CS50 Course - Statistical Inference

Minerva University

CS50: Formal Analyses

Prof. Robbins

December 10, 2022

Introduction

Dataset Summary & Purpose

The report analyzes the dataset "Sample of Youth Risk Behavior Surveillance System (YRBSS)" which consists of 100 observations of 13 variables related to physical characteristics, school grades, and behavior of youth and adults. This report aims to evaluate the relationship between the number of physically active days and students' age to make valid generalizations about the population of over 13,583 high school students who participated in the YRBSS 2013 (OpenIntro Statistics, p168). The conclusions should inform schools about suitable interventions targeting inactive student groups to promote a healthier student body.

Research Question & Hypotheses

Research question	Are older respondents ($\geq$ 16) more likely to spend more days physically active than younger respondents ($\leq$ 15)?
Null hypothesis (H_0)	The difference in the number of days physically active between students aged 16 or older and students aged 15 or younger is 0.[1] (μ days physically active | age $\geq$ 16) - (μ days physically active | age $\leq$ 15) = 0
Alternative hypothesis (H_A)	The difference in the number of days physically active between students aged 16 or older and students aged 15 or

[1] The number of days physically active for 60+ minutes in the last 7 days. In this report, this variable will be rephrased as "number of days physically active" for short.

	younger is not equal to 0. (μ days physically active | age $\geq$ 16) - (μ days physically active | age $\leq$ 15) $\neq$ 0

Variables Explanation

Independent variable **(x - x1, x2)**	Age groups: students 16 or older (**x1**) and students 15 or younger (**x2**). This is the nominal qualitative variable because labels ($x \geq 16$ and $x \leq 15$) are used as values, and it is categorical and descriptive (even though it is labeled using mathematical symbols, there are no mathematical calculations involved for this variable).
Dependent variable **(y - y1, y2)**	The number of days being physically active of students 16 or older **(y1)** and that of those 15 or younger **(y2).** The quantitative discrete variable is measured numerically by the number of days, and the values can only be taken from a strictly specified set of positive integers [0,7].[2]

[2] **#variables:** I have identified relevant variables in the report, namely the independent variable (age) and the dependent variable (number of days physically active). For the independent variable, two subgroups are divided for statistical analysis. The types of variables are also clearly defined as nominal qualitative and discrete quantitative. All of which is the first step to informing my analysis.

Descriptive Statistics & Data Visualizations - Appendix A

Table 1: Descriptive statistics for the number of days physically active for two sample groups: Group 1 ($x \geq 16$) and Group 2 ($x \leq 15$).

Statistics	**Group 1** (y | x1)	**Group 2** (y | x2)
Count	$n_1 = 67$	$n_1 = 32$
Mean	$\bar{x}_1 = 3.36$	$\bar{x}_2 = 4.47$
Median	3	4.5
Standard Deviation (STD)	$s_1 = 2.6$	$s_2 = 2.3$
Mode	0	7[3]

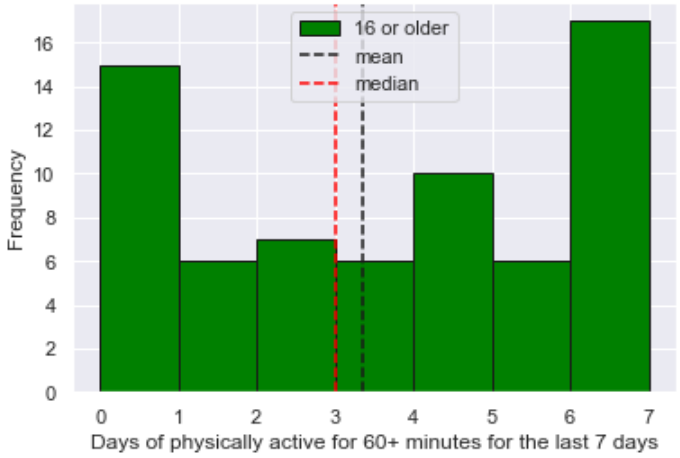


Figure 1. Histogram for group 1 ($x \geq 16$)

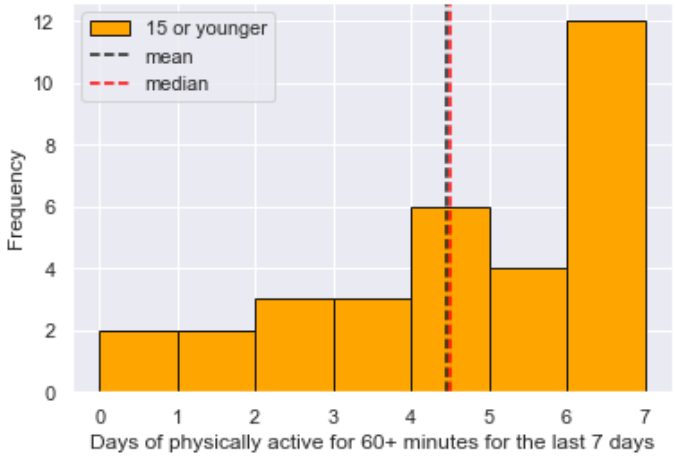


Figure 2. Histogram for group 2 ($x \leq 15$)

In figure 1, as the mean is larger than the median and the median is larger than the mode, the sample distribution is slightly right-skewed. In figure 2, even though the difference between

[3] **#descriptivestats:** I have calculated important statistics including the measures of locations (mean, mode, median) and measure of spread (standard deviation), all of which provide information about the general properties of the dataset to draw suitable histograms. Especially, the mean and median help explain the skewness of the distribution.

the mean and median is negligible, the mode of this group is 7, and the frequency of bigger values is greater than that of smaller values. This explains why the sample distribution is left-skewed. Applying the formula to quantify the degree of skewness: (Mean - Mode)/STD, the results for group 1 and group 2 are approximately 0.1 (positive) and -0.4 (negative), respectively, which is consistent with the explanation for the shapes of the two histograms. We can see that the older age group seems to have fewer average days physically active compared to the younger group in this sample. Still, the distribution of group 1 is more symmetrical than group 2. [4]

Conditions for Inference

Although the two samples' sizes are larger than 30, the standard deviation and variance of the population are unknown, so it is appropriate to use t-distribution for the sampling distribution[5]. Certain conditions of the Central Limit Theorem must be met:

1) **Randomness:** Since the respondents' information, including their ages and number of physically active, is randomly chosen from the YRBSS survey, this condition is met.
2) **Normal sampling distribution:** Because both sub-group sizes are larger than 30, the sample mean is approximately normally distributed.
3) **Independence:** Given that the sample sizes are less than 5% of the population (679) and it is sampled without replacement, the condition of independence is met, and the Finite Population Correction Factor (FPC) is not used. Besides, the number of days physically

[4] **#dataviz**: I have drawn two histograms for this dataset because it is the most useful type of visualization for this report. The histograms are appropriately labeled and lines of mean and median are included to show the skewness of the graphs. Explanations are followed to provide information about the sample as well as possible inductions.

[5] **#distribution:** the sampling distribution was explained along with the conditions for the Central Limit Theorem to justify the normality of the distribution. Relevant probabilities relating to standard error and standard deviation are also calculated throughout the study.

active in the first age group does not affect the number of days physically active in the other group when chosen randomly.

Hypothesis testing

As I have no certain predictions in the direction of difference between the two age groups, a two-tailed test will be used to increase the accuracy of the conclusion.

1) **Type I error:** There is no difference in the number of days physically active between the two age groups, but I conclude that there is a difference.
2) **Type II error:** There is a difference in the number of days physically active between the two age groups, but I conclude that there is no difference.

Since the purpose of this test is to suggest policies to promote a healthier body of students and the scale of the given population is relatively large (> 13,000), the consequence of committing the type II error is more dangerous. Therefore, I will use significance level $\alpha = 0.1$ with the equivalent confidence interval of 90%. This means I only reject the null hypothesis if it has a p-value less than or equal to 0.1. Bonferroni correction will not be used because only one test is conducted on the dataset with one dependent variable (physically active days) and one independent variable (age).

Difference of Means Test - Appendix B

Statistical Significance	The t-score = 2.17 (standard error (SE) = 0.51, degrees of freedom (df) = 31) indicates that the average of days physically active of age group 1 and age group 2 is 2.17 deviate away from each other.

	This results in a two-tailed p-value equal to $0.04 < 0.1$[6], which rejects the null hypothesis.
Practical Significance	Using Cohen's d formula in the study of comparison between different variables, the effect size is calculated (d = 0.443), which means there are about 0.4 standard deviations lie between the two means. This is considered a small effect size, indicating limited practical significance.[7]

Confidence Intervals - Appendix C

A confidence interval for the two means is constructed to examine the plausible range of values the difference could take: (y|x1) - (y|x2) = [0.09, 2.13]. Notably, the null value (0) is not within the confidence interval, which might indicate a statistical significance. This provides more evidence for the prior hypothesis testing with p-value and the direction of the difference towards the null and the size of the difference. Besides, the two 90% confidence intervals for the means of two subgroups are calculated with a standard error of 0.32 (SE_1) and 0.40 (SE_2):

- 16 and older: [3.3, 3.4] days physically active
- 15 and younger: [4.3, 4.6] days physically active

[6] **#probability:** I explained the p-value, significance test, and confidence interval in terms of probability, all of which are interpreted in terms of conditional probability.

[7] **#significance:** a measure of both statistical (sample size) and practical significance was conducted with the results of p-values and effect size, which serves as evidence for the strength of my induction.

Given that the p-value is less than α and the two intervals are not overlapping, meaning that the difference is significant and the results are not merely by chance, and an effect or relationship does exist between the variables being studied in the population.[8]

Conclusions

I can be 90% confident that the number of days physically active in older groups is likely to be 0.09 to 2.13 higher than that of younger age groups in the population participating in the YRBSS 2013 survey. The conclusion is inductive because I made statistical generalizations about the population from one given sample dataset of 100 respondents. In this test, a medium sample size (n>30) decreases the standard error, a smaller p-value would increase the probability of rejecting the null, and a medium effect size (d=0.443) would make the distance in two distributions of difference of means relatively larger. As a result, the probability of having a Type II error decreases. These render the conclusion that the population moderately strong. However, the reliability could not be ensured due to the limited practical significance. Since the study is statistically significant but not practically significant, I suggest that the size of the sample is not a problem but factors/relationships between other variables apart from age that affects the physical activity of the population should be further examined should be examined. Statistical power can also be calculated - the larger the power, the stronger the induction.[9][10]

[8] **#confidenceintervals:** three confidence intervals were calculated: one for the difference of means and two for the sample mean of each subgroup. All of this leads to conclusions about the population and provide implications for the statistical as well as the practical significance of the study.

[9] **#induction:** the generalizations of the population are followed by premises relating to statistical significance (sample size) and practical significance (effect size). The induction is relatively strong thanks to the number of premises but the diversity of evidence is suggested. This helps draw a cautious conclusion about the relationship between two chosen variables.

[10] **#organization:** I organized the report into three main parts including the introduction, descriptive statistics & visualizations, hypothesis testing, and conclusion. Particularly, I utilized tables to present information that makes it easier for the readers to follow.

Word count: 1498 words

References

Diez, D. M., Barr, C. D., & Çetinkaya-Rundel, M. (2019). OpenIntro statistics (Fourth edition). OpenIntro.

Appendix A - General Dataset, Descriptive Statistics, and Data Visualizations

```
#APPENDIX A:GENERAL DATASET (YRBSS)
#import relevant packages & libraries
import csv
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import statistics

#import the chosen dataset using pandas and read into a dataframe
pd.read_csv("https://course-resources.minerva.edu/uploaded_files/mu/00294341-4390/yrbss-samp.csv")
df.head(9)
```

	age	gender	grade	hispanic	race	height	weight	helmet_12m	text_while_driving_30d	physically_active_7d	hours_tv_per_school_day	strength_trainin
0	16.0	female	11.0	not	Black or African American	1.50	52.62	never	1-2	0	4	
1	17.0	male	11.0	not	White	1.78	74.84	rarely	0	7	1	
2	17.0	male	11.0	not	White	1.75	106.60	never	0	7	2	
3	15.0	male	10.0	hispanic	NaN	1.68	66.68	never	did not drive	3	2	
4	18.0	male	12.0	not	Black or African American	1.70	80.29	never	did not drive	0	2	
5	15.0	female	9.0	not	Black or African American	1.57	46.27	NaN	did not drive	4	NaN	
6	16.0	male	10.0	not	White	1.78	81.65	always	0	7	1	
7	16.0	male	10.0	not	Black or African American	1.63	56.70	never	0	5	3	
8	14.0	male	9.0	hispanic	White	1.63	54.43	never	0	7	5+	

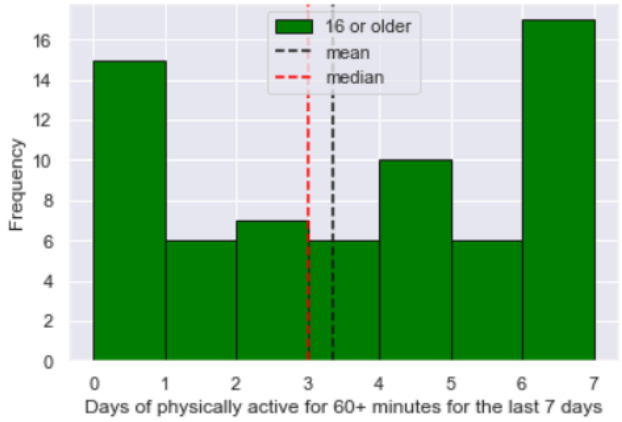
```
#APPENDIX A: DESCRIPTIVE STATISTICS

print ("The summary statistics for the number of days physically active of group 1 are:\n",group1.describe(), '\n')
print ("The summary statistics for the number of days physically active of group 2 are:\n",group2.describe())
```

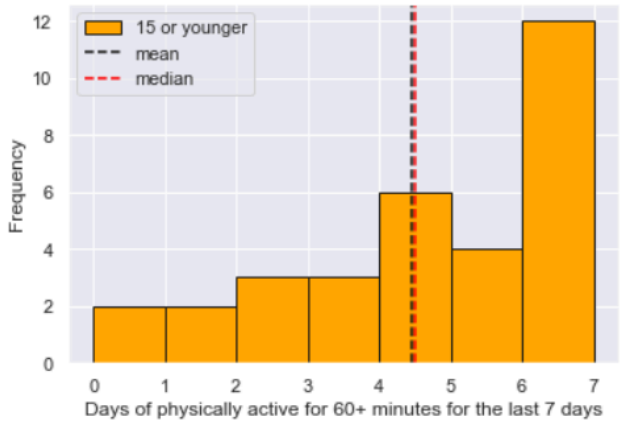
```
The summary statistics for the number of days physically active of group 1 are:
 count    67.000000
mean      3.358209
std       2.609432
min       0.000000
25%       1.000000
50%       3.000000
75%       5.500000
max       7.000000
Name: physically_active_7d, dtype: float64

The summary statistics for the number of days physically active of group 2 are:
 count    32.000000
mean      4.468750
std       2.271625
min       0.000000
25%       3.000000
50%       4.500000
75%       7.000000
max       7.000000
Name: physically_active_7d, dtype: float64
```

```
#APPENDIX A: Create two histograms for each group

import seaborn as sns
sns.set()

df[['age', 'physically_active_7d']].dropna()

#formatting the histogram using color, label, and bins
plt.hist(group1, color = 'green', bins=7, label = "16 or older", edgecolor='k')

#add mean and median lines to the histogram
plt.axvline(group1.mean(), color='k', linestyle='dashed', linewidth=1.5, label = 'mean')
plt.axvline(group1.median(), color='red', linestyle='dashed', linewidth=1.5, label = 'median')

plt.legend(loc = 'upper center')
plt.xlabel('Days of physically active for 60+ minutes for the last 7 days')
plt.ylabel('Frequency')
plt.show()

#formatting the histogram using color, label, and bins
plt.hist(group2, color = 'orange', bins=7, label = "15 or younger", edgecolor='k')

#add mean and median lines to the histogram
plt.axvline(group2.mean(), color='k', linestyle='dashed', linewidth=1.5, label = 'mean')
plt.axvline(group2.median(), color='red', linestyle='dashed', linewidth=1.5, label = 'median')
plt.legend()
plt.xlabel('Days of physically active for 60+ minutes for the last 7 days')
plt.ylabel('Frequency')
plt.show()
```

```
import csv
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import statistics

#filter relevant columns and rows
df=pd.read_csv("https://course-resources.minerva.edu/uploaded_files/mu/00294341-4390/yrbss-samp.csv")
group1 = df.loc[df['age'] >= 16, 'physically_active_7d']
group2 = df.loc[df['age'] <= 15, 'physically_active_7d']

#calculate the skewness of each histogram
print ("Skewness1:", group1.skew())
print ("Skewness2:", group2.skew())
```

```
Skewness1: 0.11073398256770325
Skewness2: -0.4322329864399809
```


16 or older
mean
median
Frequency
0
2
4
6
8
10
12
14
16
0
1
2
3
4
5
6
7
Days of physically active for 60+ minutes for the last 7 days


15 or younger
mean
median
Frequency
0
2
4
6
8
10
12
0
1
2
3
4
5
6
7
Days of physically active for 60+ minutes for the last 7 days

APPENDIX B: Difference of Means Test

```
point_estimate = 1.1105410447761193
SE =  0.5127257728445671
T-score = 2.165955182270075
p = 0.04
d = 0.4430771243871281
```

```python
from scipy import stats

def difference_of_means_test(data1,data2,tails):
    # the len function returns the sample size of two subgroups
    n1 = len(group1)
    n2 = len(group2)

    #the mean function returns the number of days physically active of two subgroups
    x1 = np.mean(group1)
    x2 = np.mean(group2)
    point_estimate = x2 - x1 #point estimate is the mean of the sampling distribution, which is the same as the population mean

    #numpy calculates the sample standard deviation of two subgroups
    s1 = np.std(data1,ddof=1) # Bessel's correction uses n-1 in denominator to take into account of small sample size
    s2 = np.std(data2,ddof=1)

    # calculate the standard error of the mean difference
    SE = np.sqrt(s1**2/n1 + s2**2/n2)

    # Tscore is a conversion of the difference between 2 means in a standardardize unit of standard error
    Tscore = np.abs(point_estimate)/SE

    # the degree of freedom = chosen sample size (n) - 1
    # choose the smaller sample size between 2 subgroup
    df = min(n1,n2) - 1

    # convert Tscore into the probability as an area under the curve.
    # 2 tailed-test requires doubling the result because the probability is equally distributed between 2 tails
    pvalue = tails * (1 - stats.t.cdf(Tscore,df))

    # Calculate the effect size based on Cohen's d formula
    SDpooled = np.sqrt((s1**2*(n1-1) + s2**2*(n2-1))/ (n1 + n2 -2))
    Cohensd = point_estimate/SDpooled

    print("point_estimate =", point_estimate)
    print ("SE = ", SE)
    print('T-score =',Tscore)
    print('p =', round(pvalue,2))
    print('d =',Cohensd)

difference_of_means_test (group1, group2, 2)
```

APPENDIX C - Confidence Intervals

```
from scipy import stats

#function to find the confidence intervals of each subgroup
def confidence_interval (data_list):
    n = len(data_list) #sample size of each subgroup
    x = np.mean(data_list) #calculate the mean of each subgroup
    s = np.std(data_list, ddof=1)/np.sqrt(n) #calculate the standard error using the formula: s/sqrt(n)

    C = 0.9 #confidence level 90%

    df = n-1 #degrees of freedom = chosen sample size - 1
    alpha = (1-C) / 2
    t = stats.t.ppf(q = 1 - 0.1/2, df = n-1) #tscore
    lower = x - t * s / np.sqrt(n) #lower bound
    upper = x + t * s / np.sqrt(n) #upper bound

    print ('Standard Error =', s)
    print ('Confidence interval =', lower, ';', upper)

confidence_interval (group1)
confidence_interval (group2)
```

```
Standard Error = 0.31879287566465864
Confidence interval = 3.2932352431079477 ; 3.423182667339814
Standard Error = 0.40157044284224414
Confidence interval = 4.348388010637793 ; 4.589111989362207
```

```
#difference_of_means_test(group1, group2, 2)
#confidence interval
#call the function

from scipy import stats
def confidence_interval(data1,data2,tails):
    # the len function returns the sample size of two subgroups
    n1 = len(data1)
    n2 = len(data2)

    #the mean function returns the number of days physically active of two subgroups
    x1 = np.mean(data1)
    x2 = np.mean(data2)
    point_estimate = x2 - x1

    #numpy calculates the sample standard deviation of two subgroups
    s1 = np.std(data1,ddof=1) # Bessel's correction uses n-1 in denominator to take into account of small sample size
    s2 = np.std(data2,ddof=1)

    # calculate the standard error of the mean difference
    SE = np.sqrt(s1**2/n1 + s2**2/n2)

    # The difference of means test requires the degree of freedom formula of both subgroups
    df_pooled = n1 + n2 - 2
    # the t critical value is used in the formular of confidence level of difference of means test
    t_critical = stats.t.ppf(0.05/tails, df_pooled)

    # confidence interval is expressed as equal intervals from t critical to the point estimate in standardardized standard erro
    Confidence_interval = [round(point_estimate + t_critical*SE,2), round(point_estimate- t_critical*SE,2)]
    print("point_estimate =", point_estimate)
    print ("t critical value =", t_critical)
    print ("SE = ", SE)
    print("Confidence interval = ", Confidence_interval)
confidence_interval(group1, group2, 2)
```

```
point_estimate = 1.1105410447761193
t critical value = -1.9847231859278835
SE =  0.5127257728445671
Confidence interval =  [0.09, 2.13]
```