Statistical Inference

Week 4 – Part 1 –Sampling Techniques

CS 457 - L1 Data Science

Zeehasham Rasheed

Sampling

- Sampling is a method that allows researchers, without having to investigate every individual, to infer information about a population based on results from a subset of the population

- Types of Sampling Methods:
  - Probability Sampling Methods
  - Non-Probability Sampling Methods

Probability Sampling Methods

1. Simple random sampling - Each individual is chosen entirely by chance and each member of the population has an equal chance, or probability, of being selected.

2. Systematic sampling - Individuals are selected at regular intervals from the sampling frame. Example: Every 10th call, Every 5th student

3. Stratified sampling - The population is first divided into subgroups (or strata) who all share a similar characteristic and then draw samples from each subgroups. Example Data Science Enthusiastic, Chess Lovers

4. Clustered sampling - The population is divided into subgroups, known as clusters, which are randomly selected to be included in the study. Example: Population divided into districts and randomly selects districsts for voting in Elections

Question

- How would you share survey questions with every 10th person who call to your call center?

Question

- How would you share survey questions with every 10th person who call to your call center?
  - **Using Systematic Sampling**

Non-Probability Sampling Methods

1. Convenience sampling - Participants are selected based on availability and willingness to take part. Example: Participation on Invitation

2. Quota sampling - Interviewers are given a quota of subjects of a specified type to attempt to recruit. Example: Research

3. Judgement Sampling - This technique relies on the judgement of the researcher when choosing who to ask to participate. Example: Any game show

4. Snowball sampling - Existing subjects are asked to nominate further subjects known to them, so the sample increases in size like a rolling snowball. Example Covid 19 Vaccine leading to Mental Health leading to Medications

Question

- 2. How would you request feedback from every person sitting in a cafeteria about the quality of food served?

Question

- 2. How would you request feedback from every person sitting in a cafeteria about the quality of food served?
  - **Using Convenience sampling**

Sampling Bias (Problem)

- Sampling bias may be introduced when:
  - Deviation in pre-agreed sampling rule
  - People in hard-to-reach groups are omitted
  - Low response rate
  - Limited knowledge of procedure (Most Dangerous)
  - Low Motivation to collect samples
  - Shortage of funds (Very Common)

End of Part 1

Statistical Inference

Week 4 – Part 2 – Central Limit Theorem

CS 457 - L1 Data Science

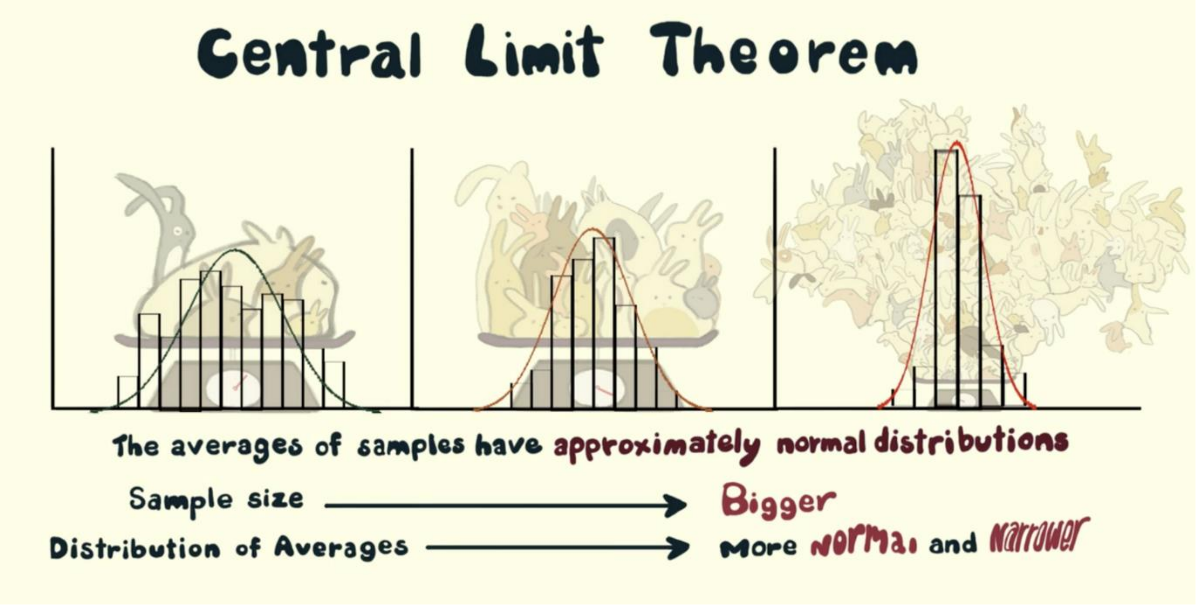
Zeehasham Rasheed

Central Limit Theorem

The sampling distribution of the sampling means approaches a normal distribution as the sample size gets larger.

“As the sample size increases, sampling distribution will get narrower and more normal.”

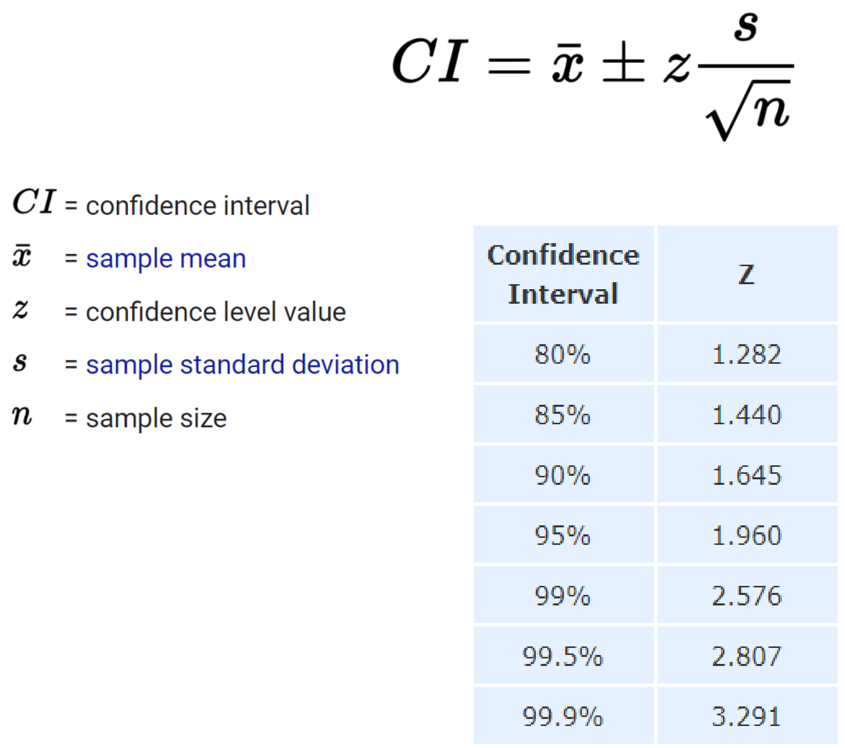
- We want to be accurate (and confident) about our estimate
- It is a good idea to give range of estimate (rather that one estimate)


Central Limit Theorem
The averages of samples have approximately normal distributions
Sample size → Bigger
Distribution of Averages → More normal and narrower

Confidence Interval

- A confidence Interval is a range of values where we are fairly sure that it will capture population parameter.

- Parameter could be anything
  - In normal distribution, it is **mean** and **standard deviation**)



$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$$

CI = confidence interval

$\bar{x}$ = sample mean

z = confidence level value

s = sample standard deviation

n = sample size

Confidence Interval	z
80%	1.282
85%	1.440
90%	1.645
95%	1.960
99%	2.576
99.5%	2.807
99.9%	3.291

integrating the probability density function in a given distribution, the cumulative distribution function helps us map z values

Confidence Interval Example

Example: Seatbelt Usage

A sample of 12th grade females was surveyed about their seatbelt usage. A 95% confidence interval for the proportion of all 12th grade females who always wear their seatbelt was computed to be [0.612, 0.668].

The correct interpretation of this confidence interval is that we are 95% confident that the proportion of all 12th grade females who always wear their seatbelt in the population is between 0.612 and 0.668.

Example: IQ Scores

A random sample of 50 students at one school was obtained and each selected student was given an IQ test. These data were used to construct a 95% confidence interval of [96.656, 106.422].

The correct interpretation of this confidence interval is that we are 95% confident that the mean IQ score in the population of all students at this school is between 96.656 and 106.422.

End of Part 2

Statistical Inference

Week 4 – Part 3 – Hypothesis Testing

CS 457 - L1 Data Science

Zeehasham Rasheed

Hypothesis Testing

- Hypothesis testing refers to the process of choosing between two hypothesis statements about a probability distribution or population parameter (for example: **mean**) of your data.

- Hypothesis testing is a step-by-step methodology that allows you to **make inferences** about a population parameter (usually mean/average)

Two Types of Hypothesis

- There are two types of Hypothesis:
  - Null Hypothesis – Old Belief (Status Quo)
  - Alternative Hypothesis – New Claim

- It refers to an assumption which is being tested, to decide whether
  - Reject the Null Hypothesis (and accept the Alternative Hypothesis)

  or

  - Can not reject the Null Hypothesis

Hypothesis Testing - Steps

- The methodology behind hypothesis testing:

1. State the null hypothesis and alternative hypothesis.
2. Select the distribution to use.
3. Determine the rejection and non-rejection regions (also called **significance level / confidence interval**)
4. Calculate the value of the test statistic (also called **p-value**).
5. Make a decision based on p-value.

Hypothesis Examples

- **Null hypothesis (H_o)** - Children who take vitamin C are no less likely to become ill during flu season.
- **Alternative hypothesis (H_a)** - Children who take vitamin C are less likely to become ill during flu season.

Other examples

- Young boys **are/are not** prone to more behavioral problems than young girls.
- Children of obese parents **are/are not** more likely to become obese themselves.
- People **are/are not** more susceptible to colds in the fall than the winter.

Hypothesis Testing Acceptance and Error

- z Score / t Score / p-value is the deciding factor of accepting or rejecting Hypothesis
- p-value is also called Probability of committing Type 1 error
  - Industry Threshold Standard is 95% confidence interval (which is also called significant level α = 0.05 (100%-95%=5%)
  - Which means that if the p-value for any hypothesis testing is less that 0.05, then we reject the Null Hypothesis and accept Alternative Hypothesis

- **Type 1 Error**
  - When we reject a true Null Hypothesis
- **Type 2 Error**
  - When we fail to reject (or accept) a false Null Hypothesis

End of Part 3

Statistical Inference

Week 4 – Part 4 – Hypothesis Testing Methods

CS 457 - L1 Data Science

Zeehasham Rasheed

Different Methods for Hypothesis Testing

- z-test and t-test
- Chi-Square Test for Independence
- Analysis of Variance (ANOVA)

  - The calculated value of the hypothesis test is converted into a z-score/t-score/p-value that explains whether the outcome is statistically significant or not.
    - Statistically significant means we reject Null Hypothesis and accept Alternative Hypothesis

z-test

- It determines how many standard deviations a data point is away from the mean of the data set.

- Two conditions to adopt Z-test:
  - The population mean and variance is known (most important)
  - Large sample size (i.e >30 observations)

Sample Problem

- The average weight of 36 young males is 185 lbs.
- This sample is compared to the general population of males in the same age range. The general population mean is 180 lbs and the standard deviation is 20 lbs.
- What is the Z statistic for this sample?

Solution

In this example, n=36, Sample Mean = 185., Population Mean=180 lbs, Population Standard deviation (σ)=20

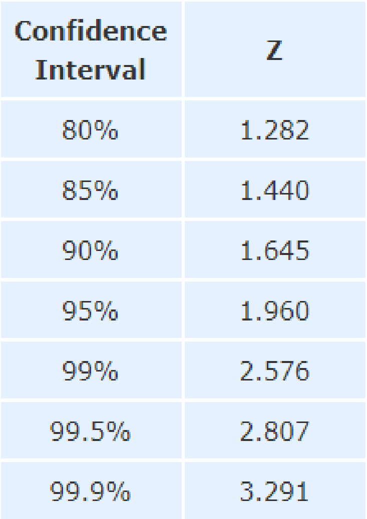
H_0: =180 lbs. , i.e. the mean weight in the population is 180 lbs.

Z = (185-180)/20/sqrt(36)

Z= ((5)*6)/20

Z= 1.5

$$Z = \frac{(\bar{X} - \mu_0)}{s}$$



Confidence Interval	Z
80%	1.282
85%	1.440
90%	1.645
95%	1.960
99%	2.576
99.5%	2.807
99.9%	3.291

- See chart for 95% confidence interval, also called 5% level of significance 100% - 95%=5%), the value is 1.96
- Since calculated value of Z statistic is less than 1.96, the sample mean is not significant at 5% level of significance.
- Therefore, H_0 **can't be rejected** which implies that mean weight of population is 180 lbs.

t-test

- t-test is used to examine how the means taken from two independent samples differ.

- Three conditions to adopt t-test
  - The population mean and variance is unknown (most important)
  - Small sample size (i.e <30 observations)
  - The **degree of freedom** implies the number of independent observations in a given set of observations

Different Samples in t-test

- One Sample t-test

$$t = \frac{m - \mu}{s/\sqrt{n}}$$

t = Student's t-test

m = mean

μ = theoretical value

s = standard deviation

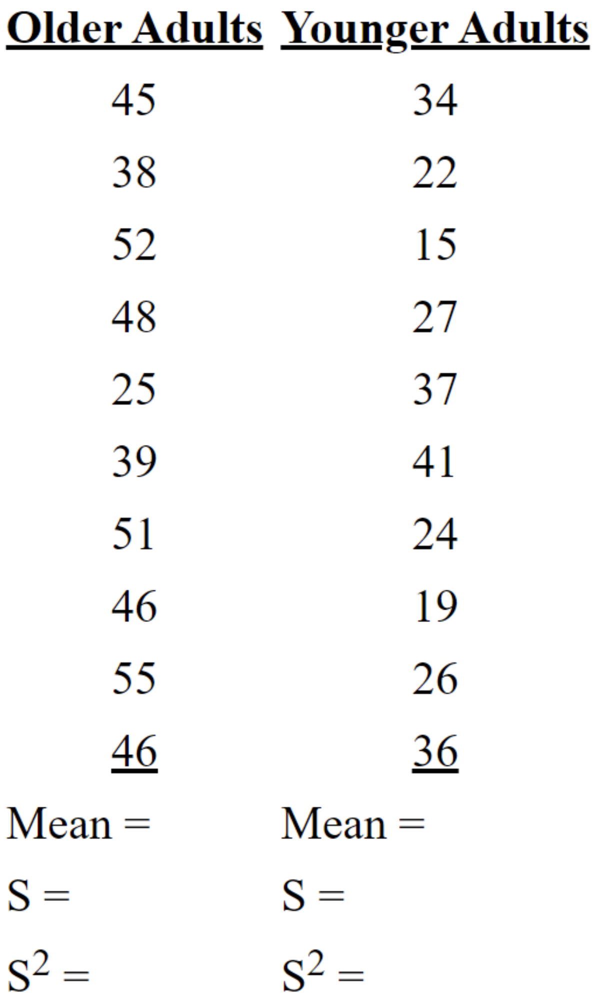
n = variable set size

- Two Sample t-test

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{(s^2(\frac{1}{n_1} + \frac{1}{n_2}))}}$$

t-test Example

- A research study was conducted to examine the differences between older and younger adults on perceived life satisfaction.
- A pilot study was conducted to examine this hypothesis. Ten older adults (over the age of 70) and ten younger adults (between 20 and 30) were given a life satisfaction test (known to have high reliability and validity).
- Scores on the measure range from 0 to 60 with high scores indicative of high life satisfaction; low scores indicative of low life satisfaction.
- The data is given. Compute the appropriate t-test.



Older Adults	Younger Adults
45	34
38	22
52	15
48	27
25	37
39	41
51	24
46	19
55	26
46	36
Mean =	Mean =
S =	S =
S^2 =	S^2 =

t-test Example (2)

What would be the null hypothesis in this study?

- The null hypothesis would be that there are no significant differences between younger and older adults on life satisfaction.

What would be the alternate hypothesis?

- The alternate hypothesis would be that life satisfaction scores of older and younger adults are different.

What probability level did you choose

- 5% or 0.05 significant level (95% confidence interval)

t-test Calculation

- Using the formula discussed earlier
- Computed t-test value is $t_{obs} = 4.257$
- $t_{critical}$ from chart is 2.228 when degree of freedom DF=10 (there were 10 samples from each group)

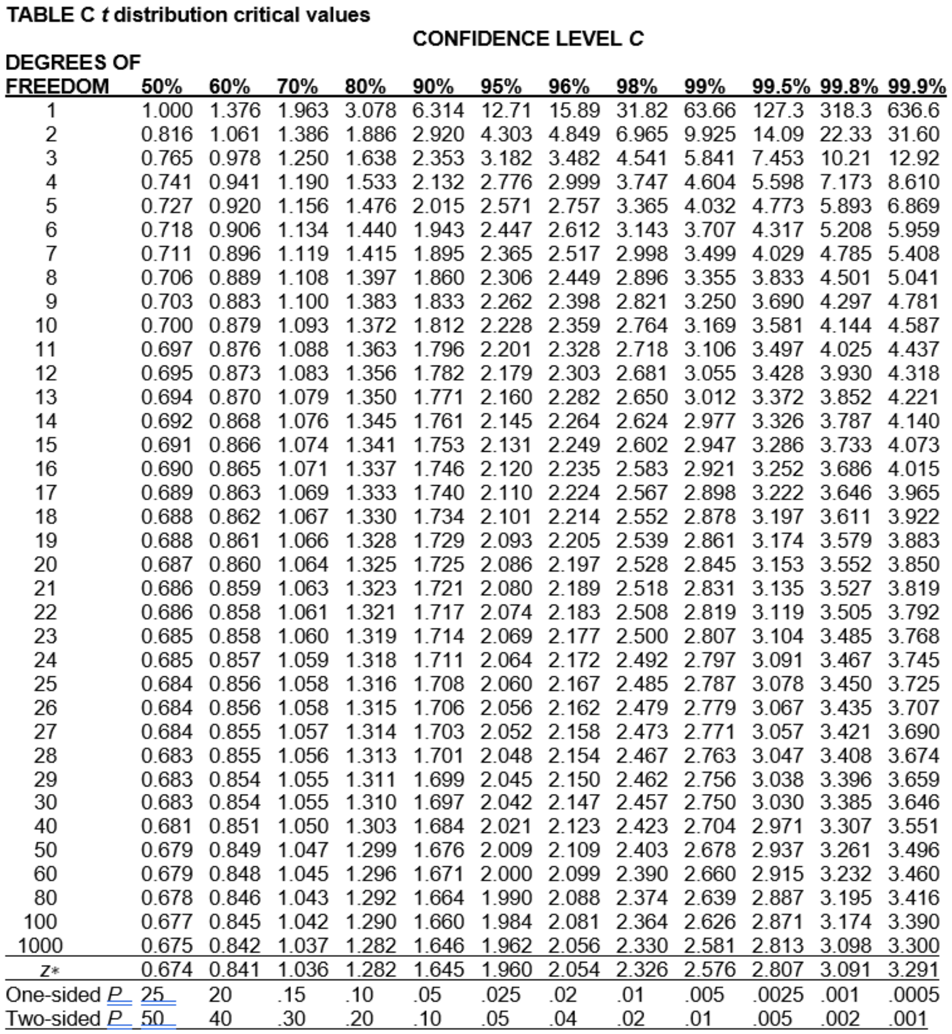


TABLE C *t* distribution critical values

DEGREES OF	CONFIDENCE LEVEL *C*											
FREEDOM	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
$z*$	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
One-sided *P*	25	20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
Two-sided *P*	50	40	.30	.20	.10	.05	.04	.02	.01	.005	.002	.001

t-test Conclusion

Is there a significant difference between the two groups?

Yes, the $t_{obs} = 4.257$ is greater than $t_{critical} = 2.228$.

Thus, we reject Null Hypothesis and accept Alternative Hypothesis and conclude that **there is a significant difference between the two groups.**

t-test using p-value

- **You can also convert t-score into p-value**

- For the previous example, we concluded that older adults in this sample have significantly higher life satisfaction than younger adults based on t-test value of t = 4.257

- The equivalent p-value for t-test value (4.257) is .0009 (we are not discussing how to convert t-score to p-values here)

  - Here p-value (0.0009) is less than significant level 0.05 (95% confidence interval)
  - Therefore, we reject Null Hypothesis and accept Alternative Hypothesis and conclude that there is a significant difference between the two groups.

- <u>Python libraries and R packages provides p-values for all hypothesis testing methods</u>

t-test vs z-test

- One of the important conditions for adopting t-test is that population variance is unknown.
- Conversely, population variance should be known or assumed to be known in case of a z-test.
- Z-test is used to when the sample size is large, i.e. $n > 30$, and t-test is appropriate when the size of the sample is small, in the sense that $n < 30$ (not very important)

Test problem - 1

Context: The average heart rate for Indians is 72 beats/minute. A group of 27 individuals participated in an aerobics fitness program to lower their heart rate. After three months the group was evaluated to identify is the program had significantly slowed their heart. The mean heart rate for the group sample was 70 beats/minute with a standard deviation of 6.3. Was the aerobics program effective in lowering heart rate?

Poll: Which Test to apply (z-Test or t-Test)

t-Test because population mean is not known

Text problem - 2

Context - The average heart rate for Indians is 72 beats/minute. A group of 50 individuals participated in an aerobics fitness program to lower their heart rate. The sample is drawn from Normal Distribution where mean and standard deviation of population was known. After three months the group was evaluated to identify is the program had significantly slowed their heart. The mean heart rate for the group was 70 beats/minute with a standard deviation of 6.3. Was the aerobics program effective in lowering heart rate?

Poll: Which Test to apply (z-Test or t-Test)

z-Test because population mean is known

End of Part 4

Statistical Inference

Week 4 – Part 5 – ANOVA and Chi-Square Tests

CS 457 - L1 Data Science

Zeehasham Rasheed

ANOVA

- Analysis of Variance (ANOVA) A statistical technique for examining the difference among mean for <u>three or more population</u>.

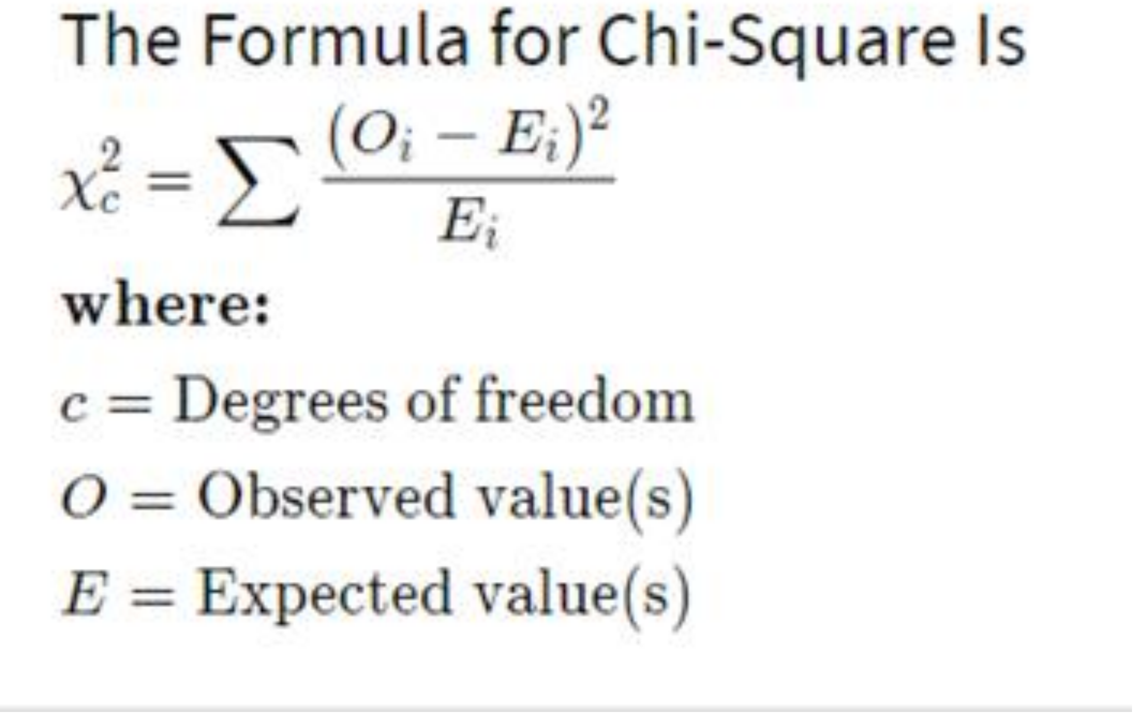
- Here the dependent variable is metric(continuous) and independent variables are categorical.
- **One-way Anova** – When there is only one factor.
  - One-way ANOVA for comparing 3(+) groups on 1 factor
  - Example: Does all people from company A, B and C have equal mean IQ scores? where (IQ score is one factor (continuous) and company A,B, and C are categories)

Anova Business Problems

1) Recommendation of movie types (sci-fi, thriller, action, romance) for the improvement of a viewer experience score.
2) Recommendations of fertilizers to increase the crop yield.
3) Recommendations of exercise to reduce weights.

Chi-Square test

- It assists us in determining whether a systematic association exists between the **two categorical variables** in some population.
- Also depends on the number of degree of freedom.
- The analysis should not be conducted if the theoretical frequencies in any cell less than 5



The Formula for Chi-Square Is

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:

$c =$ Degrees of freedom

$O =$ Observed value(s)

$E =$ Expected value(s)

Business Problems

1) A Tech company need to check "Is candidate's educational background and type of job chosen independent?"
2) A Marketing company need to submit report “Is their any association between income level and brand preference?”
3) An e-commerce company need to design its product catalogue. The company looking for an answer "Are all brands equally preferred by its customer"?

Question

- A political party needs to design its campaign to attract the voters.
- You are the data scientist and need to advise the political party to design your campaign with focus on demand from businessmen.
- A senior official said that winning elections (Yes/No) and occupation has no linkage. Which statistical test you will apply to prove your point?
    - <u>Chi-Square test as we have two categorical attributes to test</u>

End of Part 5