Understanding Media Consumption Preferences of College Students

MA4240 - Applied Statistics

Contributors :

Donal Loitam Himanshu Kumar Gupta Gunjit Mittal
Suraj Kumar Ravula Karthik Abhishek Kumar
Mannem Charan Sai Pradeep Shivanshu

May 3, 2023

Data Collection

Variable of interest

The following are the variables collected in the survey for the inference:

1. Gender
2. Age
3. Which part of the country you belong to
4. Which mode of Entertainment do you generally watch
5. Which cinema do you prefer the most
6. Preferable way of watching movie/anime/web series when not available in comfortable language

Variable of interest contd.

1. Time spent at a stretch watching movie/anime/web series
2. Hours spend weekly on these entertainments
3. Top three preferable genres
4. Which screen size do you prefer watching Movies on
5. Your preference of watching romantic Movies with
6. Your preference of watching horror Movies with
7. How much money do you spend on entertainments per month on average
8. Do you have subscription for any OTT platform (need not be yours)
9. Your preferable entertainment watching time

Data Collection and Description

The data is collected online by floating an anonymous form to people. The target audience for this survey is engineering college-going students. The primary method used for data collection is cluster sampling as the form was floated both in our college and within the colleges our friends are attending so each college forms a cluster. The total data collected over 350 samples

Data Preprocessing

In data preprocessing, the columns were renamed, the data was one hot encoded (wherever possible), making the data suitable for working. Another part was to remove troll entries.

Data Visualization

Histogram - AGE DISTRIBUTION

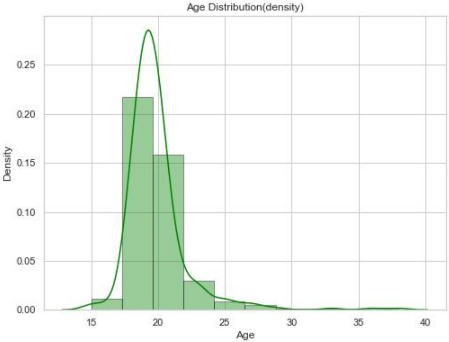

Figure: age-distribution

From the plot we can say that most of the students who filled the data are around 20 years old.

Histogram - HOURS SPENT AT A STRETCH

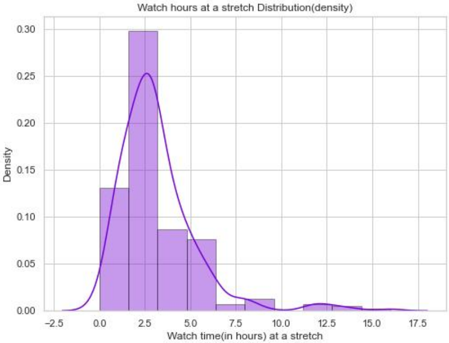

Figure: hours spent at a stretch

On average, students spend approximately 2.5 hours continuously watching entertainment based on the collected data.

Histogram - WEEKLY HOURS SPENT

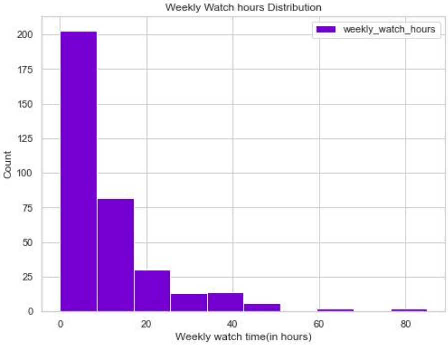

Figure: weekly hours spent

From the data collected, we can see that, the weekly time spent on entertainment by most students ranges between 0 - 10 hours.

Histogram - GENDER DISTRIBUTION

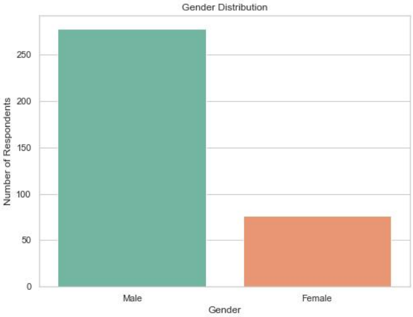

Figure: gender distribution

The dataset was primarily collected from engineering colleges, resulting in a larger representation of male respondents than female. But since the data for females are more than 70($>$ 30), we can use them for inference.

Histogram - REGION DISTRIBUTION

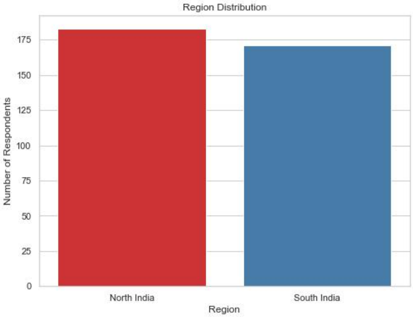

Figure: region distribution

- The dataset was collected from colleges across various regions, resulting in a balanced representation of respondents from different parts of the country
- It enables us to make region-specific conclusions that are not biased by an over-representation from any particular region

Bar Chart - Preference Distribution

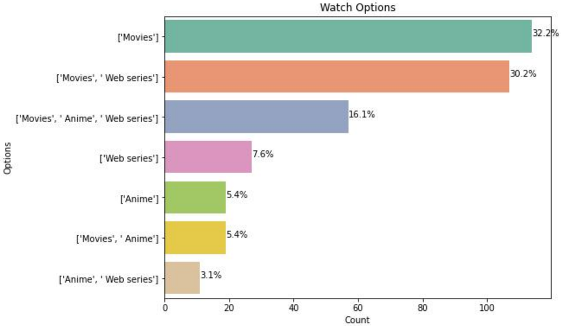

Figure: watch options - anime, movies, webseries

- As the horizontal bar chart reveals, college students still consider movies one of their top entertainment choices
- "Movies" and "web series" are the most commonly paired options.

Pie Chart - WATCH OPTIONS

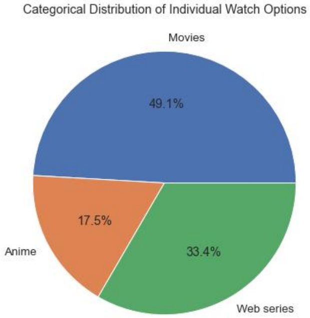

Figure: individual options - anime vs movies vs webseries

- The increasing popularity of high-quality web series and anime reflects the expanding entertainment options available to viewers
- However, movies still remain the most popular choice for entertainment.

Horizontal Bar Chart - TOP 3 GENRES DISTRIBUTION

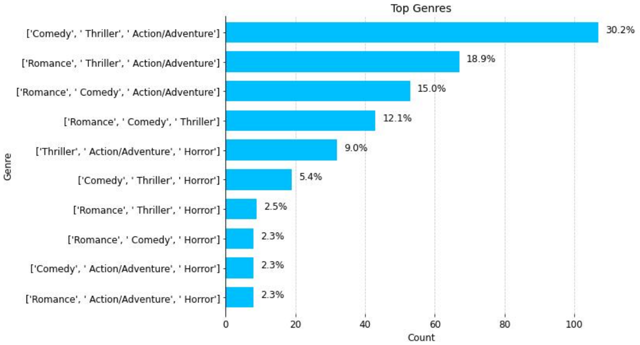

Figure: Distribution of top 3 genres

- Top 3 genres picked by the students are "comedy," "thriller," and "action/adventure," with over 100 respondents choosing them
- "Horror" is not as popular

Bar Chart - CINEMA PREFERENCES BY REGION

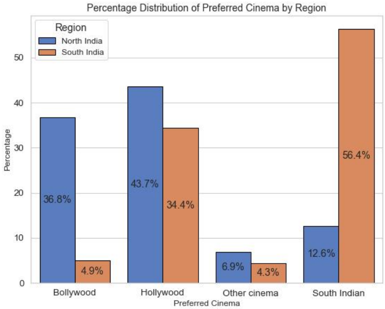

Figure: Cinema preferences vs. region

- There is a preference for Bollywood movies among respondents from North India and for South Indian movies among South respondents
- Indian audiences generally prefer Hollywood movies regardless of regional origin.

Side by side Bar Chart - OTT SUBSCRIPTION BY GENDER

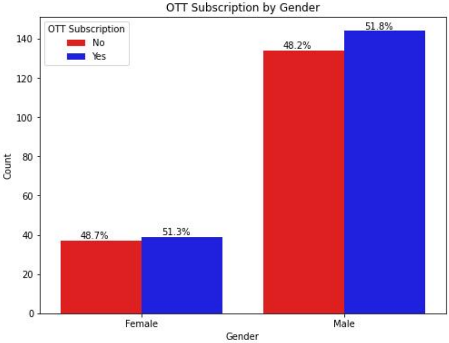

Figure: Ott-Subscription vs Gender

- Gender does not seem to have a significant impact on OTT subscriptions, as the % of subscribers is similar among both male and female respondents.

Side by Side Bar Chart -OTT SUBSCRIPTION BY REGION

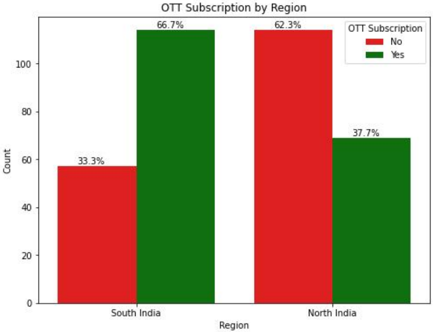

Figure: Ott-Subscription vs Region

- There is a significant difference in the % of people who own or share an OTT subscription between the North and South respondents, with a higher % in the South than in the North.

Scatter Plot - WEEKLY HOURS vs MONTHLY EXPENSES

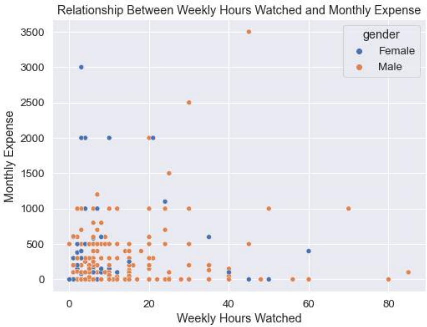

Figure: Gender Wise - Weekly Duration vs. Monthly Duration

- There is a low correlation (0.1 correlation) between the money spent on entertainment and the time spent weekly, irrespective of gender.
- They may watch entertainment on free/open sources such as YouTube etc.

Scatter Plot - WEEKLY DURATION vs HOURS SPENT AT A STRETCH

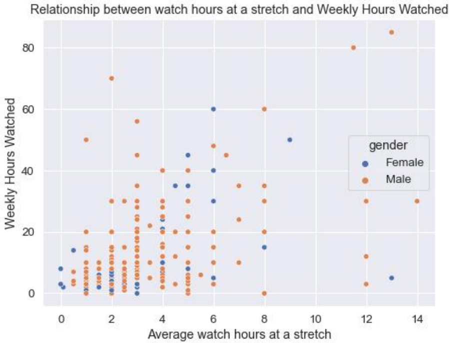

Figure: Gender Wise - weekly duration vs hours spent at a stretch

- The scatter plot shows a moderate positive correlation (0.5 correlation)
- Students who watch for longer periods at a stretch are likely to spend more time on entertainment weekly.

Side by side Bar Chart - PROPORTION OF PREFERRED TIME OF WATCHING

NOTE: However, it's important to note that correlation does not imply causation.

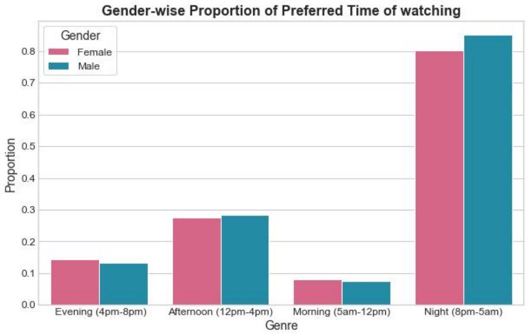

Figure: Preferred time of watching - proportion by gender

Contd.

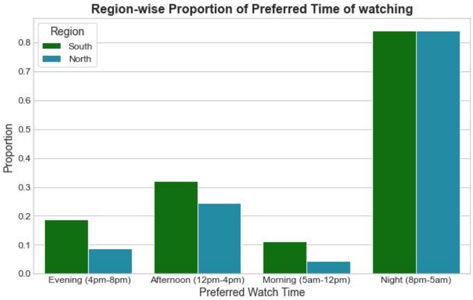

Figure: Preferred time of watching - proportion by region

- Regardless of gender and region, prefer watching movies, web series, or anime during late hours (8 pm-5 am)
- Since, classes are held during the morning and afternoon.
- Night time has become a popular and convenient time for entertainment consumption.

Side by side Bar Chart - PROPORTION OF GENRES SELECTED IN TOP-3

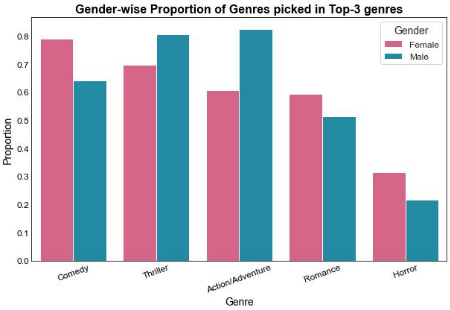

Figure: genderwise - proportion of top-3 genres selected

- There exists gender-specific differences in the genres preferred by college students
- Specifically, 80% of female respondents included "comedy" in their top three genres, compared to 60% of male respondents.

Bar Chart - TOP 3 GENRES COUNT BY (MALE + FEMALE)

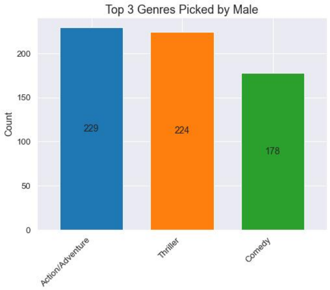

Figure: Top 3 genres (male)

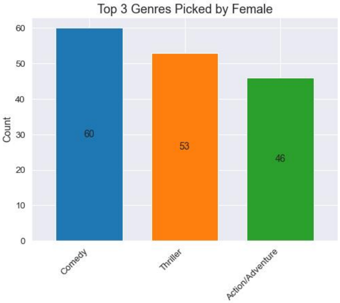

Figure: Top 3 genres (female)

Box Plot - WEEKLY HOURS SPENT vs GENDER

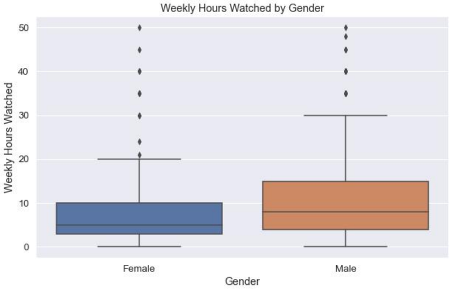

Figure: Weekly hours spent vs. Gender

- Male college students watch more entertainment with a median weekly watch time of 8 hours vs. 5 hours for females.
- Males also have more variability in their watch times
- Some students watch significantly more content than others, as indicated by outliers in the upper half of the box plots for both genders.

Box Plot - MAXIMUM WATCH DURATION AT A STRETCH vs GENDER

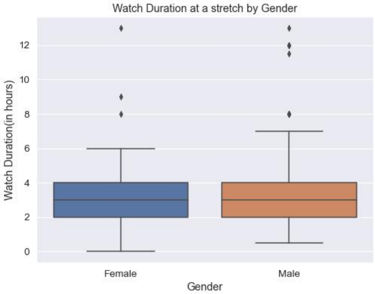

Figure: Maximum watch duration at a stretch vs gender

- Irrespective of gender, the median maximum watch duration at a stretch is around 3 hours
- While 50% of respondents reported a duration between 2 and 4 hour

Proving facts from data

Verifying Central Limit Theorem

Theorem: The central limit theorem states that the sampling distribution of the mean of samples will be normally distributed if the sample size is large enough

Objective: We are taking 5000 samples of sizes 20, 50, and 100 respectively for each of the variables, for verifying CLT.

Here, our variables of interest are monthly expenses, hours spent weekly on entertainment and hours spent in entertainment at a stretch.

Verifying Central Limit Theorem Contd.

1. Verifying CLT for Monthly expenses spent on entertainment:

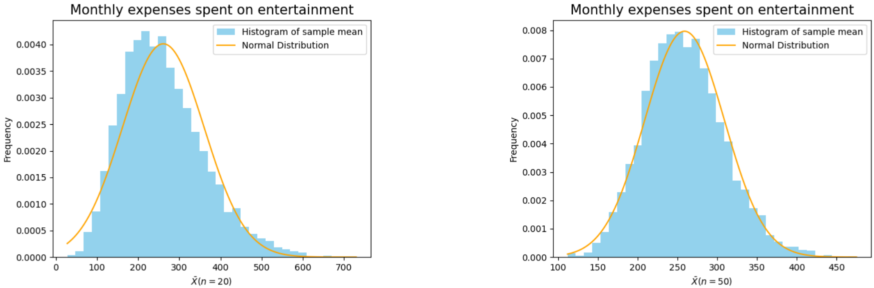

Figure: sample size = 20

Figure: sample size = 50

Figure: Sample mean plots for monthly expenses

Verifying Central Limit Theorem Contd.

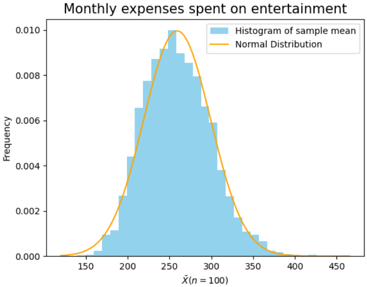

Figure: sample size $= 100$

Figure: Sample mean plot for monthly expenses

Verifying Central Limit Theorem Contd.

2. Verifying CLT for hours spent weekly on Entertainment:

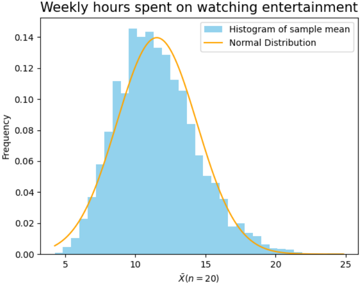

Figure: sample size = 20

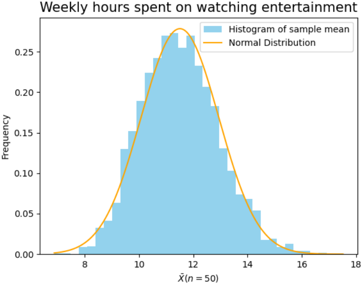

Figure: sample size = 50

Verifying Central Limit Theorem Contd.

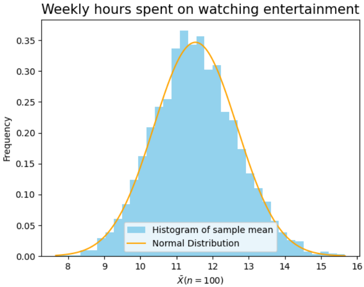

Figure: sample size $= 100$

Figure: Sample mean plots for weekly hours spent

Verifying Central Limit Theorem Contd.

3. Verifying CLT for hours spent at a stretch on Entertainment:

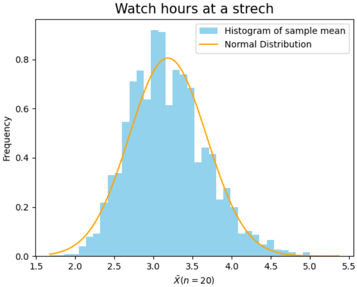

Figure: sample size $= 20$

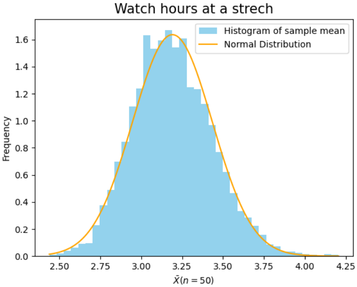

Figure: sample size $= 50$

Verifying Central Limit Theorem Contd.

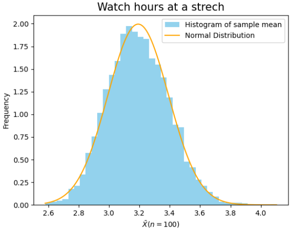

Figure: sample size $= 100$

Figure: Sample mean plots for watch at stretch

Conclusion: Here we can note that as sample size increases, the plot of sample means approaches Normal distribution. Hence, central limit theorem is verified.

Inferences from Data

Using Confidence Interval Estimation

Confidence Interval - 1

Difference of monthly mean of expenditure of North and South People on Entertainment

Here we are interested in calculating $(1-\alpha)100\%$CI for difference of population mean of expenditures of North and South people on Entertainment. From the data collected, we have

$$n_N = 172$$
$$n_S = 183$$

We will call X_i and Y_i be the monthly expenditure of randomly selected South and North person. From data we have

Cont. Confidence Interval - 1

$$\frac{S_X^2}{S_Y^2} \approx 2.040 < 4$$

So we can assume that the population variances are similar. Consider the pooled variance S_p,

$$\begin{align} S_p &= \sqrt{\frac{(n_S - 1)\, S_X^2 + (n_N - 1)\, S_Y^2}{n_N + n_S - 2}} \tag{1} \\ &= \sqrt{\frac{172 \times 128811 + 183 \times 263245}{353}} \tag{2} \\ &\approx 448.26 \tag{3} \end{align}$$

We can calculate the $(1-\alpha)100\%$ CI for difference of population means as follows

Cont. Confidence Interval - 1

$$(\overline{X} - \overline{Y}) \pm t_{\frac{\alpha}{2}, n_N + n_S - 2} \; S_p \sqrt{\frac{1}{n_S} + \frac{1}{n_N}} \tag{4}$$

For $\alpha = 0.05$ and $t_{0.025,356} = 1.98$,it evaluates to

$$(151 \pm 4.999) = (146.001, 155.999) \tag{5}$$

Inference : We can infer that, statistically with 95% confidence, the monthly mean expenditure on Entertainment of South Indians is higher than North Indians. **In other words, South Indians tend to spend more money than North Indians in entertainment**

Confidence Interval - 2

Difference of proportion of people interested in watching romantic movies alone and horror movies alone

Here we are interested in calculating $(1-\alpha)100\%$ CI for difference of population proportion of people interested in romantic movies and horror movies. From the data collected, we have

	Prefer Romantic Movies	Prefer Horror Movies
Prefer to watch alone	137	42
Total	188	85

Cont. Confidence Interval - 2

From that we can get sample proportions,

$$\hat{p}_R = \frac{137}{188} = 0.729$$
$$\hat{p}_H = \frac{42}{85} = 0.494$$

Now we get $(1-\alpha)100\%$ CI for the difference of population proportions using the following expression,

$$(\hat{p_R} - \hat{p_H}) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p_R}\left(1-\hat{p_R}\right)}{n_R} + \frac{\hat{p_H}\left(1-\hat{p_H}\right)}{n_H}}$$

Taking $\alpha = 0.05$ and $z_{\frac{\alpha}{2}} = 1.96$ we get

$$0.235 \pm 0.124 = (0.111, 0.362)$$

Inference : We can infer that,statistically with 95% confidence, proportion of people watching Romantic movies alone are higher than proportion of people watching Horror movies alone.**In other words, people who watch**

Confidence Interval - 3

Comparing of weekly watching time between people from different parts of country

Here we are interested in calculating (1-α)100%CI for difference of population mean of weekly watching time of North and South people.
From the data collected, we have

$$\begin{aligned}\text{no. of south people, } n_S &= 172\\ \text{no. of north people, } n_N &= 183\\ \text{sample mean of south people, } \overline{X}_S &= 14.20\\ \text{sample mean of north people, } \overline{X}_N &= 8.98\end{aligned}$$

Cont. Confidence Interval - 3

$$\text{sample std. deviation of south people, } S_S = 14.35$$
$$\text{sample std. deviation of north people, } S_N = 10.34$$

Since,

$$\frac{1}{2} < \frac{S_S}{S_N} = 1.388 < 2 \tag{6}$$

so we can assume that population variances are similar and can use two sample Pooled t-interval.

$$\begin{aligned} S_p &= \sqrt{\frac{(n_S - 1) S_S^2 + (n_N - 1) S_N^2}{n_N + n_S - 2}} \qquad (7) \\ &= \sqrt{\frac{172 \times 205.9 + 183 \times 106.9}{353}} \qquad (8) \\ &\approx 12.48 \qquad (9) \end{aligned}$$

Cont. Confidence Interval - 3

We can calculate the $(1-\alpha)100\%$ CI for difference of population means as follows,

$$(\overline{X}_S - \overline{X}_N) \pm t_{\frac{\alpha}{2}, n_N+n_S-2}\ S_p \sqrt{\frac{1}{n_S} + \frac{1}{n_N}} \tag{10}$$

For $\alpha = 0.05$ and $t_{0.025,353} = 1.97$,it evaluates to

$$(5.22 \pm 2.6) = (2.62, 7.82) \tag{11}$$

Inference : We can infer that, statistically with 95% confidence, the mean weekly watching time of South Indians is higher than North Indians. **In other words, South Indians tend to spend more time in watching entertainment than North Indians.**

Confidence Interval - 4

Comparison of variation in monthly expense between gender

Here we are interested in calculating $(1-\alpha)100\%$ CI for ratio of variance of monthly expense between male and female.
From the data collected, we have

$$\begin{aligned} \text{no. of male, } & n_M = 279 \\ \text{no. of female, } & n_F = 76 \\ \text{sample std. deviation of male, } & S_M = 407.1 \\ \text{sample std. deviation of female, } & S_F = 571.8 \end{aligned}$$

We can calculate $(1-\alpha)100\%$ CI for ratio of variance of monthly expense between male and female as follows,

Cont. Confidence Interval - 4

$$\left(\frac{1}{F\alpha/2(n-1,m-1)}\frac{S_F^2}{S_M^2}, F\alpha/2(m-1,n-1)\frac{S_F^2}{S_M^2}\right) \tag{12}$$

For $\alpha = 0.05$ and $F_{0.025}(75, 278) = 1.41$ and $F_{0.025}(278, 75) = 1.46$, it evaluates to

$$\left(\frac{1}{1.41} \times 1.97, 1.46 \times 1.97\right) = (1.40, 2.88) \tag{13}$$

Inference: We can infer that, statistically with 95% confidence, the variation in monthly expense is more for female than male.

Using Hypothesis Testing

Hypothesis Testing - 1

Comparison of watching preference between people from different parts of our country

Let π_1 denote the proportion of South Indians watching south cinema and pi_2 denote the proportion of north Indians watching Bollywood.

$$\begin{aligned} \text{South Indians watching south cinema}, \;\; n_{SS} &= 92 \\ \text{Total South Indians}, \;\; n_S &= 172 \\ \text{North Indians watching Bollywood}, \;\; n_{NB} &= 64 \\ \text{Total North Indians}, \;\; n_N &= 183 \end{aligned}$$

So,

$$\hat{\pi_1} = \frac{n_{SS}}{n_S} = 0.535 \qquad \hat{\pi_2} = \frac{n_{NB}}{n_N} = 0.35 \tag{14}$$

Now,

$$H_0 : \pi_1 - \pi_2 \leq 0 \qquad H_a : \pi_1 - \pi_2 > 0 \tag{15}$$

Contd. Hypothesis Testing - 1

$$n_1\hat{\pi}_1 \geq 5 \quad n_1(1-\hat{\pi}_1) \geq 5 \quad n_2\hat{\pi}_2 \geq 5 \quad n_2(1-\hat{\pi}_2) \geq 5 \tag{16}$$

So, the test statistic is

$$\begin{aligned} Z &= \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} && (17) \\ &= \frac{0.535 - 0.35}{\sqrt{\frac{0.535\times0.465}{172} + \frac{0.35\times0.65}{183}}} && (18) \\ &= 3.56 && (19) \end{aligned}$$

and

$$z_{0.05} = 1.645 \tag{20}$$

Contd. Hypothesis Testing - 1

Rejection Region approach:

We will reject H_0 if test statistic $Z > Z_\alpha$
With significance level, $\alpha = .05$

$$z_{0.05} < Z$$

Since $Z > z_{0.05}$ which means it's lying in rejection region so we will reject H_0.

Inference: South people prefer South cinema more than North people prefer Bollywood.

Hypothesis Testing - 2

Preference of watching horror movies

From the data collected, we have

$$\begin{aligned} \text{people watching horror movie in a group,} n_G &= 211 \\ \text{total people,} n &= 355 \end{aligned}$$

$$\hat{\pi} = \frac{n_G}{n} = 0.594 \tag{21}$$

Let π denote the proportion of people watching horror movie in group. Here,

$$H_0 : \pi \leq 0.5 \quad vs. \quad H_a : \pi > 0.5 \tag{22}$$

Cont. Hypothesis Testing - 2

Test statistic:

$$
\begin{aligned}
Z &= \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} && (23)\\
&= \frac{0.594 - 0.500}{\sqrt{\frac{0.594 \times 0.406}{355}}} && (24)\\
&= 3.606 && (25)
\end{aligned}
$$

p-value approach:

$$
\begin{aligned}
p &= P\left(z > Z\right) && (26)\\
&= P\left(z > 3.606\right) && (27)\\
&= 0.00016 && (28)
\end{aligned}
$$

Cont. Hypothesis Testing - 2

With significance level $\alpha = 0.05$

$$\alpha > p$$

so we can reject H_0.

Inference : We can conclude that the observations support the hypothesis that people generally prefer watch horror movies in group.

Hypothesis Testing - 3

Comparing mean watch time at a stretch among genders

From the data collected, we have mean watch time at a stretch of male ($\overline{x}_m$) and female ($\overline{x}_f$) are as follows,

$$\overline{x}_m = 3.157$$
$$\overline{x}_f = 3.310$$

Since we have $\overline{x}_f > \overline{x}_m$, we will do Hypothesis Testing with,

Cont. Hypothesis Testing - 3

$$H_0 : \mu_f - \mu_m \leq 0 \quad H_a : \mu_f - \mu_m > 0$$

And from data we got ,

$$\begin{aligned} S_f &= 2.607 \\ S_m &= 2.072 \\ \implies \frac{S_f}{S_m} &\approx 1.258 < 2.0 \end{aligned}$$

So we assume population variances to be same and then pooled variance will be equal to $(n_f = 76, n_m = 279)$,

$$S_p = \sqrt{\frac{S_f^2\,(n_f - 1) + S_m^2\,(n_m - 1)}{n_f + n_m - 2}}$$

Cont. Hypothesis Testing - 3

$$
\begin{aligned}
S_p &= \sqrt{\frac{75 \times 6.796 + 278 \times 4.297}{353}} \\
&\approx \sqrt{4.827} \\
&= 2.197
\end{aligned}
$$

The test statistic will be,

$$
\begin{aligned}
t &= \frac{(\overline{x}_f - \overline{x}_m) - 0}{S_p\sqrt{\frac{1}{n_f} + \frac{1}{n_m}}} && (29) \\
&= \frac{3.310 - 3.157}{2.197 \times 0.128} && (30) \\
&= \frac{0.159}{0.281} && (31) \\
&= 0.565 && (32)
\end{aligned}
$$

Cont. Hypothesis Testing - 3

Rejection Region Approach :

We will reject H_0 if the test statistic $t > t_{\alpha, n_f + n_m - 2}$
With $\alpha = 0.05$ (Significance Level),

$$t_{0.05,353} = 1.950$$

The observed test statistic (0.565) less than 1.95 , hence it is not in the rejection region.

Inference : There is no enough evidence to conclude that females, at a stretch, watching entertainment more than male.

Hypothesis Testing - 4

Distribution followed by different variables of interest

Part a : Distribution of money spend in a month

H_0 : Data follow normal distribution
H_a : Data doesn't follow normal distribution
From QQ plot 34 we can see that points are very far to make a straight line.
So, we can conclude that it's not following normal distribution.

Cont. Hypothesis Testing - 4

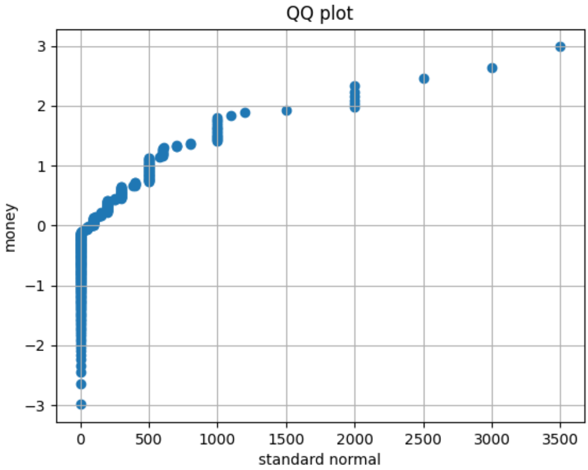

Figure: QQ plot of money spend in a month

Cont. Hypothesis Testing - 4

Part b: Distribution of hours spend at a stretch
H_0 : Data follow normal distribution
H_a : Data doesn't follow normal distribution
From QQ plot 35 we can see that points are close to a straight line.
So, we can say that its nearly following normal distribution.

Cont. Hypothesis Testing - 4

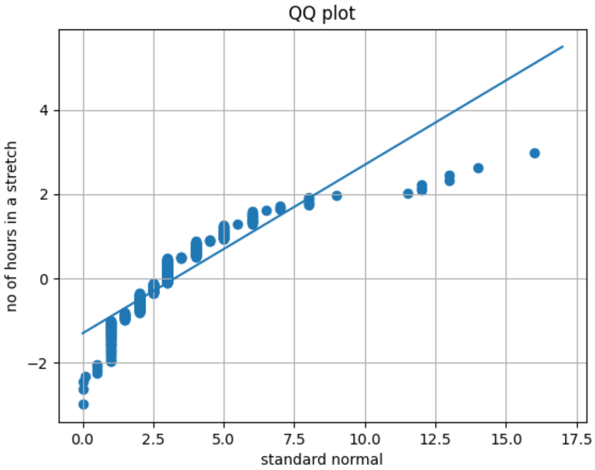

Figure: QQ plot of hours spend at a stretch

Conclusion

Summary

We were able to sample a decent population which helped us in visualizing and analyzing the various trends. We visualized the data using various plots like bar, box, pie and more. Our analysis helped reveal notable disparities in the preferences and expenditures of individuals attending engineering colleges concerning entertainment, particularly in terms of the types of content(movies/anime/web series) consumed and the duration and costs involved. These variations were found to be significantly influenced by both region and gender.