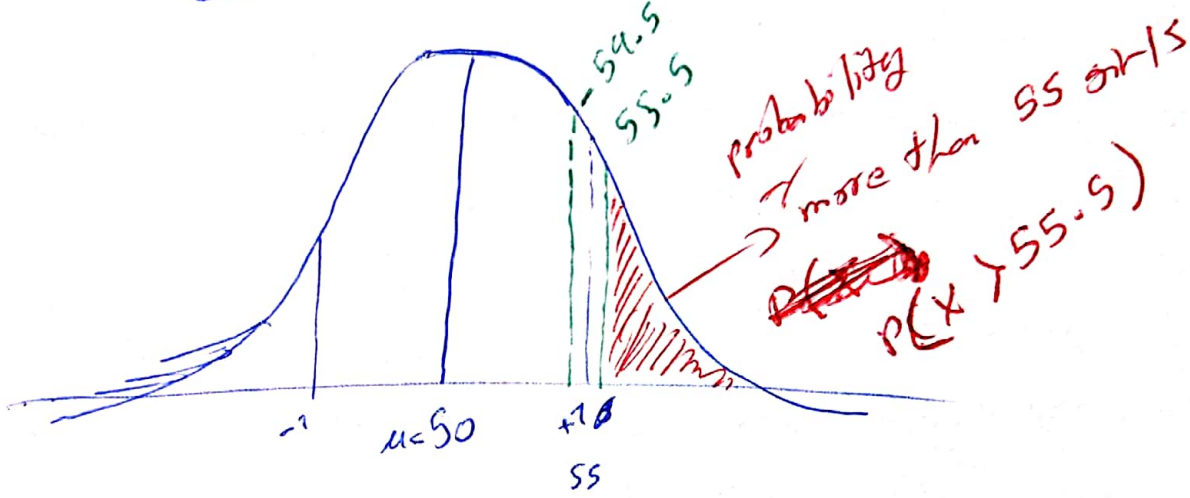
second

$$\mu = np = 100(0.50) = 50$$

$$\sigma = \sqrt{np(1-p)} = 100(0.5)(0.5) = \sqrt{25} = 5$$

third: calculate the continuity corrections:

$$\begin{cases} 55 + 0.5 = 55.5 \\ 55 - 0.5 = 54.5 \end{cases}$$



0.555 $\xrightarrow[\text{Z-Score table}]{\text{using Z-table}}$ 1.1

$$P(x > 55.5) = 1 - P(x < 55.5) = 0.1356$$

"The Central Limit Theorem"

- it states that as the sample size increases, the distribution of the sample means approaches a normal distribution, regardless of the underlying distribution of the population from which the sample is drawn.

The ~~Sample~~ variables of the sample must be:

1) independent
2) identically distributed random variables
3) the sample size must be large enough ($30 < n$)

- if the requirement is met, the sampling distribution of sample means will have the following three characteristics:

1) $\mu_{\bar{x}} = \mu$

↓ the mean of a sampling distribution

2) the standard deviation of a sampling distribution of sample mean $\longrightarrow \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

3) "The (CLT) for means"

$$Z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$\frac{\sigma}{\sqrt{n}} \rightarrow$ Standard error

"The (CLT) for Proportions"

$p = \frac{x}{N}$ → Population proportion (success)

$\hat{p} = \frac{x}{n}$ → Sample proportion

The requirements:

A) $np \geq 5$ & $n(1-p) \geq 5$

B) the population is ~~not~~ binomial distributed given by proportional data.

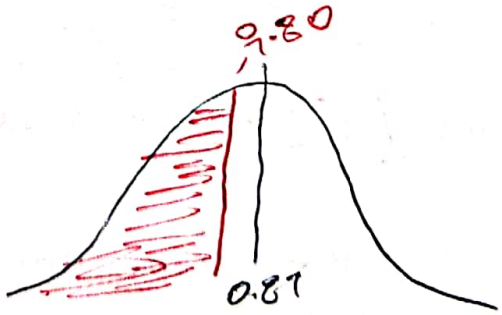
The mean of the sampling distribution of sample proportion

mean $\longrightarrow m_{\hat{p}} = p$

Std. $\longrightarrow \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

Z-score $\longrightarrow Z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$

Exp: In a certain liberal precinct across town, 81% of the votes are registered Democrates. what is the probability that, in a Random sample of 100 voters from this precinct, no more than 80 of the voters would be registered Democrates?



step 1: Identify $\hat{p}$, p and n

$n = 100$
$p = 0.81$
$\hat{p} = 0.8$

step 2: Draw the curve

step 3: Find the Z-score $Z = \frac{0.80 - 0.81}{\sqrt{\frac{0.81(0.19)}{100}}} \approx -0.2549$

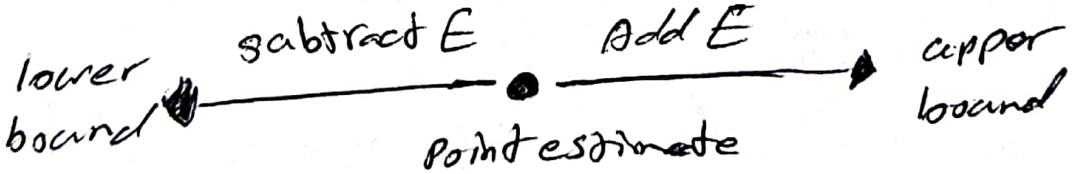
"Confidence Intervals"

- The level of confidence is represented by the letter C, thus, for a 95% confidence interval, C = 0.95

- C tells us if there is a 95% level of confidence given, that we are confident that using this statistical method will result in 95% of all possible interval estimates containing the true population parameter.

- How do we find a confidence interval for a population parameter?

- The Margin of Error or maximum error of estimate, E, is the largest possible distance from the point estimate that a confidence interval will cover.



- EXP: A college student researching study habits collects data from a random sample of 250 college students on her campus and calculates that the sample mean is $\bar{x} = 13.7$ hours per week. If the margin of error for her data using a 95% level of confidence

is $E = 0.6$ hours, construct a 95% confidence (30)
interval for her data. Interpret your results

point estimate: $\bar{x} = 15.7$

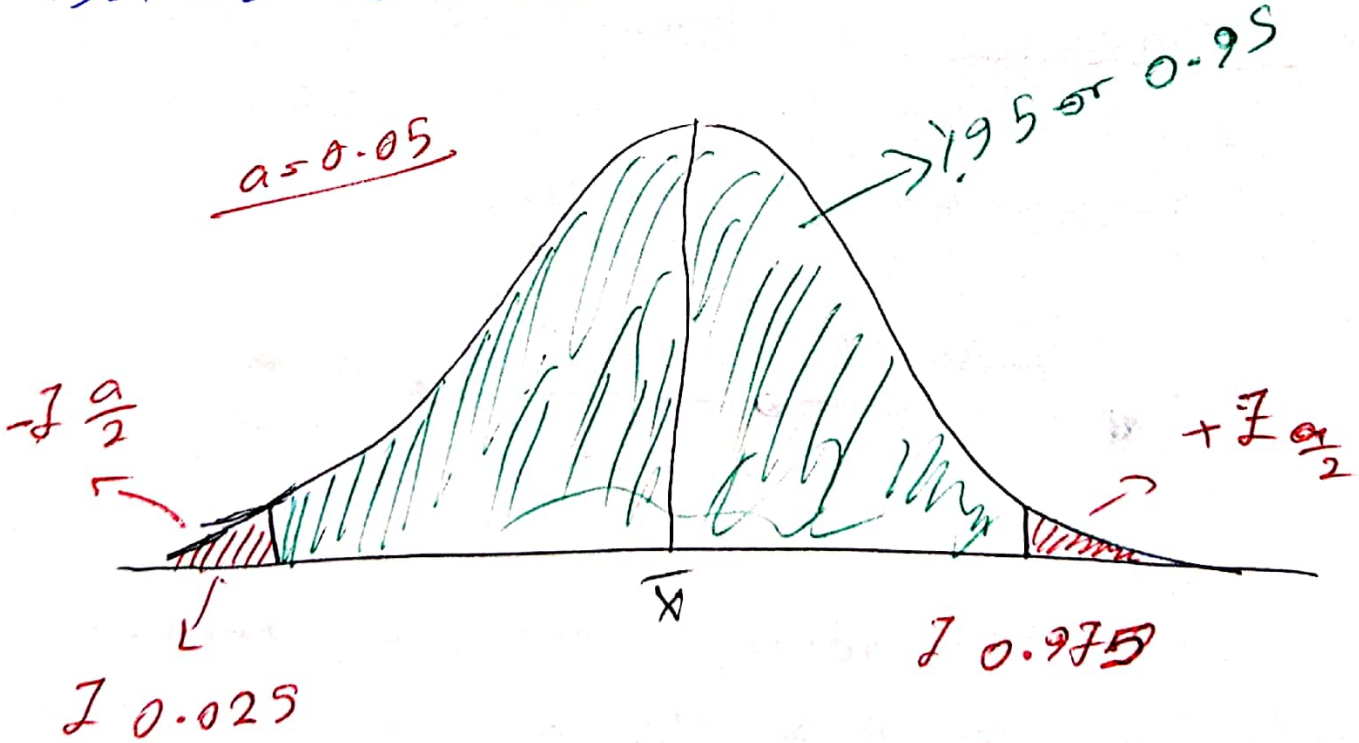
margin of Error: $E = 0.6$

lower bound: $\bar{x} - E = 15.7 - 0.6 = 15.1$ (hrs/week)

upper bound: $\bar{x} + E = 15.7 + 0.6 = 16.3$ (hrs/week)

confidence Interval: $15.1 < x < 16.3$ or $(15.1, 16.3)$

Interpretation: we are 95% confident that the true population mean for the number of hrs/week that student on this campus spend studying is between 15.1 and 16.3 hrs.



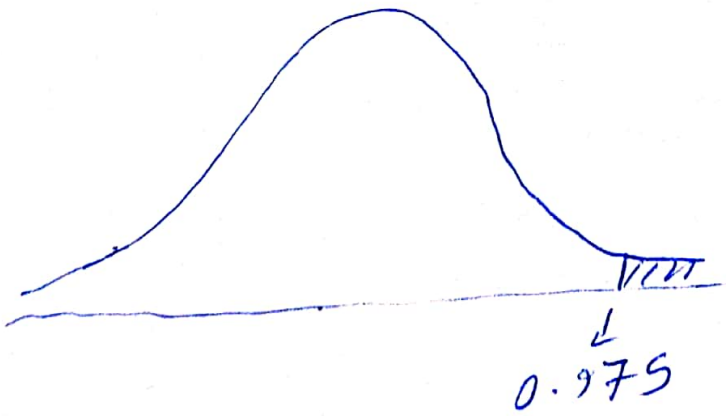
$$E = \left(Z_{\frac{\sigma}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$$

EXP: In order to estimate the number of calls to exPect at a new suicide hotline, volunteers contact a random sample of 35 similar hotline across the nation and find that the sample mean is 42.0 calls per month. construct a 95% confidence interval for the mean number of calls per month. Assume that the population standard deviation is known to be 6.5 calls per month.

1) check the conditions
- a random sample ✓
- ~~The~~ σ is known ✓
- $n \geq 30$ ✓

2) find $\bar{x}$ (point estimate). $\bar{x} = 42$

3) find $z_{\frac{\alpha}{2}}$, n and σ
- $n = 35$
- $\sigma = 6.5$
- $z_{\frac{\alpha}{2}} = 1.96$



0.975 $\xrightarrow{\text{z-Score table}}$ 1.96

4) calculate the margin of error: $E = \left(Z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$

$$E = (1.96)\left(\frac{6.5}{\sqrt{35}}\right) \simeq 2.153453$$

5) calculate the interval

lower bound: $\bar{x} - E = 42 - 2.1535 \simeq 39.8$ calls/month

upper bound: $\bar{x} + E = 42 + 2.1535 \simeq 44.2$ calls/month

(39.8, 44.2)

↙

the true population mean is in this interval.

* Find the sample size based on confidence Intervals

$$E = Z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$\frac{E}{Z_{\frac{\sigma}{2}}} = \frac{Z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)}{Z_{\frac{\sigma}{2}}}$$

$$\sqrt{n} \cdot \frac{E}{Z_{\frac{\sigma}{2}}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{n}$$

$$(\sqrt{n})^2 = \left(\frac{\sigma Z_{\frac{\alpha}{2}}}{E}\right)^2 \Rightarrow n = \left(\frac{\sigma Z_{\frac{\alpha}{2}}}{E}\right)^2$$

Exp: Determine the minimum sample size needed if we wish to be 90% confident that the sample mean is within two units of the population mean. An estimate for the population std. of 8.4 is available from a previous study.

E=2

$\sigma = 8.4$

$C = 0.90$

$E = 2$

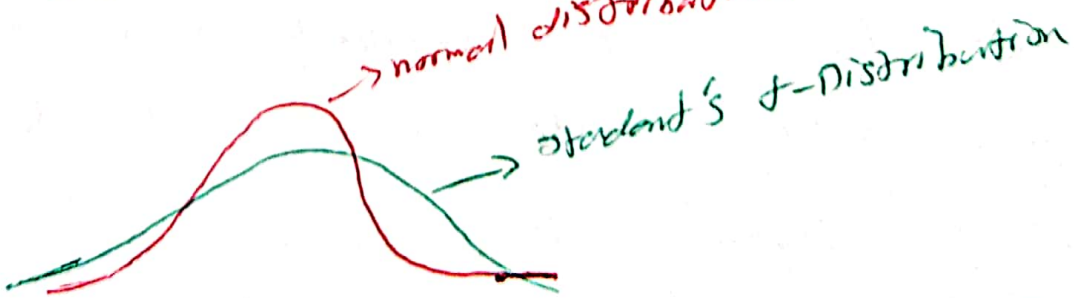
$I_{\frac{\alpha}{2}} = I_{\frac{0.10}{2}} = I_{0.05} = 1.645$

$$n = \left(\frac{I_{\frac{\alpha}{2}} \cdot \sigma}{E} \right)^2 = \left(\frac{1.645 \times 8.8}{2} \right)^2 = 47.73 \approx 48$$

the least size of Sample.

"Student's t-Distribution"

- when the population is normally distributed or the sample size is sufficiently large. in this case Student's t-distribution is a good option to calculate the margin of Error for a population mean when σ is unknown.



whereas

- standard normal distribution has two parameters 1) $\mu = 0$

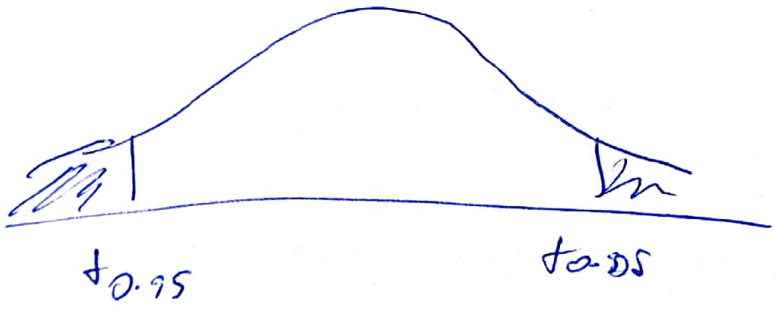
2) $\sigma = 1$, the t-distribution has only one parameter

- the number of degrees of freedom (df)

- A t-distribution with fewer degrees of freedom has more area under the tails of the curve.

Exp: find the value of t for a t-distribution with 17 degrees of freedom such that the area under the curve to the left of t is 0.05. (when we denote t, the subscript denotes the area to the right, so we would be finding $t_{0.95}$)

~~in excel = T.INV(0.05, 17)~~



- find the value of t for a t-distribution with 17 df such that the area under the curve to the right of t is 0.10. this would be indicated $t_{0.1}$. However we would calculate the area to the left; 1 - 0.1 = 0.9

in excel: t.inv(0.9, 17)

$$(\bar{x}-E, \bar{x}+E) \text{ where } E=\left(t_{\frac{\alpha}{2}}\right)\left(\frac{s}{\sqrt{n}}\right)$$

$$\left(\bar{x}-\left(t_{\frac{\alpha}{2}}\right)\left(\frac{s}{\sqrt{n}}\right), \bar{x}+\left(t_{\frac{\alpha}{2}}\right)\left(\frac{s}{\sqrt{n}}\right)\right)$$

EXP: A marketing company wants to know the mean price of new vehicles sold in an up-and-coming area of town. The simple random sample of 756 cars has a mean of $27,600 with a std. of $1300. Construct a 95% confidence interval for the mean price of new cars sold in this area.

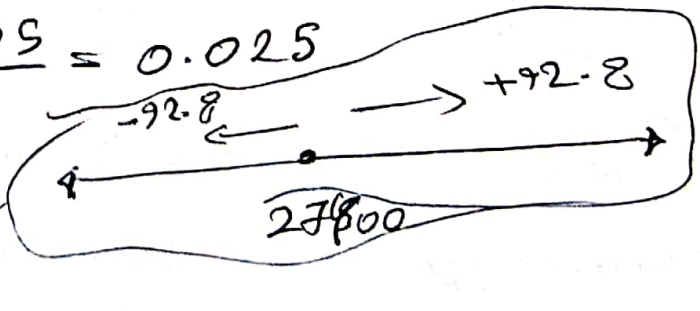
Step 1) find the point estimate: The point estimate for the population mean is the sample mean, which is $27,600.

Step 2) find the critical value: since $n=756$, is sufficiently large, we use the student's t-distribution to calculate the critical value.

$\alpha = 1-0.95 = 0.05 \rightarrow \frac{0.05}{2} = 0.025$

$df = n-1 = 755$

So, $t_{0.025} \approx 1.963$



Step3) $E=\left(t_{\frac{\alpha}{2}}\right)\left(\frac{s}{\sqrt{n}}\right)=(1.963)\left(\frac{1300}{\sqrt{756}}\right) \simeq 92.811706$

step 4) lower end point = $\bar{x} - E = 27900 - 92.8 \simeq 27304$

upper end point = $\bar{x} + E = 27,493$

step 5) ($27,307 , $27493)

Estimating population proportions

- since the p (population proportion) is often difficult to find, we use $\hat{p}$ (sample proportion) as our point estimate $\hat{p} = \frac{x}{n}$. we will calculate a margin of error using the normal model and then add it an subtracting it from a point estimate. $(n\hat{p} \geq 10 \quad n(1-\hat{p}) \geq 10)$

$(\hat{p} - E, \hat{p} + E)$ where $E = z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$$\left(\hat{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

EXP: A survey of 345 randomly selected students at one university found that 307 students think that there is not enough parking on campus. Find the 90% confidence interval for the proportion of all students at this uni who think that there is not enough parking on campus.

$$\hat{p} = \frac{x}{n} = \frac{301}{345} \simeq 0.87$$

$n = 345$

$C = 0.9$

$\alpha = 1 - 0.9 = 0.1 \Rightarrow \frac{\alpha}{2} = 0.05$

$$\left(\hat{p} - Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}} \, , \hat{p} + Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}} \right)$$

$$\left(0.87 + 1.645 \sqrt{\frac{0.87 \cdot 0.13}{345}} \, , 0.87 + 1.65 \sqrt{\frac{0.87 \cdot 0.13}{345}} \right)$$

0.00295

upper bound $= 0.87 + 0.0295 \simeq 0.843$

lower bound $= 0.87 - 0.0295 \simeq 0.902$

$$(0843, 0.902)$$

Interpret the interval: with 90% confidence, we estimate that between 84.3% and 90.2% of all students at this uni think there is not enough parking on campus.

"calculations with population proportions"

$$E = Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\Rightarrow \left(\frac{E}{Z_{\frac{\alpha}{2}}}\right)^2 = \frac{\hat{p}(1-\hat{p})}{n}$$

$$\Rightarrow n \cdot \left(\frac{E}{Z_{\frac{\alpha}{2}}}\right)^2 = \hat{p}(1-\hat{p})$$

$$\Rightarrow \boxed{n = \hat{p}\hat{q}\left(\frac{Z_{\frac{\alpha}{2}}}{E}\right)^2}$$

EXP: what is the minimum sample size needed for a 99% confidence interval for the population proportion if previous studies indicated $p \approx 0.54$ and we desire no more than a 2% margin of error?

$\hat{p} \approx 0.54$

$\alpha = 0.01$

$\frac{\alpha}{2} = 0.005$ $\Big\}$ $Z_{0.005} = ?$

$E = 0.02$

$$n = \hat{p}\hat{q}\left(\frac{Z_{\frac{\alpha}{2}}}{E}\right)^2$$

$$n = 0.54 \times 0.46 \left(\frac{2.575}{0.02}\right)^2$$

$$n = 4177.078725$$

$$n \simeq 4178$$

Comparing Population means (σ Known)

Comparing two Samples:

A) The samples are random and independent.

B) both the population std.s are known.

C) either $n_1, n_2 \geq 30$ or both are normally distributed.

- Creating the confidence interval

1) find the point estimate: $\bar{X}_1 - \bar{X}_2$

2) find the margin of error: $E = Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

3) the confidence interval,

$$\left((\bar{X}_1 - \bar{X}_2) - Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} , (\bar{X}_1 - \bar{X}_2) + Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

a) interpret the confidence Interval:

- If the CI contains 0, we must conclude that there is a possibility that the two population means are the same.
- if the CI lies completely on the positive side of the line, the first mean > second mean $\bar{X}_1 > \bar{X}_2$
- if the CI is completely on the negative side of the line

the first mean < second mean

EXP: A researcher is looking at the study habits of college students. A random sample of 92 freshmen reported a mean study time of 15 hr/week. A random sample of 39 seniors reported a mean study time of 23 hr/week. Construct a 95% CI to estimate the true difference between the amount of time per week that both groups study. the std. for freshman = 4.7 and for seniors = 5.2.

$\bar{x}_1 - \bar{x}_2 = 15 - 23 = -8$ hr/w

$\alpha = 1 - 0.95 = 0.05$

$\frac{\alpha}{2} = 0.025$

$E = t_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \Rightarrow E = 1.96 \sqrt{\frac{(4.7)^2}{92} + \frac{(5.2)^2}{39}} \approx 2.164$

$(\bar{x}_1 - \bar{x}_2) - E \; , \; (\bar{x}_1 - \bar{x}_2) + E)$

$(-10.164 \; , \; -5.836)$

Interpretations: with 95% Confidence freshmen study between 5.8 - 10.2 hr/w less than seniors.

Comparing population means (σ unknown)

- The population variances are assumed to be unequal.

1) $(\bar{x}_1 - \bar{x}_2) = ?$

2) $E = t_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

3) $\left((\bar{x}_1 - \bar{x}_2) - E \; , \; (\bar{x}_1 - \bar{x}_2) + E \right)$

Exp: misty believes that the difficulty she is having in the class is the result of an inexperienced teacher. She believes that students in her class are reciving lower scores on their exams than they in another class with a more experienced teacher.

$n_1 = 11$, $\bar{x}_1 = 75$, $s_1 = 8$ | $c = 90\%$ | $n_2 = 9$, $\bar{x}_2 = 82$, $s_2 = 5$

← both. normally distributed

$(\bar{x}_1 - \bar{x}_2) = 75 - 82 = -7$

$t_{\frac{\alpha}{2}} = \frac{1 - 0.9}{2} = \frac{0.1}{2} = 0.05$

* df = the number of degrees of freedom is the smaller of the values $n_1 - 1$ to and $n_2 - 1 = 8$

so, $df = 8$

$t_{0.05} = 1.860$

$E = t_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 1.860 \sqrt{\frac{8^2}{11} + \frac{5^2}{9}} \simeq 5.453$

$\left((\bar{x}_1 - \bar{x}_2) - E , (\bar{x}_1 - \bar{x}_2) + E\right)$

$(-12.5 , -1.5)$

Interpretations: with the 90% confidence, the mean exam score for students in another class is better than her class, ~~90~~.

Comparing two pop. means (σ unknown, Equal Var)

- The population variances are assumed to be equal (as pooled variance)

1) Point estimate : $\bar{x}_1 - \bar{x}_2$

2) Margin of error where $df = n_1 + n_2 - 2$

$$E = t_{\frac{\alpha}{2}} \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

3) Interval confidence:

$$\left((\bar{x}_1 - \bar{x}_2) - \underbrace{t_{\frac{\alpha}{2}} \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}}_{E} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \; , (\bar{x}_1 - \bar{x}_2) + E \right)$$

Exp: numbers of hours per week spent studying for a particular mathematics class were gathered from randomly selected students. One of the questions asked if the respondent usually chooses a seat at the front of the classroom or at the back. Assume that both population distributions are approx. normal. Complete a 90% confidence interval

Front of classroom Performed:
12, 14, 17, 9, 7, 11, 12, 10, 8, 9, 5

Back of classroom performed:
9, 11, 5, 8, 9, 5, 10, 7, 8, 7, 5, 2, 9, 6, 8

$n_1 = 11$

$\bar{x}_1 = \sum_{i=1}^{n} x_i \cong 9.9$

$S_1 = \sqrt{\frac{\sum (\bar{x} - x)^2}{n-1}} = 3.86$

$n_2 = 15$

$\bar{x}_2 \cong 6.27$

$S_2 \cong 2.46$

$(\bar{x}_1 - \bar{x}_2) \cong 3.64$

$df = n_1 + n_2 - 2 = 24$

$C = 90\% = 0.9$

$\alpha = 1 - 0.9 = 0.1$

$\frac{\alpha}{2} = 0.05$

$t_{\frac{\alpha}{2}} = 0.05$

$t_{0.05} \cong 1.71$

$E = 5.45$

$((\bar{x}_1 - \bar{x}_2) - E, (\bar{x}_1 - \bar{x}_2) + E)$

$(-1.8, 9.1)$

Comparing two population means (σ unknown, Paired)

- The samples are random and dependent.

our datas are dependent.

- The paired difference for any pair of data values is given by $d = x_2 - x_1$
- The mean of the paired differences (point estimate) is given by $\bar{d} = \frac{\Sigma d_i}{n}$
- The margin of Error $E = t_{\frac{\alpha}{2}} \sqrt{\frac{\frac{\Sigma (d_i - \bar{d})^2}{n-1}}{n}}$
- The Std. (of the differences) is $\frac{\Sigma (d_i - \bar{d})^2}{n-1}$. we devided it by $\sqrt{n}$ ~~for the~~ to get the standard Error.

- substract and add the E to the point estimate.

$$(\bar{d} - E, \bar{d} + E)$$

Exp: The amounts of home utility bills are given for two consecutive months for 5 different homes in a neighborhood. find a 90% confidence interval for the mean difference in utility bills from march to April.

Sample size is too small but it is just a practice!

March (x_1)	April (x_2)	diff.
119.75	121.06	1.31 → $x_2 - x_1$
68.43	79.04	10.61
202.39	189.55	-12.84
47.88	49.64	1.76
66.01	68.52	2.51

mean of Diff. = 0.07

Std of Diff = 8.96

n = 5

C-Level = 0.9

Critical value ($t_{\frac{\alpha}{2}}$) = 2.131

Standard Error = 3.78

E = 8.07359

lower point = -7.99

upper point = 8.735

"Comparing population proportions" (two pop.s)

the criteria

1) equal chance for all possible samples to be choose.
2) The samples are independent.
3) The conditions for a Binomial distribution are met (p and q)
4) The sample sizes are large enough to ensure that each distribution has at least 10 p and 10 q. (1-p)

$$n_1\hat{p}_1 \geq 10 \;,\; n_1(1-\hat{p}_1) \geq 10$$
$$n_2\hat{p}_2 \geq 10 \;,\; n_2(1-\hat{p}_2) \geq 10$$

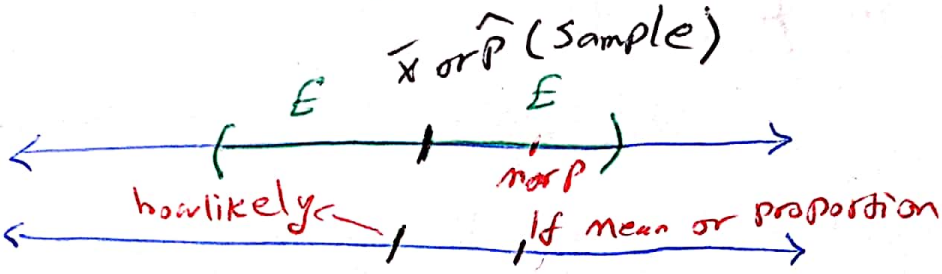
A) The point estimate = $\hat{p}_1 - \hat{p}_2$

B) margin of Error $E = z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

C) $\left((\hat{p}_1-\hat{p}_2) - E \;,\; (\hat{p}_1-\hat{p}_2) + E\right)$

"Introduction to Hypothesis Testing"

- we use confidence intervals to give us a range in wich we feel confident that the true population parameter will fall. many times, however, we want to test a claim that is being made about a parameter.



- In science, a hypothesis is usually a claim about how the world works. in statistics, a hypothesis involes a claim about the numerical value of a population parameter, such as the population mean or proportion

Exp: Based on ~~hypoth~~ historical data, the board of Education for one school district believes that the percentage of hi school sophomores considering dropping out of school is 10%. A high school counselor in the district claims that this percentage is too high. ~~He~~ She collects a random sample of 50 sophomores and finds that 8.5% of them have considered dropping out of school. perform a hypothesis test with $\alpha = 0.02$.

The null hypothesis (H_0): $H_0: \overset{(u)}{\underset{or}{p}} = 0.10$

The alternative hypothesis (H_A): $H_A: \overset{(u)}{\underset{or}{p}} < 0.10$

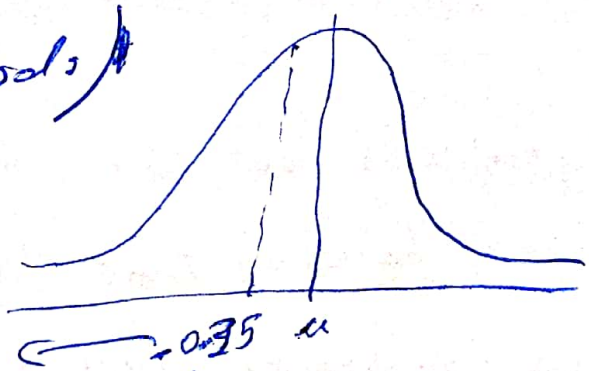
H_A = in general means $p_A \neq p_H$ [or < or >]

in this case $p < 0.10$

first steps: calculate Z or t based on $\bar{x}$ or $\hat{p}$ from sample

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \Rightarrow Z = \frac{0.085 - 0.1}{\sqrt{\frac{0.1(0.9)}{50}}} = -0.3536$$

(The first method: using Z)

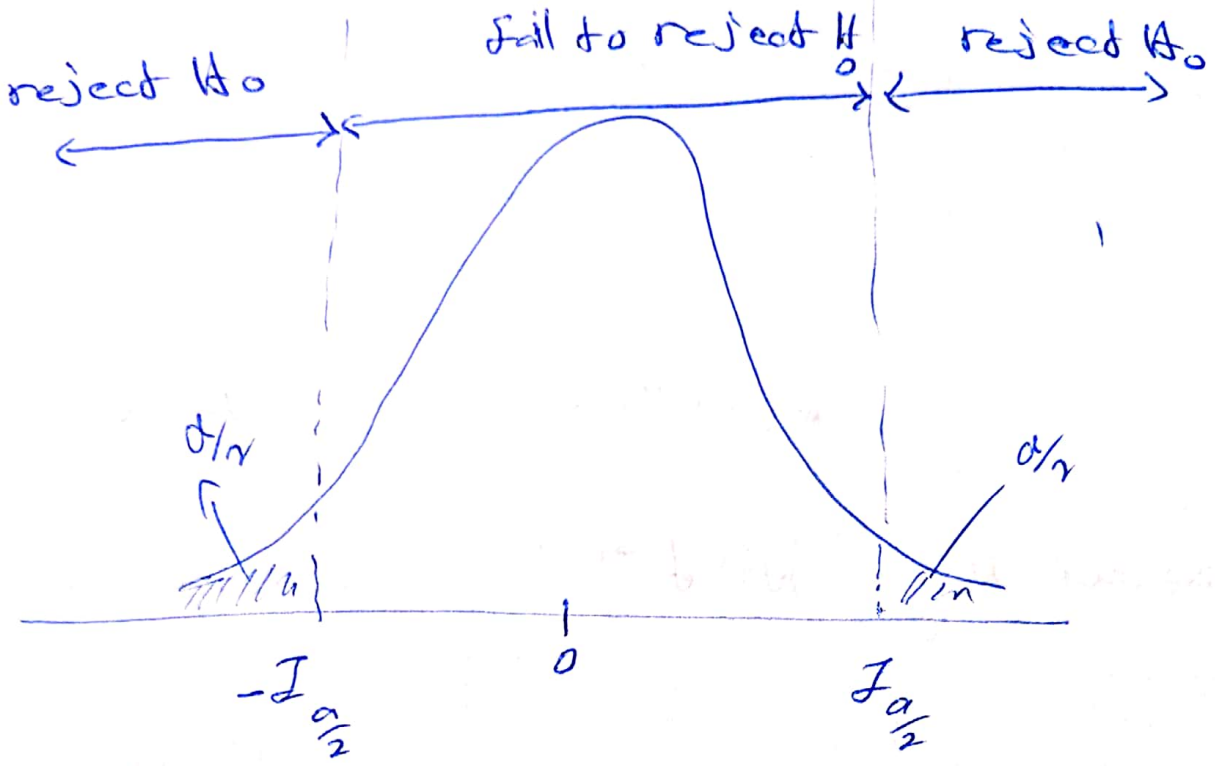


Z ← (level of significant. to the left of mean.

$\alpha = 0.02$ / based on the level of significant with the null hypothesis, then we reject the null hypothesis

if we use excel: norm.s.inv(0.02) = [-2.054]

so because our Z-score is -0.3536, so we fail to reject the null hypothesis.



The second method (using $\hat{p}$)
p_value

- if $p < \alpha$, we reject the null hypothesis
- if $p > \alpha$, we fail to reject the H_0

using $I = -0.3536$ and the fact that we are running a left-tailed test:

→ in excel = NORM.S.DIST(-0.3536)

$$P(I < -0.3536) = 0.3618$$

since $p = 0.3618$ and $\alpha = 0.02$ and $P > \alpha$, we fail to reject the null hypothesis.

"Errors in hypothesis Testing"

Decision	H_0 is true	H_0 is false
Reject H_0	Type I error	Correct decision
Fail to reject H_0	Correct decision	Type II error

a question; Technically, we can choose the level of significante "α" to be any value. So why wouldn't we choose the level of confidance to be as alaw as possible?

The probability of making a Type II error is represented by the "β" the smaller the probability of making a Type II error, the large the probability of making a type I error will be.

So you must choose an "α" small enough to control the risk of a Type I error while not making it too small, which would increase the chance of making a Type II error.

Hypothesis testing for means (σ known) (right-tailed)

Exp: The state education department is considering introducing new initiatives to boost the reading levels of fourth graders. The mean reading level of fourth graders in the state over the last 5 years was a lexile reader measure of 850 L. ($\bar{x}$) the developers of a new program claim that their techniques will raise the mean reading level of fourth graders.

To access the impact of their initiative, the developers were given permission to implement their ideas in the classrooms. At the end of the pilot study, a simple random sample of 1000 (n) fourth graders had a mean reading level of 856 L. it is assumed that the population standard deviation is 98 L. (σ) using a 0.05 (α) level of significance, should the findings of the study convince the education department of the validity of the developers' claim?

solution:

step 1) Check conditions
- random sample ✓
- σ known ✓ (σ = 98)
- n ≥ 30 or pop. dist. approx. normal. (n = 1000)

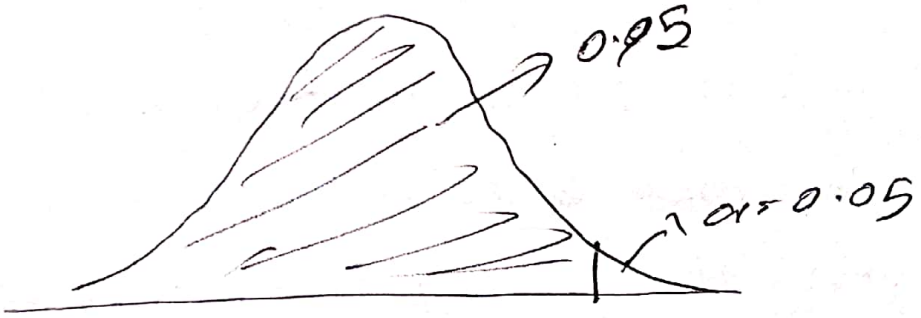
step 2) State the hypothesises
- $H_0 = \mu = 850$
- $H_a = \mu > 850$

step 3) calculate the necessary sample statistics.

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{856 - 850}{98/\sqrt{1000}} = 1.936$$

- critical value: $1 - \alpha = 1 - 0.05 = 0.95$

using excel: NORM.S.INV(0.95) = 1.645



p-value = $P(Z > 1.936) = 0.0264$

in excel: 1-NORM.S.DIST(1.936)

step 4) conclusion: with $P = 0.0264 < \alpha = 0.05$, we reject the null hypothesis. with 95% confidence, there is evidence to suggest that the mean Lexile reader measure will increase.

Hypothesis testing for means (σ-known) 2-tailed

- A recent study showed that the mean number of children for women in europa is 1.5. A global watch group claims that German women have a mean fertility rate that is different from the mean for all of europe. To test its claim, the group surveyed a simple random sample of 128 German women and found that they had a mean fertility rate of 1.4 children. The population std. is assumed to be 0.8. is there sufficient evidence to support the claim made by the global watch group at the 90% level of confidence?

$\sigma = 0.8$

$n = 128$

$\mu = 1.5$ → pop. mean

C-level = 90% = 0.90

$\bar{x} = 1.4$ → absorbed mean (sample mean)

$H_0: \mu = 1.5$

$H_1: \mu \neq 1.5$

$\alpha = 1 - 0.9 = 0.1$

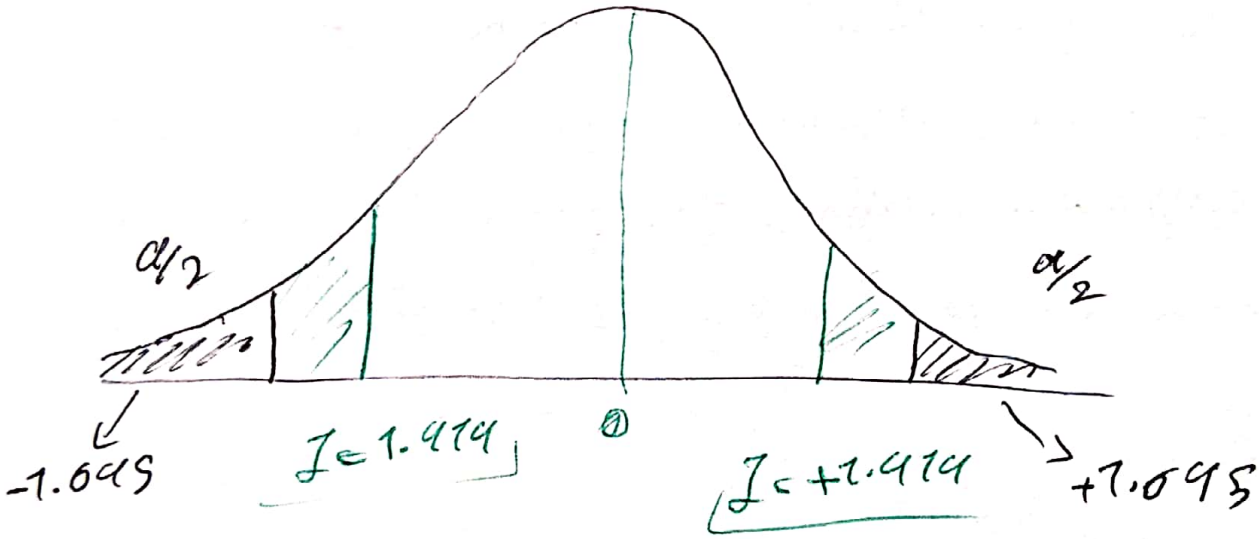
$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{1.4 - 1.5}{\frac{0.8}{\sqrt{128}}} = -1.414$$

- critical value for rejection region:

NORM.S.INV(0.95) = ±1.645

P-value: $2 \times P(Z < -1.414) = 0.1573$

2-tailed



$P = 0.1573 > \alpha = 0.1$ so we fail to reject the null hypothesis.

Hypothesis testing for population means (σ unknown)

1-tailed

Exp: Nurses in a large teaching hospital have complained for many years that they are overworked and understaffed. The consensus among the nursing staff is that the mean number of patients per nurse each shift is 8.0. The hospital administrations claim that the mean is lower than 8.0. To prove their point to the nursing staff, the administrators gather information from a simple random sample of 19 nurses' shifts. The sample mean is 7.5 patients per nurse with a standard deviation of 1.1 patients per nurse. Test the claim using $\alpha = 0.025$ and assume that the number of patients per nurse has a normal distribution.

step1) Check conditions:
- random sample ✓
- σ-unknown ✓
- $n < 30$ but pop. dist. normal ✓

$\mu = 8.0$
$n = 19$
$s = 1.1$

step2)
- $H_0 = \mu = 8$
- $H_a = \mu < 8$

$\alpha = 0.025$

Step 3) $$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{7.5 - 8}{1.7/\sqrt{19}} = -1.987$$

$df = 19 - 1 = 18$

Critical value for the rejection

excel: $T.INV(0.025, 18) = -2.101$

P-value: $P(t < -1.987) = 0.0375$

excel: T.Dist(-1.987, 18)

step 4) $P = 0.0375 > \alpha = 0.025$

Interval $\left(\bar{x} \pm t_x\left(\frac{s}{\sqrt{n}}\right)\right)$

we fail to reject the null hypothesis.
with 97.5% confidence.

"Hypothesis testing for population proportions"
(1-Tailed)

EXP: The local school board has been advertising that 65% of voters favor a tax increase to pay for a new school. A local politician believes that less than 65% of his constituents favor this tax increase. To test his belief, his staff asked a simple random sample of 50

of his constituents whether they favor the tax increase and 27 said that they would vote in favor of ~~his~~ the tax increase. if the politician wishes to be 95% confident in his conclusion, does this information support his belief?

1) - random sample ✓
   - Binomial distribution ✓
   - np: $50(0.65) = 32.5 \geq 10$ & $n(1-p) = 50(0.35) = 17.5 \geq 10$

2) H_0: $p = 0.65$

   H_a: $p < 0.65$

$p = 0.65$

$n = 50$

$x = 27$

$\hat{p} = \frac{27}{50} = 0.54$

$\alpha = 1 - C = 1 - 0.95 = 0.05$

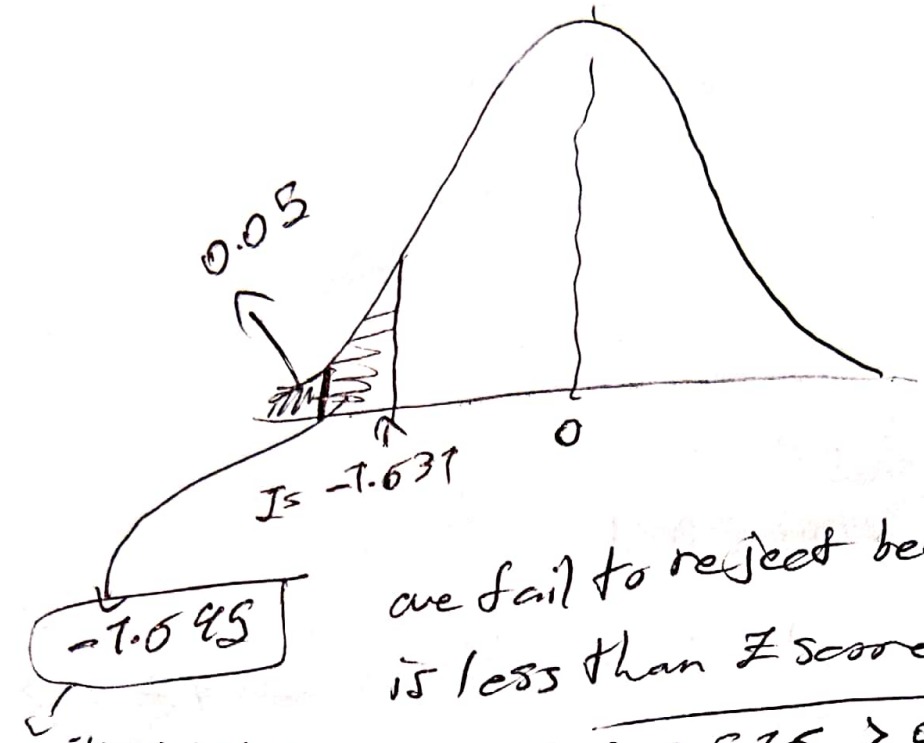
3) $Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.54 - 0.65}{\sqrt{\frac{(0.65)(0.35)}{50}}} = -1.0\overset{31}{\cancel{45}}$

critical value for rejection region

excel => NORM.S.INV(0.05) = -1.645

$P(Z < -1.031) = 0.0572$ —> NORM.S.DIST(-1.031, ?)

4)



critical value

we fail to reject because ~~Rotatue~~ is less than Z score.

and $P = 0.576 > 0.05 = \alpha$

5) finding confidence intervals:

$$\hat{p} \pm Z\sqrt{\frac{\hat{p}\hat{q}}{n}} \Rightarrow 0.94 \pm 1.645\left(\sqrt{\frac{.94(0.96)}{50}}\right)$$

$$\Rightarrow 0.94 \pm 0.116$$

$$\Rightarrow (0.824, 0.056)$$

with 95% confidence, we can claim that the true proportion of voters favor a tax increase falls between 82.4% and 05.6%. since the hypothesized value of 65% falls in our interval, tails supports failing to reject the null hypothesis.

Hypothesis testing for 2 sample means (σ known)

(1-tailed)

Test Statistic: $$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$(\mu_1 - \mu_2)$ = hypothesized difference, it is often "0"

$(\bar{x}_1 - \bar{x}_2)$ = 1st sample mean − 2nd sample mean.

Exp: A drug manufacturer claims that its new cholesterol drug, when used together with a healthy diet and exercise plan, lowers a patient's total cholesterol level by over 20 points more than simply changing a patient's diet and exercise regimen. To test the claim, a sample of 55 patients (n_1) with high cholesterol is chosen at random from a cardiologist's list of patients to take the drug in addition to changing their diet and exercise plans. Over the course of three months, this group lowers its total cholesterol level by a mean of 21.7 points ($\bar{x}_1$). The population std. for this group is 6.8 points (σ_1). Another 55 (n_2) patients with high cholesterol lowered their level by a mean of 23.1 points ($\bar{x}_2$). The population std. for this group is 5.3 points (σ_2). Test the drug manufacturer's

claim using a 0.01 level of significance.

1) - Both groups are random samples – both are randomly selected.
- Both known $\sigma_1 = 0.8$, $\sigma_2 = 5.3$
- Both $n_1, n_2 \geq 30$ $n_1 = 55$, $n_2 = 55$

2) $H_0: \mu_1 - \mu_2 = 20$

$H_2: \mu_1 - \mu_2 > 20$

$\bar{x}_1 = 44.7$ $\bar{x}_2 = 23.1$ $\alpha = 0.01$

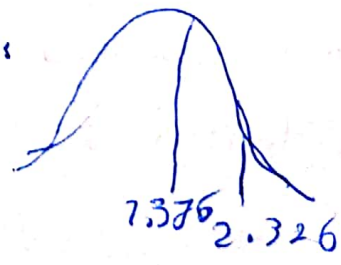
$\sigma_1 = 0.8$ $\sigma_2 = 5.3$

$n_1 = 55$ $n_2 = 55$

3) $$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(44.7 - 23.1) - (20)}{\sqrt{\frac{(0.8)^2}{55} + \frac{(5.3)^2}{55}}}$$

$$\Rightarrow Z = 7.376$$

critical value for rejection region:



CV: NORM.S.INV(0.99) = 2.326

P-Value: $P(Z > 7.376) = 0.0849$

$P > \alpha \Rightarrow (0.08 > 0.01)$

fail to reject H_0

"Determining Statistical Significance for the Pearson Correlation Coefficient."

- Is our linear relationship statistically significant?

step 1) write your hypothesis:

H_0: $P = 0$: there is no significand linear relationship.

H_1: $P \neq 0$: there is a significant linear relationship.

step 2) · Test statistics $t = \dfrac{r}{\sqrt{\dfrac{1-r^2}{n-2}}}$

P-value: $T.DIST.2T(t, df) \Rightarrow df = n-2$

step 3) Draw your conclusion:

- Reject if $P < a$, so there is a significand linear relationship
- fail to reject if $P > a$, there is not a significand linear relationship.

$$r = \frac{n\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{\sqrt{n\Sigma x_i^2 - (\Sigma x_i)^2}\sqrt{n\Sigma y_i^2 - (\Sigma y_i)^2}}$$

-> Excel

correl(x_i, y_i)

or

Pearson(x_i, y_i)

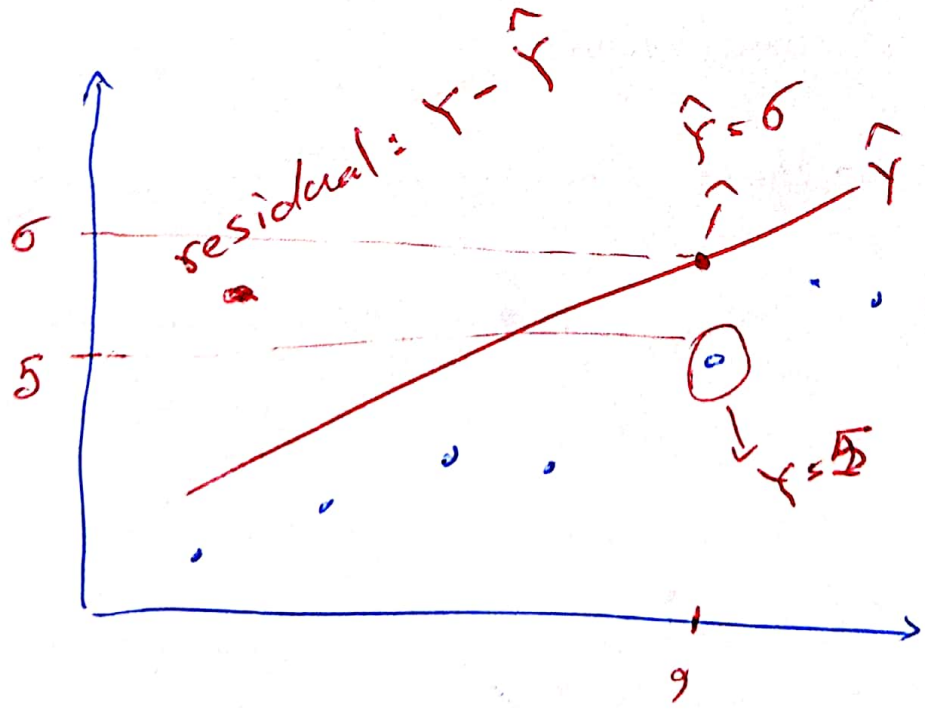
"The Least Squares Regression Line" (LSRL)

- ~~The ten~~ it is the line for which the distance between each data point and the predicted data point (located on the LSRL) is the smallest. For each value, we will round to three decimal places.

$$\hat{y} = b_0 + b_1 x$$

slope: $b_1 = \dfrac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2}$

~~Into~~ Intercept: $b_0 = \dfrac{\sum y_i}{n} - b_1 \dfrac{\sum x_i}{n}$



so: $y - \hat{y} = 5 - 6$
$= -1$

c) with $p = 0.773 > 0.05 = \alpha$ we fail to reject the null hypothesis.

point estimate (£)

- confidence interval $(\bar{x}_2 - \bar{x}_1) \pm 1.761 \sqrt{\frac{103^2}{15} + \frac{95^2}{18}}$

$\Rightarrow 44 \mp 61.23$

$(-17.23, 105.23)$

with $\alpha = 5\%$ we believe that the tax returns at smith CPA are between £17.23 less and £105.23 more on average those at Jones and company CPA. since the hypothesis value of 0 falls in our interval, this suppose failing to reject the null hypothesis.

Hypothesis testing for 2 sample (σ unknown) – equal var

$df = n_1 + n_2 - 2$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Hypothesis Testing for 2 Sample means (σ unknown)

(Paired) or dependent

- mean of pair differences: $\bar{d} = \frac{\Sigma d_i}{n}$ where each $d_i = y_i - x_i$

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} \quad \text{where } s_d = \sqrt{\frac{\Sigma(d_i - \bar{d})^2}{n-1}}$$

Hypothesis testing for 2 sample proportions

$n_1\hat{P}_1 \geq 10,\ n_1(1-\hat{P}_1) \geq 10,\ n_2\hat{P}_2 \geq 10,\ n_2(1-\hat{P}_2) \geq 10$

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (P_1 - P_2)}{\sqrt{\bar{P}(1-\bar{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \text{where } \bar{P} = \frac{x_1 + x_2}{n_1 + n_2}$$

Scatter plots and correlation

- A scatter plot is a graphical display of sets of quantitative values, typically in groups of 2. The two variables are related, as they are recorded from the same subject.

- an explanatory variable, or independent variable, is the variable whose change in value influences a change in the value of another variable.

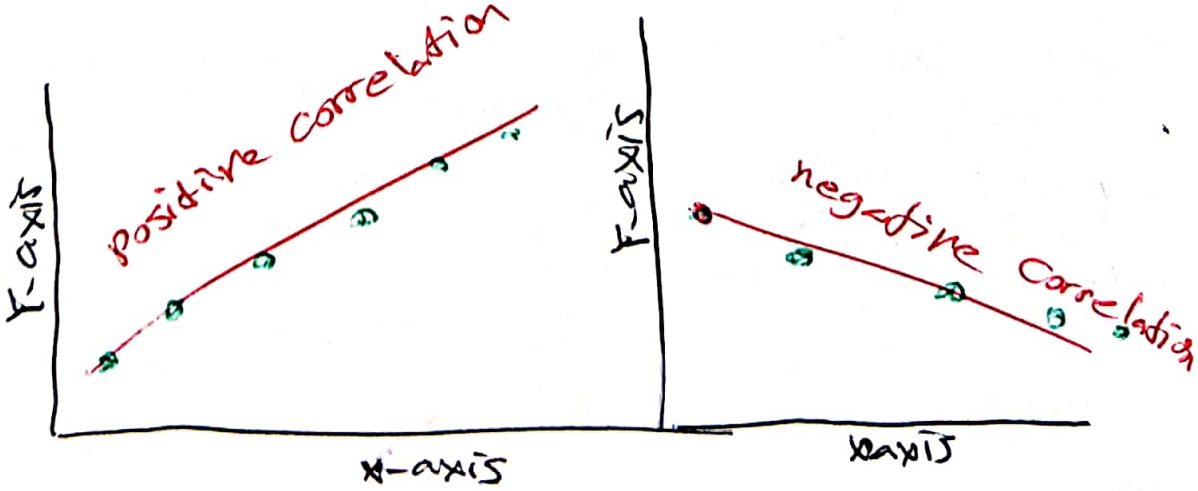
- A response variable, or the dependent variable, is the variable that has its value change as a result of a change in the value of the explanatory variable.

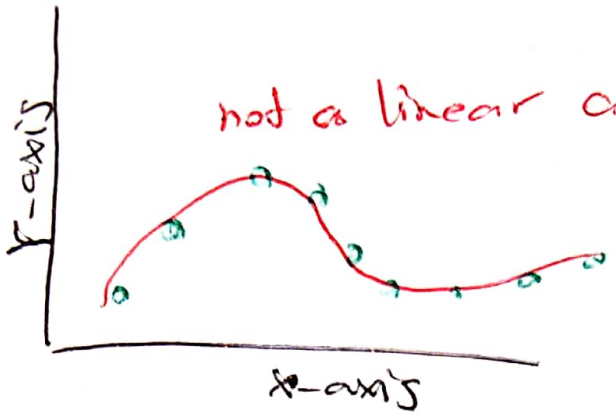
- ~~one~~ In general, in a scatter plot, as the explanatory variable increases (as are more to the right on the x-axis), what happens to the response variable (the values plotted on the y-axis) ?

- Does it increase? Then we have a positive correlation

- Does it decrease? There is a negative correlation.

- Is a relationship difficult to determine? then there is likely no linear correlation.





- The pearson correlation coefficient, ρ, is the parameter that measures the strength of a linear relationship between two quantitative variables in a population. This coefficient for a sample is denoted by r, it always takes a value between -1 and 1, inclusive $(-1 \leq r \leq 1)$

$$r = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

a) confidence interval

$$(44.7-23.1) \pm 2.326\left(\sqrt{\frac{6.8^2}{55}+\frac{5.3^2}{59}}\right)$$

$$21.6 \pm 2.298$$

$$(18.9, 24.3)$$

At the 0.01 level of significance (α), we believe the true difference in post-test cholesterol levels to be between 18.9 and 24.3 points. Since the hypothesis value of 20 falls in our interval, this supports failing to reject H_0

~~Hypothesis~~

Hypothesis Testing for 2 sample means

(σ unknown) - unequal variances

test statistic: $$t = \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

EX: Suppose that the Smith CPA firm in Chicago claims that its clients receive larger tax refounds, on average, than clients of its competitor, Jones and Company CPAs, located on the other side of Chicago. To test the claim

15 clients from the smith firm are randomly selected (n_1) and found to have a mean dark reformal of \$942 ($\bar{x}_1$) with a std. of \$103. (s_1) At Jones and Company, a random sample of 18 clients (n_2) are surveyed and found to have a mean refased \$898 ($\bar{x}_2$) with a std. of \$95. (s_2) Test the claim made by the smith firm at the 0.05 (α) level of significance. Assume that both population distributions are approx. normal.

1)
- both samples are randomly selected.
- both σ_1 and σ_2 are unknown.
- both pop. dis. are normal
- an equal variances.

2) $H_0: \mu_1 = \mu_2$

$H_A: \mu_1 > \mu_2$

$\bar{x}_1 = 942$, $s_1 = 103$, $n_1 = 15$

$\bar{x}_2 = 898$, $s_2 = 95$, $n_2 = 18$

3) $$t = \frac{(942-898)-(0)}{\sqrt{\frac{103^2}{15} + \frac{95^2}{18}}} = 1.266$$

$df_{n_1} = 15-1 = 14$ ✓

$df_{n_2} = 18-1 = 17$ ✗

14 →

critical value: $T.INV(0.95) = 1.761$

p-value: $P(t > 1.266) = 0.113$

T.DIST.RT(

- The local school board wants to evaluate the relationship between class size and performance on the state achievement test. It decides to collect data from various schools in the district, and the data from a sample of eight classes are shown in the following table. Each pair of data represents the class size and corresponding average score on the achievement test for one class.

a) Determine if there is a significant linear relationship between class size and average test score at the 0.05 level of significance.

b) if ~~there~~ the relationship is significant, find the LSRL for these data.

class size	score
15	85.3
17	86.2
18	85
20	82.7
21	81.9
24	78.8
26	79.3
29	72.1

a) ?

b) ?

"Predicting with and Interpreting Values of The LSRL"

Do not Predict if:

- The data do not fall in a linear Pattern when graphed on a scatter plot.
- The correlation coefficient is not statistically significant.
- You want to predict a value out side the range of sample data (extrapolation).
- The population is different than that from which the sample data were drawn.

"Prediction Intervals for Linear regression"

- we would like to create an interval of data in which we feel confident the True value would lie between.

(E) → margin of error

$$\hat{y} \pm \left(t_{\frac{\alpha}{2}} \times S_e\right) \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n(\sum x_i^2) - (\sum x_i)^2}}$$

$\hat{y}$ → predicted y (Plug given x into LSRL equation)

$t_{\frac{\alpha}{2}}$ → critical value (T.INV.2T(alpha, df) where df = n-2)

s_e → standard error of our sample → the root of the SSE divided by the degrees of freedom, $\sqrt{SSE/df}$

SSE → sum of squared errors (the sum of the squares of the residuals, $\Sigma(y_i - \hat{y}_i)^2$

$y - \hat{y}$ → residual (subtract the predicted value for each x (using LSRL) from the actual value for each x)

n → number of pairs data.

"multiple Regression"

- in multiple regression models, we are interested in whether multiple explanatory variables influence the response variable. for instance, we know that a student's age has a significant effect on his or her reading level. Do other factors play a role, as well?

we will let:

X_1: represent the age of the student

X_2: represent the years of experience of the child's teacher.

X_3: represent the educational level of the child's Parent(s).

The final regression equation will look like this:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

and our hypothesis will incorporate all variables:

$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$

H_a: At least one coefficient does not equal 0.