Statistical Inference

Minerva University
CS50: Formal Analyses
Prof. M. Averett
December 17, 2021

MINERVA
UNIVERSITY

Inferential Statistics Report

1. Introduction

This report will be exploring the Wii Mario Kart Auctions from Ebay dataset, specifically how the number of bids get impacted by the used condition of the game. I will be using the Wii Mario Kart auctions from Ebay dataset in this report, to explore whether the new and old games have any impact on the number of bids. I will be using the difference of means test to make any inferences on the sample for the significance. My hypothesis is: The newer condition games will result in a higher mean, hence receive a higher number of bids.

2. Dataset

This dataset is from Openintro, and is a 2009 auction data from Ebay, for the game Mario Kart Nintendo Wii. As an avid fan of the Nintendo games from watching my brother play on his Gameboy and DS Lite, I always wanted to buy a Wii when I was younger. However, the Wii was out of budget for my parents, and they weren't convinced that it was a popular game. Back in 2009, I was around 6-years old, so I am interested in finding out if 6-year-old Paulina could have presented a better case for my parents back then, by convincing them to buy an old Wii for me. Hence, I am going to be exploring whether there would have been any difference in the number of bids of the games depending on the condition of it.
The variables here are:

- Condition of the game = this can either be new or old, so it is a dummy quantitative discrete variable which will be either 0 or 1 on python (as a Boolean).
- Number of bids = this will range from 0 to however many bids the auction gets, but it is a quantitative discrete variable since you can't have 'half' of a bid. The data type on python would be just an integer. I am also assuming here that a higher number of bids means that it is more expensive.
- Confounding variables = There are other factors that could affect the number of bids in the game that is not explored in this report. For instance, the number of capital letters in the title, or shipping prices. But as these are not the main focus of my analysis they are ignored.[1]

[1] **#variables**: I identify the different types of the variables I am working with. The parameters here would be the measure of location and spread of my data once I begin to find my samples.

I used Excel for any main processing of data, such as filtering for the old and new conditions of the game. But I used python to calculate the statistics for the two sample datasets I am going to explore.

3. Analysis

*any in depth calculations can be found in the appendix

Hypotheses

$H_0 =$ There will be no difference between the means of the old and new games.
$\therefore \mu_{old} - \mu_{new} = 0$
$H_A =$ There will be a difference between the means of the old and new games.
$\therefore \mu_{old} - \mu_{new} \neq 0$

Table 1: summary statistics of both samples of old and new condition games.

	Old condition games	New condition games
Sample size	$n_{old} = 84$	$n_{new} = 59$
Mean	13.2	14.1
Median	13	13
Mode	14	16
Standard deviation	5.46	6.34
Range	27.0	28.0

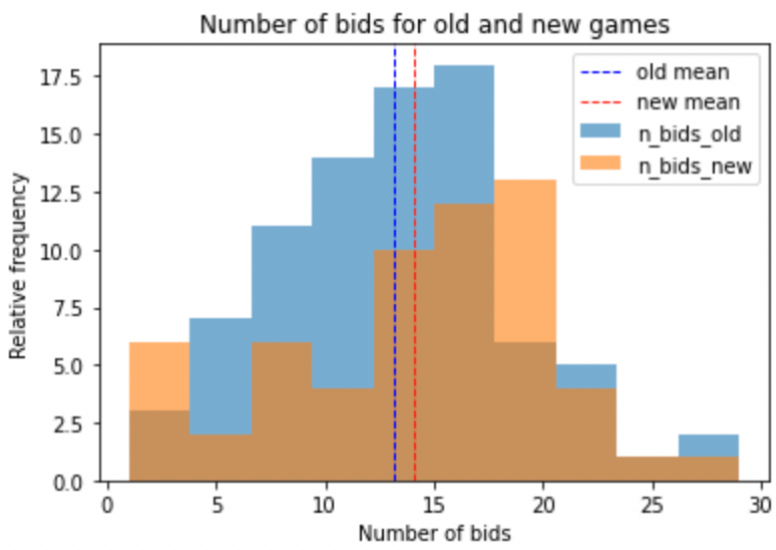


Figure 1: number of bids histogram visualized with both samples together to show the differences more clearly Bids for the old games is more narrow than the new games. Their means are plotted as well with the vertical dotted line.

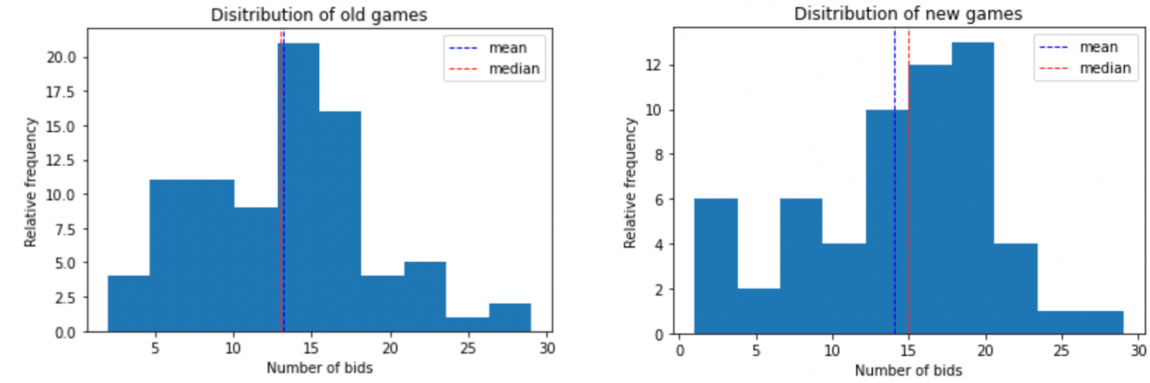


Figure 2: histograms of the 2 samples separately, with the mean and median shown with the vertical dotted line. Here we can see the skewness more easily as we can compare the mean and median together.

In both cases, they follow more of a right-skewed distribution,[2] however at first glance of the figure, it looks like the new conditions of the game have a higher mean than the older ones. Moreover, the new distribution seems to have more variability and spread than the older one. [3]Perhaps this is because of how people are more psychologically inclined to buy newer products due to the subconscious bias against older conditions (Joung, 2013)

[2] **#distributions**: I identify that it is right skewed as the 'tails' of the histogram are on the right. Perhaps if there was a larger sample size, and the sample was collected multiple times to form a sampling distribution, we could have seen it approximating to a normal curve as these sample sizes satisfy the central limit theorem (as the sample size is larger than 30).

[3] **#descriptivestats**: it is appropriate to create a histogram here as the purpose is to examine the distribution of the sample data. Furthermore, the variables given are quantitative discrete (number of bins), so it is possible to break them up into 'bins' of intervals. There are a few outliers, especially in the new games, with the left side showing more spikes. Hence this shifts the mean of the new games further right, as seen in figure 1. Python was used to create these graphs.

Difference of Means test

The difference of means test is used to see how much difference here is between two different groups in an experiment. So as my hypothesis is how newer condition games will receive more bids, the experimental group would be the new conditions, and the control group would be the old conditions.

Since the dataset's sample size n is larger than 30, I could use a z score calculation, however the population standard deviation is not given, hence I need to use a t score calculation to carry out the hypothesis test. Furthermore, the number of bids for each case and in each sample group are all independent of each other. Therefore, it is appropriate to use the t-distribution to model the difference of the two means.

The difference between the two-sample means are approximated as the point estimate.

$$point\ estimate = \overline{x_1} - \overline{x_2}$$

Then using this point estimate, we can calculate the standard error using the formula:

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Which can be approximated to:

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

as the population standard deviation is unknown, we estimate the standard error using the standard deviation from the samples. (Diez, et al., 2015)

Drawing the null hypothesis standard distribution is helpful in visualizing the p-value that we will be finding. Since it is a two-tailed test (as the alternative hypothesis does not include a range), both sides of the tail are shaded.

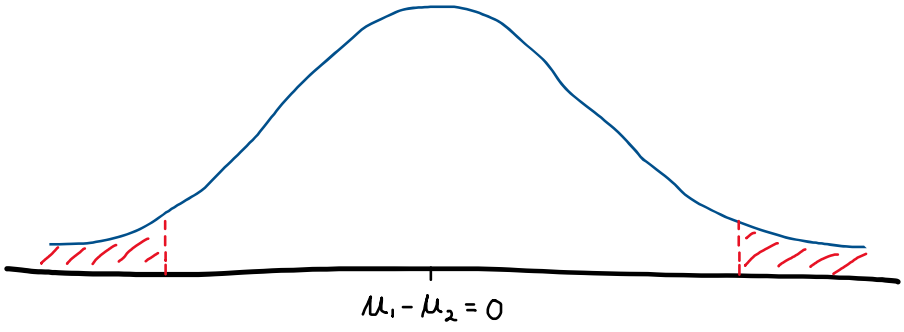


Figure 3: a sketch of the null hypothesis distribution

From here we calculate the t-score of this null hypothesis distribution with the formula:

$$T = \frac{x - mean}{standard\ error} \text{(Diez, et al., 2015)}$$

The calculated t score from python outputs: 0.946

Then we need to find the p-value from the z score table. The degree of freedom is chosen by taking the lowest sample size among the 2 samples and doing $n - 1$ on it.

This can be done through python as well.

The pooled standard deviation is a calculation that helps the t test become more accurate. This can be used when calculating for effect size. To test for the effect size, we need to calculate Cohen's d. This is done by dividing the difference of means by the pooled standard deviation.

The overall outputs are:

T score = 0.946
p value = 0.348
Cohen's d = 0.155

As I already set my significance level, alpha, as 0.05, the p value that was derived is significantly larger than the alpha value.[4] This means that there is insufficient evidence to say that there is a difference in means of the number of bids for old and new games. Hence, the null hypothesis is not rejected. However, this does not mean that the condition of a game has a strong relationship with the number of bids. There may be other external factors that were not taken into account of into the calculations. If there actually is a difference, this would be a Type 2 error. To detect Type 2 errors, we need to collect a larger sample size. Also, as Cohen's d is less than 0.2, the effect size is small, and hence not very practically significant.[5]

[4] **#probabiltiy**: The alpha value here of 0.05 means that there is a 5% chance to commit a Type 1 error. It is related to the standard 95% confidence interval of the distribution curve. The p value here is usually derived from the t score table, which is the probability that the values lie within the area under the curve. Comparing the p value to alpha helps us to check whether there is any statistical significance between the two variables we are exploring.
[5] **#significance**: I use the difference of means hypothesis test to analyze the plausibility of whether the number of bids statistically has any significant difference in their means. Practical significance shows how meaningful it is in real life, whereas statistical significance only shows its significance within the confines of the statistical analysis and dataset. Due to my low Cohen's d value, it is not very practically significant as well.

Confidence interval
The formula for constructing a confidence interval is:

$$difference\ of\ mean\ \pm t^{*}SE$$

Where the point estimate is the sample mean, t* is the critical t score, and SE is the standard error of the point estimate. There is a version of the formula with the z score, but since population standard deviation is not given, it is not used in the calculation.

Therefore, after inputting this formula into python, the confidence interval outputs as:
0.867 and 0.935.

This means that I am 95% confidence that the number of bids increase from 0.867 to 0.935 depending on the condition of the game. However, this does not seem plausible at all.[6]

4. Results and Conclusions

As my calculated p value was bigger than the significance level set, I failed to reject the null hypothesis. This usually means that there is no difference between the means of both samples, and it can be seen visually through Figure X, where the means of both samples are quite similar. However, this does not mean that there is any correlation between the two, and there are other confounding variables that affect the number of bids. Furthermore, Cohen's d showed a very small effect size. Which overall means that the number of bids and the condition of the games do not really have statistical significance from my calculations. Hence, 6-year-old Paulina would not have been able to convince my parents to buy the Wii game because the older games weren't cheaper than the newer ones, there was not much difference.

For an argument to be deductive, the premises must necessarily follow the conclusion. However, with statistical inferences, we are making conclusions about a larger population based on a small sample size. Hence, the premises do not necessarily follow the conclusion, so statistical inference is a form of induction.

Premises in the context of statistical inference is the quality of our dataset, and whether it is reliable depends on if the data satisfies the Central Limit Theorem (CLT). The CLT explains that as the sample size gets larger, the distribution approximates to a normal curve. Since my samples both had a sample size of larger than 30, it satisfies the CLT. My inference was quite

[6] **#confidenceintervals**: I set the confidence level to 95% as this means that I am confident that 95% certain the true mean is within the range of values of the confidence interval. I calculate the confidence interval using the critical t value from python, shown in the appendix. However it does not really seem to make sense as the number of bids is a discrete variable. So for it to increase in decimals does not make practical sense.

strong as the significance level was set to 0.05. There was only a 5% chance that I may have committed a type 1 error.[7]

5. References

Diez, D., Barr, C., & Cetinkaya-Rundel, M. (2015). Sections 3.2, 4.,1 and 4.4. *OpenIntro Statistics (3rd ed.)*. Retrieved from https://leanpub.com/openintro-statistics

Joung, H. M. (2013). Materialism and clothing post- purchase behavior. *Journal of Consumer Marketing*, *30*(6), 530–537.

6. Reflection

For session 19 of CS50, where we explored sampling distributions and the central limit theorem in class, the prep poll required us to calculate the mean and standard deviation of a sampling distribution. I found my TA's feedback helpful as they not only gave detailed explanation on the question, but also reminded me to check the outcome index's examples. This was helpful as I was able to use the examples for this assignment to ensure that I was applying the HC correctly.

The LBA could have been improved by exploring the statistical significance of the number of people at sky from the data I got, assuming that it is a sample. Then set up a hypothesis test with a difference of means! I could answer whether the estimation I came up with was really statistically significant or not.

Word count: 1,538

[7] **#induction**: Statistical inference's conclusions go beyond the premises of the argument, as it is making a generalization about the larger population from samples. So this process is inductive. However we need to be careful with hasty generalization fallacy, so we need to consider other external variables or factors that may contribute to our conclusions.

7. Appendix

*Session 24 of CS50 was referenced for the difference of means function

Summary statistics of dataset

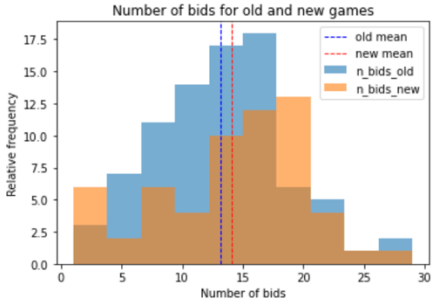
In [178]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import scipy
from scipy import stats

#I first called in my processed csv file
df = pd.read_csv('mariokart_proc.csv', sep=',')

#Then split the columns into separate lists
old = df['n_bids_old'].tolist()
new = df['n_bids_new'].tolist()
del new[-25:]      #I had to delete the last 25 elements
                   #from the 'new' list because it was nAn.

#These are the statistics for the old dataset
mean1 = np.mean(old)
median1 = np.median(old)
mode1 = stats.mode(old)
stdev1 = np.std(old)
rnge1 = np.max(old) - np.min(old)
print("Number of bids for old game summary statistics")
print('sample size: ', len(old))
print('median: ', median1)
print('mean: ',round(mean1,1))
print('mode: ',mode1)
print('standard deviation: ', round(stdev1,2))
print('range: ', rnge1)

print('\n')    #I added a line break to make the output look neater

#These are the statistics for the old dataset
mean2 = np.mean(new)
median2 = np.median(new)
mode2 = stats.mode(new)
stdev2 = np.std(new)
rnge2 = np.max(new) - np.min(new)
print("Number of bids for new game summary statistics")
print('sample size: ', len(new))
print('median: ', median2)
print('mean: ',round(mean2,1))
print('mode: ',mode2)
print('standard deviation: ', round(stdev2,2))
print('range: ', rnge2)
```

```
Number of bids for old game summary statistics
sample size:  84
median:  13.0
mean:  13.2
mode:  ModeResult(mode=array([14]), count=array([9]))
standard deviation:  5.46
range:  27


Number of bids for new game summary statistics
sample size:  59
median:  15.0
mean:  14.1
mode:  ModeResult(mode=array([16.]), count=array([6]))
standard deviation:  6.34
range:  28.0
```

Distribution of both old and new

In [106]:
```
#this plots both of the samples together to see the difference more clearly

histogram = df.plot.hist(bins = 10, alpha = 0.6)
plt.title('Number of bids for old and new games')
plt.xlabel('Number of bids')
plt.ylabel('Relative frequency')
plt.axvline(df['n_bids_old'].mean(), color='blue', linestyle='dashed', linewidth=1, label = 'old mean')
plt.axvline(df['n_bids_new'].mean(), color='red', linestyle='dashed', linewidth=1, label = 'new mean')
plt.legend()
```

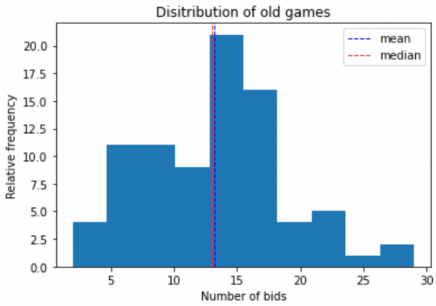
Out[106]: <matplotlib.legend.Legend at 0x7f9dd9e4a9d0>



Distribution of old games

In [107]:
```
old_hist = df['n_bids_old'].plot.hist(bins = 10, label ='')
plt.title('Disitribution of old games')
plt.xlabel('Number of bids')
plt.ylabel('Relative frequency')

#these lines denote the mean and median of the sample
plt.axvline(df['n_bids_old'].mean(), color='blue', linestyle='dashed', linewidth=1, label = 'mean')
plt.axvline(df['n_bids_old'].median(), color='red', linestyle='dashed', linewidth=1, label = 'median')
plt.legend()
```

Out[107]: <matplotlib.legend.Legend at 0x7f9dd9186fd0>

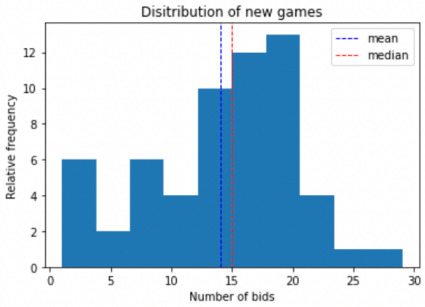


MINERVA UNIVERSITY

Distribution of new games

In [108]:
```
new_hist = df['n_bids_new'].plot.hist(bins = 10, label ='')

plt.title('Disitribution of new games')
plt.xlabel('Number of bids')
plt.ylabel('Relative frequency')

#these lines denote the mean and median of the sample
plt.axvline(df['n_bids_new'].mean(), color='blue', linestyle='dashed', linewidth=1, label = 'mean')
plt.axvline(df['n_bids_new'].median(), color='red', linestyle='dashed', linewidth=1, label = 'median')
plt.legend()
```

Out[108]: <matplotlib.legend.Legend at 0x7f9dd97cc1f0>



Difference of Means test

In [166]:
```
#getting difference of means hypothesis test

def dif_means(data1,data2,tails):
    n1 = len(data1)
    n2 = len(data2)

    #I use numpy to get the means and standard deviation
    x1 = np.mean(data1)
    x2 = np.mean(data2)

    s1 = np.std(data1)
    s2 = np.std(data2)

    SE = np.sqrt(((s1**2)/n1) + ((s2**2)/n2))

    tscore = np.abs((x2 - x1))/SE

    #this if loop checks for which sample size is smaller
    if n1 < n2:
        degfree = n1 - 1
    else:
        degfree = n2 - 1

    p = tails*stats.t.cdf(-t, degfree)

    #to make t test more accurate, we use pooled sd

    a = (s1**2) * (n1 - 1)
    b = (s2**2) * (n2 - 2)
    c = n1 + n2 - 2
    sd_pooled = np.sqrt((a + b) / c)

    #This can then be used to find our Cohen's d

    cohens_d = (x2 - x1)/sd_pooled

    print('t score = ', t)
    print('p value = ', p)
    print('cohens d  = ', cohens_d)


print(dif_means(old, new, 2))
```

```
t score =  0.9457876847373693
p value =  0.3481805590479937
cohens d  =  0.15494470046149397
None
```

Calculating confidence interval

In [238]:
```
#we use the smaller sample size for degree of freedom
if len(old) < len(new):
    degfree = len(old) - 1
else:
    degfree = len(new) - 1

#the critical t value is calclated from scipy
crit_t = (scipy.stats.t.ppf(1-(1-0.95/2),degfree))
#We need to do 1-(1-a/2) because we are trying to get the left side
#of the normal distribution curve. The divide by 2 is because it is
#two tailed.

posconf = (mean2 - mean1) + (crit_t*SE)
negconf = (mean2 - mean1) - (crit_t*SE)

print(round(posconf,3))
print(round(negconf,3))
```

```
0.867
0.935
```