collect, organize, analyze.

Statistics for Data Science:

Descriptive
- organize/summarize data
- Central Tendency
- Dispersion
- Distr. of data
- Histogram, pdf, pmf.

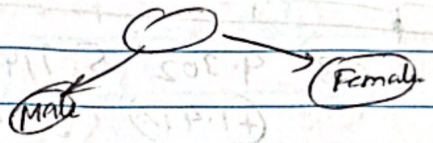
Inferential
- using data to form conclusion
- Z^{2} Test
- t-Test
- Chi Square
- ANOVA

(N) Population: group we're interested in studying.
(n) Sample: subset / representative subset of sample.

Sampling Techniques:

① Simple Random Sampling: every member of N has equal chance of being selected (n).

(Layering) ② Stratified sampling: non overlapping groups.
- divide into →



③ Systematic Sampling: Mall → outside
(Survey) ↳ every n^{th} person.

④ Convenience sampling: (Voluntary Response sampling)
- select a preferable grp of ppl for sampling.

Variables

Quantitative (Take any value w/in a range)
- Discrete: distinct, countable [1, 3, 500] values
- Continuous: [infinite values] (0.2, 238.5)

Qualitative/Categorical
Gender → M, f.

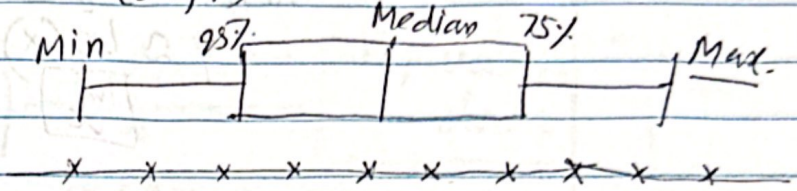
Outliers:

5 Number Summary:

① Minimum ② First quartile (25%) ③ Median } Boxplot
④ Third quartile (75%) ⑤ Maximum

Min. Low Bound = Q1 − 1.5(IQR) IQR = Q3 − Q1.
Max. Upper Bound = Q3 + 1.5(IQR)

Boxplot:



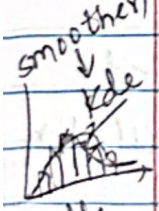
* Z-score: data points more than 3 s.d. away from mean. (Normally distributed data).

* DBScan: Points that aren't a part of Dense Clusters.

Types of Distribution:

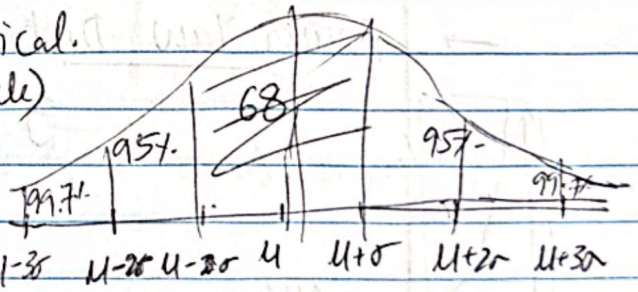
- Normal Distribution and Empirical Formula.

→ Data points that follow bell curve distr.
→ Symmetrical.



(Empirical. (3-σ rule)

68 - 95 - 99.7
(σ) (2σ) (3σ)

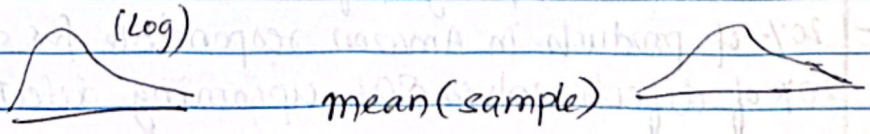


→ Size of shirt, Gym weights ⇒ Normal Distr.

- Central Limit Theorem:

n ≥ 30.

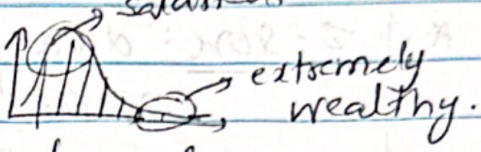
- If you take multiple samples and 1 from a populat and plot their means, the resulting distribution would be normally distributed.
- The og population need not be normally distributed



→ Log Normal Distribution:

If $X \approx$ Log Normal Distribution.
then ($\checkmark$ natural log)
$Y \approx \ln(X)$ → Normal Distribution.
$\log_e$

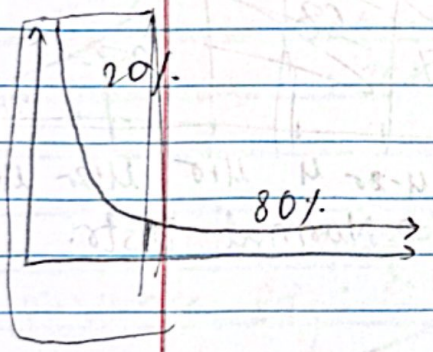
$X \approx \exp(y)$



eg. • Wealth Distribution
• Comments by Length on a channel.
(Less long paragraphs).

Data Transformation: Log Dist → Normal.
→ better to train ML models.

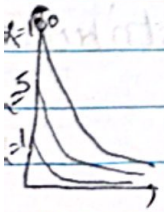
→ Power Law Distribution: (80% - 20% rule)



eg. 80% of wealth is distributed with 20% of total population
Oil Rich Nation
- 80% of total oil is with 20% of nations.

Transform to Normal Distr. using Box Cox ~~Plot~~ Transformation.

→ Pareto Distribution: follows Power Law Distribution.



α increases, height of Pareto increases.

eg. - 20% of products in Amazon responsible for 80% sales.
- 20% of defects solves 80% upcoming defects.

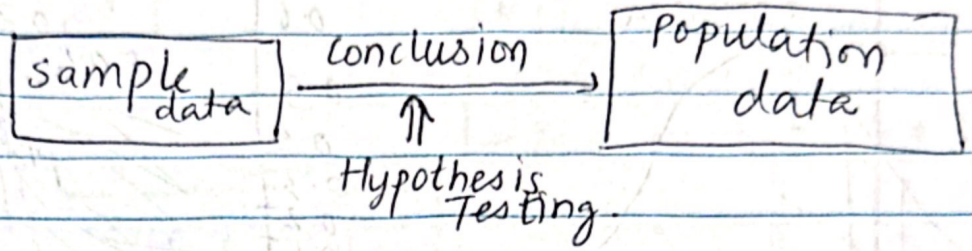
Pareto:

* Relationship between Log Normal and Pareto?

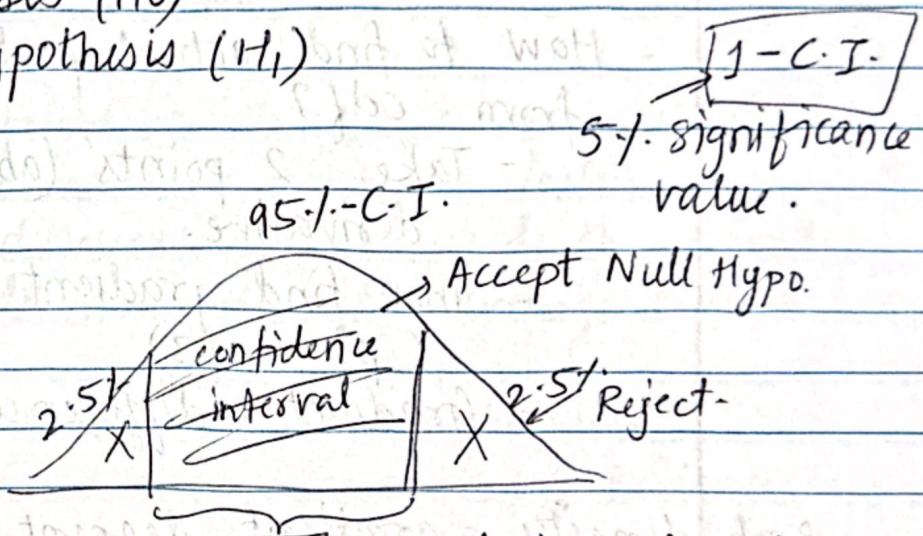
Inferential Stats:

* Hypothesis Testing:

- test an assumption regarding a population parameter.



1. Null Hypothesis (H_0)
2. Alternative Hypothesis (H_1)
3. Experiment:



Domain expert decides C.I.

Random var: map output of random process to a number.

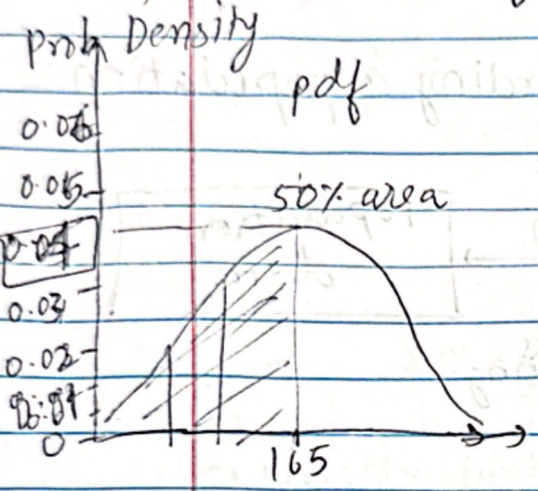
dice, coin toss etc.

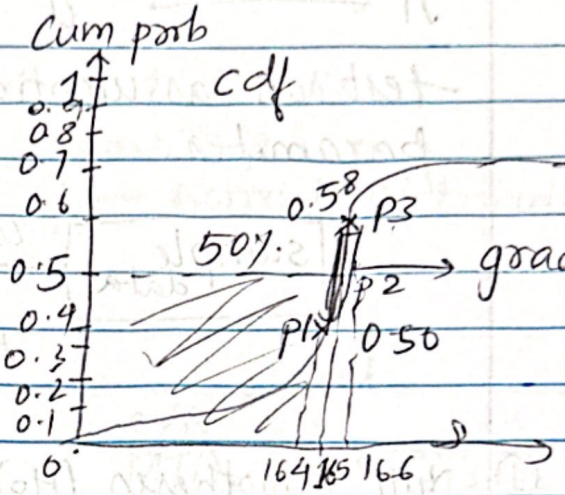
probability Distr. Function

(contn) Probability Density Function	(discr.) Prob. Mass Function
° Continuous Random var. ← check it's distribution using pdf.	Discrete random var → use pmf to check its distribution.

Area under curve (pdf)

i) Distribution of continuous random variable.



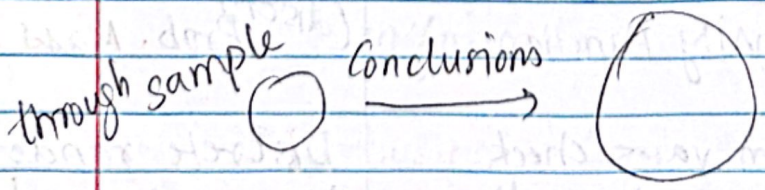


- How to find actual probability at 50%. from cdf?
  - Take 2 points' (above and below) derivative.
  - Then find gradient of the line (p1 to p3)
  - Gradient : difference : $\frac{0.58-0.50}{166-164} = \boxed{0.04}$

Prob density = gradient descent of cumulative curve.

→ Hypothesis Testing and Statistical Analysis

① z-test } Average ③ Chi-Square → Categorical Data.
② t-test } ④ ANOVA → Variance.



→ When to use t-test vs z-test?

Do you population std dev σ?

Yes ↓

Is the sample size > 30?

Yes ↓ use z-test

No ↓ Use t-test.

No ↓ Use t-test

(Sample size = 30)

* We need the area under curve (range) of the C.I.
→ This can be found using z-table.

$\mu \neq$: Two Tailed Test

$\mu >$: Right-Tail

$\mu <$: Left Tail

Q: The average heights of all residents in a city is 168cm with a population std $\sigma = 3.9$. A doctor believes the mean to be different. He measured the height of 36 individuals and found the average ≈ 169.5 cm.

a) State Null and Alternate Hypothesis
b) At a 95% C.I., is there enough evidence to reject the null hypothesis?

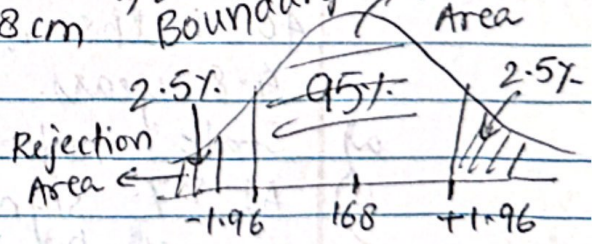
Solution:

$\mu = 168$ cm $\sigma = 3.9$ $n = 36$ $\bar{x} = 169.5$ cm

a) Null Hypothesis (H_0): $\mu = 168$ cm
b) Alternate Hypothesis (H_1): $\mu \neq 168$ cm

c) C.I. = 0.95 or 95%
$\alpha = 1 - C.I. = 1 - 0.95 = 0.05$

d) Decision Boundary



→ Decision Boundary:

$1 - 0.025 = 0.9750$

Through z-table, we find 1.96

Through decision boundary, if z-test value falls between -1.96 to $+1.96$, then we fail to reject Null Hypothesis.

$$z\text{-score} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \qquad \boxed{n=1}$$

standard error $\rightarrow \sigma/\sqrt{n}$

d) Statistical Analysis

$$z\text{-test} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{169.5 - 168}{3.9/\sqrt{36}} = 2.31.$$

e) Conclusion:

- As z-test value is greater than +1.96, we reject the null hypothesis and accept the ACCEPT the alternate hypothesis.

$2.31 > 1.96$

- The doctors' claim is right, the population mean is 169.5.

Question 2: Practice sum:

A factory manufactures bulbs w average warranty of 5 years with standard deviation of 0.50. A worker believes that the bulb w/h manufactu in less than 5 years. He tests a sample of 40 bulbs and finds the average time to be 4.8 years.

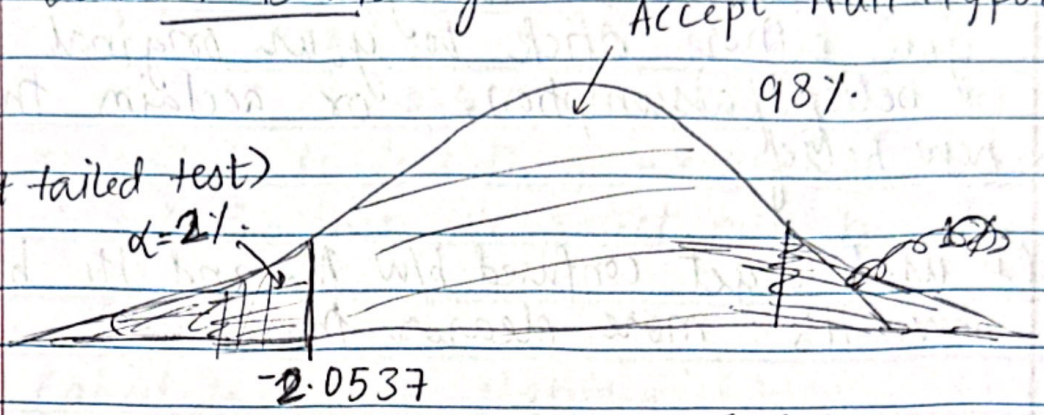
a) State null and alternate hypothesis.
b) At 2% significanu, is there enough evidence to support the idea that the warranty should be revised?

$\rightarrow$ $\sigma = 0.50$, $n = 40$, $\mu = 5$, $\bar{x} = 4.8$

a) H_0 : $\mu = 5$
b) H_a : $\mu < 5$

c) Significance level (α) = 2% or 0.02.
C.I. = 1 − 0.02 = 0.98 or 98%.

d) Decision Boundary:



* z-value: from area
Critical value: (1 − 0.02 = 0.9800)
Critical value = −2.0537

(if z-score value is below −2.0537, you reject the null hypothesis)

e) Calculate z-score:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{4.8 - 5}{0.5/\sqrt{40}} = -2.53$$

f) Conclusion:

∴ −2.53 > −2.0537
(z-value) (critical value)

- We reject the null hypothesis.
∴ The average ~~height~~ life of bulbs is 4.8 years!

How do you decide null / alternate hypothesis?

→ The claim or belief that you already have → NULL Hypothesis.

→ The claim / belief that challenges your og belief is the alternate hypothesis.

With decision boundary, critical values, you either stick to your original belief / assumption, or acclaim the new belief.

→ (I used to get confused b/w H_0 and H_1 but now it's more clearer than ever).

* T-Test:

→ Question:

In the population, avg IQ is 100. A team of researchers want to test a new medication to see if it has either a positive or negative effect on intelligence, or no effect at all. A sample of 30 participants who have taken the medication has a mean of 140 within a std deviation of 20. Did the medication affect intelligence? CI = 95%.

Given: $\mu = 100$, $n = 30$, $\bar{x} = 140$, $s = 20$

C.I. = 95% $\quad \alpha = 1 - 0.95 = 0.05$

① Null Hypothesis (H_0): $\mu = 100$

Alternate Hypothesis (H_a): $\mu \neq 100$ {2-Tailed}

[either positive or negative effect]

② Significance value: $\alpha = 0.05$

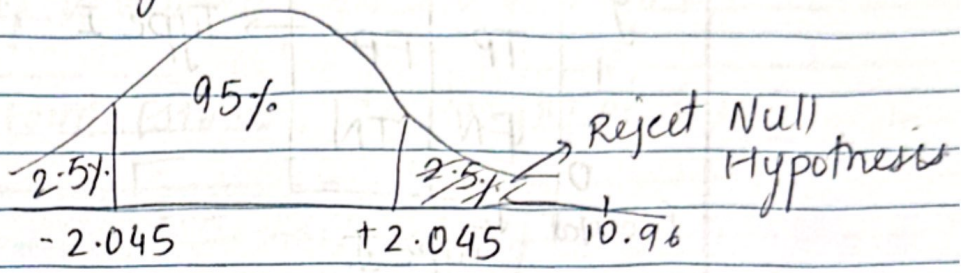
③ Degree of Freedom $= n - 1$

$= 30 - 1$

$= 29.$

Degree of Freedom: Number of choices.

④ Decision Boundary:



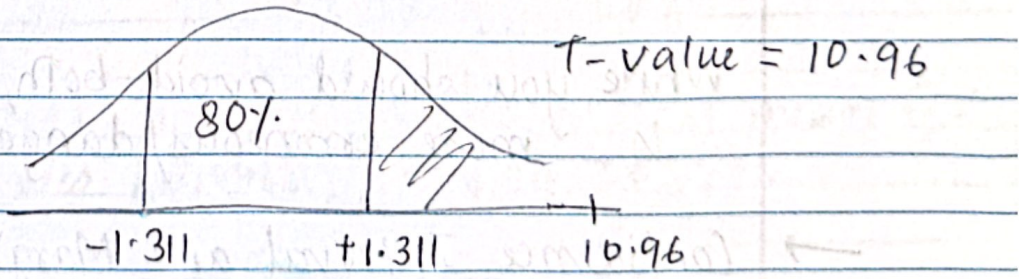
If t-test is less than -2.045 or greater than 2.045, then we reject null hypothesis.

⑤ Calculate t-test statistics:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{140 - 100}{20/\sqrt{30}} = \frac{40}{3.65} = 10.96.$$

⑥ Conclusion: Since $10.96 > 2.045$,
we reject the null hypothesis.
∴ The medication has a positive effect.

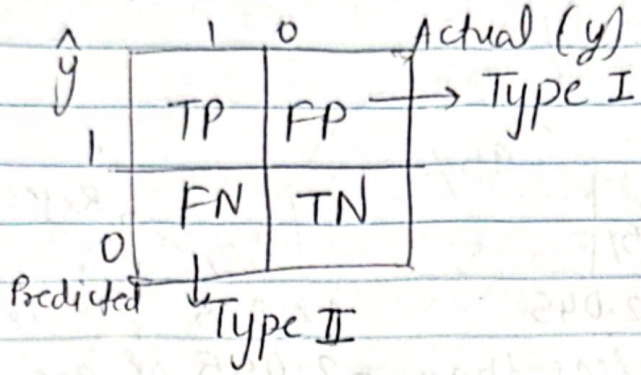
#Practice
Scenario 2: C.I. = 80% $\alpha = 1 - 0.80 = 0.20$



∴ We reject the Null Hypothesis and conclude that the medication has positive effect.

Type I and Type II Errors:

Confusion Matrix:



TP: Actual value is True, pred val is True.
TN: Actual value is False, pred val is False.
(Type I) FP: Actual value is False, pred val is True.
(Type II) FN: Actual value is False, pred val is False.

Scenario: Predicting whether a patient has cancer.

(FP) → Type I error: Will classify someone who doesn't have cancer as them having cancer. This is bad, but still better.

(FN) → Type II error: Classify person having cancer as not having cancer!
- Detrimental and harmful in long run.

While you should avoid both, Type II (FN) is more erroneous/dangerous.

→ Confidence Interval and Margin of Error:

Point Estimate: A value of any statistics that estimates value of an unknown population is called Point Estimate.

$\bar{x} \longrightarrow \mu$ {Conclusion}

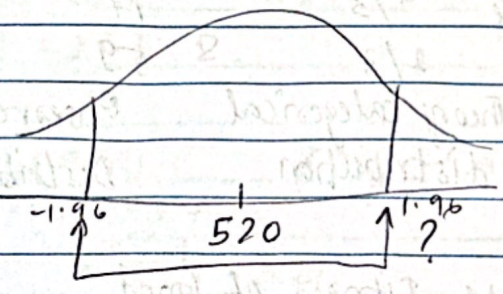
C.I.: we construct a C.I. to help estimate what actual value of unknown population mean is?

Point Estimate ± Margin of Error

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Q.) On the verbal section of CAT Exam, the std deviation is known to be 100. A sample of 25 test takers has a mean of 520. Construct the 95% C.I. around the mean?

→ $\bar{x} = 520$ $\sigma = 100$ $n = 25$ C.I. = 0.95 $\alpha = 0.05$



Lower C.I. $= \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 520 - (1.96)\frac{(100)}{\sqrt{25}} = 480.8$

Higher C.I. $= \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 520 + 1.96\frac{(100)}{\sqrt{25}} = 559.2.$

Conclusion: I am 95% confident that mean CAT score lies between 480.8 to ~~528~~ 559.2.

→ Biomedical researchers often handle categorical or nominal data.

* Chi Square Test → Inferential Stats.

→ The chi-square test for goodness of fit test claims about population proportion {categorical variables} → [ordinal, nominal]

It's a non parametric test that's performed on categorical data.

Make (a pop. categorical)

	Theory	Sample
Yellow Bike	1/3	22
Orange Bike	1/3	17
Red Bike	1/3	59
	Theory categorical distribution	Observed Categorical Distribution

→ Chi square Fitness of Good:

In 2010 census of the city, the weight of the individualists in a small city were found to be the following.

2010			$n = 500$ Observation. 2020		
Expected	< 50 kg	50-75 kg	>75	<50	50-75	>75
	20%	30%	50%	140	160	200

→ In 2020, ages of $n = 500$ individuals were sampled.

Using $\alpha = 0.05$, would you conclude the population difference of weights has changed in last 10 years?

(n = 500 2020)

→ In the sample, the expected table would be:

<50	50 - 75	>75
0.2 × 500	0.3 × 500	0.50 × 500
100	150	250

I)

H_0: Null Hypothesis: The data meets the expectation.

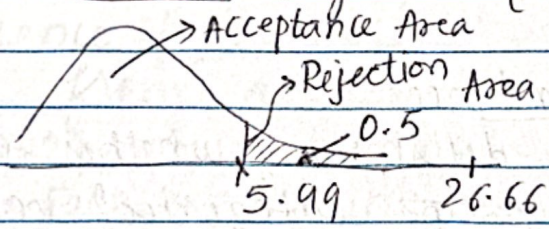
H_1: Alternate Hyp: The data doesn't meet the expectation.

II) $\alpha = 0.05$, C.I. = 95%.

III) Degree of Freedom: $df = k - 1$ (k: no. of categories)

$df = 3 - 1 = 2$.

IV) Decision Boundary: (Right-skewed Chi-square)



V) If χ^2 is greater than 5.99, Reject H_0.
Else, we fail to reject H_0.

VI) Calculate χ^2 Test Stats:

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(140-100)^2}{100} + \frac{(160-150)^2}{150} + \frac{(200-250)^2}{250}$$

$$\chi^2 = 26.66$$

Since, $26.66 > 5.99$, we reject the null hypothesis. There's a population difference and it has increase when compared to 2010.

Source: UCLA.

(categorical data)

→ Chi square Test: (3 Types)

I). goodness-of-fit test: (2 or more categories)
- whether a set of data comes from claimed distribution or not?
eg:- the frequencies I OBSERVE are consistent with my theoretical EXPECTED values?

II) test of homogeneity:
- determine whether 2 or more subgroup ∈ Pop. share the same distribution for a single categorical variable?
- eg.) Do ppl of diff races have same proportion of smokers and non smokers?

III) test of independence:
- way of determining whether two or more ~~sub-groups~~ categorical variables are associated with one another in population like race and smoking, education and political affiliation.

Diff between II and III?
- pretty similar, but diff. in matter of design.

Test of Independence	Test of Homogeneity
- Observational Unit: Data is collected randomly from pop., usme se 2 cat. variables are observed form each unit.	- Data is collected by randomly sampling from each sub group separately.
- Rep random observation me compare kare based on categorical:	- Intentionally proportionally samples utne from sub category.

ANOVA:

→ statistical analysis or method used to compare the means of 2 or more groups.

Parametric Tests: [z-test, t-test] - Mean of 2 groups.
[chi-square test] → Popn proportion.

→ Factors (variable) → Levels

Eg: Factor = Medicines

Levels: 5mg 10mg 15mg [Dosage]

Assumption in ANOVA:

1) Normality of Sample Distribution Mean - The distribution of sample mean is normally distributed
2) Absence of outliers: Outliers needs to be removed from dataset.
3) Homogenity of VARIANCE:
   - Each one of the population has same variance.

$$[\sigma_1^2 = \sigma_2^2 = \sigma_3^3]$$

4) Population variance in diff. levels of each independent variable are equal.

4) Samples are independent and random.

Types:

① 1 Way ANOVA: 1 Factor, atleast 2 Levels (levels are independent)

Eg. Dr. wants to test a new medication to decrease headache. They split participants in 3 dosages [10, 20, 30]mg. Dr. asks participant to rate the headache. [1-10]

ANOVA - F Distribution.

~~χ^2~~ χ^2 distribution.

2) Repeated Levels ANOVA:

1 Factor ; atleast 2 Levels.

Levels are dependent.

Running → Factor

Levels — Day 1 — Day 2 — Day 3

If you're learning a new skill / practicing; with incr. in the levels (duration), you'll improve a skill. So levels are dependent.

3) Factorial ANOVA: TWO or more factors (each of which w atleast 2 levels), levels can be independent and dependent.

Running Factor

Day 1 Day 2 Day 3 Level.

Gender Factor: Male, Female — Level

HYPOTHESIS TESTING IN ANOVA:

Null Hypothesis: H_0: $\mu_1 = \mu_2 = \mu_3 \dots \mu_k$

Alternate Hypothesis: H_1: Atleast one of the mean is not equal.

$\mu_1 \neq \mu_2 \neq \mu_3 \neq \dots \mu_k$.

Test statistics

F = Variance between sample / Variance within sample.

X_1	X_2	X_3	
a	c	e	$\Sigma x_3 = 220$
b	d	f	$\bar{x} = 4$
$\Sigma x_1 = 15$	$\Sigma x_2 = 19$		
$\bar{x}_1 = 3$	$\bar{x}_2 = 19/5$		

within — between

H_0: $\bar{x}_1 = \bar{x}_2 = \bar{x}_3$

H_a: Atleast one meant not equal.