Inferential

Japnit Singh

2023-12-03

Inferential Statistics

1A. Mean time of documentary and reality // TODO: Rephrase this question

Consider the null hypothesis to be as follows $H_0 : \mu_1$ - $\mu_2 \geq 0$. In the context of the question, null hypothesis states that the average runtime of a documentry show is greater than or equal to the average runtime of a reality show.

As such the alternate hypothesis would be $H_1 : \mu_1$ - $\mu_2 < 0$. In the context of the question, alternate hypothesis states that the average runtime of a documentry show is lesser than the average runtime of a reality show.

Here

μ_1 is the population mean of runtime of documentry TV shows.

μ_2 is the population mean of runtime of reality TV shows.

Significance level of the test $\alpha = 5\%$

```
filter_sample_data = sample_data[sample_data$episode_run_time > 0, ];
#print(head(filter_sample_data, 50));
#cat(subset(filter_sample_data, 1:50), file="", sep="\n")
#cat(capture.output(print(filter_sample_data[1:50, ])), sep="\n")

documentry_filter_sample = subset(filter_sample_data, type == "Documentary");
reality_filter_sample = subset(filter_sample_data, type == "Reality");
```

```
result = t.test(documentry_filter_sample$episode_run_time,
       reality_filter_sample$episode_run_time,
       var.equal = FALSE,
       alternative = "less");
print(result$statistic);
```

```
##         t
## -3.213635
```

```
print(result$p.value);
```

```
## [1] 0.0006574883
```

Result:

t_statistic = -3.213635 and p_value = 0.0006574883

Since p_value is less than α (0.05) we reject the null hypothesis H_0.

We have enough evidence to claim that the average runtime of a documentry show is not greater than or equal to the average runtime of a reality show.

1B. Average episodes per season

```r
scripted_sample_data = subset(sample_data, type == "Scripted");

filter_scripted_data = scripted_sample_data[scripted_sample_data$number_of_episodes > 0, ];
filter_scripted_data = filter_scripted_data[filter_scripted_data$number_of_seasons > 1, ];
# TODO: shouldn't scripted be enough as a filter?
filter_scripted_data = filter_scripted_data %>% mutate(episodes_per_season = number_of_episodes/numbe
```

i. Comedy and Drama // TODO: Rephrase this question

Consider the null hypothesis to be as follows $H_0 : \mu_1 - \mu_2 \leq 0$. In the context of the question, null hypothesis states that the average episodes per season of comedy shows is lesser than or equal to the average episodes per season of darma shows.

As such the alternate hypothesis would be $H_1 : \mu_1 - \mu_2 > 0$. In the context of the question, alternate hypothesis states that the average episodes per season of comedy shows is greater than the average episodes per season of darma shows.

Here

μ_1 is the population mean of episodes per season of comedy shows.

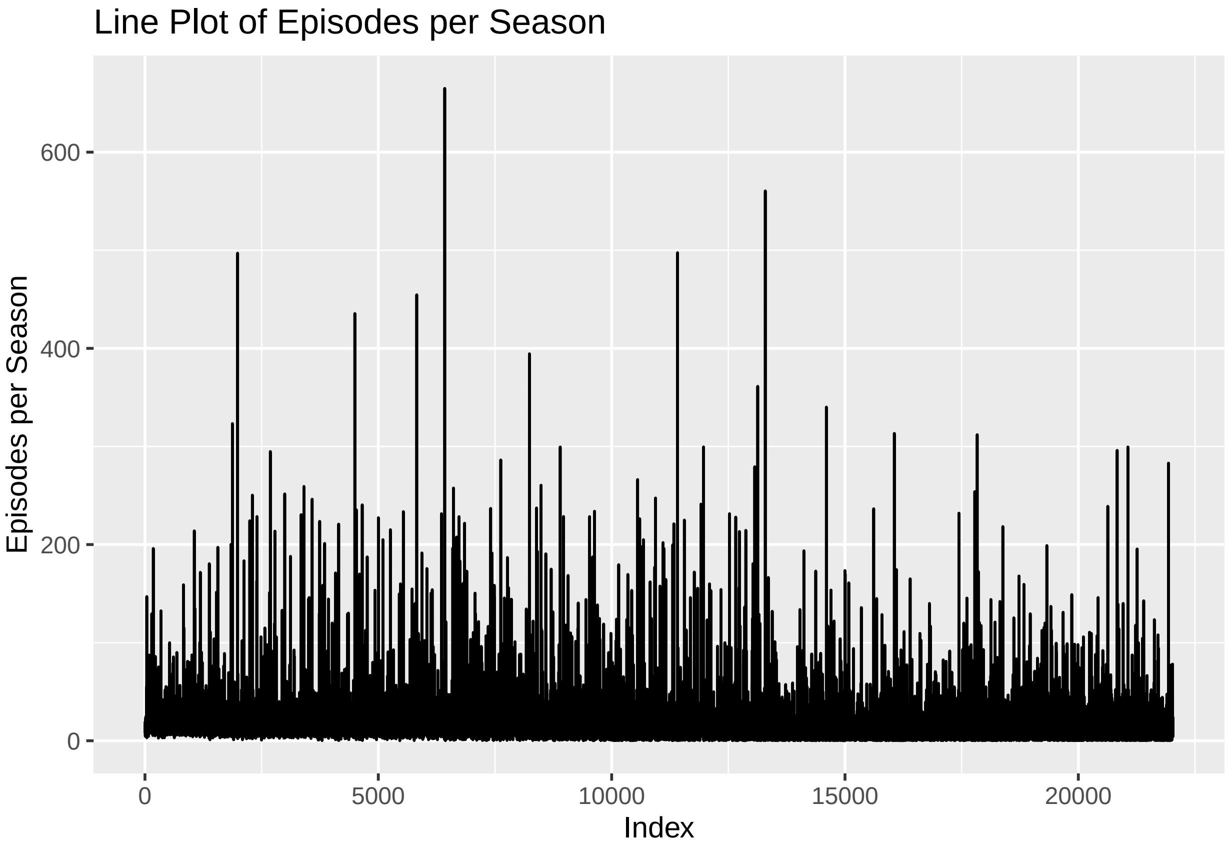
μ_2 is the population mean of episodes per season of drama shows.

Significance level of the test $\alpha = 5\%$

```r
# expanded_filter_data = expand_num_by_cat(filter_scripted_data, "genres", "episodes_per_season"); #
# comedy_filter_data =  subset(expanded_filter_data, cat_col == "Comedy");
# drama_filter_data =  subset(expanded_filter_data, cat_col == "Drama");
# colnames(expanded_filter_data);

comedy_filter_data =  subset(filter_scripted_data, grepl("Comedy", genres));
drama_filter_data =  subset(filter_scripted_data, grepl("Drama", genres));
```

```r
ggplot(filter_scripted_data, aes(x = 1:nrow(filter_scripted_data), y = episodes_per_season)) +
  geom_line() +
  labs(x = "Index", y = "Episodes per Season", title = "Line Plot of Episodes per Season");
```



```
result = t.test(comedy_filter_data$episodes_per_season,
       drama_filter_data$episodes_per_season,
       var.equal = FALSE,
       alternative = "less");
print(result$statistic);
```

```
##         t
## -3.598296
```

```
print(result$p.value);
```

```
## [1] 0.000160871
```

Result:

t_statistic = -3.598296 and p_value = 0.000160871

Since p_value is less than α (0.05) we reject the null hypothesis H_0.

We have enough evidence to claim that the average episodes per season of comedy shows is not lesser than or equal to the average episodes per season of darma shows.

ii. Comparing episodes_per_season of SciFi and Action & Adventure // TODO: Rephrase this question

Consider the null hypothesis to be as follows $H_0 : \mu_1 - \mu_2 \leq 0$. In the context of the question, null hypothesis states that the average episodes per season of SciFi shows is lesser than or equal to the average episodes per season of Action & Adventure shows.

As such the alternate hypothesis would be H_1 : μ_1 - $\mu_2 > 0$. In the context of the question, alternate hypothesis states that the average episodes per season of SciFi shows is greater than the average episodes per season of Action & Adventure shows.

Here

μ_1 is the population mean of episodes per season of SciFi shows.

μ_2 is the population mean of episodes per season of Action & Adventure shows.

Significance level of the test $\alpha = 5\%$

```
scifi_filter_data =  subset(filter_scripted_data, grepl("Sci-Fi & Fantasy", genres));
adv_filter_data =  subset(filter_scripted_data, grepl("Action & Adventure", genres));
```

```
result = t.test(scifi_filter_data$episodes_per_season,
       adv_filter_data$episodes_per_season,
       var.equal = FALSE,
       alternative = "less");
print(result$statistic);
```

```
##         t
## -1.460528
```

```
print(result$p.value);
```

```
## [1] 0.0721231
```

Result:

t_statistic = -1.460528 and p_value = 0.0721231

Since p_value is greater than α (0.05) we fail to reject the null hypothesis H_0.

We have enough evidence to claim that the average episodes per season of SciFi shows is lesser than or equal to the average episodes per season of Action & Adventure shows.

iii. Comparing episodes_per_season of Family > Crime // TODO: Repharase this question

Consider the null hypothesis to be as follows H_0 : μ_1 - $\mu_2 \geq 0$. In the context of the question, null hypothesis states that the average episodes per season of Family shows is greater than or equal to the average episodes per season of Crime shows.

As such the alternate hypothesis would be H_1 : μ_1 - $\mu_2 > 0$. In the context of the question, alternate hypothesis states that the average episodes per season of Family shows is lesser than the average episodes per season of Crime shows.

Here

μ_1 is the population mean of episodes per season of Family shows.

μ_2 is the population mean of episodes per season of Crime shows.

Significance level of the test $\alpha = 5\%$

```
  family_filter_data =  subset(filter_scripted_data, grepl("Family", genres));
  crime_filter_data =  subset(filter_scripted_data, grepl("Crime", genres));
```

```
  result = t.test(family_filter_data$episodes_per_season,
         crime_filter_data$episodes_per_season,
         var.equal = FALSE,
         alternative = "greater");
  print(result$statistic);
```

```
##        t
## 11.55692
```

```
  print(result$p.value);
```

```
## [1] 1.619423e-30
```

Result:

t_statistic = 11.55692 and p_value = 1.619423e-30

Since p_value is lesser than α (0.05) we reject the null hypothesis H_0.

We have enough evidence to claim that the average episodes per season of Family shows is lesser than or equal to the average episodes per season of Family shows.

1C Independent : we can consider the popularity rating of genre in english (particular language) / country or origin, eg weather the popularity of animation is same in english to that in japanese. // TODO: Rephrase this question

Consider the null hypothesis to be as follows H_0 : μ_1 - $\mu_2 \leq 0$. In the context of the question, null hypothesis states that the average popularity of English Animated shows is lesser than or equal to the average popularity of Japanese Animated shows.

As such the alternate hypothesis would be H_1 : μ_1 - $\mu_2 > 0$. In the context of the question, alternate hypothesis states that the average popularity of English Animated shows is greater than the average popularity of Japanese Animated shows.

Here

μ_1 is the population mean of popularity of English Animated shows.

μ_2 is the population mean of popularity of Japanese Animated shows.

Significance level of the test $\alpha = 5\%$

```
#anime_sample_data = subset(sample_data, genres == "Animation"); #FALSE, this is strict matching
anime_sample_data = subset(sample_data, grepl("Animation", genres)); # This is 'contains' matching

q1 = quantile(anime_sample_data$popularity, probs = 0.25);
q3 = quantile(anime_sample_data$popularity, probs = 0.75);
iqr = q3 - q1;
upper_whisker_limit = q3 + (1.5 * iqr);
lower_whisker_limit = q1 - (1.5 * iqr);
```

```
anime_sample_data = subset(anime_sample_data, popularity > lower_whisker_limit); # This should not be 0
anime_sample_data = subset(anime_sample_data, popularity < upper_whisker_limit);

english_anime_data = subset(anime_sample_data, original_language == "en"); # strick matchin is correct
japanese_samples = subset(anime_sample_data, original_language == "ja");
```

```
  result = t.test(english_anime_data$popularity,
         japanese_samples$popularity,
         var.equal = FALSE,
         alternative = "less");
  print(result$statistic);
```

```
##         t
## -2.510185
```

```
  print(result$p.value);
```

```
## [1] 0.006047949
```

Result:

t_statistic = 9.019615 and p_value = 1 (In case of strict matching lol)

t_statistic = -2.510185 and p_value = 0.006047949 (In case of contains matching. We are using this.)

Since p_value is less than α (0.05) we reject the null hypothesis H_0.

We have enough evidence to claim that the average popularity of English Animated shows is not lesser than or equal to the average popularity of Japanese Animated shows.

1D Comparing Rating (Vote Average) of Animated English and Animated Japanese Shows

Consider the null hypothesis to be as follows H_0 : μ_1 - $\mu_2 \leq 0$. In the context of the question, null hypothesis states that the average Rating of English Animated shows is lesser than or equal to the average Rating of Japanese Animated shows.

As such the alternate hypothesis would be H_1 : μ_1 - $\mu_2 > 0$. In the context of the question, alternate hypothesis states that the average Rating of English Animated shows is greater than the average Rating of Japanese Animated shows.

Here

μ_1 is the population mean of rating of English Animated shows.

μ_2 is the population mean of rating of Japanese Animated shows.

Significance level of the test $\alpha = 5\%$

```
  result = t.test(english_anime_data$vote_average,
         japanese_samples$vote_average,
         var.equal = FALSE,
         alternative = "less");
  print(result$statistic);
```

```
##         t
## -14.29599
```

```
print(result$p.value);
```

```
## [1] 8.377412e-46
```

Result:

t__statistic = -7.357608 and p__value = 1.533366e-13 (In case of strict matching)

t__statistic = -14.29599 and p__value = 8.377412e-46 (In case of contains matching. We are using this.)

Since p__value is less than α (0.05) we reject the null hypothesis H_0.

We have enough evidence to claim that the average Rating of English Animated shows is not lesser than or equal to the average popularity of Japanese Animated shows.

1E # Comparing runtime of pre and post covid // TODO: Rephrase this question

Consider the null hypothesis to be as follows $H_0 : \mu_1 - \mu_2 = 0$. In the context of the question, null hypothesis states that the average runtime of shows before covid is same as the average runtime of shows after covid.

As such the alternate hypothesis would be $H_1 : \mu_1 - \mu_2 \neq 0$. In the context of the question, alternate hypothesis states average runtime of shows before covid is significantly different to average runtime of shows after covid.

Here

μ_1 is the population mean of runtime of pre-covid shows.

μ_2 is the population mean of runtime of post-covid shows.

Significance level of the test $\alpha = 5\%$

```
filter_scripted_data = scripted_sample_data[scripted_sample_data$episode_run_time > 0, ]; # scripted_
filter_scripted_data$last_air_date = as.Date(filter_scripted_data$last_air_date, format = "%m-%d-%y")
filter_scripted_data$first_air_date = as.Date(filter_scripted_data$first_air_date, format = "%m-%d-%y
# filter_scripted_data = add_pre_post_covid(filter_scripted_data);
# condition_1 <- with(df, year(start_date) >= 2021 & year(start_date) <= 2023)
# condition_2 <- with(df, year(end_date) <= 2019 & year(end_date) >= 2015)


precovid_filter_data =   filter_scripted_data[(filter_scripted_data$last_air_date > as.Date("01-01-20
postcovid_filter_data =   filter_scripted_data[(filter_scripted_data$first_air_date > as.Date("01-01-2
#Intentionally missed 2020. Was the year that bad?
# postcovid_filter_data =  subset(filter_scripted_data, Covid_Type == "Post Covid");
# print(head(precovid_filter_data,5))
```

```
#result = t.test(precovid_filter_data$episode_run_time,
 #      postcovid_filter_data$episode_run_time,
 #      var.equal = FALSE,
  #     alternative = "two.sided");
#print(result$statistic);
#print(result$p.value);
```

2A Same pre processing issue as 1E

2B Amazon rating variance vs Netfllix rating variance (rating = voting_average) // TODO: rephrase the question

Consider the null hypothesis to be as follows $H_0 : \sigma_1^2$ - $\sigma_2^2 = 0$. In the context of the question, null hypothesis states that the variance in rating of shows on Netflix is the same as variance in rating of shows on Amazon Prime

As such the alternate hypothesis would be $H_1 : \sigma_1^2$ - $\sigma_2^2 \neq 0$. In the context of the question, alternate hypothesis states that the variance in rating of shows on Netflix is significantly different from variance in rating of shows on Amazon Prime

Here

σ_1 is the standard deviation in rating of shows on Netflix.

σ_2 is the standard deviation in rating of shows on Amazon Prime.

Significance level of the test $\alpha = 5\%$

```
filter_sample_data = sample_data[sample_data$vote_average >0, ];
netflix_sample_data = subset(sample_data, grepl("Netflix", networks));
amazon_sample_data = subset(sample_data, grepl("Prime Video", networks));
```

```
f_statistic = var(netflix_sample_data$vote_average) / var(amazon_sample_data$vote_average);
df1 = nrow(netflix_sample_data) - 1;
df2 = nrow(amazon_sample_data) - 1;
p_value = 2 * pf(f_statistic, df1, df2);

print(f_statistic);
```

```
## [1] 0.5180385
```

```
print(p_value);
```

```
## [1] 6.529343e-32
```

Result:

f_statistic = 0.5180385 and p_value = 6.529343e-32

Since p_value is less than α (0.05) we reject the null hypothesis H_0.

We have enough evidence to claim that the variance in rating of shows on Netflix is not the same as variance in rating of shows on Amazon Prime

Including Plots

You can also embed plots, for example:

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.