Statistics in Everyday Life

Valerie Kim, Sam Aslanoiwcz, Yadi Su, Tejaswi Lavanya

Abstract:

Statistics can be applied to various fields of life. Throughout this paper, there was an exploration of three unique research questions inquiring from three different datasets to show how statistics is applicable almost everywhere. Our first research question explores statistics in the world of baseball, and compares the league-wide OPS' among infielders vs outfielders, using data from Baseball Savant. To answer this question, we needed a two-sided hypothesis test, as well as a t-test, and a permutation test. After finding a p-value of 0.8321 from the t-test and then comparing it with the p-value of 0.8385 from the permutation test, we did not have enough evidence to reject the null hypothesis. Therefore it is still a possibility that there is no difference in mean OPS between positions. As for the second research question, it explored data from the World Happiness Report from 2016. A two-sample hypothesis test was conducted, more specifically a permutation test was performed in order to compare the mean happiness scores across Eastern Europe and the Asian regions. It was found that, after obtaining a p-value of 0.673, there was not enough evidence to reject the null hypothesis in support of the alternative. As for the third research question, it explores the business operating statistics of an Airline Company by comparing the mean fuel efficiency of small narrow body and wide body types of aircrafts. In order to answer that question a one-way hypothesis testing was done and using a formula to calculate the fuel efficiency we realized that the null hypothesis was rejected as the fuel efficiency of Wide body types was significantly greater. Hence, making us accept the alternate hypothesis being the mean values are not equal to each other.

Question 1

Did MLB infielders with more than 502 plate appearances in 2022 have an equal mean OPS as MLB outfielders with more than 502 plate appearances in 2022?

Introduction

To answer this question, I needed to find a dataset that contained individual player statistics from the 2022 MLB season. I found this information on Baseball Savant, a website that has tracked all kinds of baseball sabermetrics for years. I have always wondered whether infielders or outfielders provide more value for their team in terms of hitting. The OPS stat is the best judge of this, as it considers all aspects of hitting, from a player's ability to get on base, to their ability to hit for power. OPS is tracked on a decimal point scale, as it is the sum of the probability that a player gets on base, and the average amount of bases each player hits for per at-bat. The highest OPS of all time was 1.422, and it was set by Barry Bonds in 2004. This is an extreme case however, as any OPS above .850 is considered to be very good.

Data

Using the filter() function, I separated the data into two new data sets, one containing only infielders, and one containing only outfielders. I also used the filter function to ensure that only players with more than 502 plate appearances were included in each data set. I then arranged the players in each data set in descending order of their OPS, using the arrange(desc()) function, and then used the head(10) to see the 10 infielders and outfielders with the highest OPS. Lastly, I used the select() function so each tibble only displayed the player's name, team position, and OPS. There were no missing values in the data set, so I didn't need to do any cleaning. The ten players with the highest OPS in each data set, infielders and outfielders, are displayed below.

Player <chr>	Team <chr>	Pos <chr>	OPS <dbl>
Paul Goldschmidt	STL	IF	0.982
Jose Altuve	HOU	IF	0.920
Freddie Freeman	LAD	IF	0.918
Manny Machado	SD	IF	0.897
Nolan Arenado	STL	IF	0.891
Rafael Devers	BOS	IF	0.879
Austin Riley	ATL	IF	0.877
Pete Alonso	NYM	IF	0.870
Jose Ramirez	CLE	IF	0.869
Nathaniel Lowe	TEX	IF	0.850

Player <chr>	Team <chr>	Pos <chr>	OPS <dbl>
Aaron Judge	NYY	OF	1.111
Yordan Alvarez	HOU	OF	1.019
Mookie Betts	LAD	OF	0.873
Julio Rodriguez	SEA	OF	0.854
Taylor Ward	LAA	OF	0.833
Kyle Schwarber	PHI	OF	0.827
Starling Marte	NYM	OF	0.815
George Springer	TOR	OF	0.814
Kyle Tucker	HOU	OF	0.808
Teoscar Hernandez	TOR	OF	0.807

1-10 of 10 rows

Methods

To answer this research question, I needed to conduct a two-sided hypothesis test, as well as a t-test to compute the p-value. After visualizing the data in tibbles, my next step was to find the mean of each data set. I used the summarize() function as well as the group_by() function to create a summary table that displayed the mean OPS of each position, as well as the median, min, max, and standard deviation, for my own interest. I then plotted each data set using the ggplot() and geom_boxplot() functions, to compare the spread of OPS in each data set. Next, I had to run a t-test. Using the t.test() function, I was able to find a p-value, and well as a t-value and a 95% confidence interval. My last step was to complete a permutation test, and find observed test statistic for each of 500 trials. I then graphed all of these test statistics together to form a sampling distribution. Below is the summary table that displays the mean OPS of the two groups.

A tibble: 2 × 6

Pos <chr>	mean_OPS <dbl>	median_OPS <dbl>	min_OPS <dbl>	max_OPS <dbl>	sd_OPS <dbl>
IF	0.7614789	0.757	0.561	0.982	0.08197523
OF	0.7653333	0.756	0.564	1.111	0.09538822

2 rows

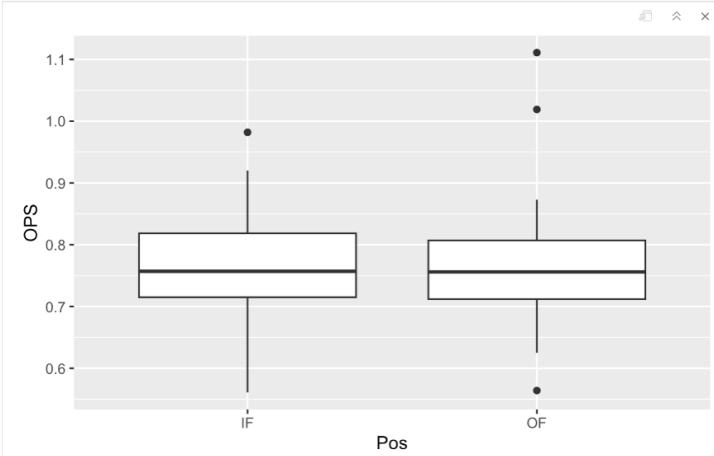
Hypothesis and Observed Test Statistic

Null hypothesis: **The difference in OPS between MLB infielders and outfielders with more than 502 Plate appearances is 0**

Alternative hypothesis: **The difference in OPS between MLB infielders and outfielders with more than 502 Plate appearances is not 0**

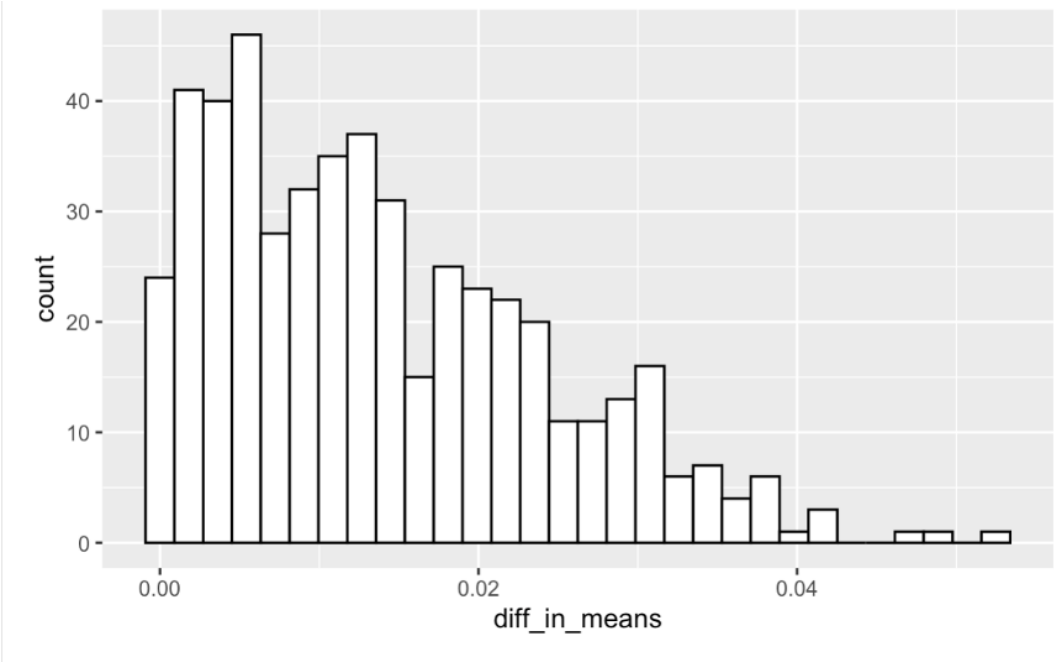
Test statistic: **0.0038544** (The difference between the mean OPS of outfielders and infielders)

Results and Data Visualizations




```
        Welch Two Sample t-test

data:  infielders$OPS and outfielders$OPS
t = -0.21284, df = 68.927, p-value = 0.8321
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.03998252  0.03227360
sample estimates:
mean of x mean of y
0.7614789 0.7653333
```



The p-value that was computed in the t-test in 0.8321, and the p-value that was computed in my permutation test was 0.8385. The boxplot displays that the median OPS of infielders is very similar to that of the outfielders, however, the spread among infielders is much larger, and there is only one outlier, whereas the outfielders box plot contains a very narrow spread and has 3 outliers. The spread of test statistics is right skewed, and the tail extends out to about 0.05. It is a unimodal distribution, and the peak is at roughly 0.005.

Discussion

Both of these p values are much higher than the standard alpha-level of 0.05. This means that there is an 83-84% chance that we would obtain a test statistic as or more extreme than 0.0038544 if we were to conduct this test again assuming the null hypothesis is true. Since both p-values are greater than the alpha-level, we do not have enough evidence to reject the null hypothesis. The histogram above represents different computed test statistics calculated during the permutation test, and about 83% of the results are greater than our observed test statistic (0.0038544), which further strengthens the argument for our resulting p-value

Conclusion

Since the two p-values were very similar to each other, we can be more certain that the true p-value is within that range. With a p-value this high, there is strong data to support the null hypothesis that there is no difference between the mean OPS of qualified MLB infielders and outfielders in 2022, even though we are not able to accept the null. We can also determine that either of these types of tests would be a valid way to answer the research question. In the future, I would like to explore this same research question among MLB seasons prior to 2022, and see if the results are consistent. I would also like to separate both the infielders and outfielders by position, and compare the OPS of each individual position. This way, I could truly see which exact position provides the most value on average to the team in terms of hitting.

Question 2

Introduction:

Each year, the United Nations produces a World Happiness Report, which takes individual responses describing their personal quality of life from a variety of countries across the world. Then, a ranking is created placing each participating country on a scale of most to least "happy" (World Happiness Report, n.d.). Upon looking at this data set, it was clear where the most and least happy regions appeared in the final ranking, however for the countries in the middle there was a combination of many different regions rather than a single prominent one. From this observation, we wanted to determine whether or not there was a distinct difference between the average happiness score among the countries in Central Eastern Europe and Asia. Both of these regions frequently occurred in the same areas of the final ranking, and so we inquired whether there actually existed a difference in overall happiness. Evidently, lifestyles and cultures in either continent greatly differ, as a result it would be very interesting to investigate how individual happiness may compare between them. Ultimately, we came up with the research question, *Is the average happiness score of countries in Central Eastern Europe greater than the average happiness score of countries in Southern, Southeastern, and Eastern Asia in 2016?*

Data:

The data set used to answer the research question was from the 2016 World Happiness Report. This annual report is released by the Sustainable Development Solutions Network, who seeks to investigate the state of global happiness (World Happiness Report, n.d.). Research is conducted through a survey, also known as the Gallup World Poll, to determine the overall happiness of those living in each country. Citizens from 156 countries across the world were instructed to envision their form of an ideal and nonideal life and place their current quality of living situations on a scale of 0-10, 1 being an unideal life and 10 being the ideal (World Happiness Report, n.d.). This score can often be reflected on the extent to which a country scores on 6 different variables including Economy, Family, Health, Trust, Freedom, and Generosity. As a result, the data set we will be using for our own investigations will include the actual happiness score of each country in addition to each country's score on each variable. After data was collected, countries were ranked by their final happiness score. In the dataset, each country is categorized into their respective region, and will be what we are going to be taking a closer look at.

To prepare the dataset for analysis, it was necessary to filter the data to include observations from countries that were categorized in the regions: Central Eastern Europe, Eastern Asia, Southeastern Asia, and Southern Asia. In addition, the research question only requires us to look at the happiness score, and so the rest of the variables were not included. Since the dataset splits the Asian region into three subregions, it was necessary to find a way to categorize them under a single region for the analysis. To do this, a new variable called 'Sub_Region' was mutated into the data set which acted as a binary variable classifying all the countries as either being in the CEE (Central Eastern Europe sub-region or Asia sub-region. For all the countries not considered as Europe, this would make up the entire Asian region.

Shown below is the cleaned and organized data that was used for the permutation test:

Region	Happiness Score	Sub_Region
<chr>	<dbl>	<chr>
Central and Eastern Europe	6.596	CEE
Central and Eastern Europe	6.078	CEE
Central and Eastern Europe	5.987	CEE
Central and Eastern Europe	5.919	CEE
Central and Eastern Europe	5.897	CEE
Central and Eastern Europe	5.856	CEE
Central and Eastern Europe	5.835	CEE
Central and Eastern Europe	5.813	CEE
Central and Eastern Europe	5.802	CEE
Central and Eastern Europe	5.768	CEE
Central and Eastern Europe	5.658	CEE
Central and Eastern Europe	5.560	CEE
Central and Eastern Europe	5.528	CEE
Central and Eastern Europe	5.517	CEE
Central and Eastern Europe	5.488	CEE
Central and Eastern Europe	5.401	CEE
Central and Eastern Europe	5.291	CEE
Central and Eastern Europe	5.185	CEE
Central and Eastern Europe	5.177	CEE
Central and Eastern Europe	5.163	CEE
Central and Eastern Europe	5.161	CEE
Central and Eastern Europe	5.145	CEE
Central and Eastern Europe	5.121	CEE
Central and Eastern Europe	4.996	CEE
Central and Eastern Europe	4.655	CEE
Central and Eastern Europe	4.360	CEE
Central and Eastern Europe	4.324	CEE
Central and Eastern Europe	4.252	CEE
Central and Eastern Europe	4.217	CEE
Southeastern Asia	6.739	Asia

Region	Happiness Score	Sub_Region
<chr>	<dbl>	<chr>
Southeastern Asia	6.474	Asia
Eastern Asia	6.379	Asia
Southeastern Asia	6.005	Asia
Eastern Asia	5.921	Asia
Eastern Asia	5.835	Asia
Eastern Asia	5.458	Asia
Southeastern Asia	5.314	Asia
Southeastern Asia	5.279	Asia
Eastern Asia	5.245	Asia
Southern Asia	5.196	Asia
Southern Asia	5.132	Asia
Southeastern Asia	5.061	Asia
Eastern Asia	4.907	Asia
Southeastern Asia	4.876	Asia
Southern Asia	4.793	Asia
Southern Asia	4.643	Asia
Southern Asia	4.415	Asia
Southern Asia	4.404	Asia
Southeastern Asia	4.395	Asia
Southeastern Asia	3.907	Asia
Southern Asia	3.360	Asia

Methods/Analysis:

The extent of this research question must be examined through a two-sample hypothesis test. The chosen method will allow us to form inferences about the difference in average happiness score between Central Eastern Europe and Asia for the entire population, by looking at the simulated differences in averages from many random samples and comparing it to the observed statistic.

For this investigation, the parameter of interest would be the difference between average happiness scores in Central Eastern Europe and Asia and the hypotheses are as follows:

> H0: The mean happiness score in Central Eastern Europe is equal to the mean happiness score in Asia.
>
> H1: The mean happiness score in Central Eastern Europe is greater than the mean happiness score in Asia.

The observed test statistic was calculated by finding the difference between the mean happiness score for all the countries in the Central Eastern Europe region (n = 29) and the mean score from all the Asian regions (n = 22). This would be the value, in which we compare all of the simulated test statistics with for the hypothesis test. Under the assumption that the null hypothesis is true, a permutation test was conducted. There were a total of 1000 trials, where each trial consisted of reshuffling the random sample, calculating the mean happiness score for each sample, and finding the test statistic which was the difference between the means. The collection of sampled test statistics that were calculated formed the sampling distribution. The results from the permutation test are shown in the histogram located in the Data Visualization section.

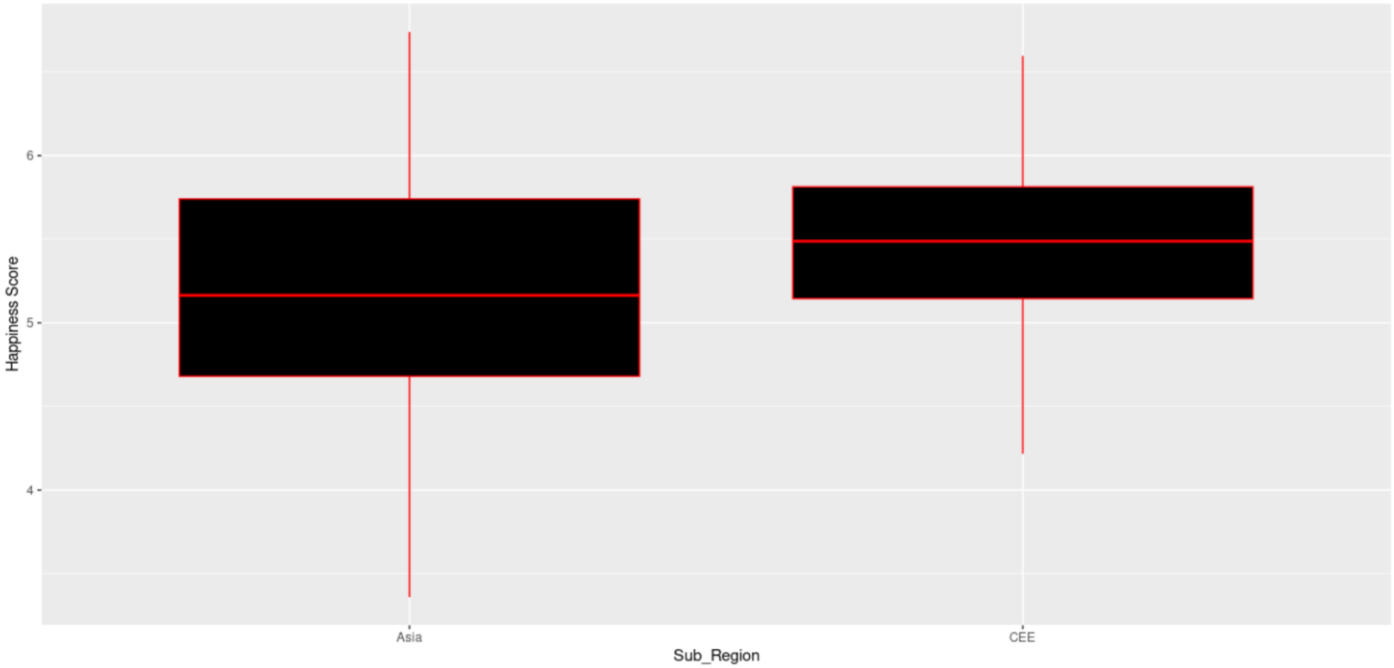
The observed test statistic was then compared in relation to the sampling distribution and the p-value was calculated which tells us the probability of obtaining a value this extreme given that the null hypothesis is true. Using a significance level of 0.05, a conclusion of whether to reject the null hypothesis was made.

Results:

We obtained a p-value of 0.673 when comparing the observed bottom test statistic with the sampling distribution, and with the chosen significance level to be 0.05. Obviously, we obtained a p-value much higher than the chosen significance level, which ultimately means that we fail to reject the null hypothesis because there is not enough evidence to support the alternative hypothesis. Our p-value tells us that the probability of receiving this extreme value of the test statistic given that the null hypothesis is true is 67.3%. This is a relatively high percentage of support for the null hypothesis that there is no difference in the mean happiness scores between the selected regions.

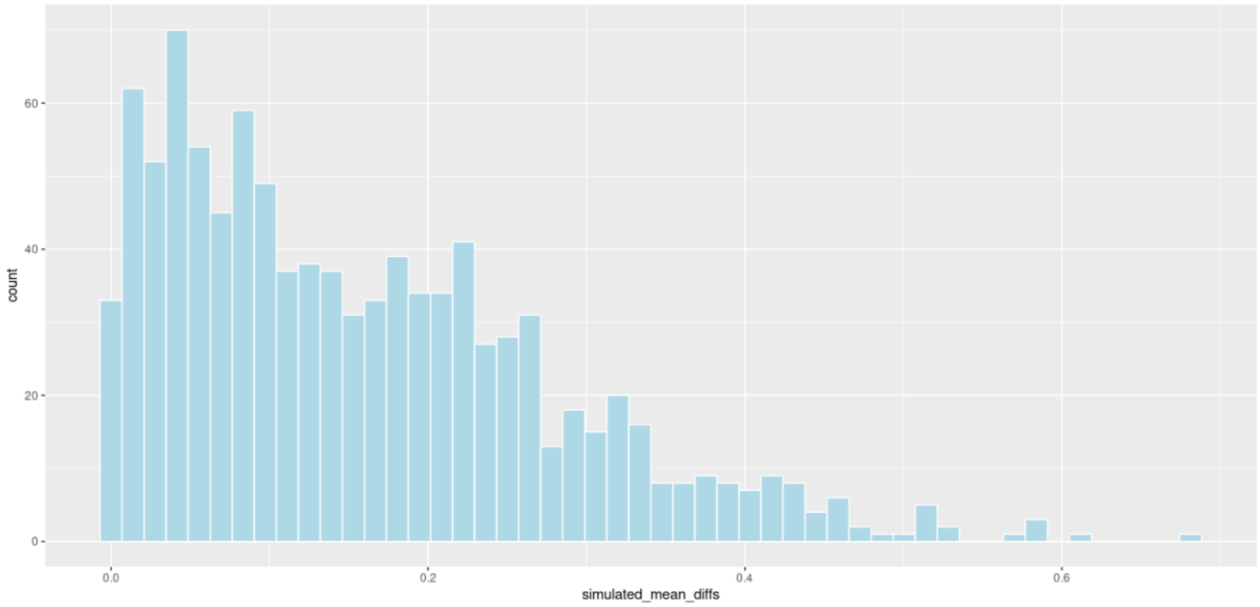
Data visualizations:

For this chart we have used boxplot to help us visualize the distribution of happiness scores across both regions. A box plot would allow us to see whether a similarity existed between the happiness scores for both regions and give us an idea of what to expect for the hypothesis test. The left part represents the Asian region and the right part represents Central Eastern Europe



(CEE). It is clear from the graph that the median happiness index of Central Eastern Europe is slightly higher than the median happiness index of the Asian region. The chart also shows that the IQR of happiness in Asia is higher than the IQR of happiness in Central Eastern Europe, which means that the dispersion of the data for the happiness index in Asia is higher than the dispersion of the data for the happiness index in Central and Eastern Europe. In this way we can see that there is a large difference between the high and low scores of the happiness index in the Asian region, however for the most part the distributions are relatively similar.

The graph above shows the output from the permutation test, a histogram of the sampling distribution. From the graph we can see that it is right--skewed, meaning most of the values cluster on the lower end of the scale. There is only one peak in the graph (unimodal) and this peak is on the left. From this we can also see that we have a small portion of EXTREME value.



If we were to place our observed test statistic (0.200781) on the sampling distribution, it would also appear on the left side of the graph where most of the values appear.

Discussion:

Our goal in studying this question is to understand whether people in the world are happy living in the countries where they live. In the 2016 World Happiness Report, there is a great deal of attention to the areas rated by Drunken Fish Gum Cake or the lower rated areas, but it is difficult to pay attention to the middle part of the world. The two regions we studied are the ones that belong to the more intermediate happiness index obtained to. Comparing the average happiness indices of the two areas will reveal which area needs more attention or improvement in people's lives. It is not only in the lower happiness areas that people are unhappy, but also in the middle areas where there may be a small percentage of unhappy people who need our attention.

Conclusion:

We find that the p-value is significantly larger than the significance level found by the two-sample hypothesis test. There is not enough data to support the conclusion that the average happiness score in Central and Eastern Europe is greater than the average happiness score in Asia. Evidently, given how the data was collected, there was a lot of subjectivity in each respondents' definition of an ideal and non ideal life. Therefore, people in these two regions may lead completely different lifestyles, but it appears that people rate each other's lives similarly based on the results of the happiness report. We have no way of knowing what respondents' ratings of happiness look like, and we cannot turn all respondents' ratings of happiness into consistency. As a result, we find that there is no operational definition of happiness. Therefore based on our analysis alone, we find that the average happiness scores in Central and Eastern Europe and across Asia are likely to be very similar, however there are still several factors that are still up for question.

Question 3: Judging from the span of 20 years of 1995-2015, is the mean fuel efficiency of widebodies greater than of small narrowbodies?

Introduction:

The data used to answer this research question is taken from the data sets of the analytics of the operations of the United Airlines. This data set was used in the ADP project which was established by the MIT Global Air Industry Program. It contains data from the years of 1995 till 2015 and the goal of the project was to grasp a better understanding of the growth, opportunities and challenges faced by this industry. In this set there are 3 types of airplane bodies United Airlines own of which we are going to look at 2 specific ones. Related to the operating costs the data set mentions everything about how exactly is the financial budget distributed and where primarily are the costs going towards from the pilots to the fuel consumption rates. In our research question, since we are calculating the fuel efficiency we are going to be looking at the specific fields of total Miles, Gallons, and the cost of Fuel/Oil of the Small Narrow Body Types and the Wide Body Types.

		1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
1	United Airlines Aircraft Operating Statistics- Actuals										
2	United Airlines										
4	**Small Narrowbodies**	**$2,326,550**	**$2,666,989**	**$2,710,156**	**$2,679,390**	**$2,854,374**	**$3,312,501**	**$3,452,306**	**$2,762,638**	**$2,401,097**	**$2,557,097**
6	**Pilots (000)**	**$583,404**	**$627,762**	**$669,258**	**$689,224**	**$728,240**	**$838,740**	**$999,164**	**$897,116**	**$634,961**	**$556,401**
7	Salaries and Wages (000)	$432,613	$469,059	$515,538	$508,323	$533,546	$605,325	$649,540	$563,728	$370,899	$344,352
8	Pilot Training (000)	$28,235	$22,388	$23,838	$22,037	$25,315	$25,163	$29,921	$22,831	$11,309	$9,293
9	Benefits and Payroll Taxes (000)	$77,752	$87,128	$77,679	$105,045	$108,242	$143,949	$247,814	$257,181	$213,327	$166,083
10	Per Diem/ Personnel (000)	$44,804	$49,187	$52,203	$53,819	$61,137	$64,303	$71,889	$53,376	$39,426	$36,673
12	**Purchased Goods (000)**	**$627,471**	**$792,582**	**$772,448**	**$662,620**	**$663,255**	**$962,185**	**$935,681**	**$646,796**	**$770,532**	**$1,032,877**
13	Fuel/Oil (000)	$559,698	$684,007	$670,673	$567,226	$559,818	$837,938	$855,035	$574,475	$698,838	$956,759
14	Insurance (000)	$7,483	$5,449	$4,200	$2,560	$2,445	$2,851	$7,264	$16,507	$9,446	$6,164
15	Other (inc. Tax) (000)	$60,290	$103,126	$97,575	$92,734	$100,992	$121,396	$73,382	$55,814	$62,248	$69,954
17	**Maintenance (000)**	**$615,007**	**$672,665**	**$807,724**	**$854,160**	**$960,789**	**$947,817**	**$949,448**	**$704,516**	**$594,208**	**$582,333**
18	Labor (000)	$139,373	$140,549	$145,996	$149,664	$164,053	$166,942	$148,830	$156,926	$96,316	$78,252
19	Materials (000)	$95,149	$114,794	$143,919	$171,557	$197,205	$175,974	$154,068	$81,388	$66,075	$73,795
20	Third Party (000)	$34,261	$33,546	$75,406	$110,889	$153,840	$156,635	$196,397	$122,092	$157,145	$211,043
21	Total Direct (000)	$268,767	$288,889	$365,321	$428,600	$512,748	$491,585	$489,601	$359,038	$317,272	$356,938
22	Burden (000)	$346,240	$383,776	$442,403	$425,560	$448,041	$456,232	$459,847	$345,478	$276,936	$225,395
24	**Aircraft Ownership (000)**	**$496,317**	**$469,981**	**$449,816**	**$459,125**	**$497,285**	**$555,950**	**$570,138**	**$507,536**	**$399,283**	**$377,093**
25	Rentals (000)	$367,411	$329,030	$327,893	$316,320	$313,635	$315,140	$315,176	$330,539	$238,917	$209,656
26	Depreciation and Amortization (000)	$128,906	$140,951	$121,923	$142,805	$183,650	$240,810	$254,962	$176,997	$160,366	$167,437
28	**Other (000)**	**$4,351**	**$3,999**	**$10,910**	**$14,361**	**$4,805**	**$7,809**	**-$2,125**	**$6,674**	**$2,113**	**$8,393**

Data:

In the beginning we used basic visualization functions such as glimpse in order to see the raw data. Then I used functions such as filter() and slice() to show me the specific rows and data types for the body types I need. Using these functions it will show all the observations in the raw format and the data set in a more clean and relevant format allowing us to take a proper look at the data. I have also used colnames() to rename the column names in order to prepare the data

more properly. Then I have used the functions such as glimpse() and head() again to visualize the cleaned data.

Filter

	year	Small Narrowbodies	fuelsnb	milessnb	gallonssnb	Widebodies	fuelwb	mileswb	gallonswb
...3	1995	2326550	559698	401510305	1026301421	2873557	829949	271851019	1439660164
...4	1996	2566989	684007	414535134	1066688619	3135415	988897	274457861	1440332508
...5	1997	2710156	670673	419974756	1071180655	3275725	980063	290602638	1499205972
...6	1998	2679390	567226	419764630	1058185737	3222970	836976	309791201	1553290817
...7	1999	2854374	559818	439361183	1097702604	3204827	821684	315416649	1545940406
...8	2000	3312501	837938	444981631	1138703961	3817373	1183038	320851359	1545677910
...9	2001	3452306	855035	425713355	1041841612	3725005	1159531	302604723	1380294297
...10	2002	2762638	574475	363368085	804542288	3535081	944562	285891356	1264238761
...11	2003	2401097	698838	371588253	719084191	2898261	943977	250020471	910530528
...12	2004	2557097	956759	386583608	795864053	3210999	1374360	259989162	1177485953
...13	2005	2623971	1228693	349186347	719039213	3707720	1950803	248521068	1155416091
...14	2006	3023429	1472334	352220970	717106308	4014907	2315494	249894476	1158383646
...15	2007	3139163	1492542	340752609	695146747	4232577	2504636	256777246	1200855919
...16	2008	3721438	2224542	308465451	625325920	5774390	3993223	256645804	1164137938
...17	2009	1989693	827557	239878261	489635018	3509225	1806139	241632442	1081934087
...18	2010	2188492	1034620	223507104	452003458	4231365	2519583	243098607	1109622003
...19	2011	2536421	1297652	223616760	449302985	4895834	3097350	239178172	1084370237
...20	2012	3544397	1893299	295036339	580217102	7355226	4749674	338468631	1462933619
...21	2013	3296735	1621367	272689062	535828758	7045375	4411252	326834763	1442635837
...22	2014	3200531	1462161	255325000	506014602	7158573	4275457	334415816	1462819941
...23	2015.0	2409721.7	913457.1	241168236.0	483020605.0	5930869.4	2899576.1	351316299.0	1508363069.0

	tabledata.year	tabledata.fuelconsumptionsnb	tabledata.fuelconsumptionwb
1	1995	736237393322	471562810001
2	1996	646455240390	399748992336
3	1997	670772245514	444535923067
4	1998	783089887211	574921894655
5	1999	861508409297	593434145504
6	2000	604701476472	419202813434
7	2001	518722494428	360217685773
8	2002	508890709765	382648183698
9	2003	382353618855	241161883680
10	2004	321573141289	222746286408
11	2005	204346143534	147193356233
12	2006	171550666762	125015946586
13	2007	158704457012	123112690117
14	2008	86710631642	74819542276
15	2009	141927138122	144745435166
16	2010	97645496796	107060399767
17	2011	77425748786	83735351529
18	2012	90416320718	104250763161
19	2013	90118178927	106886557803
20	2014	88361116358	114418207043
21	2015.0	127525668430	182755172713

Methods and Analysis:

In order to answer that question I have conducted a one-sided hypothesis test to compare the mean values of the fuel consumption and efficiency of the small narrow body types and wide

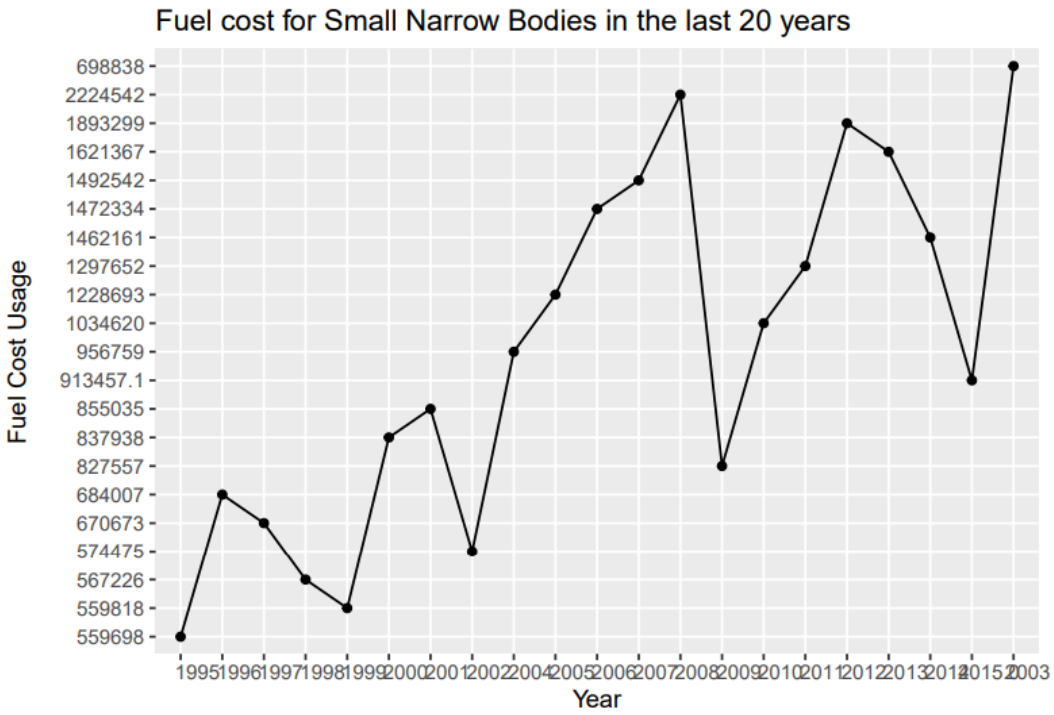
body types. After visualizing all the tibbles, I needed to calculate the mean for the body types. In order to calculate the variables I created a new tibble with the necessary information for the formula using data.frame() .We use other functions such as mutate() to create the new variables for the fuel consumption and efficiency. Then as a test to see the general comparison between the two bodies I created a graph for the fuel costs for both Small Narrow Body and Wide Body types by using the fuel costs on the y-axis and the type of craft on the x-axis, these are represented as line graphs. Then once the fuel consumption value is calculated for both categories it will be represented as a box-plot having the fuel efficiency as the y-axis and the category as the x-axis. Once the box-plot is constructed we can conclude on a few statements helping us come closer to answering the research question.

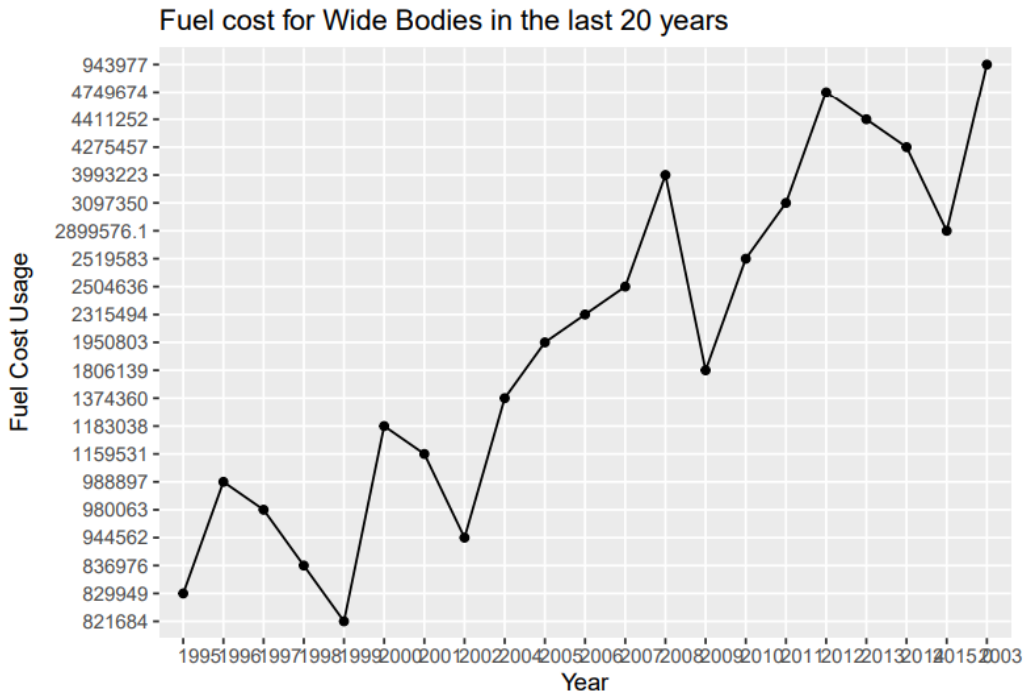
Hypotheses and Observed Test Statistic:

HO: The mean fuel efficiency of Small Narrow Bodies is equal to the mean fuel efficiency of Wide Bodies

H1: The mean fuel efficiency of Small Narrow Bodies is lesser to the mean fuel efficiency of Wide Bodies

Test Statistic: 92612482985 (mean fuel consumption of Small Narrow Bodies - mean fuel consumption of Wide Bodies





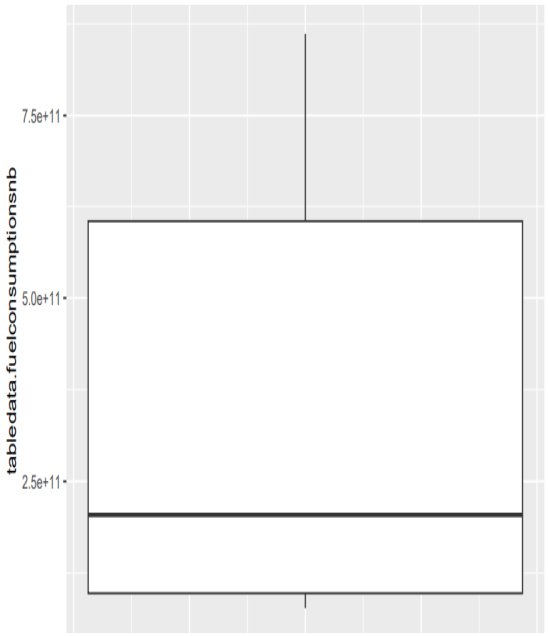
Results:

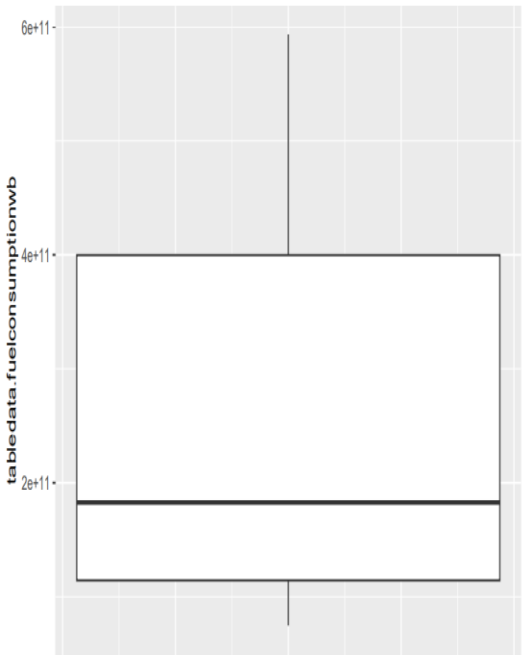
Based on the results from the box plot graphs above, we need to notice one thing is that the formula is for the calculation of the fuel consumption of the aircraft as there is no way to compare 2 different units. However since we are able to create a relationship between fuel efficiency for fuel consumption where the higher the fuel consumption the lower the fuel efficiency then we can use the means of the graphs in order to help us answer the research question. Therefore we have arrived at the fuel consumption values of:

Small Narrow Body Mean: 350906484935

Wide Body Mean: 258294001950

Since we can see that the mean fuel consumption of Small Narrow body types is significantly larger than compared to Wide body types by an amount of 92612482985 (gallons*miles*$^{-1}$). Therefore we reject the first null hypothesis where they are equal and accept the alternate hypothesis where the mean values aren't equal.





Fuel Consumption for Small Narrow Bodies ***Fuel Consumption for Wide Bodies***

Conclusion:

Given that the fuel costs for the wide body types was immensely higher the reason as to why the fuel consumption for small narrow types was higher was because the miles of those types of aircrafts were way higher. This is probably due to their small size as they are easier to fly and can fly short distances very easily allowing them to make more flights per day earning the Airlines more money. However if we are to see this from a cost perspective, then we can say that wide body types are more efficient than small narrow body types.

References

How to perform T-tests in R. (n.d.). Retrieved April 11, 2023, from https://datascienceplus.com/t-tests/

Statcast custom leaderboards. (n.d.). Retrieved April 11, 2023, from https://baseballsavant.mlb.com/leaderboard/custom?year=2022&type=batter&filter=&sort=4&sortDir=desc&min=q&selections=xba%2Cxslg%2Cxwoba%2Cxobp%2Cxiso%2Cexit_velocity_avg%2Claunch_angle_avg%2Cbarrel_batted_rate%2C&chart=false&x=xba&y=xba&r=no&chartType=beeswarm

World Happiness Report. (n.d.). *About*. Retrieved from World Happiness Report: https://worldhappiness.report/about/

United Airlines Aircraft Operating Statistics. (n.d.). Retrieved from ADP Project: https://data.world/adamhelsinger/united-airlines-data/workspace/file?filename=United+Airlines+Aircraft+Operating+Statistics-+Actuals.xls

Individual Contribution

Tejaswi Lavanya: The overall management and completion of the 3rd research question was done by me. Including the presentation of the question and answering the question related with the 3rd question during the presentation were all done by me.