BIO606: BIO STATISTICS

→ MTWT Lectures
→ Quiz every week Thursday

by Prof. N.G. Prasad

KUNAL VERMA
MS18148

Lecture - 1
(06-01-2022)

Data in Biostatistics:

Data variables are properties of a sample population w.r.t. which they can be differentiated. Some of these variables can be "measured" and assigned a number. (continuous $\in \mathbb{R}$ or discrete $\in \mathbb{Z}$)

① Nominal variable: variables which can't be measured and must be expressed qualitatively, i.e. we have some categories but no order to the categories. Example - Eye colors. variable, gender,

variable → Eye color	Data
→ Black	40
→ Blue	0
→ Brown	8
→ Green	2

"Categorical Variable".

Questions like mean eye color are ill-defined.

(ordered)
② Ordinal variable: the order b/w the categories have some particular meaning i.e. the categories are ordered but the distance b/w the categories is unknown. Example - Ranks.

Rank	Data
1 →	100 marks
2 →	99 marks
3	
⋮	
n	

diff. categories ordered

Based on scores of rank 1 & rank 2, one can't say anything about the score of rank 3. Hence, the distance b/w data is unknown.

③ Interval scale: The categories are ordered & the difference/distance b/w the categories is constant.

Interval variable (Temp °C)

0 - - - the zero of this scale is arbitrary.
1 0°C ⇏ no temperature.
2
3
4
⋮

Another Example

Heights (cm)		Heights on new scale (cm')
170	redefine →	0 → arbitrary 0
171	− 170cm	1
172		2
⋮		3
179		⋮
		9

0cm in new scale ⇏ no height

0cm in original scale actually means no height.

We can't even say the height 2cm' is twice the height of 1cm'.

∴ The zero is arbitrary ⇒ { although diff. b/w categories is defined, we can't say anything about ratios. Ratios only make sense when 0 is defined absolutely.

For example, Kelvin scale has an absolute 0, so the ratios make sense.

④ Ratio scale: categories are ordered, differences b/w categories is constant, and the zero is an absolute zero. Therefore, defining ratios is meaningful.

Another common classification of data is continuous v/s discontinuous

Continuous - Height, weight

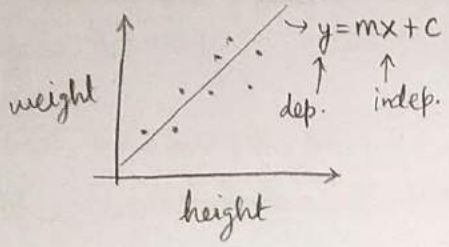
Discontinuous - Ranks, Eye color (roughly)

	Nominal	Ordinal	Interval	Ratio
Categories	✓	✓	✓	✓
Rank order		✓	✓	✓
Equal spacing			✓	✓
True zero				✓

- Categorical
  - Nominal
  - Ordinal
- Quantitative
  - Interval
  - Ratio

Independent v/s dependent variables.

For example, we can ask if there's some correlation b/w heights & weights of individuals (statistically).



If we find a relation, then we say that the weight is linearly "dependent" on the height variable.

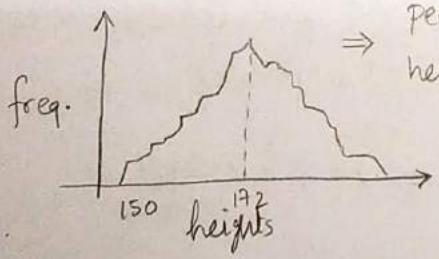
(not necessarily biologically, just according to the statistical model which explains the data)

Do not confuse statistical independence & dependence with biological causation.

Frequentist view of statistics.

Let's say we measure the heights of a bunch of students $(x_1, \ldots x_{1800})$ ← population data

To make sense of these 1800 numbers isn't easy. We need to do something with that data to summarize the knowledge. One of the ways is creating graphs.



⇒ peak height = 172
heights ∈ [150, 190]

The frequency distribution of population has an interpretation

frequentist view of statistics ⟶ i.e. freq. of population ≡ probability distribution

Probability of a randomly collected person's height to be b/w 150 & 151 cm
= Area under the freq. distrib. graph from h = 150 cm ⟶ h = 151 cm.

Lecture-2

(10-01-2022)

Heights of IISERM students $\{x_1, \cdots x_{1300}\}$

However, as remarked earlier, raw data in itself is confusing and we need to do something to it to extract out useful info.

One of the ways is to use other numbers which represent the raw data. Those numbers should summarize the core knowledge of the raw data.

One of the ways is to use - Percentiles, Quartiles, Deciles.

Percentile (P) is a number associated to a particular percentile rank.

$$\text{A percentile rank} \equiv \frac{P}{100}(N+1)$$ (N → no. of observations ("ordered"))

Ex- 157, 159, 160, 161, 164, 165, 166, 167, 168, 169, 175, 181, 181
(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13 → ranks (ordered))

30th percentile? $P = 30 \Rightarrow$ Percentile Rank (PR) $= \frac{30}{100}(13+1)$

$= 0.3 \times 14 = 4.2$ Rank

So, 30th percentile = Rank 4 data + 0.2 × (Rank 5 − Rank 4 data)

$\Rightarrow$ 30th percentile $= 161 + 0.2(164-161) = 161 + 0.6 = 161.6$

We are usually interested in calculating P_{25}, P_{50}, P_{75}.

$P_0 \longleftrightarrow P_{25}$ (160.5) $\longleftrightarrow P_{50}$ (166) $\longleftrightarrow P_{75}$ (172) $\longleftrightarrow P_{100}$

Q_1. Q_2. Q_3.

For this data, P_{50} = Rank 7 data = 166. , P_{75} = Rank 10.5 data = 169 + 3 = 172

P_{25} = Rank 3.5 data = 160 + 0.5 × 1 = 160.5

So, $P_n = \text{rank } \lfloor PR \rfloor \text{ data} + \text{frac}(PR) \cdot [\text{rank}(\lfloor PR \rfloor + 1)\text{ data} - \text{rank}(\lfloor PR \rfloor)\text{ data}]$

where $PR \equiv \frac{n}{100}(N+1)$ ← no. of observations.

Why is this N+1? What is P_0 & P_{100}?

Another representative of data are Central Tendency measures and Dispersion measures.

The central tendency & dispersion measures give an idea of where the distribution is centred and how much is it spread around its central tendency.

CENTRAL TENDENCY.

A central tendency representing the raw data must be a number which is closest to all the numbers simultaneously in the given data.

Say we have a set of nos. $\{X_1, X_2 \ldots\ldots X_n\}$. Say Y is a good representative for central tendency. One of the ways of saying this is to do a sum over absolute differences.

$$|X_1 - Y| = \Delta X_1$$
$$|X_2 - Y| = \Delta X_2$$
$$\vdots \qquad \vdots$$
$$|X_n - Y| = \Delta X_n$$

Summed diff. $= X_\Delta$

We want Y to be such that X_Δ is minimised.

It turns out, after doing the math, $Y = P_{50}$ (or Q_{12}) which is also known as the median.

The median gives the minimum sum of abs. differences!

Another way to get a central tendency measure is to find a number μ which minimizes the sum of squared differences. (instead of abs diff.)

$(X_1 - \mu)^2 = \Delta X_1^2$

$(X_2 - \mu)^2 = \Delta X_2^2$

$\vdots$

$(X_n - \mu)^2 = \Delta X_n^2$

Summed sq. diff = X_Δ^2

So, now, we want a number μ which minimizes the sum of squared differences X_Δ^2.

Consequently, again doing the math, one finds out that μ = mean/avg. (arithmetic) of the distribution. $= \sum_{i=1}^{N} \frac{X_i}{N}$

In summary,

> Mean ≡ minimizes the sum of squared differences.
>
> Median ≡ minimizes the sum of absolute differences.

Another common measure of the central tendency can be obtained by observing which number appears the most frequently (might even be a qualitative category like eye color), then this is called the mode.

> Mode ≡ the most common observation.

Can't use mean and median for categorical variable, or for ordinal variables.
Can use mean & median for interval & ratio scale variables.
Mode is generally used for categorical & ordinal variables

Lecture - 3

(11-01-2022)

Meaning of k^{th} percentile? Three different def^{n}s -

(1) Smallest value that is larger than k % of values.

(2) smallest value that is equal to or larger than k % of values.

(3) value associated with k^{th} percentile rank. → we use this def^n. Also valid for non-integer ranks.

Using def^n (1), P_{100} is ill defined because $\nexists$ any value > all the values.

∴ (1) can give slightly diff. values of percentile compared to (2) or (3).

Whatever the case, $P_{50} \stackrel{!}{=}$ Median. (value for which sum of abs. differences is minimized)

And turns out we get the correct median when we define percentile with N+1.

Given $\{x_1, x_2 \ldots\ldots x_n\}$

$$\text{Mean} = \frac{x_1 + x_2 + x_3 \ldots\ldots + x_n}{n}$$

Median = value at the centre position of ordered list $= P_{50}$

Mean is more susceptible to extreme values (outliers) than median.

1, 2, 3, 4, 5

mean = 3

median = 3

1, 2, 3, 4, 50

mean = 12 → most values of distribution aren't even close to 12.

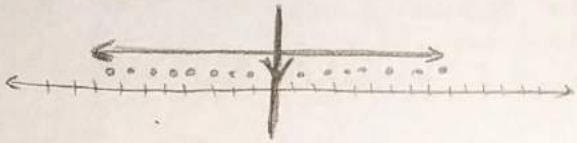
median = 3

3 is much closer to most of the distribution.

⇒ Mean is a bad representative when distribution has outliers!

Most of the times, we work with a small sample instead of the entire population. On an avg., we see that the sample mean is a better estimate of Population mean μ compared to sample median acting as an estimate of Population median. (Efficiency of an estimate)

Say we know where the central tendency of a distribution is anchored. How do we estimate its spread about the centre?



150, 151, 153, 157, 160, 165, 167, 180

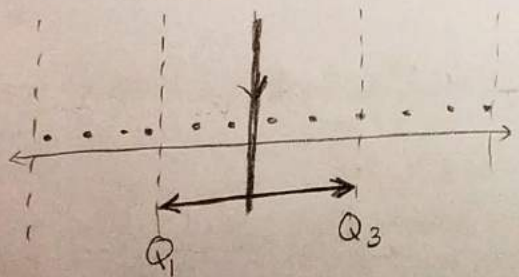
- The simplest estimate of dispersion is the Range.

$$\text{Range} \equiv \max(x_i) - \min(x_i) = 180 - 150 = 30.$$

However, the range is extremely susceptible to outliers & hence not a good representation for dispersion.

- A better alternative is the Inter-Quartile Range (IQR).

$$IQR \equiv Q_3 - Q_1 = P_{75} - P_{25}.$$



The outliers don't affect $Q_3 - Q_1$ much since they dependent on intermediate values.

- **Median absolute deviation. (MAD)**

150, 151, 153, 157, | 160, 165, 167, 180

158.5
Median ≡ M

To calculate median absolute deviation, we calculate $|x_i - M| \; \forall \; x_i$, and then find the median of this ordered list.

$|150 - 158.5| = 8.5 = |D_1|$
$|151 - 158.5| = 7.5 = |D_2|$
$\vdots$
$5.5 = |D_3|$
$1.5 = |D_4|$
$1.5 = |D_5|$
$6.5 = |D_6|$
$8.5 = |D_7|$
$|180 - 158.5| = 11.5 = |D_8|$

$\Rightarrow$ Median $(\{1.5, 1.5, 5.5, 6.5 | 7.5, 8.5, 8.5, 11.5\})$

$= \underline{\underline{7}}$

$\therefore$ Median absolute dev. = Median of differences $\{|x_i - M| : \forall x_i\}$

- **Mean squared deviation.**

150, 151, 153, 157, 160, 165, 167, 180

$\mu \approx 160.4$

(minimizes the $\sum$(squared deviation))

To calculate mean sq. deviation, calculate $(x_i - \mu)^2 \; \forall \; x_i$ and then calculate the mean of this list.

$(150 - 160.4)^2 = D_1^2$
$\vdots$
$(180 - 160.4)^2 = D_8^2$

So, mean squared dev. (MSD) $= \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n} = \sum_{i=1}^{n} \frac{D_i^2}{n} =$ "Variance" (σ^2)

However, MSD is in squared units of the data, so we generally take a $\sqrt{\;}$ to interpret the dispersion.

$\sqrt{\text{Variance}} = \sigma \equiv$ Standard deviation $= \sqrt{\sum_{i=1}^{n} \frac{D_i^2}{n}}$

∴ A distribution can be described by the mean & the std. dev. (or variance) or via median and M.A.D. (median absolute dev.) as the central tendency and dispersion measures respectively.

Imp. One important diff. b/w variance & std. dev. is that variances can simply be added whereas std. dev. can't be. (in what respect though?)

Population & sample variance.

$$\text{Population variance} \equiv \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n}$$

$x_i \in$ Population data
μ = population mean

However since we don't really have access to population data, we estimate the variance using a small sample.

$$\text{Sample variance} \equiv \sum_{j=1}^{m} \frac{(x_j - \bar{X})^2}{m-1}$$

$\bar{X}$ = sample mean
$x_j \in$ sample data

$\bar{X}$ is an estimate of μ which we obtain from a sample $\subset$ population.

The reason why we have a $(m-1)$ in the denominator instead of m is that sample variance, if defined with just m, will be an underestimate of population variance. (when $m \ll n$)

This is because by def^n, $\bar{X}$ minimizes the sum $\sum_{i=1}^{m} (x_i - \bar{X})^2$, and in general, $\bar{X} \neq \mu \Rightarrow \sum_{i=1}^{m} \frac{(x_i - \bar{X})^2}{m} \leq \sum_{i=1}^{m} \frac{(x_i - \mu)^2}{m}$. $\Rightarrow$ sample variance is always an underestimate so we use the $m-1$ correction factor. CHECK THIS

Lecture-4
(12-01-2022)

Idea: A distribution can be summarized by specifying the central tendency and a measure of dispersion.

Given some population $\{x_1, x_2 \dots x_n\}$, we can calculate μ & σ^2 (or Var[x])

$$\sigma^2 = \text{Var}[x] = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n}$$

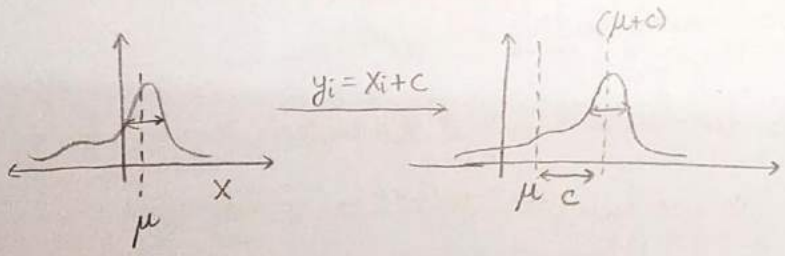
Let's define a new variable $y_i \equiv x_i + c$. Calculating Var[y]

$$\text{Var}[y] = \text{Var}[x+c] = \sum_{i=1}^{n} \frac{(x_i + c - \tilde{\mu})^2}{n} \quad \text{where } \tilde{\mu} = \sum_{i=1}^{n} \frac{(x_i + c)}{n} = c + \mu$$

$$\Rightarrow (x_i + c - \tilde{\mu})^2 = (x_i + c - \mu - c)^2 = (x_i - \mu)^2$$

$$\therefore \text{Var}[x+c] = \sum_{i=1}^{n} \frac{(x_i + c - \tilde{\mu})^2}{n} = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n} = \text{Var}[x]$$

$$\Rightarrow \underline{\text{Var}[x+c] = \text{Var}[x]}, \text{ but } \underline{\mu[x+c] = \mu[x] + c}$$



If we multiply by a scalar, i.e. $y_i \equiv c x_i$

$$\tilde{\mu} = \mu[cx] = c \sum_{i=1}^{n} \frac{x_i}{n} = c\mu \Rightarrow \text{Var}[cx] = \sum_{i=1}^{n} \frac{(cx_i - c\mu)^2}{n} = c^2 \text{Var}[x]$$

$$\Rightarrow \underline{\text{Var}[cx] = c^2 \text{Var}[x]}, \text{ and } \underline{\mu[cx] = c\mu[x]}$$

So, • $\text{Var}[c+x] = \text{Var}[x]$, $\mu[x+c] = c + \mu[x]$

• $\text{Var}[cx] = c^2 \text{Var}[x]$, $\mu[cx] = c\mu[x]$

$\Rightarrow$ $\boxed{\begin{aligned} \text{Var}[cx+b] &= \text{Var}[cx] = c^2 \text{Var}[x] \\ \mu[cx+b] &= b + \mu[cx] = b + c\,\mu[x] \end{aligned}}$

Now, let's see what happens if we add variances of 2 variables.

$$\text{Var}[x \pm y] = \sum_{i=1}^{n} \frac{(x_i \pm y_i - \mu_x \mp \mu_y)^2}{n} = \sum_{i=1}^{n} \frac{[(x_i - \mu_x) \pm (y_i - \mu_y)]^2}{n}$$

$$= \sum_{i=1}^{n} \left(\frac{(x_i - \mu_x)^2}{n} \right) + \left(\frac{(y_i - \mu_y)^2}{n} \right) \pm 2 \left(\frac{(x_i - \mu_x)(y_i - \mu_y)}{n} \right)$$

$$= \text{Var}[x] + \text{Var}[y] \pm 2\,\text{Cov}[X,Y]$$

where $\text{Cov}[X,Y] \equiv \sum_{i=1}^{n} \frac{(x_i - \mu_x)(y_i - \mu_y)}{n}$

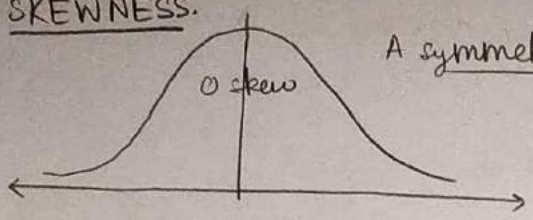
If X & Y are independent random variables, $\text{Cov}[X,Y] = 0$.

and $\text{Var}[X \pm Y] = \text{Var}[X] + \text{Var}[Y]$

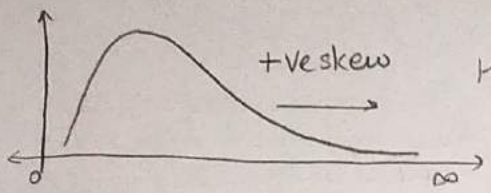
Quite often, describing the distribution via just the central tendency & the dispersion isn't enough. We use two more quantities - { SKEWNESS and KURTOSIS

Describe the tails of the distribution.

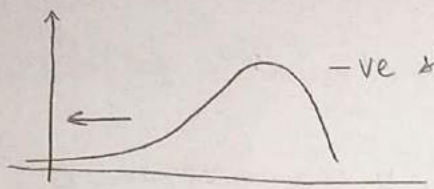
SKEWNESS.



A symmetric distribution (about mean) has zero skewness.

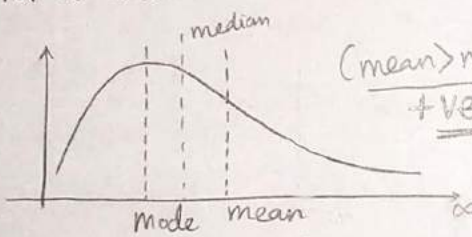


However, if one of the tails extends longer than the other, we have skewness. The graph on the left has +ve skewness



For a symmetric distribution, mean = median = mode.

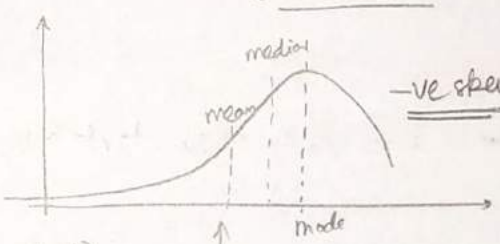
For a skewed distribution,



(mean > median)
+ve skew

(large values pull mean towards it)

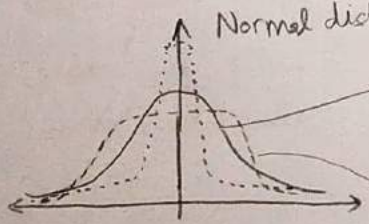
(median > mean)



-ve skew

why here?
(read wiki page on skewness)

KURTOSIS (measure of tail thickness)



Kurtosis (normal dist.) = 3

Kurtosis (⎍) < 3 shorter/thinner tails.

Kurtosis (∧) > 3 longer/thicker tails.

	Excess kurtosis.
$K=3 \Rightarrow$ meso kurtic	0
$k>3 \Rightarrow$ leptokurtic	+ve
$k<3 \Rightarrow$ platykurtic	-ve

Therefore, we use the following to describe a distribution -

(1) mean
(2) variance
(3) skewness
(4) kurtosis

} moments of a distribution. (Read about moments.)

The higher moments are important because there might exist data sets which have the exact same mean & variance, however they still might be drastically different distributions.

(Look up - Anscombe's quartet)

Given a population, the population mean μ, the population variance σ^2 are called the parameters of the population. (and constant)

If we instead collect a sample, and calculate sample mean $\bar{x}$ and sample variance s^2 are called a statistic. They are variable in the sense that they may depending upon the sample set we choose.

Lecture - 5

(13-01-2022)

BOX - WHISKER PLOT

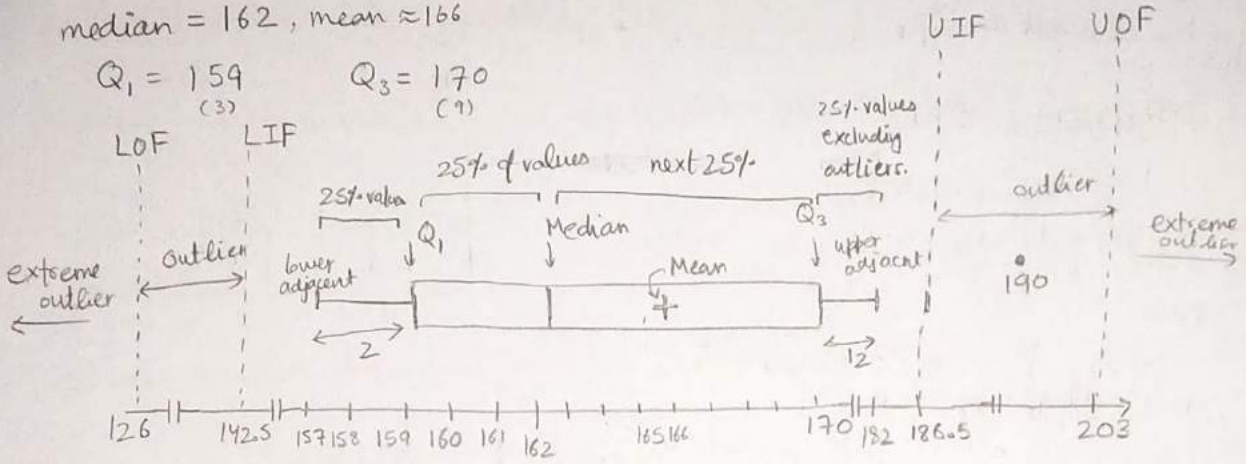
157, 158, 159, 159, 159, 162, 165, 167, 170, 182, 190

1 2 3 4 5 6 7 8 9 10 11

To make this plot, we need the median, Q_1 & Q_3

median = 162, mean ≈ 166

Q_1 = 159 (3) Q_3 = 170 (9)



Q_1 → lower hinge

Q_3 → upper hinge

$Q_3 - Q_1$ → Hspread (I.Q.R.)

Step ≡ Hspread × 1.5

Hspread = 11

Step = 11 × 1.5 = 16.5

UIF = 186.5 = 170 + 16.5

UOF = 203 = 170 + 33

LIF = 142.5 = 159 − 16.5

LOF = 126 = 159 − 33

Upper inner fence ≡ Upper hinge + step

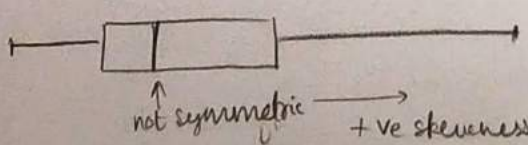
Upper outer fence ≡ Upper hinge + 2 · step.

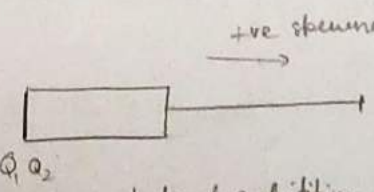
Lower inner fence ≡ lower hinge − step

Lower outer fence ≡ lower hinge − 2 · step.

Upper adjacent = value in the list ≤ UIF but closest

Lower adjacent = value in the list ≥ LIF but closest.





Lots of repititive values in beginning

Interpreting the plot:

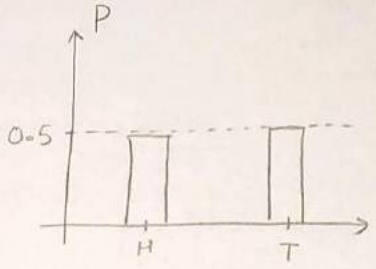
- There are two measures of skewness.
  - if Q_1 is closer to Q_2 than Q_3 is, then distrib^n is positively skewed. 
    if Q_1 is further to Q_2 than Q_3, then negatively skewed 
  - If mean > median ⇒ +ve skew. else -ve.

THESE TWO MEASURES NEED NOT BE CONSISTENT AND ARE DIFFERENTLY DEFINED.

PROBABILITY.

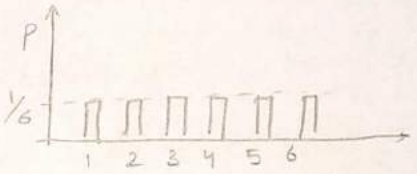
For the classical example of coin toss, the possible outcomes — {H, T}

$$P(H) = \frac{1}{2} = P(T)$$



For a die, {1, 2, 3, 4, 5, 6} is the sample space

$$p(i) = \frac{1}{6} \quad \forall i \in \text{sample space}$$



Let's now discuss the probabilities of occurances of more than one events.
We'll start with the simpler case of "independent" events.

Two independent events –

Probability of getting H on coin AND 6 on die = ?

Since H (coin) AND 6 (die) are independent events.

$$\Rightarrow p(H \text{ AND } 6) = p(H) \cdot p(6) = \frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}$$

Similarly $p(1 \text{ AND } 1) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$

$p(\text{even no. on die 1 AND odd no. on die 2}) = \frac{3}{6} \cdot \frac{3}{6} = \frac{1}{4}$

Lecture - 6

(17-01-2022)

$p(\text{event 1 AND event 2}) = p(\text{event 1}) \cdot p(\text{event 2})$ if 1 & 2 are independent.

AND → multiplicative.

$p(\text{event 1 OR event 2}) = p(\text{event 1}) + p(\text{event 2}) - p(\text{both})$

example- $p(H \text{ or even}) = p(H) + p(\text{even}) - p(\text{both}) = \frac{1}{2} + \frac{1}{2} - \frac{1}{2} \cdot \frac{1}{2} = \frac{3}{4}$

$p(\text{both}) = p(H) \cdot p(\text{even})$ since H & even are independent events.

Let's now consider dependent events.

We draw an A of spades. $p(A_{\text{spades}}) = \frac{1}{52}$

Now if we draw an A of hearts without putting A spades back, $p(A_{\heartsuit}) = \frac{1}{51}$

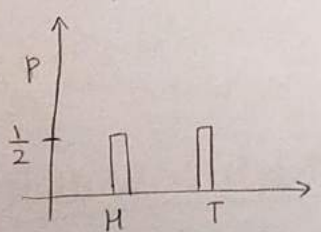
The probability of event 2 is dependent on event 1. This is known as conditional probability. $P(B|A)$ = prob. of B given A has happened.

$\sum_{i=1}^{N} i = \frac{N(N+1)}{2}$

$\sum_{i=1}^{N} i^2 = \frac{N(N+1)(2N+1)}{6}$

Uniform distributions

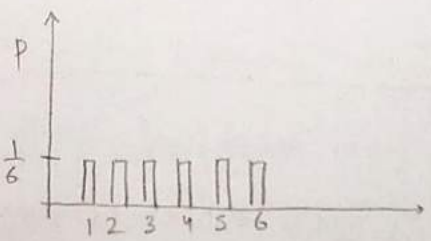
Toss of a fair coin



$\mu = \frac{a+b}{2}$

a, b are the endpts.

Throw of a die



$\sigma^2 = \frac{n^2-1}{12}$ for the throw of die

$\mu = \frac{n+1}{2}$

Technically, ∄ a mean for something like a coin toss, but we can assign #s to coin heads & tails. for ex- assign 0 to H & 1 to T.

$$\mu = \frac{0+1}{2} = 0.5 \qquad \sigma^2 = \frac{0.5^2 + 0.5^2}{2} = 0.25$$

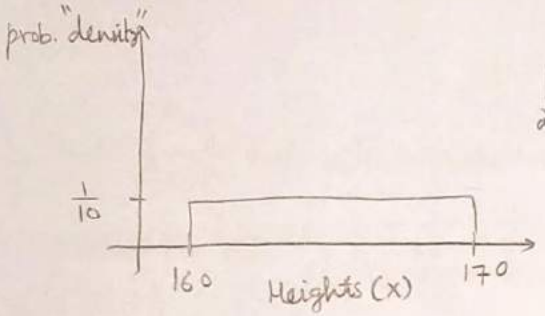
Lecture - 7

(18-01-2022)

For a uniform discrete pdf, $p(X = \text{any outcome}) = \frac{1}{n}$ where n = # of possible outcomes

$$\mu = \frac{a+b}{2} \quad \text{and} \quad \sigma^2 = \frac{n^2-1}{12}$$

Similarly, we can also have a uniform continuous probability density f^n.



$$\int_a^b dx\, p(x) \stackrel{!}{=} 1$$

$$\mu = \langle x \rangle = \int_a^b dx\; x p(x) = \left(\frac{b+a}{2}\right)$$

$$\sigma^2 = \langle x^2 \rangle - \langle x \rangle^2 = \int_a^b dx\; x^2 p(x) - \left[\int_a^b dx\; x p(x)\right]^2 = \frac{(b-a)^2}{12}$$

Bernoulli Trials / Binomial distribution.

Given a fixed no. of trials with a success & failure outcome which has a fixed prob of success / failure (p/q)

$$p(X = r \text{ successes}) = {}^nC_r\, p^r q^{n-r} \qquad {}^nC_r \equiv \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

n = no. of trials
r = no. of successes
p = prob. of success
q = prob. of failure = $1-p$

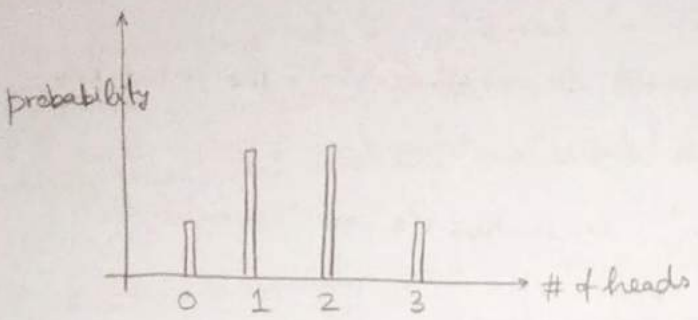
For $n=3$ trials,

$$P(X=0 \text{ heads}) = \binom{3}{0} 0.5^0\, 0.5^3$$

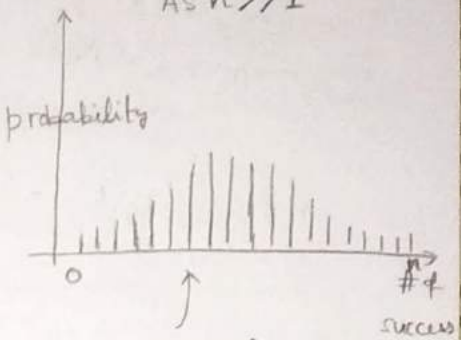
$$P(X=1 \text{ heads}) = \binom{3}{1} 0.5^1\, 0.5^2$$

$$P(X=2 \text{ heads}) = \binom{3}{2} 0.5^2\, 0.5^1$$

$$P(X=3 \text{ heads}) = \binom{3}{3} 0.5^3\, 0.5^0$$



As $n \gg 1$



$$\mu = np \quad , \quad \sigma^2 = npq$$ for Binomial distribⁿ

Lecture-8
(19-01-2022)

If $p \neq q$, the binomial distribution isn't symmetric about the mean anymore. There is a non-zero skewness for $p \neq q$.

In general, binomial distribution is just a special case of multinomial distribution.

Multinomial distribution – n trials (independent)
Each trial has k different outcomes
$k=2$ case is called binomial distribution.

$p_i \rightarrow$ probability of each outcome $\quad i = 1, 2, 3 \ldots k.$

Multinomial distribution gives the probability of a particular combination of outcomes of the n trials.-

Example.- Say for a certain cricket match of IND v/s SA, we have 4 outcomes

- Win $\rightarrow p_1 = 0.3$
- Loss $\rightarrow p_2 = 0.1$
- Draw $\rightarrow p_3 = 0.2$
- Tie $\rightarrow p_4 = 0.4$

5 matches

Now, if we want to answer what's the probability that IND will win 2 matches, lose 1 match, draw 0 & tie 2 matches? We do this via multinomial dists.

So,
$$P(r_1 \to 1, r_2 \to 2, \ldots, r_k \to k) = \frac{n!}{r_1!\, r_2! \ldots r_k!}\, p_1^{r_1} p_2^{r_2} \ldots p_k^{r_k}$$

where $r_1, r_2 \ldots r_k$ are the # of times outcomes $1, 2, \ldots k$ occur respectively.

$$\sum_{i=1}^{k} r_k \overset{!}{=} n$$

$p_1, p_2 \ldots p_k$ are the probabilities of each outcome $1, 2, \ldots k$ respectively

$$\sum_{i=1}^{k} p_i = 1$$

In the context of above problem, $r_1 = 2, r_2 = 1, r_3 = 0, r_4 = 2$

$$r_1 + r_2 + r_3 + r_4 = 5 = n \quad \Rightarrow \quad P = \frac{5!}{2!\,1!\,0!\,2!} (0.3)^2 (0.1)^1 (0.2)^0 (0.4)^2$$

Poisson Distribution.

Possibility of the # of events when they are rare.

Say we have k number of events happening in an interval (space/time), given that they occur at a fixed constant rate on average.

i.e. the average rate of an event is a fixed probability for a given interval, & scales with the size of the interval. These events are independent. And more than one event can't happen at the same instance.

$\Rightarrow$ Poisson distribution gives the probability of such X number of events happening in a given interval-

$$p(X) = \frac{e^{-\mu}\mu^X}{X!}$$

μ = mean rate of the event happening

X = # of events of interest.

For example, if we encounter 2 delivery guys at the gate every 10 mins, then what's the prob. that we see 4 zomato guys in that 10 min interval?

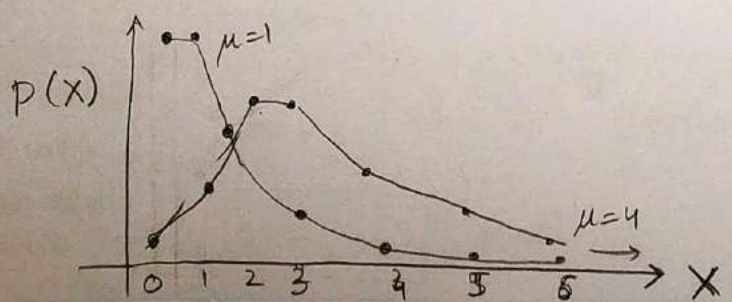
$$\mu = 2, \quad X = 4 \Rightarrow p(X=4) = \frac{e^{-2}2^4}{4!}$$

$\mu = 2$ ↑ avg rate 2 guys/10 mins

$X = 4$ ↑ # of guys of interest

For a Poisson distribution, $\boxed{\mu = \sigma^2}$

The skewness ↓ as μ ↑.



MEAN OF A BINOMIAL DISTRIBUTION.

Say, the total no. of trials is N. The probability of event E_1, which we label as success is p_1 & probability of E_2 (failure) $= p_2 = 1 - p_1$.

The # of successes (E_1) in a series of trials is given the notation n_1 & # of failures (E_2) is n_2.

$$\sum_{i=1}^{2} p_i = 1 \quad \text{and} \quad \sum_{i=1}^{2} n_i = N$$

Now here we have 2 random variables, n_1 & n_2, which represent the # of successes (E_1) & # of failures (E_2) respectively.

$$\Rightarrow \text{The average \# of successes} = \langle X_1 \rangle = \langle n_1 \rangle = \sum_{n_1=0}^{N} n_1 \cdot \mathbb{P}(X_1 = n_1, X_2 = n_2)$$

$$\langle n_1 \rangle = \sum_{n_1=0}^{N} n_1 \, {}^{N}C_{n_1} \, p_1^{n_1} p_2^{n_2} = \sum_{n_1=0}^{N} n_1 \frac{N!}{n_1!\,(N-n_1)!} p_1^{n_1} (1-p_1)^{N-n_1}$$

Relabelling $n_1 = n$ & $p_1 = p$ for now,

$$\langle n \rangle = \sum_{n=1}^{N} \frac{N!}{(n-1)!\,(N-n)!} p^n (1-p)^{N-n} = \sum_{n=1}^{N} \frac{N!}{(n-1)!\,(N-n)!} p^n (1-p)^{N-n}$$

(the $n=0$ term $=0$)

$$= \sum_{n=1}^{N} Np \frac{(N-1)!}{(n-1)!\,(N-n)!} p^{n-1} (1-p)^{N-n} \qquad \left| \begin{array}{l} n-1 \equiv \tilde{n} \\ N-1 \equiv \tilde{N} \end{array} \right.$$

$$= Np \sum_{\tilde{n}=0}^{\tilde{N}} \frac{\tilde{N}!}{\tilde{n}!\,(\tilde{N}-\tilde{n})!} p^{\tilde{n}} (1-p)^{\tilde{N}-\tilde{n}} = Np$$

$$\because (p + (1-p))^{\tilde{N}} = 1^{\tilde{N}} = 1$$

$\therefore$ mean no. of successes (E_1): $\boxed{\langle n_1 \rangle = N p_1}$

Similarly, mean # of failures (E_2) though we often don't talk about this: $\boxed{\langle n_2 \rangle = N p_2}$

For a multinomial distribution, i.e. k outcomes (instead of 2 as in binomial) with N trials, the mean number of times outcome E_i occcurs $(i=1,2,3\cdots k)$ is given by -

$$\langle n_i \rangle = N p_i$$

where n_i is the # of times E_i occurs and p_i is the probability of each outcome

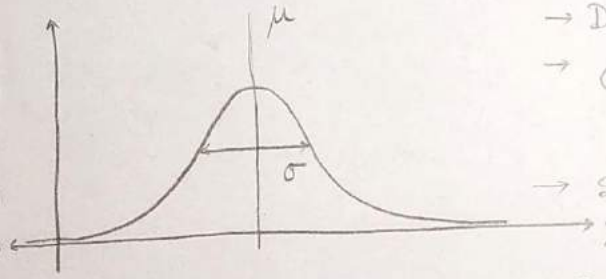
$$\text{Var}(n_i) = n p_i (1-p_i)$$

Lecture - 9

(20-01-2022)

Gaussian / Normal distribution / Bell curve.

For np & $nq \geq 5$, Binomial ⟶ Normal.



→ Described by μ, σ.

→ CONTINUOUS probability distribution
$X \in (-\infty, \infty)$

→ Symmetric about $\mu \Rightarrow$ Skewness = 0
Excess Kurtosis = 0

→ Centre is heavy, tails are light.

→ 68-95-99.7 rule ($\mu \pm \sigma \approx 68\%$ of area, $\mu \pm 2\sigma \approx 95\%$ of area, $\mu \pm 3\sigma \approx 99.7\%$ of area)
3σ rule

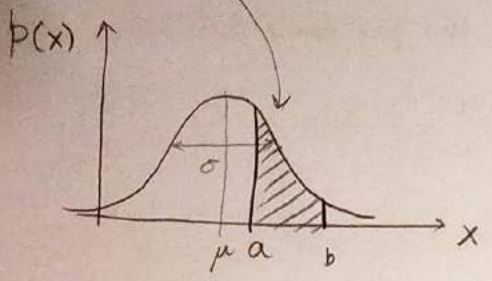
$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$\int_{-\infty}^{\infty} dx\ p(x) = 1$$

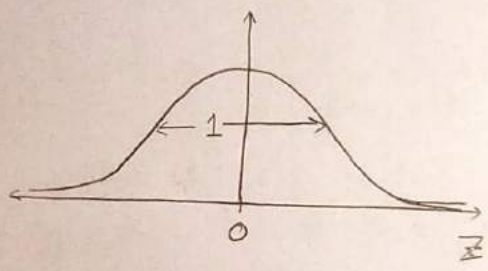
$$\langle x \rangle = \int_{-\infty}^{\infty} dx\ x\, p(x)$$

Probability that a normally distributed random variable (say, height) has values b/w a & b -

$$P(X=a \rightarrow X=b) = \int_a^b dx \; \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}}$$



The standard normal distribution is defined as $N(0,1)$ i.e. $\mu = 0$. $\sigma = 1$.
The random variable is called a z-variate.



$$p_{standard}(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

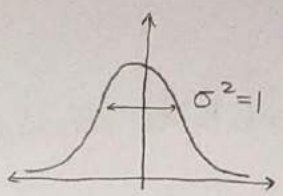
Any normal distribution can essentially be converted into a standard normal distr.

$$X \sim N(\mu, \sigma^2) \longrightarrow Z \sim N(0,1)$$

This is achieved by $\left(\frac{x_i - \mu}{\sigma}\right) = z_i \quad \forall i$

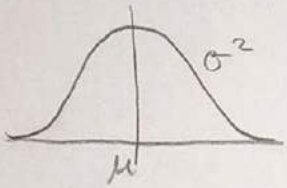
Lecture-10
(31-1-22)

Standard Normal distribution $\equiv N(\mu=0, \sigma^2=1)$



standard normal distribution

Given any normal distribution with $\mu \neq 0$ & $\sigma^2 \neq 1$, we can convert it into a standard normal distribution



Say $X \sim N(\mu, \sigma^2)$

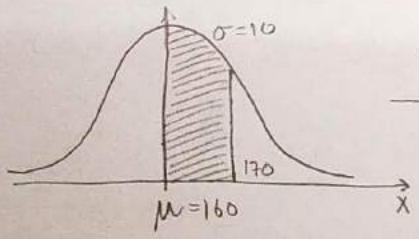
Then $\frac{X-\mu}{\sigma}$ gives a new variable z

$z \equiv \frac{X-\mu}{\sigma}$ s.t. now $z \sim N(0, 1)$.

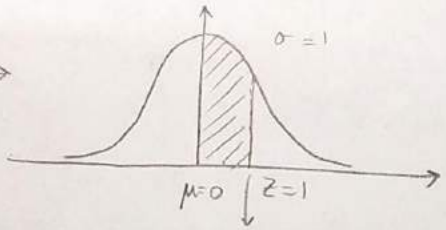
This can be useful for calculating probabilities given z-variate tables.

Say heights are distributed as $N(160, 100)$ i.e. $\mu = 160$ & $\sigma = 10$

Then what's the probability of finding a height b/w 160 & 170cm.

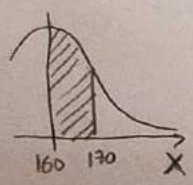


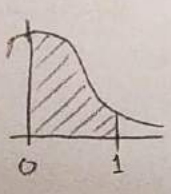
$z \equiv \frac{X-\mu}{\sigma}$



$z = \frac{170-\mu}{\sigma} = \frac{170-160}{10} = \frac{10}{10}$

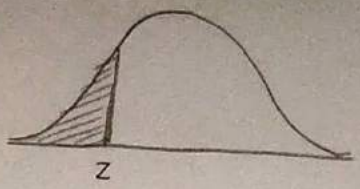
$= 1$

$\therefore$ 

$\equiv$ 

= Find from z-table

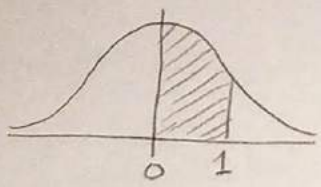
Negative z-score table calculates



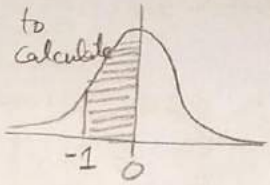
Positive z-score table calculates

But since gaussian is symmetric,

to calculate



=



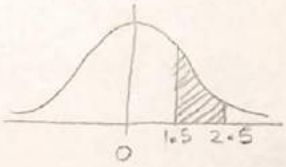
Given a negative z-score table, $P(160cm - 170cm) = z(0) - z(-1)$

Given a positive z-score table, $P(160cm - 170cm) = z(1) - z(0)$

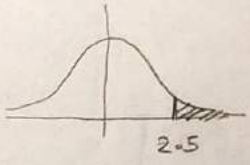
Similarly, let's now find $P(175cm - 185cm) = ?$

$\mu = 160$, $\sigma = 10$ $\Rightarrow$ $\frac{175-160}{10} = 1.5$ $\quad$ $\frac{185-160}{10} = 2.5$

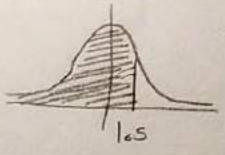
$$\therefore P(175cm - 185cm) = \int_{1.5}^{2.5} dz\, N_z(0,1) =$$



$$P(\text{atleast } 185) = \int_{185}^{\infty} dx\, N_x(160, 10^2) = \int_{2.5}^{\infty} dz\, N_z(0,1) =$$



$$P(\text{atmost } 175) = \int_{-\infty}^{175} dx\, N_x(160, 10^2) = \int_{-\infty}^{1.5} dz\, N_z(0,1) =$$



Inferential Statistics

Distribution of heights in IISERM students? Possible to take data of all the students, but samples may give some estimates.

We'll have to estimate a measure of → Central tendency (sample mean)
→ Dispersion (sample variance)

$\bar{X}$ would be an estimate of μ.

S^2 would be an estimate of σ^2.

However, they might or might not be close to μ & σ^2.
(depend on the sample)

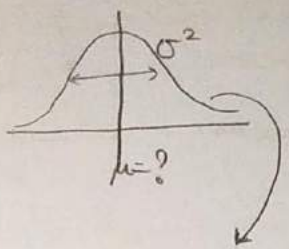
Lecture - 11

(01/02/2022)

Say we have a population whose data X is distributed normally

$$X \sim N(\mu, \sigma^2)$$

$\mu \rightarrow$ unknown, $\sigma^2 \rightarrow$ known



From a sample of population (size n), we calculate the sample mean $\bar{X}$

$$\bar{X} \equiv \sum_{i=1}^{n} \frac{X_i}{n} \quad \text{(an estimate of } \mu \text{ pop}^n \text{ mean)}$$

The quantity which we want to estimate (here μ) is referred to as the estimand.

The rule for calculating the estimand (here $\bar{X} \equiv \sum_i \frac{X_i}{n}$) is referred to as the estimator.

The numerical value of the $\bar{X}$ is known as the estimate.

Properties of the estimator –

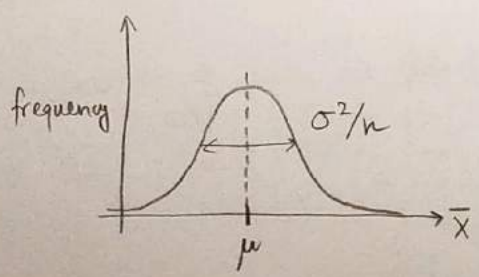
→ should be unbiased. (the mean of all $\bar{X}$ estimates $\longrightarrow \mu$) and accurate.

→ should be precise (there shouldn't be too much variation in $\bar{X}$ estimates).

→ as $n \rightarrow$ pop^n size, then $\bar{X} \longrightarrow \mu$.

But how are the estimates of $\bar{X}$ distributed?

If we calculate $\bar{X}$ infinitely many times & plot it, then $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$



where n = sample size.

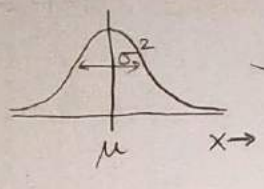
Lecture-12

(02-02-2022)

Sampling distribution of mean.

$\mu \rightarrow$ unknown

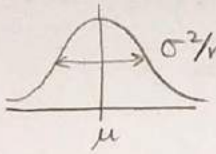
$\sigma^2 \rightarrow$ known



$n \longrightarrow$ take samples of size n & estimate $\bar{x}$

$\bar{X}_1, \bar{X}_2 \cdots\cdots \bar{X}_m$

The distribution of $\bar{X}$'s looks like



$$X \sim N(\mu, \sigma^2) \Longrightarrow \bar{X} \sim N(\mu, \sigma^2/n)$$

$\frac{\sigma}{\sqrt{n}} \equiv$ standard error of the mean $= S_{\bar{x}}$

We can also normalize the original normal distribution of the data $X \sim N(\mu, \sigma^2)$ to make it a standard normal distribution

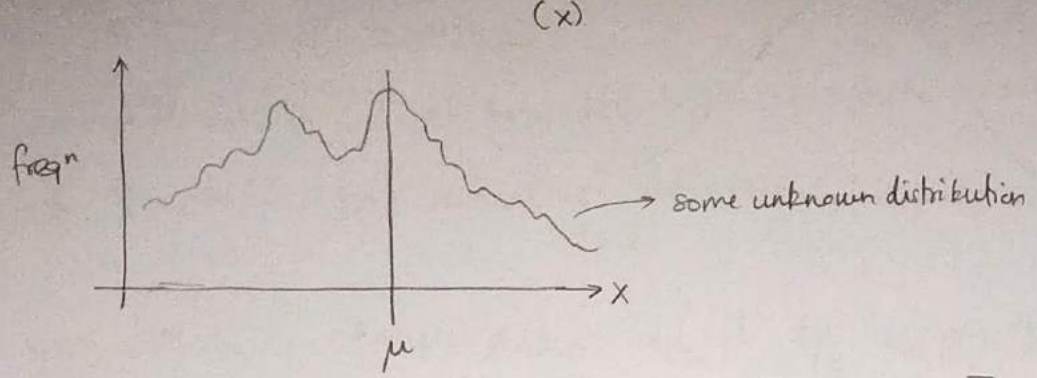
$$Z \equiv \frac{X-\mu}{\sigma} \Rightarrow Z \sim N(0,1)$$

Similarly, we can also renormalize the sample distribution of means & convert it into a standard normal distribution, but this will require the change of variables

$$z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \Rightarrow Z \sim N(0,1)$$

A standard normal distribution $\int_{-\infty}^{\infty} dz\ N_z(0,1) = \int_{-\infty}^{\infty} dz\ \frac{e^{-x^2/2}}{\sqrt{2\pi}} = 1$

Let's now take the case when Heights ~ unknown distribution.



Now take samples of size n (n ~ Large) - $\bar{X}_1, \bar{X}_2, \bar{X}_3, \ldots \bar{X}_m \ldots$

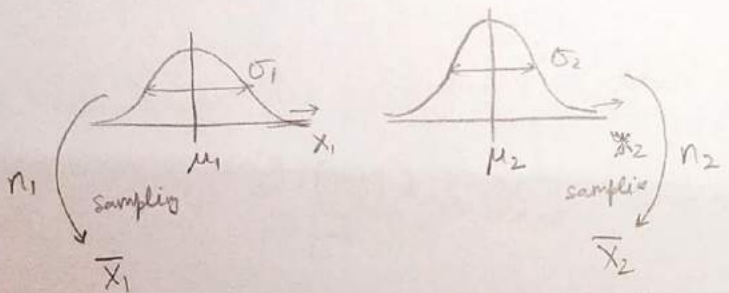
What about the sample distribution of $\bar{X}$'s?

As long as your sampling size is large i.e. $n >> 1$, then the $\bar{X}$ distribution will approximate a normal distribution $N(\mu, \sigma/\sqrt{n})$ where μ & σ are the mean & std. dev. of our unknown distribution.

The above is a consequence of the Central Limit Theorem.

Sampling Distribution of "difference b/w means"

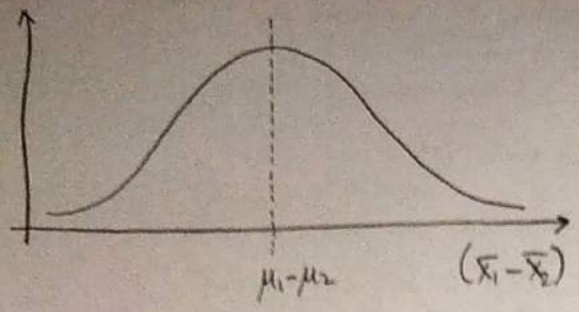
Say $X_1 \sim N(\mu_1, \sigma_1)$ and $X_2 \sim N(\mu_2, \sigma_2)$ are two distributions.



diff b/w sample means $\equiv (\bar{X}_1 - \bar{X}_2)$

Doing this a number of times, we get the distribution

$$(\bar{X}_1 - \bar{X}_2)_1, (\bar{X}_1 - \bar{X}_2)_2, (\bar{X}_1 - \bar{X}_2)_3 \ldots\ldots (\bar{X}_1 - \bar{X}_2)_\infty$$



mean of $(\bar{x}_1-\bar{x}_2) = \mu_1-\mu_2$

$Var(\bar{x}_1-\bar{x}_2) = Var(\bar{x}_1) + Var(\bar{x}_2)$

Now $Var(\bar{x}_1) = \frac{\sigma_1^2}{n_1}$ and $Var(\bar{x}_2) = \frac{\sigma_2^2}{n_2}$

So, $Var(\bar{x}_1-\bar{x}_2) = \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \Rightarrow$ std-dev of $(\bar{x}_1-\bar{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

So, $\Rightarrow (\bar{x}_1-\bar{x}_2) \sim N\left(\mu_1-\mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$

To convert this into a normal distribution, we define

$$Z = \frac{(\bar{x}_1-\bar{x}_2) - (\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \Rightarrow Z \sim N(0,1)$$

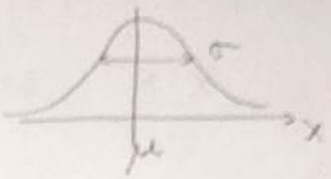
Lecture - 13

(03-02-2022)

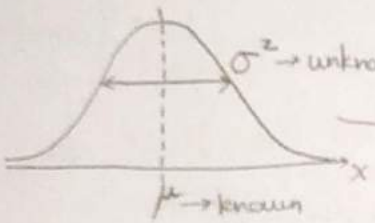
Till now, we've mainly explored how distributions of sample means approach $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$ where n = sample size.

Let's move onto the next sample statistic and ask, let's say we have a population with $X \sim N(\mu, \sigma)$.

Say $\mu \rightarrow$ known, but $\sigma^2 \rightarrow$ unknown.



Now we need an estimator function to estimate the variance.



n →

take a sample of size n

$(x_1, x_2 \dots\dots x_n)$

↓

Calculate $s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$

Let's say we repeat this sampling process multiple times.

$s_1^2, s_2^2, s_3^2 \dots\dots, s_m^2, \dots\dots$

Here, the sampling distribution of s^2 doesn't follow any distribution, but

$$\frac{s^2 \cdot (n-1)}{\sigma^2} \sim \chi^2_{n-1}$$

But what if our initial distribution is some unknown distribution?

Unfortunately, it turns out that for the variance, there is no equivalent central limit th^m. If the original population is NOT normally distributed then $\frac{s^2(n-1)}{\sigma^2}$ does NOT follow a χ^2 distribution.

Let's now take the case σ is unknown, μ is known but we'd like to figure out what's the distribution of a certain estimand.

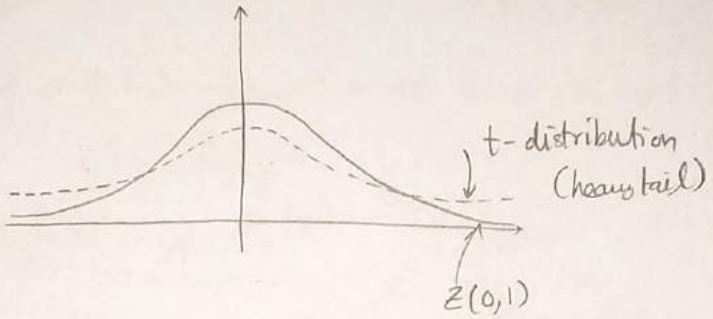
So, $X \sim N(\mu, \sigma) \Rightarrow \bar{X} \sim N(\mu, \sigma/\sqrt{n})$

$$\Rightarrow Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

But since σ is unknown, we replace it by the sample stddev. S

i.e. $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ which ends up following a student t-distribution with $(n-1)$ d.o.f.s.

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$



Low d.o.f. ($n \sim$ small) $\Rightarrow$ Heavy tails of t-distribution.

High d.o.f. ($n \sim$ large) $\Rightarrow$ tends to a Z-distribution

---

Lecture-14

(14-02-2022)

Student's t-distribution.

Say we start with a normal distribution (standard) Z, and we divide that by $\frac{\chi^2}{\text{d.o.f}}$, then

$$\frac{Z}{\chi^2/\text{d.o.f.}} \sim t\text{-distribution } (t_{\text{d.o.f.}})$$

We know that the distribution $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ with $\bar{X}$ as the random variable follows a Z-distribution.

Similarly, we know $\frac{s^2(n-1)}{\sigma^2}$ follows a χ^2 distribution with s as the (with $n-1$ d.o.f.) random variable

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim Z \quad \text{and} \quad \frac{s^2(n-1)}{\sigma^2} \sim \chi^2_{n-1}$$

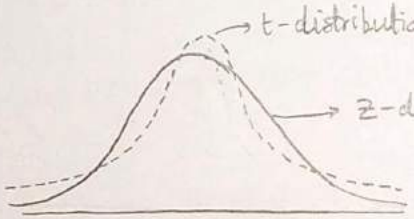
So, $$\frac{Z}{\sqrt{\left(\frac{\chi^2_{n-1}}{n-1}\right)}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{s^2/\sigma^2}} = \frac{(\bar{x}-\mu)}{s/\sqrt{n}} \sim t_{n-1}$$

→ A t-distribⁿ is obtained ONLY if our original data is normally distributed.

$\Rightarrow$ The t-distribution with $(n-1)$ d.o.f.s is given by

$$\boxed{t_{n-1} = \frac{\bar{X}-\mu}{s/\sqrt{n}}}$$

$\bar{x}$ = sample mean
s = sample std. dev.
n = sample size.



The t-distribution depends on the # of d.o.f.s $= n-1$

The t_{n-1} is our new random variable now. The probability density function is a bit complicated but not needed much

$$p_{n-1}(t) = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{(n-1)\pi}\,\Gamma\left(\frac{n-1}{2}\right)} \cdot \left(1 + \frac{t^2}{n-1}\right)^{-n/2}$$

Why does the t-table find values of t- corresponding to some area instead of the area corresponding to some t-value & d.o.f.s ?

Now earlier we looked at the random variable "difference of means" and how that followed a Z-distribution when renormalized properly.

Similarly, if we look at the distribution of "difference of means" when σ is unknown, we can see another t-distribution emerging

Say $X_1 \sim N(\mu_1, \sigma_1)$ & $X_2 \sim N(\mu_2, \sigma_2)$, and eventually

With σ_1 & σ_2 unknown, $t_1 = \dfrac{\bar{X}_1 - \mu_1}{S_1/\sqrt{n_1}}$ and $t_2 = \dfrac{\bar{X}_2 - \mu_2}{S_2/\sqrt{n_2}}$

follow t-distributions

Difference b/w means = $(\bar{X}_1 - \bar{X}_2)$

Now $\mu_{\bar{X}_1 - \bar{X}_2} = E[\bar{X}_1 - \bar{X}_2] = E[\bar{X}_1] - E[\bar{X}_2] = \mu_1 - \mu_2$

$S^2_{\bar{X}_1 - \bar{X}_2} = S.Var[\bar{X}_1 - \bar{X}_2] = S.Var[\bar{X}_1] + S.Var[\bar{X}_2]$

$S^2 = \dfrac{S_1^2}{n_1} + \dfrac{S^2}{n_2} \Rightarrow S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}$

So, $t \equiv \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} \sim t_{n_1+n_2-2}$

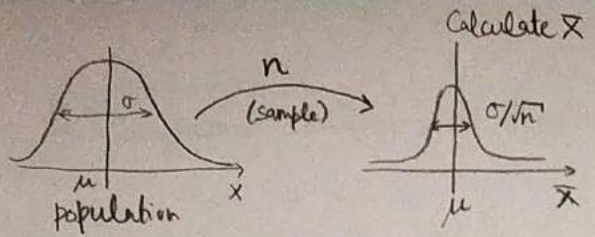
We can also write this as

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

↑ t-distribution with $(n_1 + n_2 - 2)$ d.o.f.

→ pooled variance

$$S_p^2 = \frac{S_1^2(n_1 - 1) + S_2^2(n_2 - 1)}{n_1 + n_2 - 2}$$



However, estimating μ from $\bar{x}$ generally has a lot of error. $\bar{x}$ is known as a point estimate of μ.

We will now discuss INTERVAL ESTIMATES.

Instead of giving a simple point estimate i.e. a single answer, we can specify an interval in which μ will lie (with some probability). This idea is also known as a "Confidence Interval".

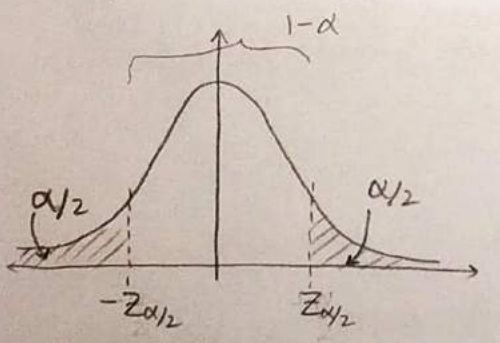
Confidence interval takes the following form – $(\bar{x} \pm \text{margin of error})$

We can demand the probability with which we want to have μ in some interval. If we have an error α, then the prob. of μ lying in the confidence interval $= 1-\alpha$.

So, a confidence interval $\equiv (1-\alpha)\cdot 100$ % C.I.

Confidence intervals for Z-distribution:

Say our error is α. Compute $\alpha/2$



Calculate Z value ($z_{\alpha/2}$) corresponding to area of $\alpha/2$ b/w $z_{\alpha/2}$ and $+\infty$ on $Z(0,1)$.

Then

$$P(z \in [-z_\alpha, z_\alpha]) = 1-\alpha$$

Since the $\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right)$ distribution is also z-distributed, the probability of this number $\in [-z_{\alpha/2}, z_{\alpha/2}]$ is

$$P\left(-z_{\alpha/2} \leq \frac{\bar{x}-\mu}{\sigma}\sqrt{n} \leq z_{\alpha/2}\right) = 1-\alpha$$

$$\Rightarrow P\left(-\frac{\sigma}{\sqrt{n}} \cdot z_{\alpha/2} - \bar{x} \leq -\mu \leq \frac{\sigma}{\sqrt{n}} z_{\alpha/2} - \bar{x}\right)$$

$$= \boxed{P\left(\bar{x} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq \mu \leq \bar{x} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right) = 1-\alpha}$$

So, with a probability of $(1-\alpha)$, the mean μ lies in the confidence interval of $\left[\bar{x} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \bar{x} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right]$!!!

$$\boxed{\text{Margin of Error} \equiv \frac{\sigma}{\sqrt{n}} z_{\alpha/2}}$$

The width of the confidence interval remains the same $\frac{2\sigma}{\sqrt{n}} z_{\alpha/2}$ independent of the sample chosen. The ranges of confidence intervals only change because of the sample mean $\bar{x}$, so the confidence interval may or may not capture μ. The probability of μ-capture $= (1-\alpha)$

Lecture-15

(15-02-2022)

The confidence interval is always calculated for a population parameter.

For the popⁿ mean μ, we found out that the confidence interval was of the form-

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

The length of C.I. depends upon $\sigma \rightarrow$ std-dev. of popⁿ

$n \rightarrow$ sample size

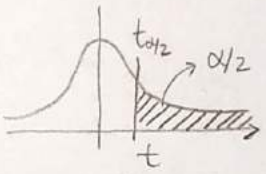
$\alpha \rightarrow$ level of error

$\Rightarrow 1-\alpha \rightarrow$ level of confidence.

Higher the $(1-\alpha)$, the longer is the length of C.I.

Now let's say one also doesn't know what the popⁿ stddev σ is. Then we estimate $\sigma \approx s$ (sample stddev.), and we replace the Z-distribution with the t-distribution.

$$\Rightarrow \text{C.I. if } \sigma \text{ is unknown} = \bar{X} \pm t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}}$$



∴ We can also see that the length of the C.I. changes depending on the value of sample std-dev. s!

Let us now discuss Confidence intervals for the parameter of difference in means b/w 2 distributions X_1 & X_2.

Assuming σ is known, then the $(1-\alpha)\cdot 100\%$ C.I. of difference b/w 2 means $\mu_1 - \mu_2$ will be-

$$\text{C.I. of } (\mu_1-\mu_2) \underset{(\sigma\text{-known})}{=} (\bar{x}_1 - \bar{x}_2) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

But in the case when σ is unknown, we write the C.I. as

$$\text{C.I. of } (\mu_1-\mu_2) \underset{(\sigma\text{-unknown})}{=} (\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2}, n_1+n_2-2} \sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where $S_p^2 \equiv \dfrac{S_1^2(n_1-1) + S_2^2(n_2-1)}{n_1+n_2-2}$

Since we calculate C.I.s for population parameters, we can do a similar thing for pop^n variance σ^2.

Say $X \sim N(\mu, \sigma)$ where σ or σ^2 is unknown.

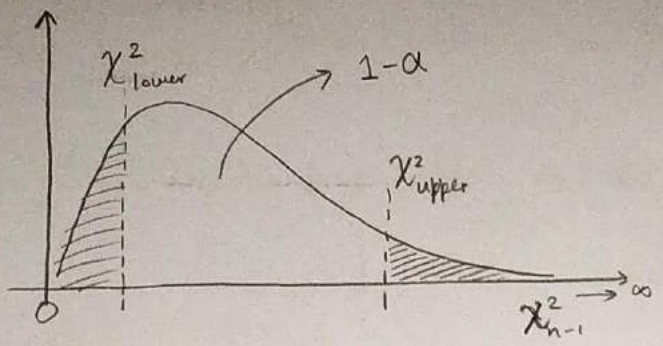
Just like $\bar{x}$ was a point estimate for μ, the sample variance s^2 is a point estimate for σ^2.

$$s^2 \approx \sigma^2$$

For the distribution where σ was known, and our parameter of interest was $\bar{x}$, we knew that $\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right)$ followed a Z-distribution & we used that to get a C.I. for μ.

For a distribution where σ was unknown, and parameter of interest was s, we know that

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$



χ^2_{n-1} is a normalized distribution!

We now try to find χ^2_{lower} s.t. the area before it $= \alpha/2$.

Similarly, we find a χ^2_{upper} s.t. the area beyond it $= \alpha/2$

∴ We can now say that if our error is α, then the probability of lying in the confidence of $(1-\alpha)\cdot 100\%$ is

$$P\left(\chi^2_{lower} < \chi^2_{n-1} < \chi^2_{upper}\right) = 1-\alpha$$

So, $$P\left(\chi^2_{lower} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{upper}\right) = 1-\alpha$$

$$\Rightarrow P\left(\frac{s^2(n-1)}{\chi^2_{upper}} < \sigma^2 < \frac{s^2(n-1)}{\chi^2_{lower}}\right) = 1-\alpha$$

∴ The value of σ^2 lies between $\left[\frac{s^2(n-1)}{\chi^2_{upper\,(n-1)}}, \frac{s^2(n-1)}{\chi^2_{lower\,(n-1)}}\right]$ with a probability of $(1-\alpha)$.

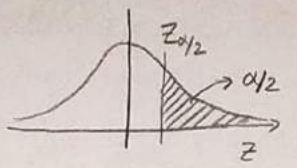
↳ level of confidence.

Example 1

$X \sim N(\mu, \sigma = 20)$

Sample size $n = 10$. Say we take a sample & get $\bar{x} = 25$.

Find the 95% C.I. for μ.



95% C.I. $\Rightarrow \alpha = 0.05$ error.

Also, σ is known but μ isn't, so the sample will follow a normal distribution.

$$P\left(z_{-\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{+\alpha/2}\right) = 1 - \alpha$$

$$\Rightarrow \text{C.I. of } \mu = \bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 25 \pm z_{0.025} \cdot \frac{20}{\sqrt{10}}$$

$$= 25 \pm 2 z_{0.025} \cdot \underbrace{\sqrt{10}}_{\approx \pi} = 25 \pm 2 \cdot (1.96) \cdot 3.14$$

$$= 25 \pm 12.31$$

Example 2 Let's say even σ is unknown now.

$X \sim N(\mu, \sigma)$

$n = 10$, $\bar{x} = 25$, $s = 5$.

Let's now calculate the 95% C.I.

Now we'll use the t-distribution to figure this out.

$$\text{C.I. of } \mu = \bar{x} \pm t_{\alpha/2}^{(n-1)}\frac{s}{\sqrt{n}} = 25 \pm t_{0.025}^{\text{d.o.f} = 9} \cdot \sqrt{\frac{5}{2}}$$

Now $t_{0.025}^{d.o.f=9} \approx 2.26216$ and $\sqrt{\frac{5}{2}} \approx 1.58$

So, C.I. (95%) of $\mu = 25 \pm 3.574$

Example 3.

Say $X \sim N(\mu, \sigma = 20)$ like before.

$\bar{X} = 25$, and the Margin of Error for 95% C.I. is 5.

Find n=?

So, $M.o.E. = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \Rightarrow n = \left\lceil \left(\frac{Z_{\alpha/2} \cdot \sigma}{M.o.E.} \right)^2 \right\rceil = \left\lceil \left(\frac{1.96 \times 20}{5} \right)^2 \right\rceil$

$= \lceil 61.46 \rceil = 62$

Lecture - 16.

(17-02-2022)

HYPOTHESIS TESTING.

(1) Let's say we have a data set of heights X. We take a sample & calculate the sample mean $\bar{x}$.

We now "hypothesize" that $\mu_{Height} = \mu_H$

Assuming this popⁿ is normally distributed, is there any relation of hypothesized mean μ_H with $\bar{X}$ i.e does your data support your hypothesis?

Model ≡ Hypothesized distribution (or distribution parameters)

So, we want to calculate the probability that given some data, our model is some XYZ

$$P(\text{Model XYZ} \mid \text{data})$$

However, most of the times we know $P(\text{data} \mid \text{model})$

(2) Taking another example where we have 2 means $\bar{X}_1$ & $\bar{X}_2$, we calculate the difference b/w means $\bar{X}_1$ & $\bar{X}_2$. If there is a significant difference, we say $\bar{X}_1$ & $\bar{X}_2$ don't come from the same distribution. If the difference is not significant, then we say they do come from the same one.

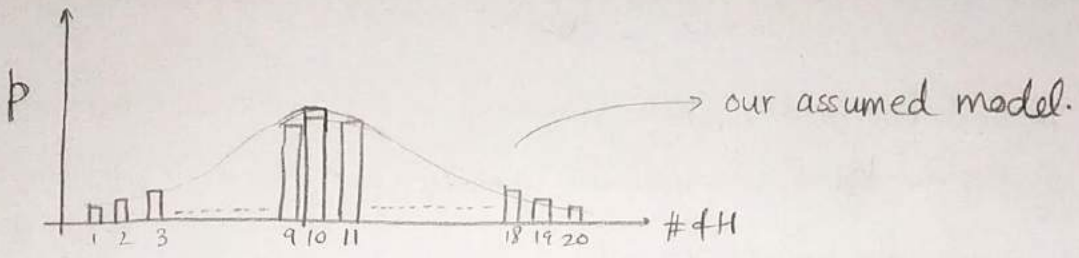
Again, we want to know $P(\text{Model} \mid \text{Data})$

same distribⁿ? ↓ $\bar{X}_1, \bar{X}_2$

Let's now take a concrete example we can work with-

We have a coin. Is the coin fair or not? We only toss it 20 times.

We begin with the assumption that the coin is fair. Then the distrib^n for the # of heads looks like-



$\Rightarrow$ we carry forward the assumption that $p = q = 0.5$ $\Rightarrow$ Binomial distb^n with $p = q = 0.5$

$\therefore$ If our assumption is true, then our most likely event is # of H = 10.

As our data starts deviating from the expected # of Heads (=10), the $p(\text{data}|\text{model})$ decreases further, & our model becomes more and more unlikely.

Now say we get 19 Heads out of 20 flips, an extremely rare event if our model is correct (the coin is fair).

We can argue that the rarity of this event $\Rightarrow$ our assumption was wrong & our model was incorrect.

But we could also have disregarded it as a genuine rare event since $p \neq 0$.

Where do we draw the line?

We decide that boundary.

So in this experiment, we assume the model of a fair coin (20 flips), take the data & calculate $\bar{x}$ from the data, and now we ask,

what's the probability that we get such a data outcome given our model?

So we end up calculating $P(\text{data} \mid \text{model})$ instead of ~~$P(\text{model} \mid \text{data})$~~

To take the final decision, we have to define 'what is rare?' before rejecting the hypothesis. However, finally, we take a SUBJECTIVE decision to decide the bounds of rarity.

Frequentist statistics ⟶ calculate $P(\text{data} \mid \text{model})$

Bayesian statistics ⟶ calculate $P(\text{model} \mid \text{data})$

NOTES

- If the sample obtained ($\bar{x}$) has a probability of occurrence $P(\bar{x})$ less than the pre-specified threshold $P(\bar{x}_a)$ i.e. $P(\bar{x}) < P(\bar{x}_a)$, given H_0 is true, then difference between sample & null hypothesis is statistically significant.

We then reject H_0 & accept H_A.

- Hypothesis Tests based on statistical significance are another way of expressing Confidence Intervals!

Lecture - 17

(21-02-2022)

Hypothesis Testing (contd.)

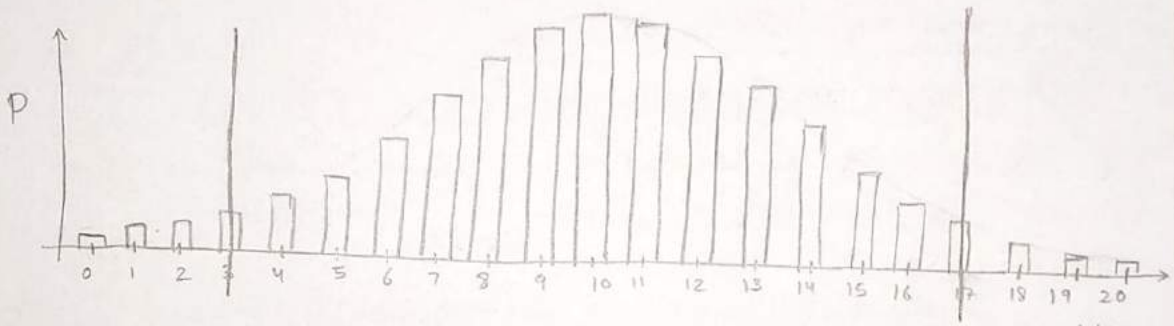
Ideally, we want to calculate $p(\text{Model} \mid \text{data})$ where model ~ distrib^n (parameters)
i.e. we want to find certain parameters of a distrib^n given the data.

However, actually, we end up calculating $p(\text{data} \mid \text{model})$ in frequentist statistics. We assume that the model is known & try to see if the data fits the model.

Example - Is the coin fair? (we can only flip n=20 times before making a decision.

We say, assume that the coin is fair $p = q = 0.5$. This is our assumed / hypothesized model.

$\Rightarrow$



$$P(\#\text{ of heads} = r) = {}^{n}C_{r}\,(0.5)^{r}\,(0.5)^{n-r}$$

We put bars at # of H = 3 and 17, and say that if # of H > 17 or < 3, then the events are so rare that upto some error, we can say that the coin is not fair.

However, # of H = 0, 1, 2, 18, 19, 20 are still valid possibilities, although extremely unlikely. So, we go forward with the hypothesis with an error of

$$\varepsilon = P(0) + P(1) + P(2) + P(18) + P(19) + P(20)$$

If our data says $3 \leq$ # of Heads ≤ 17, then we cannot reject the Hypothesis that the coin is fair. However, we reject our hypothesis if # of Heads > 17 or < 3, but our decision can still be wrong with an error

$$\mathcal{E} = \sum_{i \in \{0,1,2,18,19,20\}} P(i) \longrightarrow \text{extremely low possibility. Acceptable error.}$$

So, three main steps in hypothesis testing are-

① Assume a model.

② Make the boundaries of acceptability of a Hypothesis. These are decided by choosing what level of error is okay with us.

③ Reject or not reject the hypothesis depending on your error & data.

Z-Test for a single mean.

Say heights $\sim N(\mu, \sigma = 20)$ μ is unknown.

Make a hypothesis that $\mu = 170$ cm.

Take a sample of $n = 16$ students, and we get $\bar{X} = 175$ cm.

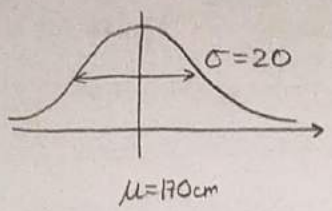
Now "ideally" we'd have wanted to find what's the probability that $\mu = 170$ cm IF $\bar{X} = 175$cm i.e. $P(\mu = 170\text{cm} \mid \bar{X} = 175\text{cm}, n = 16) = ?$

However, we end up calculating the probability that $\bar{X}$ can be as extreme as 175 cm when $\mu = 170$cm, i.e. $P(\bar{X}$ is as extreme as $175\text{cm} \mid \mu = 170\text{cm})$

We start by taking a null hypothesis H_0.

H_0: $\mu = 170\,cm$ is what we hypothesize. (with a normal distrib^n)

So, hypothesized distribution

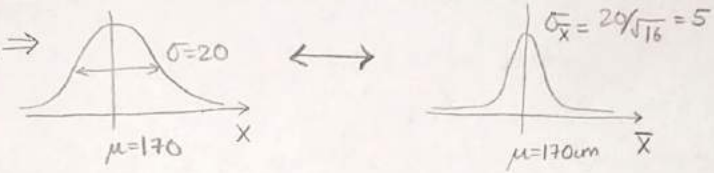


We also have an alternative hypothesis H_A.

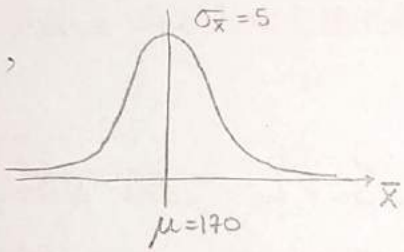
H_A: $\mu \neq 170\,cm$ is our alternate hypothesis.

We accept the type I error $\alpha = 0.05$. (5%)

Assuming that $\mu = 170\,cm$ is true, $\Rightarrow$



Now given,

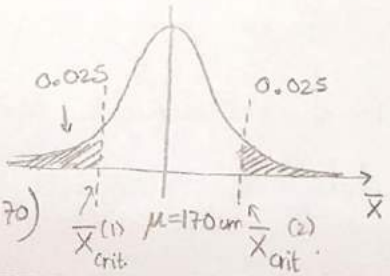


how do we decide the boundaries on this where we say a certain value of $\bar{x}$ obtained is within our acceptable range & we can't reject H_0?

We divide $\alpha = 0.05$ → 0.025, → 0.025

~~(since distrib^n is symmetric)~~

(since $\mu \neq 170$ is our H_A, so it can be < 170 or > 170)



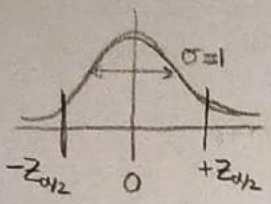
- If our calculated $\bar{x}$ from the sample $\in [\bar{x}^{(1)}_{crit}, \bar{x}^{(2)}_{crit}]$, then we do not reject the hypothesis that $\mu = 170cm$!
- However, if $\bar{x} > \bar{x}^{(2)}_{crit}$ or $\bar{x} < \bar{x}^{(1)}_{crit}$, then we reject H_0.

Now we make the switch to Z-distrb^n to figure out $\bar{X}^{(1,2)}_{critical}$.

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = Z \sim N(0,1) \Rightarrow$$

$$\bar{X}^{(1)}_{critical} = \mu - \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}$$

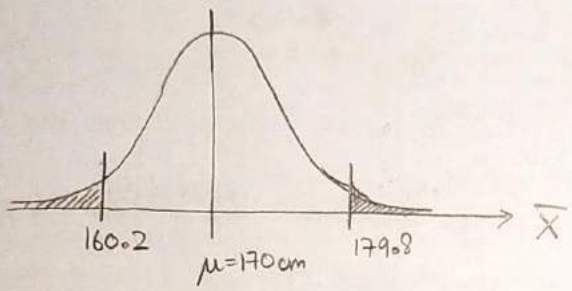
$$\bar{X}^{(2)}_{critical} = \mu + \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}$$



Usually, $\alpha = 0.05$ so $Z_{\alpha/2} = Z_{0.025} = 1.96$

For $\mu = 170$ cm (hypothesized) and $\frac{\sigma}{\sqrt{n}} = 5$, $\bar{X}^{(1)}_{critical} = 160.2$ cm

$\bar{X}^{(2)}_{critical} = 179.8$ cm



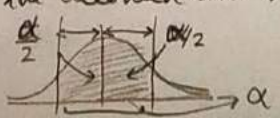
Now if our $\bar{X} \in [160.2, 179.8] \Rightarrow H_0$ can't be rejected & is rejected if it is outside the critical extremes.

Another way to check if our data, and hence $\bar{X}_{obs}$ rejects or doesn't reject H_0 is to calculate Z_{obs} corresponding to $\bar{X}_{obs}$ and see if $Z_{obs} \in [-1.96, +1.96]$ $([-Z_{\alpha/2}, +Z_{\alpha/2}])$

Isn't "error" a slightly misleading word for α here?

Shouldn't the allowed error be this?

i.e. how much error do we allow that $\bar{X}$ is some error away from μ?



Lecture-18

(22-02-2022)

We label "α" as an error because $100.\alpha$ % of the times, we might make an error in our decision EVEN IF our Hypothesis is TRUE.

In the whole game of Hypothesis testing, we're comparing $\bar{X}_{calc}$ with $\bar{X}^{(1,2)}_{critical}$ (or Z_{calc} with $Z^{(1,2)}_{critical}$) and seeing if it's less or more extreme.

In the beginning we stated that in frequentist stats, we calculate

$$P(\text{data being atleast as extreme as critical pts.} \mid \text{model } H_0) = \text{"p" value}$$

↳ when H_0 is true

However, we haven't yet calculated it in our heights $X \sim N(\mu, \sigma=20)$ example where we took $H_0: \mu = 170$ as the null hypothesis with $\alpha = 0.05$.

There are two ways of making decisions about H_0 -

(I) Check if Z_{calc} is as extreme as $Z^{(1,2)}_{critical}$. (which we did)

(II) Compare α vs p.

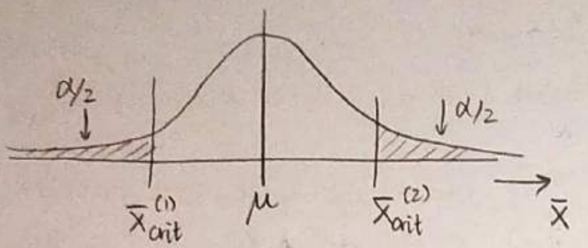
If $p < \alpha \Rightarrow$ reject H_0.

If $p > \alpha \Rightarrow$ do not reject H_0.

Let us calculate the p-value for the same example with $X \sim N(\mu, \sigma=20)$ and $H_0: \mu = 170$ with $\alpha = 0.05$

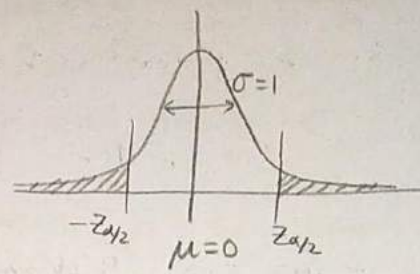
If we take a sample and $\bar{X} = 175$. $(n=16)$

$$Z_{calc} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{175 - 170}{20/\sqrt{16}} = \frac{5}{5} = 1$$



$\bar{X}$-distribution

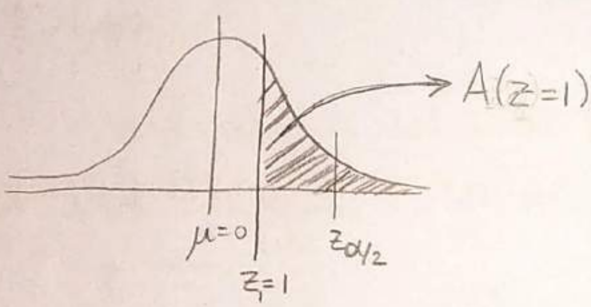
$\equiv$

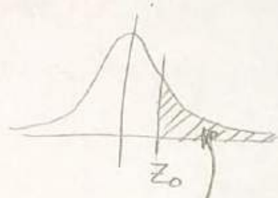


z-distribution

$z \equiv (\bar{x} - \mu)/(\sigma/\sqrt{n})$

Now, $Z_{calc} = +1$. So if we calculate the area b/w Z_{calc} and $+\infty$





area-value corresponding to z_0.

$A(z=+1) = 0.1587$

<u>p-value</u> is the area associated with $(Z_{calc}$ to $\infty) \times 2$ when we put our error α on both the tails. (TWO-TAILED TEST)

So, $p = 0.1587 \times 2 = \underline{0.3174}$

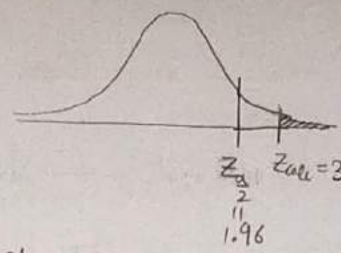
Now $p = 0.3174 > \alpha = 0.05 \Rightarrow p > \alpha$.

The reason why $p > \alpha$ is because $Z_{calc} < Z_{critical} \Rightarrow A(Z_{calc}) > A(Z_{crit.})$

$A(Z_{calc}) = p/2$, $A(Z_{crit.}) = \alpha/2$

Since $p > \alpha \Rightarrow$ we can't reject H_0.

However, suppose that $\bar{X} = 185 \Rightarrow Z_{calc} = 3$



Now $A(Z=3) = 0.0013$

$\Rightarrow p = 0.0013 \times 2 = 0.0026$

But $p = 0.0026 < \alpha = 0.0050 \Rightarrow p < \alpha$

Since $p < \alpha \Rightarrow H_0$ is rejected.

The reason why $p < \alpha$ is b/c $Z_{calc} > Z_{crit} \Rightarrow A(Z_{calc}) < A(Z_{crit.})$

$$A(Z_{calc}) = p/2, \quad A(Z_{crit.}) = \alpha/2$$

If $Z_{calc} < 0$, then take $|Z_{calc}|$ and calculate area from $|Z_{calc}|$ to $+\infty$ and compare that to $\alpha/2$. (or calculate area b/w Z_{calc} & $-\infty$)

Ques: why did we distribute $\alpha/2$ & $\alpha/2$ in both the tails?

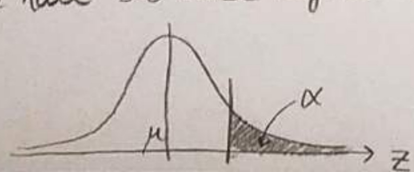
Ans: Not because $N(\mu, \sigma)$ is symmetric, but because our alternate hypothesis H_A was defined as $H_A : \mu \neq 170$ (not < or > 170)

So our alternate hypothesis works even if $\mu > 170$ or $\mu < 170$.

Since we are okay with μ being < 170 or > 170, we put the error in both the tails.

So, $H_A : \mu \neq 170 \Rightarrow$ Two tailed test!

However, if we had $H_A : \mu > 170$ cm. In that case, we put all the error in a single tail s.t. area beyond $Z_{critical}$ is α.



$Z = (\bar{x} - \mu)/(\sigma/\sqrt{n})$

In such a case, if we see $Z_{calc} > Z_\alpha$, then we say H_0 is rejected.

If $Z_{calc} < Z_\alpha$, then H_0 is NOT rejected.

∴ The directionality of alternative hypothesis H_A carries forward to rejecting or not rejecting H_0, where we put all the error in one tail.

Same procedure if $H_A: \mu < 170$ cm, with all α in left-tail.

When all the error α is sent into a single tail ⇒ ONE-TAIL TEST!

★ For a one-tail test, we won't multiply the area by 2 to find p.

Lecture -19

(23-02-2022)

Discussion on α-error:

Say heights $\sim N(\mu, \sigma = 20)$ where μ is unknown. The size of the sample is known. $(n = 16)$. And we calculate the $\bar{x}$.

We are interested in testing the hypothesis that $\mu = 170$. So we start with -

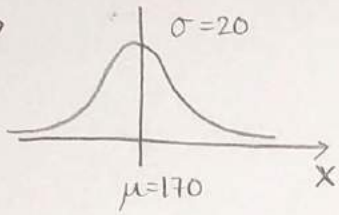
Let H_0 be true where $H_0: \mu = 170$.

assumption

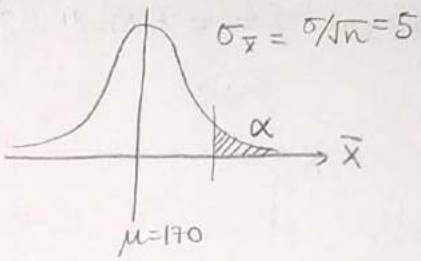
We are analyzing the one-tail example here.

$H_0: \mu = 170. \Rightarrow$

$H_A: \mu > 170.$



$n = 16$ $\Longrightarrow$



Distribution if H_0 is true

(1) IF H_0 is really TRUE -

- then we don't reject H_0 if $\bar{x}$ lies less than the $\bar{x}_{critical}$.
- we reject it if it lies beyond $\bar{x}_{critical}$.

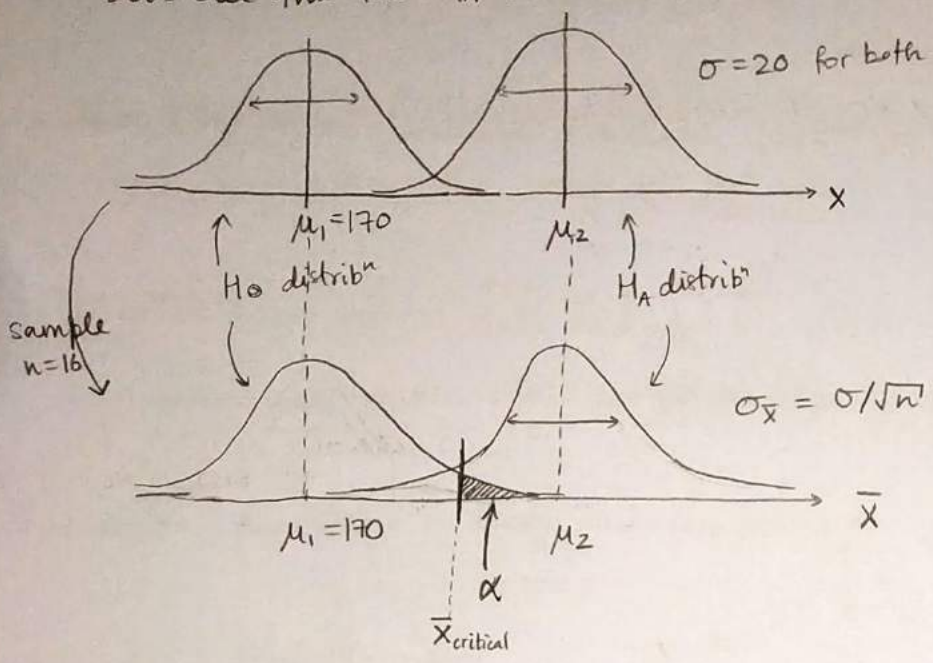
However, those values of $\bar{x}$ are still a valid outcome of the distribution.

So, with probability α, we make the error of rejecting H_0 EVEN WHEN IT IS REALLY TRUE. This is known as type I error.

Let us now consider the scenario that in reality, H_0 is false (or that $\mu \neq 170$ cm), but we as statisticians do not have that intel.

What do we do in this case?

Say the "actual" value of μ was μ_2 (which we are not aware of).
Let's call this the H_A distrib^n.

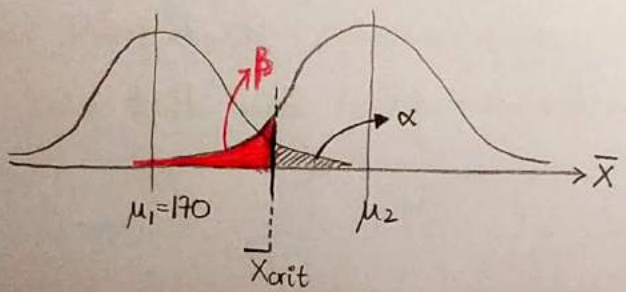


② Now, when H_0 is really FALSE -

- if $\bar{x} > \bar{x}_{critical}$ i.e. lies in the shaded band with area α, then we reject null hypothesis H_0 because it's too extreme.

  So, we get $\bar{x} > \bar{x}_{critical} \Rightarrow H_0$ is rejected, which is good because in reality H_0 was false.

- However, if $\bar{x} < \bar{x}_{critical}$ i.e. if it lies outside the α-area band, then we do not reject H_0. This is a problem (or an error) however, since in reality H_0 is false, but we still didn't reject it



Since our actual distribution from which the samples are drawn is the

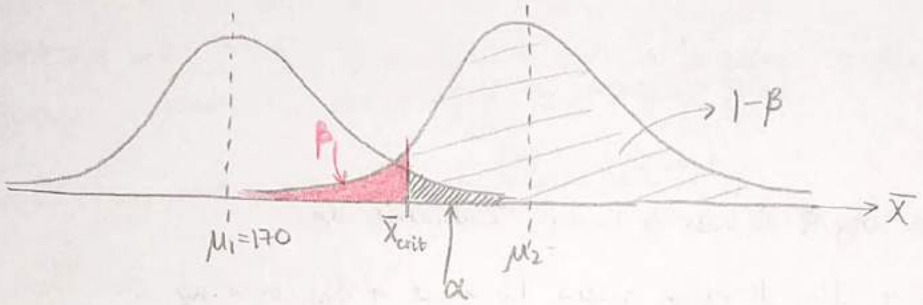
H_A distribution, then the probability that we DO NOT REJECT H_0 DESPITE H_0 being false in reality is -

$$\int_{-\infty}^{\bar{x}_{crit.}} d\bar{x}\, N_{\bar{x}}(\mu_2, \sigma/\sqrt{n}=5) \equiv \beta$$

And we define it as the β-error.

So, $H_0: \mu = 170$
$H_A: \mu > 170$ ⟶

	H_0 TRUE	H_0 FALSE
Reject H_0	$\alpha \cdot 100\%$ times Type I error.	✓ $1-\beta$
Do not reject H_0	✓ $1-\alpha$	$\beta \cdot 100\%$ times Type-II error



β = probability that you don't reject H_0 when H_0 is actually false.

$1-\beta$ = probability that you reject H_0 when H_0 is actually false.

$\beta \equiv p$ (not reject H_0 | H_0 is actually FALSE)

$\Rightarrow$ $1-\beta = p$ (reject H_0 | H_0 is actually FALSE) ⟶ POWER OF A TEST

$\alpha \equiv p$ (reject H_0 | H_0 is actually TRUE)

$\Rightarrow 1-\alpha = p$ (not reject H_0 | H_0 is actually TRUE)

Now we come back to the question which always haunted me.

Why can't I make α as small as possible if it's an error?

Because reducing $\alpha \Rightarrow \bar{X}_{critical}$ becomes more and more extreme.

The more extreme $\bar{X}_{critical}$ becomes, then in cases where H_0 is actually FALSE, the probability that H_0 is NOT REJECTED even when $\bar{X}$ was extreme as hell will increase.

Geometrically, as $\alpha \downarrow$, $\bar{X}_{crit}$ becomes more extreme, and the value of $\beta \uparrow \Rightarrow (1-\beta) \downarrow$ which is a measure of the "power of the test".

$\therefore$ Decreasing α comes at a cost of increasing β. (when n is fixed.)

Qn: Is there a way of decreasing α & β simultaneously?

Ans: Yes! Since the β error arises because of the overlap b/w the hypothesized (H_0) and the actual distribution (H_A), if we increase n, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ decreases, and so does the overlap. Hence we can decrease both α & β errors simultaneously $\Rightarrow$ larger $(1-\beta) \Rightarrow$ larger power of test.

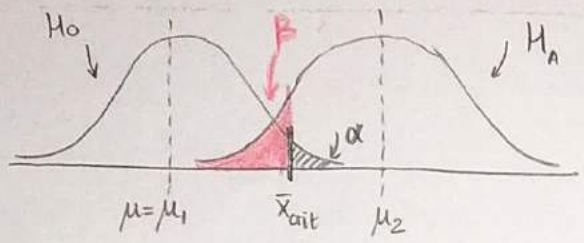
$\therefore$ "Power of the test" is directly dependent on sample size n.

Lecture - 20

(24-02-2022)

As discussed earlier, for a one-tailed test with $H_0: \mu = 170$ & alternative hypothesis as $\mu > 170$, we get all the α in the right tail.

In the case when our H_0 is actually false (although we don't know it), there might exist a different $\mu = \mu_2$ of the actual distribⁿ.



Corresponding to a given n, decreasing $\alpha \Rightarrow$ increasing β.

However, changing n can give us a handle on both α & β.

Converting to a z-distribⁿ

$z_\alpha, z_\beta > 0$

$$z_\alpha = \frac{\bar{X}_{crit} - \mu_1}{\sigma/\sqrt{n}} \qquad z_\beta = -\frac{(\bar{X}_{crit} - \mu_2)}{\sigma/\sqrt{n}}$$

Now, $z_\alpha + z_\beta = \underbrace{\left(\frac{\mu_2 - \mu_1}{\sigma}\right)}_{\text{effect size / Cohen's D}} \cdot \sqrt{n}$ = Non-centrality parameter δ.

(standardized difference b/w 2 means)

$$\Rightarrow n = \frac{(z_\alpha + z_\beta)^2}{[(\mu_2 - \mu_1)/\sigma]^2}$$

For one-tailed test.

HW - do this for a two-tailed test. ($\alpha \to \alpha/2$)

But how do I know μ_2 beforehand? That itself is the quantity of interest.

The quantity $(\mu_2 - \mu_1)$ has to be addressed subjectively.
We have to ask, "what to me is a meaningful difference?"

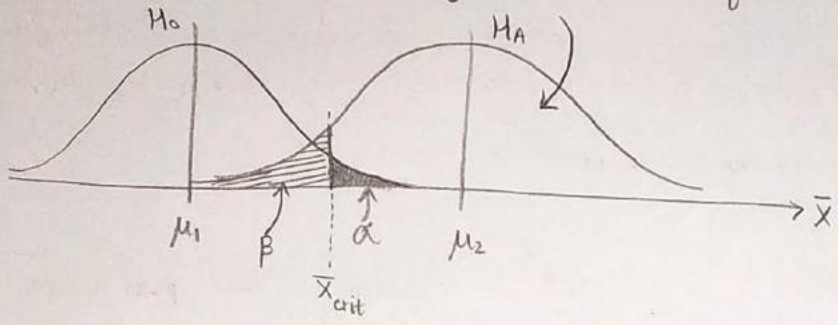
So, $(\mu_2 - \mu_1)$ is not the exact difference b/w hypothesized & actual mean. It's a meaningful difference, a difference which is significant enough for us to test the hypothesis for.

In general, we want to go for high power (high $1-\beta$) tests.

Lecture - 21
(28-02-2022)

The errors α, β, "p-value", all of these are inter-related by the parameter $(1-\beta)$ or the power of the test.

$(1-\beta)$ = probability that we reject H_0 when it's false.



we want to be in the $(1-\beta)$ area when our null hypothesis is false. This area depends on the amount of overlap. The overlap here is in our control b/c $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, so if $n \uparrow$, $\sigma_{\bar{x}} \downarrow$, and the overlap goes $\downarrow$

$\Rightarrow$ power of test $(1-\beta) \uparrow$.

So we can decide the optimum "n" for our desired α, β values.

Say we want • $(1-\beta) = 0.9$ or $\beta = 0.1$

• $\alpha = 0.01$

Now $$n = \left[\frac{Z_\alpha + Z_\beta}{\frac{(\mu_2 - \mu_1)}{\sigma}}\right]^2$$ for a one-tailed test.

$(\mu_2 - \mu_1)$ is NOT the difference b/w actual & hypothesized mean. We wouldn't be doing Hypothesis testing if we knew μ_2. So, $\mu_2 - \mu_1$ is a minimum "meaningful difference".

Let's say we keep our minimum meaningful difference at $\mu_2 - \mu_1 = 5$ and find out a sample size "n" value corresponding to it.

If in reality $\mu_2 - \mu_1 = 7$ now, then $Z_\alpha + Z_\beta \propto n \cdot (\mu_2 - \mu_1)^2$

$\Rightarrow$ $Z_\beta \propto (\mu_2 - \mu_1)^2$. So if $(\mu_2 - \mu_1) \uparrow$ for fixed n, $Z_\beta \uparrow \Rightarrow \beta \downarrow$ and the power of test $(1-\beta) \uparrow$.

So, once our $(\mu_2 - \mu_1)$ is fixed to a min. difference, having $\mu_2 - \mu_1 >$ diff.$_{min}$ only improves the power of the test!

However, by choosing a large enough n according to above prescription, we can make any difference statistically significant, which apparently is bad.

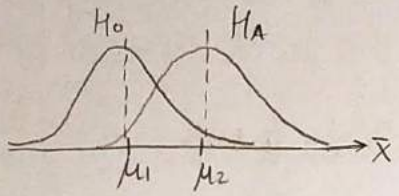
There is a difference b/w statistical significance vs Biological significance!

Misusing statistical significance

- **Case - 1**

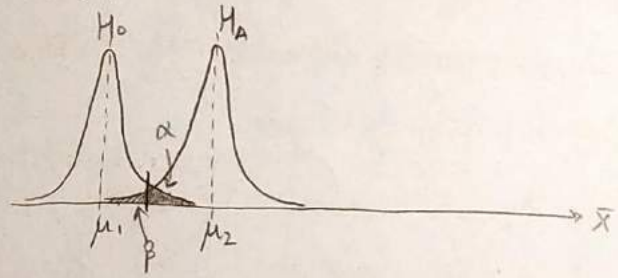
Let's say the hypothesized mean is $\mu = \mu_1$ & the actual mean (unknown to us) is $\mu = \mu_2$. Say $|\mu_2 - \mu_1| \ll 1$ or maybe it is extremely small.

In that case, our distribution would look like this for intermediate "n" -



since μ_1 & μ_2 are not very different, a null hypothesis $H_0 : \mu = \mu_1$ won't be rejected since there is a large overlap b/w the values of both the distributions. So, the result $\bar{x}$ drawn from H_A distrib" ($H_A : \mu > \mu_1$) won't have a statistically significant difference.

However, by ramping up the value of n, we can make the distributions sharply peaked & reduce overlap.



Since $n = \left[\frac{Z_\alpha + Z_\beta}{(\mu_2 - \mu_1)/\sigma}\right]^2$

If $n \uparrow \Rightarrow Z_\beta \uparrow \Rightarrow \beta \downarrow \Rightarrow (1-\beta)$ power $\uparrow$

So even though the difference b/w μ_2 & μ_1 was NOT significant, by choosing a high enough n & a high powered $(1-\beta)$ test, one can make it look like the $\bar{x}$ generated from the sample to have a statistically significant difference from μ_1. (i.e. we consider it as a rare event assuming H_0 is true)

$\bar{x} > \bar{x}_{crit}$ or $p < \alpha \Rightarrow$ rejecting H_0.

∴ Mathematically, a high-power test is always better because higher the power, the higher the accuracy of discerning small differences.

But it can be used to create a biased impression depending on how it's presented.

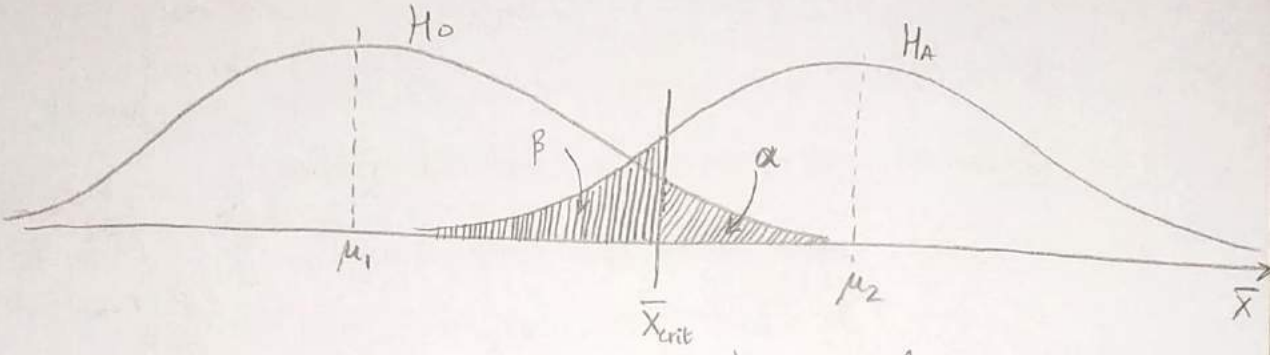
- Case-2

Let's say $|\mu_2 - \mu_1| \gg 1$ or a significant difference exists b/w hypothesized & actual mean.

If we choose a very low n, it is easy to see that since $n = \left\lceil \frac{(z_\alpha + z_\beta)}{(\mu_2 - \mu_1)/\sigma} \right\rceil^2$

$n \downarrow \Rightarrow z_\beta \downarrow \Rightarrow \beta \uparrow \Rightarrow (1-\beta)$ power of test $\downarrow$.

As the power $(1-\beta) \downarrow$, the overlap b/w distrib^{n}s increases



$\beta \equiv P$(not rejecting H_0 even when it's wrong) is very high.

If β is high, it's very likely that we find a sample $\bar{x}$ (sampled from H_A) and mistake it to be sampled from H_0 & not rejecting H_0 hypothesis since

$\bar{x} < \bar{x}_{crit}$ or $p > \alpha$ is a likely possibility.

This would make $\bar{x}$ look like a statistically INsignificant difference.

In all of this, we use the following def^n of statistical significance

Def^n: A result is said to be statistically significant if the chances of that particular event happening, assuming H_0 is true, is extremely unlikely (or in simple terms, $p < \alpha$).

Therefore-

① If $|\mu_2 - \mu_1| \ll 1$, we can make $\bar{x}$ look like a statistically significant result (assuming H_0 is true) i.e. an extremely unlikely event in H_0 distrib^n, by increasing the sample size n.

② If $|\mu_2 - \mu_1| \gg 1$, we can make $\bar{x}$ look like a statistically INsignificant result (assuming H_0 is true) i.e. as if it came from H_0 distrib^n, by decreasing the sample size n.

⇒ A high power test and a statistically significant difference
A low power test and a statistically INsignificant difference
} can be used to fudge the inferences of the data.

Coming up with a measure of "biologically" meaningful difference.

$$\text{Effect size} = \frac{\mu_2 - \mu_1}{\sigma} = \text{Cohen's } D/\delta$$

(μ_2: actual; μ_1: hypothesized)

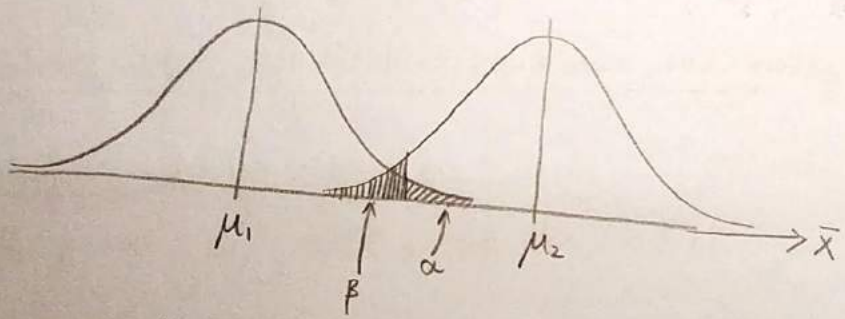
However, when we don't have access to actual mean, we find $\bar{x}$ and replace μ_2 with it (and write $\mu_1 = \mu_H$)

$$\text{Effect size} \equiv \boxed{\frac{\bar{x} - \mu_H}{\sigma} = \text{Cohen's } d} \quad \text{(doesn't depend on } n\text{)}$$

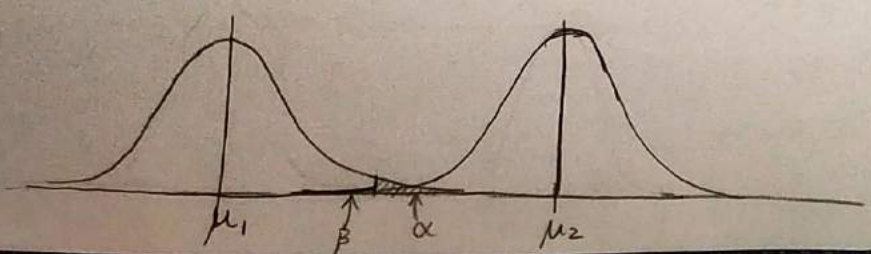
Thumb rules
- $d = 0.2$ → low diff.
- $d = 0.2$ to 0.5 → intermediate diff.
- $d > 0.6$ → large diff.

Still very subjective. Not valid everywhere.

Let's take an example where H_0 is false.



To increase power $(1-\beta)$, we increase the sample size n.



Now if we do high powered test (given H_0 is false), what happens to the p-value v/s α-value comparision?

Ques Are we more likely to see $p < \alpha$ ($\bar{x} > \bar{x}_{crit.}$) i.e. H_0 being rejected, when we have a high powered test compared to a low powered test?

Ans: Yes. We are more likely to see $p < \alpha$ because as $n \uparrow$, the $\beta \downarrow$ and $(1-\beta) \uparrow$ hence decreasing the overlaps b/w the H_0 & H_A distributions. If the overlap decreases, then we are less likely to make the mistake of not rejecting H_0 even when it's false.

So, we prefer high powered tests.

Reproducibility Crisis.

Reproducing results of an experiment is likely when the tests we do are high powered. If $(1-\beta)$ is low however, then it's very likely that a hypothesis was not rejected purely by chance.

Need to repeat analysis for TWO-TAILED TEST.

Typically, we need a higher n (for the same power) in a TWOTAILED compared to a ONE TAILED TEST.

Lecture-22

(02-03-2022)

All of our prior discussion involved one sample hypothesis tests, where σ is a known quantity. We used this to test our hypotheses concerning the means of the distribution.

But what if σ is unknown?

Say Heights $(x) \sim N(\mu, \sigma)$. We want to check the hypothesis that

$H_0: \mu = 175\text{cm}$.

The only thing we know about x is that it's normally distributed.

Take a sample of size $n = 5$.

If we take $H_A: \mu \neq 175\text{cm}$. (Two-tailed test.)

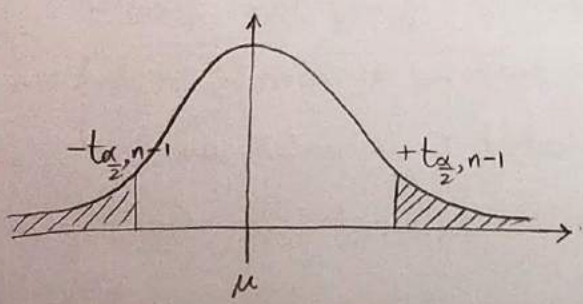
Fix $\alpha = 0.05$. $\rightarrow \alpha/2 = 0.025$, $\rightarrow \alpha/2 = 0.025$ } to be put in both the tails.

σ known $\Rightarrow \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}} = z \sim N(0,1)$

for $n \geq 30$, $t \approx z$.

σ unknown $\Rightarrow \dfrac{\bar{x} - \mu}{s/\sqrt{n}} = t \sim t_{n-1}$

$\pm t_{0.025,4} = \pm 2.78$



If calculated t is more extreme than $\pm t_{\frac{\alpha}{2}, n-1} = \pm 2.78$, then we say that there is a significant difference & we can reject H_0.

Say the sample gives $\bar{x} = 168$ and $s = 5$.

$$\Rightarrow t = \frac{168 - 175}{5/\sqrt{5}} = -3.13$$

Now -3.13 is more extreme than $\pm t_{crit}$, which is a significantly different value. $\therefore$ we reject H_0 (which means that at $\alpha = 0.05$, there is a significant difference b/w observed & hypothesized mean).

But here we were blindly given n (the sample size). Usually, we decide on a α, β which gives a sample size estimate.

For the case when σ was known (one-tailed), $n = \left[\frac{Z_\alpha + Z_\beta}{\left(\frac{\mu_2 - \mu_1}{\sigma}\right)}\right]^2$ ⟶ For the case when σ is unknown (two-tailed) $\sigma \to s$, $z \to t$ $\quad n = \left[\frac{t_{\alpha, n-1} + t_{\beta, n-1}}{(\mu_2 - \mu_1)/s}\right]^2$

For a two-tailed test though, $\alpha \to \alpha/2$ $\quad n = \left[\frac{t_{\alpha/2, n-1} + t_{\beta, n-1}}{(\mu_2 - \mu_1)/s}\right]^2$

However, this is a problem since calculating $t_{\frac{\alpha}{2}, n-1}$ & $t_{\beta, n-1}$ requires knowledge of n, which we are supposed to estimate. So this is a recursive eqn. We also can't know what s is unless we know the sample size n.

So, we usually need some prior estimate of s based on some previous data to estimate n. Once we have that, we start with any random value for n, calculate $t_{\alpha/2, n-1}$ and $t_{\beta, n-1}$ and use it to find $\left(\frac{t_{\alpha/2, n-1} + t_{\beta, n-1}}{(\mu_2 - \mu_1)/s}\right)^2 = n$.

Now our input & output would generically be different. Use this new calculated value of n again in the formula, find the new n & compare with the old one. Repeat till you converge to a n s.t. input $n \approx$ output n.

So, usually, we report a # of such parameters for our statistical tests.

$t_{\alpha, df} = ?$ $\quad \alpha = ?$ $\quad d_{realised} = \frac{\bar{x} - \mu_H}{s} = ?$

$p = ?$ $\quad \beta = ?$

$\bar{x}$ = sample mean

s = sample std dev

However, all of this includes a big assumption, that the x is normally distributed $x \sim N(\mu, \sigma)$. This is because a t-variable is defined as follows-

$$t = \frac{z}{\sqrt{\frac{\chi^2}{d.f.}}}$$

(z → z-dist., χ^2 → χ^2-dist.)

since $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ is a z-distrib^n [for any distrib^n]

and $\frac{s^2}{\sigma^2}(n-1) = \chi^2$ is a χ^2-distrib^n [for $x \sim N(\mu, \sigma)$] CRUCIAL!

$$\Rightarrow \frac{\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{s^2}{\sigma^2}\frac{(n-1)}{(n-1)}}} = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

so, $\frac{s^2}{\sigma^2}(n-1) \sim \chi^2_{n-1}$ only for $x \sim N(\mu, \sigma) \Rightarrow \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{n-1}$ only for $x \sim N(\mu, \sigma)$

Lecture - 23

(03-03-2022)

Assumptions in One sample mean Hypothesis testing with the t-test-

- samples are independent.
- The original population should be normally distributed $X \sim N(\mu, \sigma)$

t-test is "robust" in the sense that tiny perturbations from normal distribution don't affect the results of t-test too much. Infact, more precisely, as long as the distrib^n remains symmetric under perturbations, t-test is fairly accurate.

Severe skew is bad for a t-test. But as long as distrib^n is symmetric, like even a uniform distrib^n, t-test is robust.

One sample test = determine, using your sample mean, if that sample came from a hypothesized distrib^n with hypothesized mean μ_H.

However, we don't usually do that in experiments, and most of the times we compare two different samples. (two sample test)

Two independent sample hypothesis test.

Say height of males $\sim N(\mu, \sigma_1 = 10)$ $\quad \sigma_1 = \sigma_2 = 10.$

height of females $\sim N(\mu, \sigma_2 = 10)$

We want to check if mean height of males & females is same or different.

i.e. if $\mu_1 \overset{?}{=} \mu_2$

Say we take samples of size n_1 & n_2 from pop^n of males & females respectively. From these, we calculate $\bar{X}_1$ & $\bar{X}_2$.

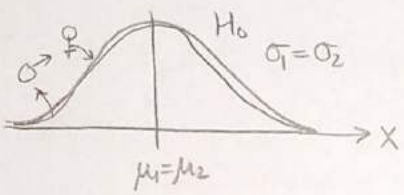
Our null hypothesis is $H_0: \mu_1 - \mu_2 = 0$ (diff. of means = 0)

Alternate hypothesis $H_A: \mu_1 - \mu_2 \neq 0$ (two-tailed) → other options are $\mu_1 - \mu_2 > 0$ or $\mu_1 - \mu_2 < 0$ (one-tailed tests)

$\alpha = 0.05$

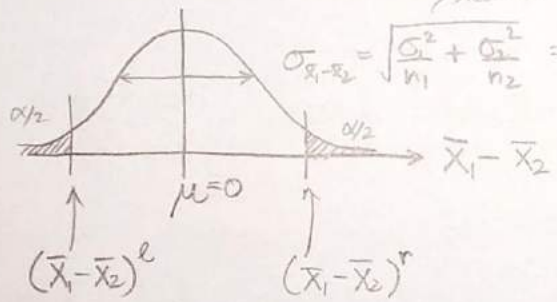
Choosing a value of β, we can calculate n_1 & n_2 & use that as sample size.
(think about how to do this?)

If H_0 is true, then the distrib^n for ♂ (M) and ♀ (F) is the same.



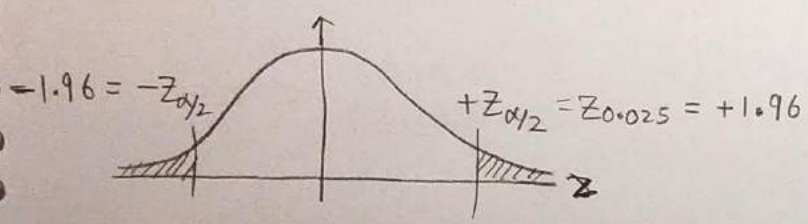
⇒ distrib^n of $(\bar{X}_1 - \bar{X}_2)$ should have $\mu = 0$ if H_0 is true.

$\sigma_1 = \sigma_2$
$n_1 = n_2$



renormalize into z-distrib^n

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{2}\,\sigma/\sqrt{n}}$$



Say $\bar{X}_1 = 175\,cm$ $\bar{X}_2 = 172\,cm$ $n_1 = n_2 = 9$
$\sigma_1 = \sigma_2 = 10$

$\Rightarrow Z_{calc} = \dfrac{(175-172)-(0)}{\sqrt{2}\cdot 10/\sqrt{9}}$ → value of z-distrib^n assuming H_0 true

$= \dfrac{(3-0)}{\sqrt{2}\cdot 10/3} = \dfrac{9}{\sqrt{2}\cdot 10} \approx 0.64$

Now $0.64 \in [-1.96, +1.96]$ i.e. it is not as extreme as $Z_{critical}$ values.

$\Rightarrow$ The difference b/w observed $(\bar{x}_1 - \bar{x}_2)$ (= 3) and hypothesized $(\mu_1 - \mu_2)$ (= 0) is not significant enough. $\therefore$ We do not reject H_0.

Thus, given the data, it is a valid possibility that $\mu_1 = \mu_2$. But we're not sure.

The samples must be independent for an independent sample test to be valid!

This assumption affects the calculation of $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$

Now say we have a case where we don't know what μ_1 & μ_2 are AND we don't know what σ_1 & σ_2 are (but $\sigma_1 = \sigma_2$).

Now how do we setup the test to check if the two pop^{n}s are identical?

So, $H_0 : \mu_1 - \mu_2 = 0$

$X_1 \sim N(\mu_1, \sigma_1)$
$X_2 \sim N(\mu_2, \sigma_2)$ $\quad \sigma_1 = \sigma_2 = ?$

So, estimate $\sigma_1 \approx s_1$, and $\sigma_2 \approx s_2$

Males $\xrightarrow{n_1}$ $\bar{x}_1, s_1$ Females $\xrightarrow{n_2}$ $\bar{x}_2, s_2$

So given $\sigma_1 = \sigma_2 = \sigma$, what should we use as $s_1 = s_2 = s$?

So, we use an average $s_p \equiv \dfrac{s_1^2(n_1-1) + s_2^2(n_2-1)}{n_1+n_2-2}$

(s_p: pooled variance)

For two sample test,

$$d_{\text{effect size}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p}$$

Lecture - 24

(07-03-22)

Two independent sample testing.

Say X_1 & X_2 are random variables representing the two samples.

(X_σ) $X_1 \sim N(\mu_1, \sigma)$

(X_φ) $X_2 \sim N(\mu_2, \sigma)$

Samples are assumed to be independent.

$n_1 = n_2$

$\sigma_1 = \sigma_2 = \sigma$ but unknown

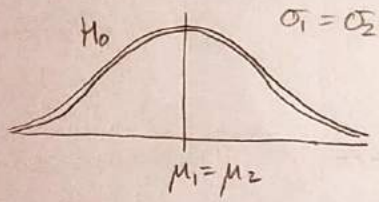
We want to test whether the means μ_1 & μ_2 of these distrib^{n}s are same or not.

H_0: $\mu_1 - \mu_2 = 0$ (no difference b/w X_1 & X_2 distrib^{n}s)

H_A: $\mu_1 - \mu_2 \neq 0$ (different distrib^{n}s) $\longrightarrow$ two-tailed.

Choose $\alpha = 0.05$

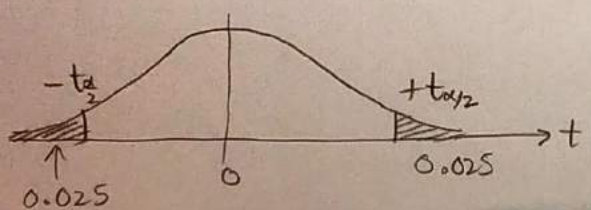
According to the H_0 distribution, X_1 & X_2 distrib^{n}s overlap



$\xrightarrow[n_2]{n_1}$ $\bar{X}_1, s_1$; $\bar{X}_2, s_2$ $\Big\}$ $(\bar{X}_1 - \bar{X}_2) \sim ?$

$\Rightarrow$ $\bar{X}_1 - \bar{X}_2$ has a mean of $\mu_1 - \mu_2 = 0$ and $s = \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

$\therefore$ $$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2} \quad (t\text{-distrib}^n)$$



If $n_1 = n_2 = 5$, then $\dfrac{(\bar{X}_1 - \bar{X}_2) - 0}{\sqrt{S_p^2\left(\frac{1}{5} + \frac{1}{5}\right)}} \sim t_8$ d.o.f. = 8

$t_{8,\frac{\alpha}{2}=0.025} = 2.306$

$\therefore$ $t_{critical}$ is ± 2.306

Knowing this, if we pull out samples & get $\bar{X}_1 = 175, s_1 = 10$
$\bar{X}_2 = 168, s_2 = 8$

$$t_{calc} = \frac{(175 - 168)}{\sqrt{82 \cdot \frac{2}{5}}} = 1.22$$

$$S_p^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2} = 4 \cdot \frac{164}{82} \; 82$$

$\therefore$ $t_{calc} \in [-t_{\frac{\alpha}{2}}, +t_{\frac{\alpha}{2}}] \Rightarrow$ We do not reject H_0 (or we say that the difference of means isn't statistically significant!)

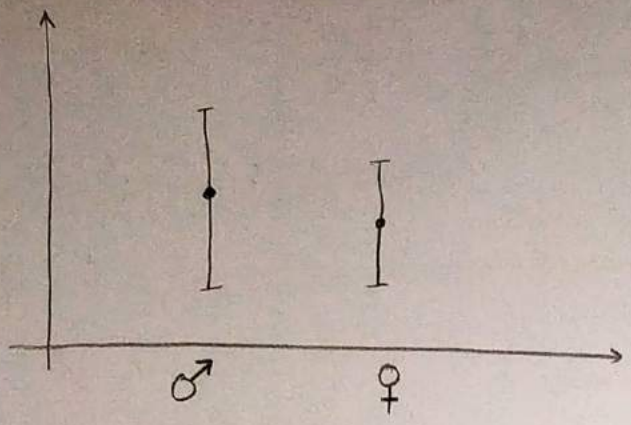
$$\text{Cohen's } d = \frac{(\bar{X}_1 - \bar{X}_2) - \overbrace{\mu_{H_1}}^{0}}{S_p} = \frac{(\bar{X}_1 - \bar{X}_2)}{S_p} = \frac{7}{\sqrt{82}} \approx 0.883$$

$$\left(S_p^2 = \frac{\sum_i (\bar{X}_1 - X_{1,i})^2 + \sum_j (\bar{X}_2 - X_{2,j})^2}{n_1 + n_2 - 2} \right)$$ → new type of variance of the entire sample set of 182

Good way to plot the results of this data analysis?

Mean, 95% C.I.

Contd. on next page.

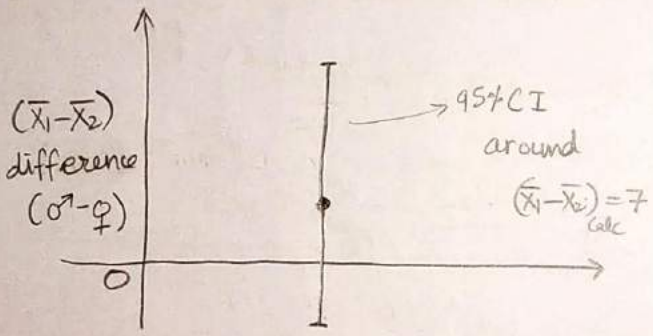


→ But this plot doesn't give any info. on the statistical test to determine differences in means.

For ♂, 95% C.I is ⇒ $\bar{X}_1 \pm t_{\frac{\alpha}{2}, n_1-1} \cdot \frac{S_1}{\sqrt{n_1}}$

For ♀, 95% of C.I. is ⇒ $\bar{X}_2 \pm t_{\frac{\alpha}{2}, n_2-1} \cdot \frac{S_2}{\sqrt{n_2}}$

A much better way is to plot the difference (♂ − ♀)



VISUAL HYPOTHESIS TESTING!

The 95% CI around $(\bar{X}_1 - \bar{X}_2)$ will be given by

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\frac{\alpha}{2}, n_1+n_2-2} \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$= 7 \pm (2.306) \cdot (5.78)$

$= 7 \pm 13.21$

SINCE $\bar{X}_1 - \bar{X}_2 = 0$ IS CONTAINED IN THE 95% C.I.

⇒ H_0 CANNOT BE REJECTED!

Lecture - 25.
(08-03-22)

Once the samples become correlated and aren't independent anymore, our assumptions are violated as well as the variance $s \neq \sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$ anymore since we'll also have covariance terms now.

In such cases, we do <u>paired sample t-tests</u>.

<u>Example</u> - a Before-After experiment.

Use the same set of individuals for the sample. Take data from them <u>before</u> taking a drug and <u>after</u> taking a drug.

Before	After		Difference (B−A)
X_{11}	X_{21}	take a difference →	Δ_1
X_{12}	X_{22}		Δ_2
X_{13}	X_{23}		Δ_3
⋮	⋮		⋮ → mean = $\overline{X}_\Delta$, std. dev. = s_Δ

To do the Hypothesis, take $H_0 : \mu_{Before} - \mu_{After} = 0$ (meds have no effect)

∴ restating it, H_0 : mean$(\Delta) = \overline{X}_\Delta = 0$ (or $\mu_B - \mu_A = 0$)

H_A : mean$(\Delta) \neq 0$ (or $\mu_B - \mu_A \neq 0$)

$\alpha = 0.05$

Now, $\dfrac{\overline{X}_\Delta - 0}{s_\Delta/\sqrt{n}} \sim t_{n-1}$ (0 → expected diff. b/w μ_A & μ_B assuming H_0 is true.)

Find $t_{calc} = \dfrac{\overline{X}_\Delta - 0}{s_\Delta/\sqrt{n}}$ from the sample and compare with the critical value $t_{\frac{\alpha}{2}, n-1}$.

Three samples.

$A \sim N(\mu_1, \sigma_1)$

$B \sim N(\mu_2, \sigma_2)$

$C \sim N(\mu_3, \sigma_3)$

$\sigma_1 = \sigma_2 = \sigma_3 = \sigma$ (unknown)

Take samples of sizes $n_1 = n_2 = n_3 = n$

$n_1 \longrightarrow \bar{X}_1, \; s_1$

$n_2 \longrightarrow \bar{X}_2, \; s_2$

$n_3 \longrightarrow \bar{X}_3, \; s_3$

Is atleast one of these means different from the other?

One possibility is to do pair-wise t-tests

A vs. B	$H_0^{(1)}: \mu_1 - \mu_2 = 0$	(α)
B vs. C	$H_0^{(2)}: \mu_2 - \mu_3 = 0$	(α)
C vs. A	$H_0^{(3)}: \mu_3 - \mu_1 = 0$	(α)

} 3 different pair-wise tests.

Total error in this pairwise test $= 3\alpha$???

Bon Ferroni correction $= \dfrac{\alpha}{\#\text{ of pairwise tests}} = \alpha'$

So, to keep the total error $= \alpha$, we put α value for all $H_0^{(i)}$ to be $\alpha' = \frac{\alpha}{3}$ in our example.

But decrease in $\alpha \Rightarrow$ increase in $\beta \Rightarrow$ Power $(1-\beta)$ decreases.

So, we'll need to increase sample sizes "n" to compensate.

But this method isn't very realistic as pairwise test for 4 samples would demand $\alpha' = \alpha/6$ & 6 separate tests. Better method - ANOVA

Lecture-26.

(21-03-2022)

A	B	C
5	5	7
5	7	8
6	7	8
7	6	9
6	5	9
$\bar{X}$ = 5.8	6	8.2

One way to compare the measure & compare the yields from fertilizers A, B, and C is to do pairwise tests

A vs. B
B vs. C
C vs. A

} Is atleast one of them different?

A better method is ANOVA (Analysis of variance)

ANOVA allows us to partition the variances in the problem, but it addresses questions about the means.

Assumptions-

1. Populations are normally distributed. $A \sim N(\mu_1, \sigma)$, $B \sim N(\mu_2, \sigma)$, $C \sim N(\mu_3, \sigma)$
2. Variances are equal.
3. Samples are independent.

Based on the # of factors (explainatory variables), we can have-

- One-Factor (One-way) ANOVA
- Two-Factor (Two-way) ANOVA
- ⋮
- n-Factor (n-way) ANOVA

Each factor can have multiple levels.

Say our yield was only dependent on one factor – fertilizers.
(one-way ANOVA)

And we have 3 different fertilizers → A, B, C

So under fertilizers, we have 3 levels.

Let's now add another factor, say concentration of fertilizers along with the type.

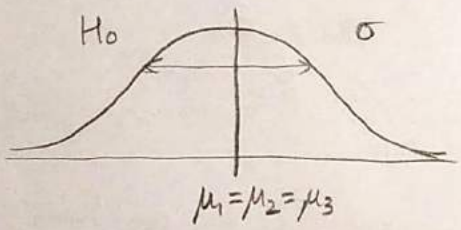
① Types of fertilizers → A, B, C
↑
factor-1

② Concentration → High, Low
↑
factor-2

This would be a two-factor ANOVA.

For the one-factor example (fertilizers A, B, C), the problem is setup as follows:

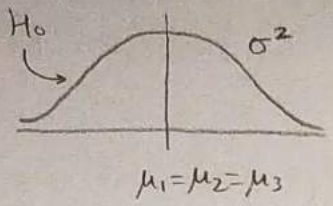
$H_0: \mu_1 = \mu_2 = \mu_3 = \mu$ (σ is same, but unknown)



$H_A: \mu_i \neq \mu_j$ for atleast one $i \neq j$

$\alpha = 0.05$

We'll come back to this setup in a while.



How can we estimate value of σ^2?

σ^2 can be estimated in two ways -

① First method estimates σ^2

② Second method estimates σ^2 only if H_0 is TRUE.
won't give σ^2 if H_0 is FALSE. Infact, the estimate $> \sigma^2$.

- **Method 1 : Variance Within**

From our data on first page, we have the data from a sample for fertilizers A, B & C. We pick up sample & calculate $\bar{X}$ and s^2.

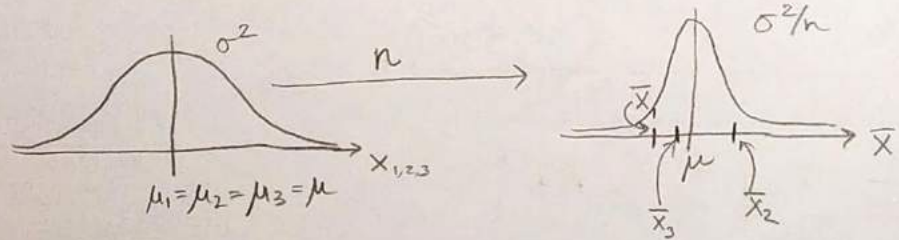
$n_1 \longrightarrow \bar{X}_1, s_1^2$
$n_2 \longrightarrow \bar{X}_2, s_2^2$
$n_3 \longrightarrow \bar{X}_3, s_3^2$

Now all three should estimate σ^2.
So which one do we choose?

$\Rightarrow$ We use s^2_{pooled} using s_1^2, s_2^2, s_3^2 to estimate σ^2.

- **Method 2 : Between Variance**

Let's say H_0 is true. Then taking samples of size $n_1=n_2=n_3=n$ for all three



Now $\frac{\sigma^2}{n}$ = between variance

$\therefore$ (b/w variance) $\cdot n \approx \sigma^2$

b/w variance = variance of sample means

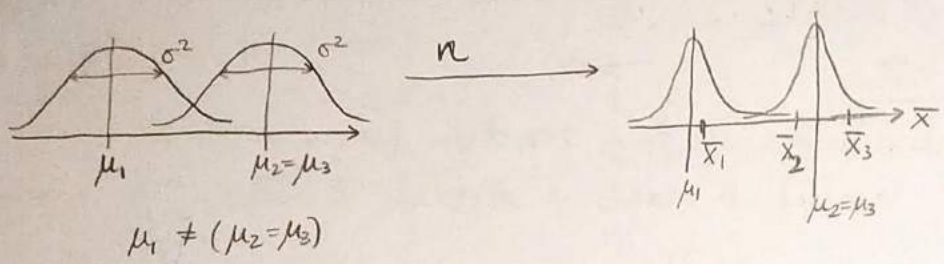
$= \text{Var}(\bar{X}_1, \bar{X}_2, \bar{X}_3)$

$= S_{\bar{X}}$

$\bar{X}_{1,2,3}$ = sample mean for fertilizer A, B, C.

$$\Rightarrow \underbrace{n S_{\bar{X}}^2}_{\text{variance between}} \approx \sigma^2$$

However if H_0 is FALSE,



$\mu_1 \neq (\mu_2 = \mu_3)$

The variance between $\equiv \text{Var}(\bar{X}_1, \bar{X}_2, \bar{X}_3) = S_{\bar{X}}^2 > \frac{\sigma^2}{n}$ always in this case because the $\bar{X}_1$ will be further away from $\bar{X}_2$ & $\bar{X}_3$ than if it would have come from the same distribution.

∴ If H_0 is true, then $n S_{\bar{X}}^2$ is a good estimate for σ^2

If H_0 is false, then $n S_{\bar{X}}^2$ overestimates σ^2

So, in the end, we'll have $S^2_{\text{pooled}}(S_1^2, S_2^2, S_3^2) \equiv V_W$

and $n \cdot S_{\bar{X}}^2 \equiv V_B$

	H_0 is TRUE	H_0 is FALSE
and $\frac{V_B}{V_W}$	$\frac{\sigma^2}{\sigma^2} = 1$	$\delta + \frac{\sigma^2}{\sigma^2} > 1$

↑ b/c V_B is an overestimate of σ^2 when H_0 is false.

But when is $\frac{V_B}{V_W} > 1$ due to H_0 being FALSE or just by chance?

Soln: ratio of variances ~ F-distribⁿ

The F-test here will be one-tailed test because one of the tails close to $\frac{V_B}{V_W} = F = 0$ will always support H_0 but we begin to doubt H_0 as $\frac{V_B}{V_W} = F > 1$.

Lecture-27

(22-03-2022)

Variance within - an average of the variance within each group i.e. for fertilizers A, B, C we get s_A^2, s_B^2, s_C^2 and $s_{pooled}^2 = s_{pooled}^2 (s_A^2, s_B^2, s_C)$

Variance between. - a measure of variance across/between the group means. This gives us an accurate measure of variance only when H_0 is true.
($V_B \geq \sigma^2$ if H_0 isn't true)

A	B	C
5	5	7
5	7	8
6	7	8
7	6	9
6	5	9
$\bar{X}_A = 5.8$	$\bar{X}_B = 6$	$\bar{X}_C = 8.2$

$\bar{X}_{Grand} = 6.67$

ANOVA Table

Source	s.s.	d.o.f.	MS	F	p
Factor (between)					
Error (within) variance					
Total					

S.S (sum of squares) for within variance

$$= (5-5.8)^2 + (5-5.8)^2 + \dots + (6-5.8)^2 +$$
$$+ (5-6)^2 + (7-6)^2 + \dots + (5-6)^2 + \quad = 9.6$$
$$+ (7-8.2)^2 + (8-8.2)^2 + \dots + (9-8.2)^2$$

$$SS_{within} = SS_{error} = \sum_{i \in \{A,B,C\}} \sum_{j=1}^{5} (x_{ij} - \bar{x}_i)^2$$

Now, $SS_{between}$ = sum of squares for b/w variance $= n \cdot \sum_{i \in \{A,B,C\}} (\bar{x}_i - \bar{x}_{Grand})^2$

$$= 5 \cdot \left[(5.8-6.67)^2 + (6-6.67)^2 + (8.2-6.67)^2 \right] = 17.7$$

Here $n_1 = n_2 = n_3$

If that wasn't the case, $SS_{between} = \sum_{i \in \{A,B,C\}} n_i (\bar{x}_i - \bar{x}_{Grand})^2$

Also, $SS_{total} = \sum_{i \in \{A,B,C\}} \sum_{j=1}^{5} (x_{ij} - \bar{x}_{Grand})^2$

$$SS_{total} = (5-6.67)^2 + (5-6.67)^2 + (6.67-6)^2 + \dots$$
$$+ \dots (9-6.67)^2 + (9-6.67)^2$$

It turns out that

$$SS_{total} = SS_{between/factor} + SS_{error}$$

$\therefore SS_{tot} = 27.3$

The next thing in the table is d.f. (degrees of freedom)

Each group has $n-1$ d.f.

∴ Total d.o.f for error is a sum of $n-1+n-1+n-1$ since we are summing over all of them

$$df_{error} = n-1+n-1+n-1 = 3(n-1) = 3\cdot(5-1) = 12 \text{ here}$$

In general, $\boxed{df_{error} = k\cdot(n-1)}$

k = # of groups (or levels) in the factor F

n = # of entries in a group in the factor

degrees of freedom for "between" the "factors"

$\boxed{df_{factor} = df_{between} = k-1}$ $\quad df_{factor} = 2$ here.

$\boxed{df_{total} = k\cdot n - 1}$

and $d.f._{total} = d.f._{error} + df_{factor}$

Roughly speaking, $\frac{\sum_i (x_i-\bar{x})^2}{n-1}$ ($\sum_i (x_i-\bar{x})^2 \to$ s.s., $n-1 \to$ d.f.) $\approx$ variance measure

∴ $\boxed{\text{Variance} \approx MS \text{ (Mean square)} \equiv \frac{s.s.}{d.f.}}$

$MS_{error} = \frac{SS_{error}}{df_{error}}$ $\qquad MS_{total} = \frac{s.s.\ total}{d.f.\ total}$

$MS_{factor} = \frac{SS_{factor}}{df_{factor}}$ $\qquad$ but $MS_{total} \neq MS_{error} + MS_{factor}$

Observe that $MS_{error} = s^2_{pooled}$

$$MS_{error} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2}{k \cdot (n-1)}$$ and this is $\equiv s^2_{pooled}$

Now $MS \approx \sigma^2$

So, $MS_{error} \approx \sigma^2 \approx s^2_{error}$

↑ also written as $s^2_{\bar{x}}$ or s^2_{within}

and $MS_{factor} \approx \sigma^2 \approx n\, s^2_{factor}$

↑ also written as S.E. or $s^2_{b/w}$

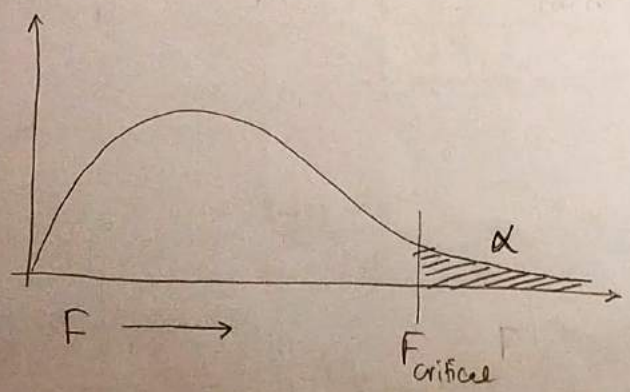
So, the F-value is calculated as

$$\frac{MS_{factor}}{MS_{error}} = F_{num\,df,\,den\,df} = F_{f\,d.f,\,e\,d.f.}$$

In our case $d.f._{factor} = 2$, $d.f._{error} = 12$

$MS_{factor} = 8.85$, $MS_{error} = 0.8$

$\Rightarrow F_{2,12} = \frac{MS_{factor}}{MS_{error}} \approx 11$



Now ideally $F \doteq 1$ for H_0 to not be rejected. But it occasionally might be > 1. But how much > 1 is acceptable before we consider A, B, C to be different distrib^{n}s?

We accept an error of α & say that with probability α, we have rejected H_0 even when if it's true, so we generally keep it small.

For $df_{num, den} = 2, 12 \Rightarrow F_{critical} = 3.89$ for $\alpha = 0.05$

Now $F_{calc} > F_{crit}$ $(11 > 3.89) \Rightarrow p < \alpha$

$\therefore$ We reject our null hypothesis H_0 i.e. atleast one of the data groups are different from the others.

We put all our α in the right tail only because $F < 1$ or $\sigma^2_{factor} < \sigma^2_{error}$

$\Rightarrow$ the variance b/w group means $\bar{x}_A, \bar{x}_B, \bar{x}_C$ is less than the variance within the groups themselves $\Rightarrow$ the distrib^{n}s must be fairly close to each other. Infact, it is a much stronger evidence for our H_0.

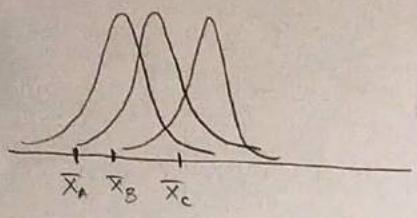
The MS_{error} & MS_{factor} measures effectively partition the variance of the system. This can be used to calculate the effect size

$$\text{Effect size} = \eta^2 = \frac{SS_{factor}}{SS_{total}}$$

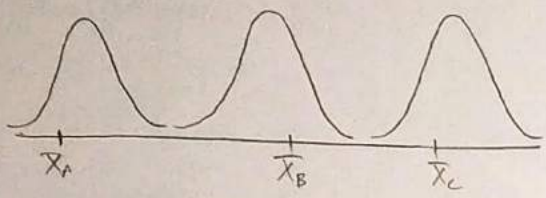
The higher the η^2, the higher are the chances of rejecting H_0.

SS_{factor} = ss value calculated <u>between</u> the factor means.

SS_{error} = ss value considered as an error present <u>within</u> a factor's group.



Low between variance (factor)
$\Rightarrow$ low η^2



High between variance (factor)
$\Rightarrow$ high η^2

In summary,

m = # of groups
n_i = # of entries in group i

Source	S.S.	d.f.	M.S.
Factor (between)	$\sum_{i=1}^{m} n_i (\bar{X}_i - \bar{X}_{grand})^2$	$m-1$	$MS_{factor} = \frac{SS_{factor}}{df_{factor}}$
Error (within)	$\sum_{i=1}^{m} \sum_{j=1}^{n} (X_{ij} - \bar{X}_i)^2$	$m \cdot (n-1)$	$MS_{error} = \frac{SS_{error}}{df_{factor}}$
Total	$\sum_{i=1}^{m} \sum_{j=1}^{n} (X_{ij} - \bar{X}_{grand})^2$	$m \cdot n - 1$	$MS_{total} = \frac{SS_{total}}{df_{total}}$
	$SS_{total} = SS_{factor} + SS_{error}$	$df_{total} = df_{factor} + df_{error}$	

$F_{calc} = \frac{MS_{factor}}{MS_{error}}$. compare with $F^{crit(\alpha)}_{df_{factor}, df_{error}}$

Lecture - 28

(23-03-2022)

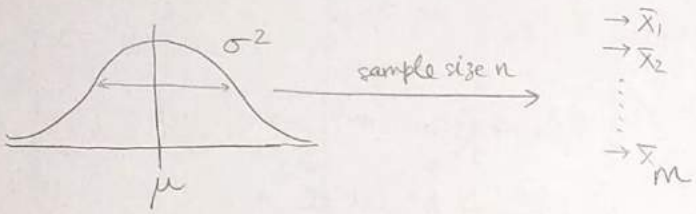
In ANOVA test, if we reject H_0, we do Post-Hoc tests.

Post hoc tests -

- Tukey's honest significant difference test. (HSD test)

The central idea is to figure which of the pairs (out of AB, BC, AC) are significantly different.

Tukey's HSD test depends on Tukey's studentized range. (Q-distrib")



m = # of groups in data table

$\bar{X}_{max} - \bar{X}_{min} \rightarrow$ range $(\bar{x})$

Then $$\frac{\bar{X}_{max} - \bar{X}_{min}}{\sqrt{\frac{MS_{within}}{n}}} \sim Q_{m,\, df_\omega}$$

	A	B	C	
	5	5	7	
	5	7	8	
	6	7	8	
	7	6	9	
	6	5	9	$\bar{X}_9$
$\bar{X}$	5.8	6	8.2	6.67

here, $df_\omega = 3.(5-1) = 12$

$k/m = 3$

$n = 5$

At $\alpha = 0.05$, we find the critical value Q_{crit} at $(3, 12)$

$$Q^{Crit(\alpha=0.05)}_{3,12} = 3.773$$

We define $HSD = Q^{crit(\alpha)}_{k, df_\omega} \cdot \sqrt{\frac{MS_\omega}{n}}$

$$Q^{crit(\alpha=0.05)}_{3,12} = 3.733 \quad \text{and} \quad MS_\omega = 0.8 \quad \text{and} \quad n = 5$$

$$\Rightarrow \underset{(\bar{x}_{max} - \bar{x}_{min})}{HSD} = 3.733 \sqrt{\frac{0.8}{5}} = 1.5$$

Now, we do pairwise differences & see if it's $>$ or $<$ HSD.

$\Delta_{AB} = |5.8 - 6| = 0.2 < HSD$ } A-B difference isn't significant.

$\Delta_{BC} = |6 - 8.2| = 2.2 > HSD$ } BC and AC differences are significant.

$\Delta_{AC} = |5.8 - 8.2| = 2.4 > HSD$ }

ANOVA Models (ways of interpreting the observed data)

- Factor effects → $X_{ij} = \mu + \tau_i + \varepsilon$ (τ_i: factor term; $i \in \{A, B, C\}$; μ: $\bar{x}_{grand}$; ε: error)

$\varepsilon \sim N(0, \sigma)$

In our case $\mu \approx \bar{x}_{grand} = 6.67$

$$\tau_i = \bar{x}_i - \bar{x}_{grand}$$

and $\varepsilon \sim N(0, \sigma)$ where $\sigma = V_i = s_i^2$

Due to this way of defining things, we can write null hypothesis as

$H_0: \tau_i = 0 \ \forall i$ and $H_A: \tau_i \neq 0$ for any i

INTRO. TO TWO-FACTOR ANOVA

We have two factors in our example now -

Factors	# of levels
Fertilizers	3 → A, B, C
Concentration	2 → High, Low

Fertilizers / Concentration	A	B	C	$\bar{X}_{\cdot j \cdot}$
Low Conc.	5	5	7	
	5	7	8	
	6	7	8	
	7	6	9	
	6	5	9	
$\bar{X}_{ij \cdot}$	5.8	6	8.2	6.67
High conc.	7	7	11	
	7	9	12	
	8	9	12	
	9	8	13	
	8	7	13	
	7.8	8	12.2	9.33
$\bar{X}_{i \cdot \cdot}$	6.8	7	10.2	8

X_{ijk}

$i \in \{A, B, C\}$

$j \in \{\text{low, high}\}$

$k \in \{1, 2, 3, 4, 5\}$

$\bar{X}_{\hat{i}\cdots}$ = mean over all $i = \hat{i}$.

$\bar{X}_{\cdot \hat{j} \cdot}$ = mean over all $j = \hat{j}$

$\bar{X}_{\hat{i}\hat{j}\cdot}$ = mean over all values with $i = \hat{i}$ & $j = \hat{j}$.

$\bar{X}_{i \cdot \cdot}$: 6.8 (mean of all A's), 7 (mean of all B's), 10.2 (mean of all C's)

$\bar{X}_{grand} = \bar{X}_{\cdots}$ = 8

$$\therefore \quad \bar{X}_{ijk} = \mu + \tau_i + \beta_j + \underbrace{\tau_i \beta_j}_{\text{interaction effect}} + \varepsilon$$

$\mu \approx \bar{X}_{grand} = \bar{X}_{...}$

$\tau_i \approx (\bar{X}_{i..} - \bar{X}_{...})$

$\beta_j \approx (\bar{X}_{.j.} - \bar{X}_{...})$

Lecture-29

(24-03-2022)

Two-factor ANOVA. (refer to the data on prev. page)

ANOVA table in this case looks like.

Source	S.S.	d.f.	MS	F	P
Factor 1 (fertilizer)	72.8	2	36.4		
Factor 2 (concentration)	53.06	1	53.06		
Interaction $F_1 \times F_2$	6.67	2	3.345		
Error	19.2	24	0.8		
Total	151.7	29	5.23		

For n-factor ANOVA with $n \geq 2$, we don't have a single null hypothesis

We have more than one H_0, each one for F_1, F_2 and $F_1 \times F_2$

① Let's start with the error row.

$i \in \{A, B, C\}$
$j \in \{low, high\}$
$k \in \{1, 2, 3, 4, 5\}$

$$SS_{error} = \sum_{i \in \{A,B,C\}} \cdot \sum_{j \in \{l,h\}} \sum_{k=1}^{5} (X_{ijk} - \bar{X}_{ij\cdot})^2$$

$$SS_{error} = (5-5.8)^2 + (5-5.8)^2 + \dots + (5-6)^2 + (7-6)^2 + \dots + (7-8.2)^2 + (8-8.2)^2$$
$$+ \dots + (7-7.8)^2 + (7-7.8)^2 + \dots + (7-8)^2 + (9-8)^2 + \dots + (11-12.2)^2$$
$$+ (12-12.2)^2 + \dots + (13-12.2)^2 = \underline{19.2}$$

$SS_{error} \equiv$ sum of squared errors for all groups.

② For Factor 1, the S.S. can be calculated as

$$SS_{F_1} = \sum_{i \in \{A,B,C\}} (\bar{X}_{i\cdots} - \bar{X}_{\cdots})^2 \cdot n_i \cdot (\#\text{ of levels in } F_2)$$

(n_i: sample size = 5)

$$= \underbrace{5 \cdot 2}_{10} \left[(6.8-8)^2 + (7-8)^2 + (10.2-8)^2 \right] = \underline{72.8}$$

③ For Factor 2,

$$SS_{F_2} = \sum_{j \in \{l,h\}} (\bar{X}_{\cdot j\cdot} - \bar{X}_{\cdots})^2 \cdot n_j \cdot (\#\text{ of levels in } F_1) = 5 \cdot 3 \left[\begin{matrix} (6.67-8)^2 \\ + (9.33-8)^2 \end{matrix} \right]$$

(n_j: sample size = 5)

$$= \underline{53.06}$$

$n_j \cdot (\#\text{ of levels in } F_1) = \#$ of entries over which the average $\bar{X}_{\cdot j\cdot}$ is performed

$n_i \cdot (\#\text{ of levels in } F_2) = \#$ of entries over which the average $\bar{X}_{i\cdots}$ is done

④ What about the interaction $F_1 \times F_2$?

observed — expected

$$\bar{X}_{ij\bullet} = \underbrace{\bar{X}_{\bullet\bullet\bullet}}_{\text{central mean}} + \underbrace{(\bar{X}_{i\bullet\bullet} - \bar{X}_{\bullet\bullet\bullet})}_{\text{deviation due to fertilizer } i\ (\tau_i)} + \underbrace{(\bar{X}_{\bullet j\bullet} - \bar{X}_{\bullet\bullet\bullet})}_{\text{deviation due to concentration } j\ (\beta_j)} + \text{interaction effects}$$

$$\therefore \text{Interaction effect} = \bar{X}_{ij\bullet} - [\mu + \tau_i + \beta_j] = \bar{X}_{ij\bullet} - [\bar{X}_{i\bullet\bullet} + \bar{X}_{\bullet j\bullet} - \bar{X}_{\bullet\bullet\bullet}]$$

and a measure of interaction variance would be

$$SS_{F_1 \times F_2} = n \sum_{ij} (\text{obs}_{ij} - \text{exp}_{ij})^2$$

$$= n \sum_{i \in \{A,B,C\}} \sum_{j \in \{L,M\}} \left[\bar{X}_{ij\bullet} - (\bar{X}_{i\bullet\bullet} + \bar{X}_{\bullet j\bullet} - \bar{X}_{\bullet\bullet\bullet})\right]^2$$

For $i=A$, $j=L$

$\bar{X}_{\bullet\bullet\bullet} = 8$

So, $\bar{X}^{(\text{expected})}_{A,L,\bullet} = 8 + \tau_A + \beta_L = 8 + (6.8-8) + (6.67-8)$

$= 5.47$

$\bar{X}^{(\text{observed})}_{A,L,\bullet} = 5.8$

$\Rightarrow \Delta\bar{X}_{A,L} \equiv \text{interaction}_{A,L} = (\bar{X}^{\text{obs}}_{A,L} - \bar{X}^{\text{exp}}_{A,L})$

$\Rightarrow \boxed{SS_{\substack{F_1 \times F_2 \\ (\text{int})}} = n \sum_{i \in \{A,B,C\}} \sum_{j \in \{L,M\}} \Delta\bar{X}_{ij}^2} = 6.67$

⑤ Total measure

$$SS_{\text{total}} = \sum_{i,j,k} (X_{ijk} - \bar{X}_{\bullet\bullet\bullet})^2 = 151.7$$

Let's move onto # of d-of-s now

① $df_{error} = \underbrace{(5-1) + (5-1) + \cdots + (5-1)}_{6 \text{ times}} = 6 \cdot (5-1) = 24$

$df_{error} = (n-1) \cdot l \cdot m$

n = sample size
l = # of levels for F_2 (conc.)
m = # of levels for F_1 (fertilizer)

② $df_{F_1} = m - 1 = 2$

$m-1$ → # of levels of F_1

③ $df_{F_2} = l - 1 = 1$

$l-1$ → # of levels of F_2

④ $df_{F_1 \times F_2} = df_{F_1} \times df_{F_2} = 1 \cdot 2 = 2$

⑤ $df_{total} = df_{error} + df_{F_1} + df_{F_2} + df_{F_1 \times F_2} = 29$

Mean squared values $MS = \frac{SS}{df}$

$MS_{F_1} = 36.4$
$MS_{F_2} = 53.06$
$MS_{F_1 \times F_2} = 3.345$
$MS_{error} = 0.8$
$MS_{total} = 5.23$

Now we calculate F values

$$F_{F_1} = \frac{MS_{F_1}}{MS_{error}} \qquad F_{F_2} = \frac{MS_{F_2}}{MS_{error}}$$

There are two different types of factor → Fixed factors
→ random factors

Lecture - 30
(28-03-2022)

For a $n \geq 2$ factor ANOVA, $\exists$ a series of hypotheses.

$H_0^{(1)}$: Factor 1 has no effect.
$H_0^{(2)}$: Factor 2 has no effect.
$H_0^{(1\times2)}$: No interaction exists b/w the two factors.

	F	df	Critical F ($F^{\alpha}_{df_{Fi}, df_{error}}$)		p
Factor 1	$\frac{MS_{F_1}}{MS_e} = 45.5$	2	$F^{0.05}_{2,24} = 3.4$		$p < \alpha$ Reject
Factor 2	$\frac{MS_{F_2}}{MS_e} = 66.3$	1	$F^{0.05}_{1,24} = 4.26$	$\Rightarrow$	$p < \alpha$ Reject
$F_1 \times F_2$	$\frac{MS_{F_1\times F_2}}{MS_e} = 4.16$	2	$F^{0.05}_{2,24} = 3.4$		$p < \alpha$ Reject.
Error		24			

Factors → Fixed
→ Random

description of the levels

This gives rise to different types of ANOVA tests.

Expected MS table	Model I (Fixed)	Model II (Random)	Model III (Mixed) $F_1 \to$ fixed, $F_2 \to$ random
F_1	$\sigma^2_{F_1} + \sigma^2_{error}$	$\sigma_e^2 + \sigma_{F_1}^2 + \sigma^2_{F_1 \times F_2}$	$\sigma_e^2 + \sigma^2_{F_1} + \sigma^2_{F_1 \times F_2}$
F_2	$\sigma^2_{F_2} + \sigma^2_{error}$	$\sigma_e^2 + \sigma_{F_2}^2 + \sigma^2_{F_1 \times F_2}$	$\sigma_e^2 + \sigma^2_{F_2}$
$F_1 \times F_2$	$\sigma^2_{F_1 \times F_2} + \sigma^2_{error}$	$\sigma_e^2 + \sigma^2_{F_1 \times F_2}$	$\sigma_e^2 + \sigma^2_{F_1 \times F_2}$
Error	σ^2_{error}	σ_e^2	σ_e^2

- **FIXED (MODEL-I)**

If H_0 is true, then $\sigma^2_{F_i} \to 0 \Rightarrow F_{F_i} = \dfrac{\sigma^2_{F_i} + \sigma_e^2}{\sigma_e^2} \approx 1$

If H_0 is false, then $\dfrac{\sigma^2_{F_i} + \sigma_e^2}{\sigma_e^2} > 1 \Rightarrow F_{F_i} > 1$

} For factor i, $i \in \{1, 2\}$

$$F_{F_1 \times F_2} = \frac{\sigma^2_{F_1 \times F_2} + \sigma_e^2}{\sigma_e^2}$$

So, for fixed (model-I) ANOVA, our F-values are calculated as

$$F_I = \frac{\sigma^2_{F_1/F_2/F_1 \times F_2} + \sigma_e^2}{\sigma_e^2} = \frac{M.S._{F_1/F_2/F_1 \times F_2}}{M.S._{Error}}$$

- **RANDOM (MODEL-II)**

For model-II (random), the F-values are measured a little differently since MS_{F_1} & MS_{F_2} also have an interaction component now.

For Factors 1 & 2

$$F_{F_i} = \frac{MS_{F_1/F_2}}{MS_{F_1 \times F_2}} = \frac{\sigma_e^2 + \sigma^2_{F_1 \times F_2} + \sigma^2_{F_i}}{\sigma_e^2 + \sigma^2_{F_1 \times F_2}}$$

If $H_o^{(i)}$ is true, $\sigma^2_{F_i} \to 0$ & $F_{F_i} \approx 1$.

If $H_o^{(i)}$ is false, $\sigma^2_{F_i} \neq 0$ & $F_{F_i} > 1$

For the interaction though,

$$F_{F_1 \times F_2} = \frac{MS_{F_1 \times F_2}}{MS_{error}} = \frac{\sigma_e^2 + \sigma^2_{F_1 \times F_2}}{\sigma_e^2}$$

- MIXED (MODEL-III) ANOVA

Say $F_1 \to$ fixed & $F_2 \to$ random

F_1 – $\sigma_e^2 + \sigma^2_{F_1} + \sigma^2_{F_1 \times F_2}$

F_2 – $\sigma_e^2 + \sigma^2_{F_2}$

$F_1 \times F_2$ – $\sigma_e^2 + \sigma^2_{F_1 \times F_2}$

error – σ_e^2

$\Rightarrow$

$$F_{F_1} = \frac{MS_{F_1}}{MS_{F_1 \times F_2}} = \frac{\sigma_e^2 + \sigma^2_{F_1} + \sigma^2_{F_1 \times F_2}}{\sigma_e^2 + \sigma^2_{F_1}}$$

$$F_{F_2} = \frac{MS_{F_2}}{MS_e} = \frac{\sigma_e^2 + \sigma^2_{F_2}}{\sigma_e^2}$$

$$F_{F_1 \times F_2} = \frac{MS_{F_1 \times F_2}}{MS_e} = \frac{\sigma_e^2 + \sigma^2_{F_1 \times F_2}}{\sigma_e^2}$$

Since the numerator & denominator in F depend upon the type of factors, the critical F's dependence on # of d.f.'s also changes!

F-values

	Model-I (fixed)	Model-II (random)	Model-III (F_1 → fixed, F_2 → random, Mixed)
F_1	$\frac{MS_{F_1}}{MS_e}$	$\frac{MS_{F_1}}{MS_{F_1 \times F_2}}$	$\frac{MS_{F_1}}{MS_{F_1 \times F_2}}$
F_2	$\frac{MS_{F_2}}{MS_e}$	$\frac{MS_{F_2}}{MS_{F_1 \times F_2}}$	$\frac{MS_{F_2}}{MS_e}$
$F_1 \times F_2$	$\frac{MS_{F_1 \times F_2}}{MS_e}$	$\frac{MS_{F_1 \times F_2}}{MS_e}$	$\frac{MS_{F_1 \times F_2}}{MS_e}$

Critical-F values.

	Model-I (fixed)	Model-II (random)	Model-III (mixed) (F_1-fixed, f_2-random)
F_1	$F^{\alpha}_{df_{F_1}, df_e}$	$F^{\alpha}_{df_{F_1}, df_{F_1 \times F_2}}$	$F^{\alpha}_{df_{F_1}, df_{F_1 \times F_2}}$
F_2	$F^{\alpha}_{df_{F_2}, df_e}$	$F^{\alpha}_{df_{F_2}, df_{F_1 \times F_2}}$	$F^{\alpha}_{df_{F_2}, df_e}$
$F_1 \times F_2$	$F^{\alpha}_{df_{F_1 \times F_2}, df_e}$	$F^{\alpha}_{df_{F_1 \times F_2}, df_e}$	$F^{\alpha}_{df_{F_1 \times F_2}, df_e}$

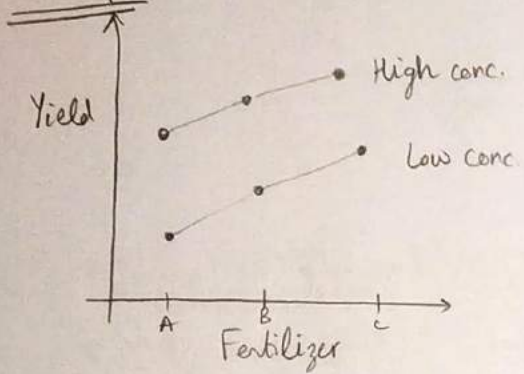
In our example, we rejected $H_0^{(F_1 \times F_2)}$, $H_0^{(F_1)}$ and $H_0^{(F_2)}$

Rejection of $H_0^{(F_1)}$ $\Rightarrow$ Atleast one of the three fertilizers gives a different mean yield.

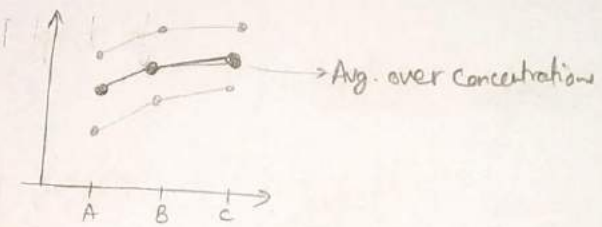
Rejection of $H_0^{(F_2)}$ $\Rightarrow$ Atleast one of the concentration has a different mean yield.

Rejection of $H_0^{(F_1 \times F_2)}$ $\Rightarrow$ $\exists$ an interaction b/w the factors.

Example 1



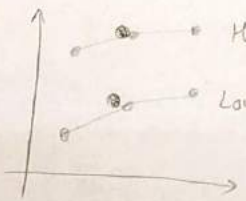
Q Effect of fertilizer? Average over concentrations.



Dependence on Fertilizers? YES!

If the lines are parallel, then $\nexists$ any interaction b/w F_1 & F_2.

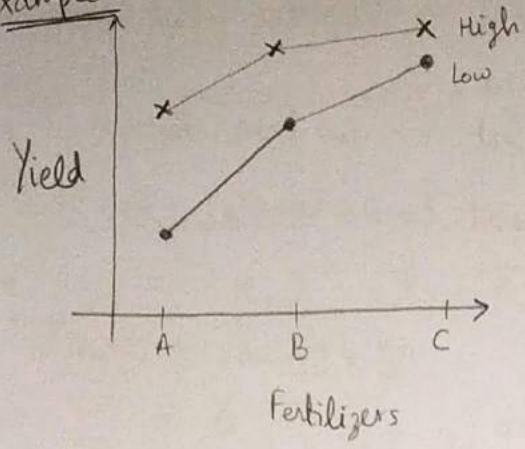
Q Effect of conc.? Avg. over fertilizers

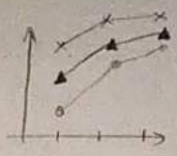


Dependence on concentration? YES!

Effect of F_1? Average over F_2.

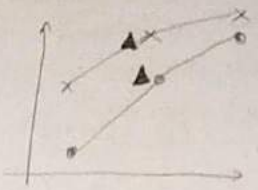
Example-2





Average over conc.

⇒ Effect of fertilizer? YES!



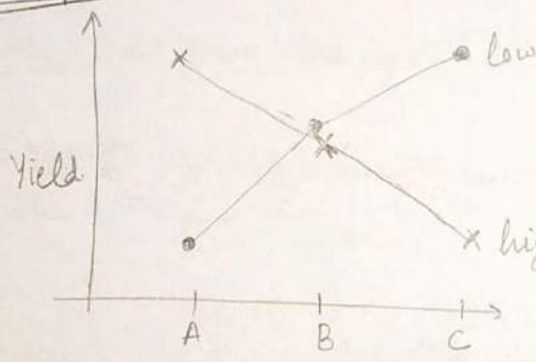
Average over fertilizer.

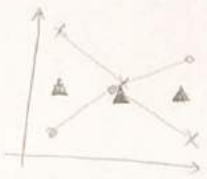
⇒ Effect of conc.? YES!

But lines aren't parallel.

⇒ ∃ an interaction between concentration and fertilizer.
(F_2) (F_1)

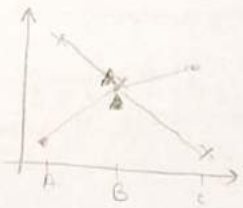
Example-3





Avg. over conc.

⇒ Effect of fertilizer? NO!



Avg. over fertilizer

⇒ Effect of conc.? NO!

Lines aren't at all parallel!

⇒ ∃ a significant interaction.

So, by effect of fertilizer averaged over the levels of other factor (conc.) doesn't exist. Same for concentration's effect

The reason is that the strong interaction cancels the average effect of each other

Lecture - 31
(29/03/2022)

Let's consider this dependent sample test i.e. we take observations for the same group of people but at different time intervals.

Subjects	Pre-exercise	After 3 months	After 6 months	Subject Mean
1	45	50	55	50
2	42	42	45	43
3	36	41	43	40
4	39	35	40	38
5	51	55	59	55
6	44	49	56	49.7
Time mean	42.8	45.3	49.7	

Grand mean $\mu_g = 45.9$

Due to dependent samples, we don't use standard ANOVA. We use **Repeated Measures ANOVA.**

Assumption
(1) samples are normally distributed
(2) The variance remains the same.

→ ANOVA table

Source	S.S.	d.f.	M.S.	F	p	
Time						Time - Factor
Subjects						
Error						
Total						

In one-factor ANOVA, $SS_{within} = SS_{error}$

But in repeated measures ANOVA, $SS_w = SS_{error} + SS_{subjects}$

S.S calculations.

- $SS_T = n_s \sum_t (\bar{X}_t - \bar{X}_G)^2$

  (n_s = # of subjects)

  Dummy variable t takes a sum over all time intervals (pre, 3 month, 6 month)

$$= 6 \cdot \left[(42.8 - 45.9)^2 + (45.3 - 45.9)^2 + (49.7 - 45.9)^2 \right]$$

$$= 143.44$$

- $SS_{subject} \equiv SS_s = n_T \sum_s (\bar{X}_s - \bar{X}_G)^2$

  (n_T = # of time points.)

$$= 3 \left[(50 - 45.9)^2 + (43 - 45.9)^2 + \dots + (49.7 - 45.9)^2 \right]$$

$$= 658.3$$

- $SS_{error} = SS_w - SS_{sub.}$

Remember,

$$SS_w = \sum_{j \in subj.} \sum_{i \in time} (x_{ij} - \bar{X}_j)^2$$

$$= (45 - 42.8)^2 + (42 - 42.8)^2 + \dots + (44 - 42.8)^2$$
$$+ (50 - 45.3)^2 + (42 - 45.3)^2 + \dots + (49 - 45.3)^2$$
$$+ (55 - 49.7)^2 + (45 - 49.7)^2 + \dots (56 - 49.7)^2$$

$$= 715.5$$

$$\Rightarrow SS_{error} = 715.5 - 658.3 = 57.2$$

$$\Rightarrow SS_{total} = SS_T + \underbrace{SS_s + SS_e}_{SS_w} = SS_T + SS_w = 143.44 + 715.5$$

$$= 858.9$$

degrees of freedom (d.f. calculation)

- $d.f._T$ = # of time points (# of time columns) -1
  $= 3-1 = \underline{2}$

- $d.f._{subj}$ = # of subjects -1 $= 6-1 = \underline{5}$

- $d.f._{error} = \underbrace{(n-1)}_{df_T} \cdot \underbrace{(k-1)}_{df_{subj}} = 2 \cdot 5 = \underline{10}$

looks like an interaction from 2-factor ANOVA

$\Rightarrow df_{Total} = 2+5+10 = 17$

M.S. calculations.

$MS_i \equiv \frac{SS_i}{df_i}$

- $MS_T = \frac{143.4}{2} = \underline{71.7}$

- $MS_{subj} = \frac{658.3}{5} = \underline{131.6}$

$MS_{error} = \frac{57.2}{10} = \underline{5.72}$

$MS_{total} = \frac{858.9}{17} = \underline{50.53}$

F-ratio calculations.

$$F_T = \frac{MS_T}{MS_{error}} = 12.5$$

$$F_{Subj} = \frac{MS_{subj}}{MS_{error}} = 23$$

and $F_T^{crit} = F^{\alpha}_{2,10} = 4.1$ $(\alpha = 0.05)$

$F_S^{crit} = F^{\alpha}_{5,10} = 3.33$ $(\alpha = 0.05)$

$\therefore$ In both cases, we reject the null hypothesis H_0.

$\Rightarrow$ We conclude that there is a significant effect of time as well as subjects!

(data for atleast one time is different from the others
data for atleast one of the subjects is diff. from the others)

Lecture - 32

(30-03-2022)

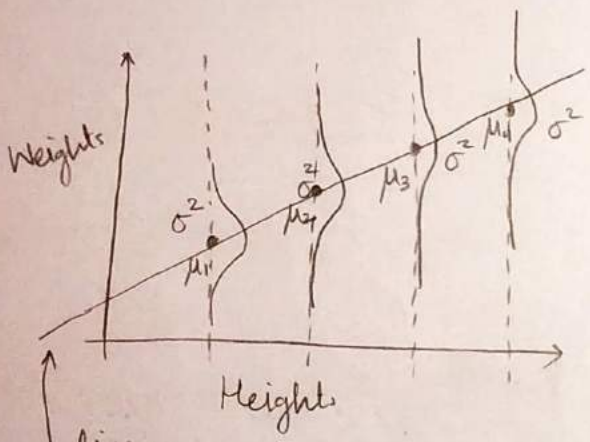
Correlation

We want to test if height & weight are correlated.

⟶ how to quantify correlation?

⇒ Pearson's product moment correlation coefficient.

(Pearson's "r")



linear relation b/w height & weight.

If we look at the entire pop^n, at every value of height, we have a normal distrib^n of weights.

However, the σ^2 of the weight distrib^n at each value of height is same.

Assumptions in correlation calcs.-

1. relationship is linear ($y = m \cdot x + c$)
2. $\forall x \; \exists y$ s.t. $y \sim N(\mu(x), \sigma^2)$
3. σ^2 is the same $\forall x$.
4. Samples are independent.

However, we don't have the access to means $\mu(x)$ for every value of x (height). We only have access to samples from the distribⁿ $y \sim N(\mu(x), \sigma^2)$, and we have to estimate the best fit line using this. This is called regression.

Let's do the regression and correlation on the following x-y dataset.

x	y
2	8
3	10
4	10
5	12
6	13
7	14
$\bar{x} = 4.5$	$\bar{y} = 11.16$

$$\sum_i (x_i - \bar{x})^2 = 17.5$$

$$\sum_i (y_i - \bar{y})^2 = 24.86$$

$$\sum_i (x_i - \bar{x})(y_i - \bar{y}) = 20.5$$

$$\Rightarrow r = 0.98$$

Pearson's coeff

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_j (y_j - \bar{y})^2}}$$

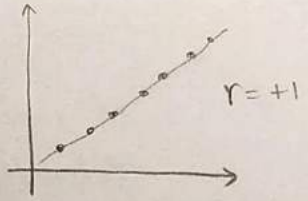
$r \in [-1, 1]$

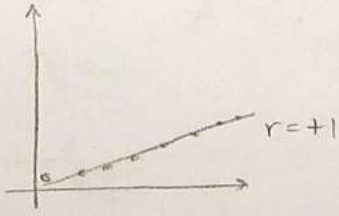
$r = -1 \Rightarrow$ Perfect negative relation

$r = 0 \Rightarrow$ No relation.

$r = +1 \Rightarrow$ Perfect positive relation

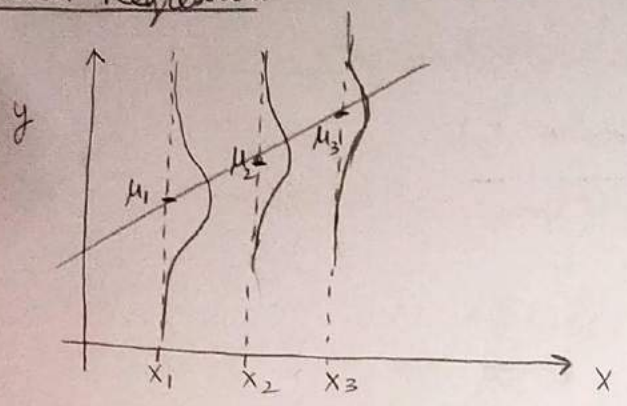
Disclaimer: $r \neq$ slope. r represents the strength of correlation. (if datapts. sit on the line or not.)





Lecture - 33
(31-03-2022)

Linear Regression



For every value x, the corresponding $y \in N(\mu(x), \sigma^2)$.
This means that the data is not necessarily linear when we look at a sample & is generally scattered.

We wish to get the line joining means. It is given by

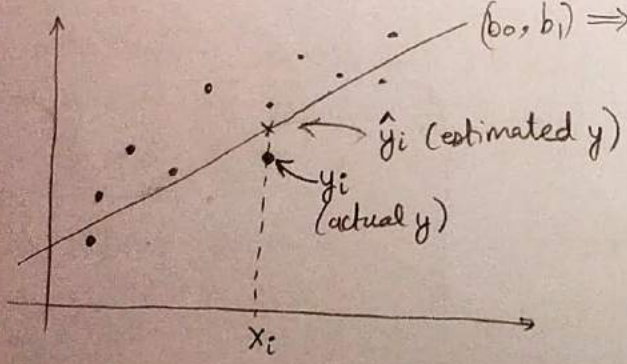
$$\mu_y = \beta_0 + \beta_1 X$$

(β_0: y-intercept, β_1: slope)

But we don't know what μ_y's are. So our goal is estimating β_0 & β_1 without knowing μ_y's.

Best-Fit line

Let's say our guess for (β_0, β_1) is (b_0, b_1)



We want to reduce the difference for all y's & $\hat{y}$'s

$\Rightarrow$ minimize $\sum_i (y_i - \hat{y}_i)^2$

The best fit line is defined to give least squared deviation.

One can derive that the estimands b_0 & b_1 are -

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{\sum_i (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

For yesterday's dataset, $\sum_i (x_i - \bar{x})(y_i - \bar{y}) = 20.5$

$$\sum_i (x_i - \bar{x})^2 = 17.5$$

$$\Rightarrow b_1 = \frac{20.5}{17.5} = \underline{1.17}$$

$$\Rightarrow b_0 = (11.16) - (1.17) \cdot (4.5) = \underline{5.89}$$

$$b_0 \approx \beta_0$$

$$b_1 \approx \beta_1$$

Lecture - 34

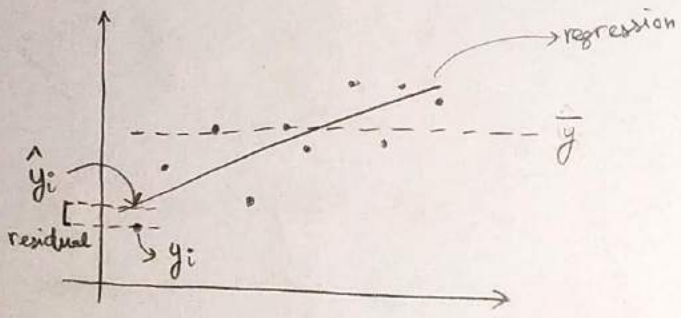
(04-04-2022)

Ultimately, we are using linear regression to find the relationship b/w the x & y variables. We can pose this question in terms of a hypothesis test, whether $\beta_1 = 0$ or $\beta_1 \neq 0$
(independent) (dependent)

We are estimating β_1 by b_1. $\beta_1 \approx b_1$

So, H_0: $\beta_1 = 0$

H_A: $\beta_1 \neq 0$



F - table.

Source	SS	df	MS	F
Regression				
Residual (error)				
Total				

SS calculation

$$\bar{y} = \langle y \rangle = \sum_{i=1}^{n} \frac{y_i}{n}$$

$\hat{y}$ → predicted y (by regression)

y → observed y values.

- $SS_{res} = \sum_i (y_i - \hat{y}_i)^2$
- $SS_{reg} = \sum_i (\hat{y}_i - \bar{y})^2$
- $SS_{total} = SS_{reg} + SS_{res}$.

degrees of freedom.

df for regression is always 1.

- $df_{res} = n-2$
- $df_{reg} = 1$

$\Rightarrow df_{tot} = n-2+1 = n-1$

MS calculation

$$MS_{reg} = \frac{SS_{reg}}{df_{reg}}$$

$$MS_{res} = \frac{SS_{res}}{df_{res}}$$

$$MS_{total} = \frac{SS_{total}}{df_{total}}.$$

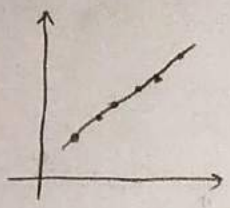
F-values

- $F_{reg} = \frac{MS_{reg}}{MS_{res}}$ and $F^{crit.} = F^{\alpha}_{1,n-2}$

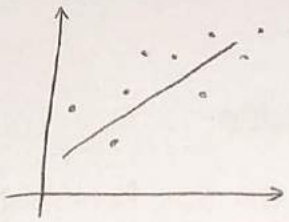
In ANOVA, your independent variable is a categorical variable.
In regression, the independent variable is continuous.

We define something analogous to effect size for regression as well

$$r^2 = \frac{SS_{reg}}{SS_{total}}$$



$SS_{residue} = 0$
$\Rightarrow r^2 = 1$



$r^2 < 1$

The notation is suggestive & r is literally the Pearson's coefficient!

$$r = \sqrt{r^2} = \sqrt{\frac{SS_{reg}}{SS_{tot}}}$$

(r (Pearson's coeff))

All our tests till now have been parametric tests which try to estimate a parameter of the population assuming some distribution.

Non-parametric tests - ??
(distrib^n-free tests)

Lecture - 35

(05/04/2022)

A	B
122	
123	
135	
150	
	153
	160
	173
175	
	180
	185
	190

Say we don't have any reason to believe that the pop^n distrib^n $\sim N$.

$\Rightarrow$ we do distribution-free tests. (non-parametric tests)

Based upon ranking of data.

→ A & B are two independent samples & the pop^n distrib^n is unknown!

The first test which dealt with this was the Wilcoxin's rank sum test. It was improved by Mann & Whitney $\Rightarrow$ Mann-Whitney U-test

This is a non-parametric equivalent of 2 independent-sample t-test.

A	B	Rank A	Rank B
122		1	
123		2	
135		3	
150		4	
	153		5
	160		6
	173		7
175		8	
	180		9
	185		10
	190		11
		$R_1 = 18$	$R_2 = 48$

H_0: Medians are same. (median A = median B)

H_A: Median are different.

OR

H_0: distrib^{n}s are same.

H_A: distrib^{n}s are different.

$\alpha = 0.05$

R_1, R_2 = Rank sums

Calculate U_1, U_2

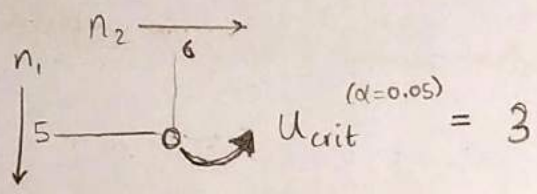
$n_1, n_2 \rightarrow$ sample sizes of A, B
$R_1, R_2 \rightarrow$ rank sums.

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 27$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 = 3$$

$$U_{\underset{calc}{\equiv}} \min(U_1, U_2) = 3$$

U_{crit} is calculated from a Mann-Whitney U-table. ($\alpha = 0.05$)



$U_{calc} > U_{crit}^{\alpha=0.05} \Rightarrow$ Do not reject H_0.

$U_{calc} \leq U_{crit}^{\alpha=0.05} \Rightarrow$ Reject H_0.

Here $U_{calc}(3) \leq U_{crit}(3) \Rightarrow$ We reject H_0.

Also notice, $\boxed{U_1 + U_2 = n_1 \cdot n_2}$

Let's take a simple example where $n_A = n_1 = 3$ & $n_B = n_2 = 3$.

Both groups have the same # of ranks.

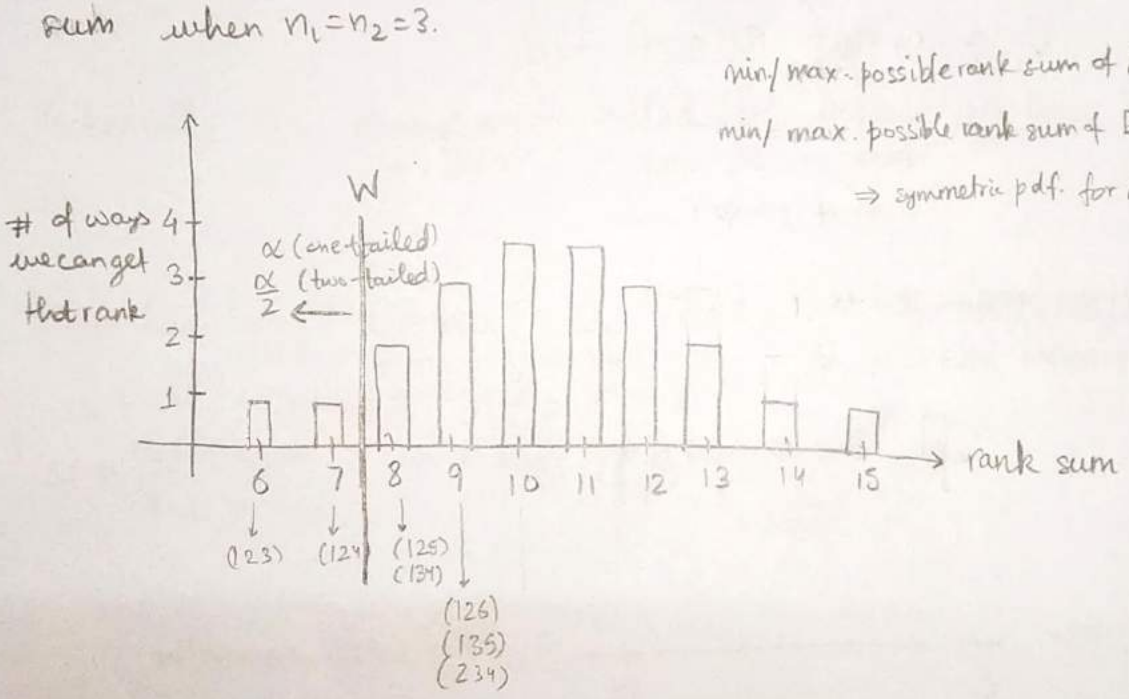
Depending upon the ranks in A or B, the total ranks R_1 & R_2 can vary.

Example-

A	B
1	4
2	5
3	6
6	15

A	B
1	3
2	5
4	6
7	16

Let's now find the # of ways we can get a particular rank sum when $n_1 = n_2 = 3$.



Probability distrib" of rank sums assuming H_0 is true.

If we have $n_1 = 3$ and $n_2 = 4$ $\quad (n_1 \neq n_2)$

A	B
1	4
2	5
3	6
	7

min. rank sum of A = 6 (123)
min. rank sum of B = 10 (1234)

max. rank sum of A = 18 (5 6 7)
max. rank sum of B = 22 (4 5 6 7)

for A, ranks $\in [6, 18]$
for B, ranks $\in [10, 22]$

In such a case, we can't have a W.

The U values

$$U_i = \underbrace{n_1 n_2 + \frac{n_i(n_i+1)}{2}}_{\text{max possible rank sum of group } i} - \underbrace{R_i}_{\text{the observed rank sum}}$$

minimum possible value of $U_i = 0$

max. possible value of U_i = max. rank-sum of group i − min. rank sum of group i

= range of rank-sums = $18 - 6$ (for A) = $22 - 10$ (for B) = 12

So,

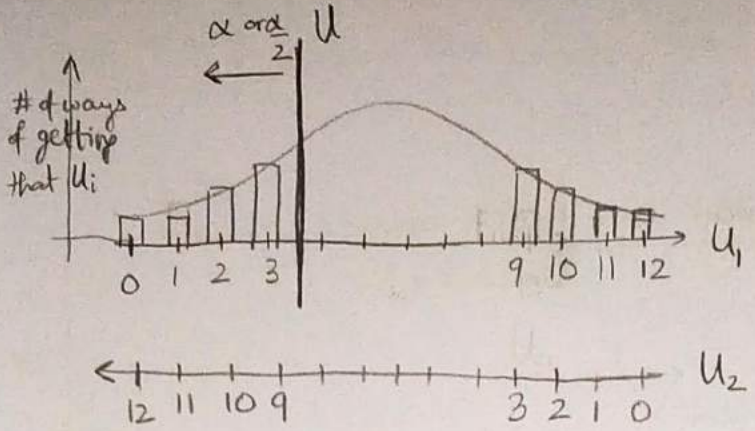
U_i $\quad i = A, B$

The scales are now the same now.

If $U_1 = 0$, $U_2 = n_1 n_2 - U_1 = 12 - 0 = 12$.

If $U_1 = 12$, $U_2 = n_1 n_2 - U_1 = 12 - 12 = 0$

$$U_1 + U_2 = n_1 n_2$$



The drawback of Wilcoxon was exactly that it needed $n_1 = n_2$.

Mann-Whitney renormalizes both groups to the same scale hence making hypothesis testing achievable.

Technically, U is always a one tailed test because we have $U = \min(U_1, U_2)$

If we had defined $U \equiv \max(U_1, U_2)$ instead of min, the only change is that the condition for rejecting H_0 gets reversed!

Lecture - 36

(06 - 04 - 2022)

Non-parametric test equivalent to a paired sample t-test. (dependent)

- Wilcoxon's signed rank test

Before	After	$\lvert\Delta\rvert$	Ranks	sign (B-A)
85	75	10	3	+
70	50	20	6	+
40	50	10	3	−
65	40	25	7	+
80	20	60	8	+
75	65	10	3	+
55	40	15	5	+
20	25	5	1	−

$n = 8$

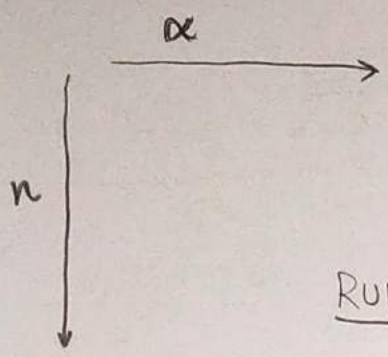
If we get a tie, we do an average of the ranks the numbers were supposed to take, so here we have 10 at rank 2, 3 & 4, so we take $\frac{2+3+4}{3} = \frac{9}{3} = 3$ ← avg. rank. The next rank to be taken would be 5.

Signed rank sum → $-\ (3+1) = 4$

→ $+\ (3+6+7+8+3+5) = 32$

$W = \min_{\text{Calc}} (R+, R-) = \underline{4}$

$W_{crit}^{\alpha=0.05}$ gives us the critical value from a W-table



$W_{crit}\ (n=8,\ \alpha=0.05) = 3$

RULE:

$W_{calc} \leq W_{crit} \Rightarrow$ Reject H_0.

$W_{calc} > W_{crit} \Rightarrow$ Do not reject H_0.

Here $W_{calc}\ (4) > W_{crit}\ (3) \longrightarrow$ we don't reject H_0.

The null hypothesis here is -

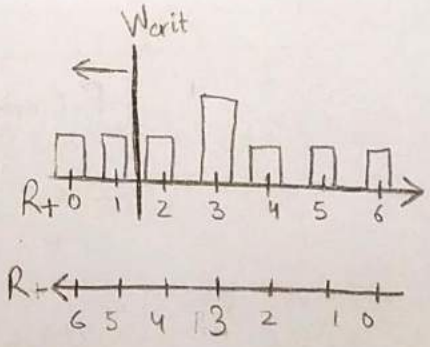
H_0: The distrib^{n}s are the same.

H_A: The distrib^{n}s are NOT the same

Why does this work?

Assume $n=3$ & the Δ values are distinct.

Ranks	sign possibilities (B−A)							
1	+	−	+	−	+	−	+	−
2	+	+	−	−	+	+	−	−
3	+	+	+	+	−	−	−	−
R_+	6	5	4	3	3	2	1	0
$R-$	0	1	2	3	3	4	5	6



Non parametric equivalent of 2-sample ANOVA

- **Kruskal-Wallice ANOVA**

A	B	C
120		
	121	
122		
123	124	
125	126	
	127	
128		
	129	
		130
		131
		132
		133
		134

Ranks

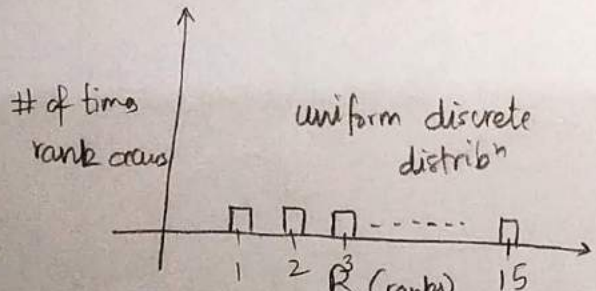
A	B	C
1	2	11
3	5	12
4	7	13
6	8	14
9	10	15
Rank sums: 23	32	65
$\bar{R}_j$: 4.6	6.4	13

$j = A, B, C$

$\bar{R}_{..} = 8$

$\bar{R}_{\cdot j} \equiv \frac{\sum_i R_{ij}}{n_j}$ → mean ranks

$\bar{R}_{..} = \frac{\sum_{ij} R_{ij}}{n_i \times \#\text{ of gps.}}$ = grand mean $\bar{R}_G$



$\Rightarrow \bar{X} = \frac{N+1}{2} = \frac{15+1}{2} = 8$

$S^2 = \frac{n^2-1}{12}$

The ratio $\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$ where S^2 = sum of squares

$$\frac{\overbrace{\sum_j n_j(\bar{R}_{\cdot j} - \bar{R}_{\cdot\cdot})^2}^{S^2}}{\underbrace{\frac{N^2-1}{12}}_{\sigma^2}} \sim \chi^2_{k-1}$$

$$\Rightarrow \frac{\overbrace{\sum_j n_j \left(\bar{R}_{\cdot j} - \frac{N+1}{2}\right)^2}^{SS_{b/w\ groups}}}{\frac{N^2-1}{12}} \sim \chi^2_{k-1}$$

k = # of groups

$\frac{N^2-1}{12}$ = variance of the discrete uniform distribⁿ

$$\boxed{\frac{\overbrace{\sum_j n_j \left(\bar{R}_{\cdot j} - \frac{N+1}{2}\right)^2}^{SS_{b/w\ groups}}}{\frac{N^2-1}{12}} \cdot \left(\frac{N-1}{N}\right) = H \sim \chi^2_{k-1}}$$

A simplified form looks like-

$$H = \frac{12}{N(N+1)} SS_{R\ b/w\ gp.}$$

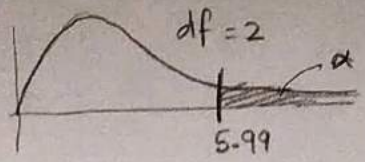
$N = n \cdot k$
$= 5 \cdot 3 = 15$

Here, $SS_{R\ b/w\ group} = 5\cdot(4.6-8)^2 + 5\cdot(6.4-8)^2 + 5\cdot(13-8)^2$

$= 5[(4.6-8)^2 + (6.4-8)^2 + (13-8)^2]$

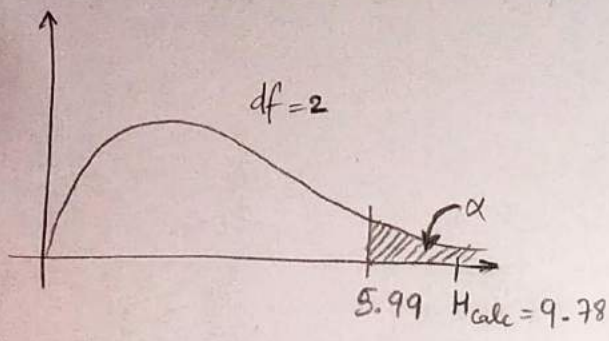
$= 195.6$

$$\Rightarrow H = \frac{12}{15 \cdot 16} \times 195 \cdot 6 = \underline{9.78}$$



Now $H \sim \chi^2_{k-1}$ where $k = 3 \Rightarrow \chi^2_{crit} = \chi^2_{2}{}^{(\alpha = 0.05)}$

$= 5.99.$



$H_{calc}(9.78) > \chi^2_{crit}(5.99)$

$\therefore$ We reject H_0.

where

H_0: All three groups have the same ranks.

H_A: The ranks of atleast one of the groups is different from others.

Lecture - 37

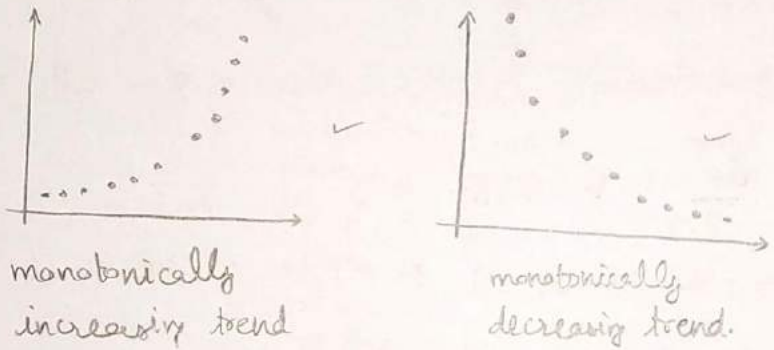
(07 - 04 - 2022)

Non-parametric equivalent of correlation & regression.

The reason for doing this test is because we don't know if the underlying probability distrib^n $\sim N$.

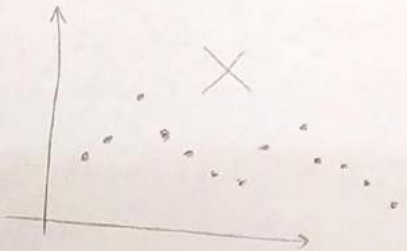
- **Spearman rank correlation test**

Here we assume that the two variables we're testing for are related monotonically



monotonically increasing trend

monotonically decreasing trend.



not monotonic

Height	Weight	R_H	R_W	Diff. b/w ranks (D)	D^2
150	45	1	1	0	0
151	55	2	3	-1	1
160	52	3	2	+1	1
165	57	4	5	-1	1
167	56	5	4	+1	1
172	60	6	6	0	0
175	65	7	7	0	0
180	70	8	8	0	0

$$\sum_i D_i^2 = 4$$

Spearman rank correlation $$\rho = 1 - \frac{6 \sum_i D_i^2}{N(N^2-1)}$$

N = # of paired samples = 8

$$\rho = 1 - \frac{6 \cdot 4}{8 \cdot 63} = 0.95$$

Just like Pearson's correlation coeff, $\rho \in [-1, 1]$

strong negative correlation

strong positive correlation

Logic:

$\min \left(\sum_i D_i^2\right) = 0 \Rightarrow$ perfect correlation

R_W	R_H
1	1
2	2
⋮	⋮
8	8

$\max \left(\sum_i D_i^2\right)$ is obtained when

R_H	R_W
1	8
2	7
3	6
⋮	
7	2
8	1

$\Rightarrow$ perfect negative correlation

$= 2 \cdot [7^2 + 5^2 + 3^2 + 1^2]$

$= 168$

For a given N,

$$\max\left(\sum_i D_i^2\right) = 2 \cdot \sum_{i=1}^{\lfloor N/2 \rfloor} (2n)^2 = \frac{N(N^2-1)}{3}$$

So, lowest $\sum_i D_i^2 = 0 \Rightarrow$ perfect positive correlation.

highest $\sum_i D_i^2 = \frac{N(N^2-1)}{3} \Rightarrow$ perfect negative correlation

So we can define a measure of correlation

$$\tilde{\rho} = \frac{\sum D_i^2 \leftarrow \text{calculated}}{\left(\frac{N(N^2-1)}{3}\right) \leftarrow \max}$$

- $= 0$ for perfect +ve correlation
- $= 0.5$ for no correlation
- $= 1$ for perfect −ve correlation

We map this into a scale similar to Pearson's coeff by following

$$\rho = 1 - 2\tilde{\rho}$$

- $= 1$ for perfect +ve corr.
- $= 0$ for no corr.
- $= -1$ for perfect −ve corr.

Hence, $\boxed{\rho = 1 - \frac{6\sum_i D_i^2}{N(N^2-1)}}$

Statistical Tests for Nominal (named) data.

Chi-squared tests-

① Goodness of fit.
② Independence
③ Homogeneity.

$\approx \chi^2_{d.f.}$ $(n \geq 5)$

n ↑ sample size

- **Goodness of fit**

used to check whether an observed variable fits a pre determined proportion.

	AA	Aa	aa
expected proportion	0.25	0.5	0.25
observed values	30	50	24

↳ are these deviations significant?

In the sense of significance of these deviations, we do a goodness of fit.

- **Independence**

Take a single pop^n & measure 2 variables – x & y.

Are these independent ?

- **Homogeneity**

measure the same variable in two populations.

Are the pop's homogeneous w.r.t. that variable ?

Lecture-38

(11-04-22)

χ^2 tests for categorical / nominal data.

$n_i \geq 5$

- → Goodness of fit
- → Independence
- → Homogeneity

} all of them use the same test statistic

$$\sum_i \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{df}$$

① Goodness of fit.

To check that the observed data fits well with the predicted values or not.

Typical monohybrid cross

F_0 AA × aa

↓

F_1 Aa

↓

F_2 AA : Aa : aa ← expected ratio

1 : 2 : 1

↓ ↓ ↓

But we get observed data as 30 65 20 ← observed distrib^n

The null hypothesis is described as: ($\alpha = 0.05$)

H_0: The observed data fits well with the predicted model.

H_A: The goodness of fit is bad!

To get the expected distrib^n, find total sample size = 30+65+20 = 115

	AA	Aa	aa	
Observed	30	65	20	→ total = 115
Expected	$\frac{1}{4} \times 115$	$\frac{2}{4} \times 115$	$\frac{1}{4} \times 115$	
	≈ 28	≈ 59	≈ 28	→ to keep the sum 115

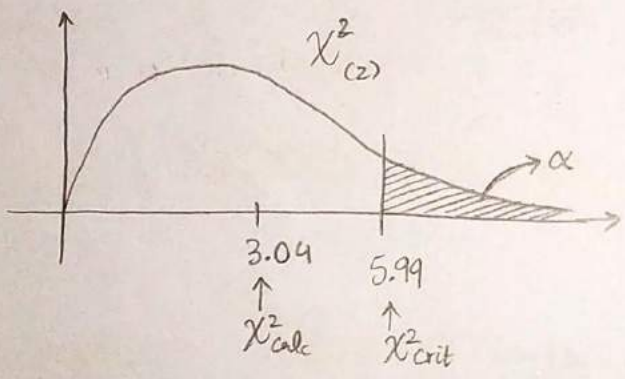
(each group has $n_i \geq 5$ so we can use χ^2 test)

$$\chi^2_{calc} = \frac{(30-28)^2}{28} + \frac{(65-59)^2}{59} + \frac{(20-28)^2}{28} \approx 3.04$$

$df = k - 1 = 3 - 1 = 2$

(k ↑ # of grps.)

$\chi^2_{(2)}$ $(\alpha = 0.05) = 5.99$



Since χ^2_{calc} (3.04) $< \chi^2_{crit}$ (5.99) $\Rightarrow$ We can't reject H_0.

∴ The level of fit is good enough.

- X^2 test for Independence

We want to see if 'smoking' and 'cancer' are linked or independent of each other.

[H_0: Smoking & Cancer are independent of each other.
H_A: They are dependent.]

	No Cancer.	Cancer.	total
Non-smoker.	50	10	60
Previous smokers	50	25	75
Current smokers	50	28	78
	150	63	213

This is the observed data.
What are the expected values though?

Start by assuming the two factors are independent. Then we should get a 1:1 ratio b/w both factors.

Individuals without any cancer in the popⁿ $= \frac{150}{213}$

Individuals who are non-smokers in the popⁿ $= \frac{60}{213}$

Ideally, the proportion of non-smokers without cancer (assuming they are independent) $= \frac{150}{213} \times \frac{60}{213}$

$\Rightarrow$ Expected # of non-smokers without cancer $= \frac{150 \times 60}{213} \approx 42$

Similarly, we can do this for each of them

Expected # of non-smokers with cancer $= \underbrace{\frac{63}{213}}_{\text{prob. of cancer}} \times \underbrace{\frac{60}{213}}_{\text{prob. of non-smoker}} \times 213 \approx 18$

Expected # of prev. smokers without cancer $= \underbrace{\frac{150}{213}}_{\text{prob. of no cancer}} \times \underbrace{\frac{75}{213}}_{\text{prob. of prev. smokers}} \times 213 \approx 53$

Expected # of prev. smokers with cancer $= \underbrace{\frac{63}{213}}_{\text{prob. of cancer}} \times \underbrace{\frac{75}{213}}_{\text{prob. of prev. smoker}} \times 213 \approx 22$

Expected # of smokers without cancer $= \underbrace{\frac{150}{213}}_{\text{prob. of no cancer}} \times \underbrace{\frac{78}{213}}_{\text{prob. of smoker}} \times 213 \approx 55$

Expected # of smokers with cancer $= \underbrace{\frac{63}{213}}_{\text{prob. of cancer}} \times \underbrace{\frac{78}{213}}_{\text{prob. of smoker}} \times 213 \approx 23$

	No cancer		Cancer		
	Obs.	Exp.	Obs.	Exp.	
Non-smoker	50	42	10	18	
Previous smoker	50	53	25	22	
Current smoker	50	55	28	23	
Total	150		63		213

$$\chi^2_{calc} = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(50-42)^2}{42} + \frac{(50-53)^2}{53} + \frac{(50-55)^2}{55} + \frac{(10-18)^2}{18} + \frac{(25-22)^2}{22} + \frac{(28-23)^2}{23}$$

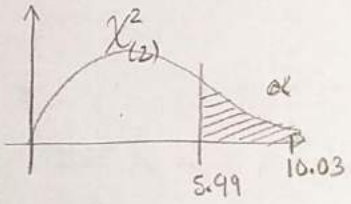
$$= \underline{10.03}$$

degrees of freedom $= (r-1)\cdot(c-1) = (3-1)(2-1) = \underline{2}$

(r: # of rows, c: # of columns)

$\Rightarrow \chi^2_{crit}\ (\alpha = 0.05) = \chi^2_{(2)}\ (\alpha = 0.05) = 5.99$

Since $\chi^2_{calc}\ (10.03) > \chi^2_{crit}\ (5.99) \Rightarrow$ Reject H_0.



∴ Smoking & Cancer are related and not independent variables.

- χ^2 test for Homogeneity.

Ex - Risk of ankle injury in two pop^{n}s → those who play sports; → those who don't play sports.

H_0 : The pop^{n}s are homogeneous w.r.t. no. of injuries

# of injuries	Sportsperson	Not sportsperson (normies)	total
0-1	50	10	60
2-3	50	25	75
>3	50	28	78
total	150	63	213

We have the same dataset as before, and because the test runs in exactly the same way, we again reject the H_0.

No cancer, Cancer ⟶ sportsperson, non sportsperson

smoker categories ⟶ # of injury categories

The only difference b/w test of homogeneity & independence lies in-

Independence

measure 2 variables (smoking, cancer) in same popⁿ.

Homogeneity.

measuring same variable (injuries) in 2 popⁿs (sportsperson, not sportsperson)