Python Inferential Statistics

Lucas Hasting

University of North Alabama

DA 380 - Programming for Data Analytics

Dr. Michael Floren

2/24/2026

Python Inferential Statistics

Contents

List of Tables

Excel Inferential Statistics Practice Using BRFSS

First, an independent-samples t-test was conducted to compare average general health between people who exercise and people who don't. People who did not exercise ($\bar{x} \approx 3.11$, $s \approx 1.08$), on average, had a higher score than people who did exercise ($\bar{x} \approx 2.46$, $s \approx 0.96$). This difference was statistically significant, $t(587) \approx 10.47$, $p < 0.01$. These results suggest that the general health differs between people who exercise and people who do not.

Next, a one-way analysis of variance (ANOVA) was conducted to examine differences in mean general health across 11 groups defined by annual household income. Mean and standard deviation of general health values for each group are approximately shown in Table 1. The overall effect was statistically significant, $F(10, 1170) \approx 20.00$, $p < 0.01$. These results indicate that average general health differs among annual household income groups.

Then, a chi-squared test of independence was conducted to examine the association between exercise participation and median annual household income. The test indicated that the association was statistically significant, $\chi^2(1, N = 1180) = 79.06$, $p < 0.01$. Examination of the contingency table, shown in Table 2, showed that people who did not exercise tend to be below (or equal to) the median annual household income; however, more people tend to participate in exercise in general. This pattern suggests that people who tend to not participate in exercise are more likely to have a lower annual household income than those who do, and vice versa.

Finally, a multiple linear regression analysis was conducted to examine whether exercise participation and disability status were associated with BMI. The overall regression model was statistically significant, $F(2, 1344) \approx 17.97$, $p < 0.01$, and explained 2.60% of the variance in BMI ($R^2 \approx 0.026$). Exercise participation was a significant predictor of BMI ($\beta = 2.03$, $p < 0.01$), and disability status was significant ($\beta \approx 1.14$,

$p < 0.01$) as well. These results indicate that although there is not a strong linear relationship of exercise participation and disability status with BMI, the variables do have some relationship to BMI, and thus are associated with BMI.

Group	$\bar{x}$	s
1	3.58	1.24
2	3.24	1.10
3	3.12	1.18
4	3.31	1.14
5	3.01	1.03
6	2.69	1.01
7	2.61	1.00
8	2.44	0.91
9	2.18	0.86
10	2.17	0.84
11	2.00	0.75

Table 1

Mean and Standard Deviation of General Health by Annual Household Income

Median Annual Household Income	Participated in Exercise	Did not Participate in Exercise	Total
Less than or equal median household income	420	235	655
More than median household income	457	68	525
Total	877	303	1180

Table 2

Contingency Table for Median Annual Household Income and Exercise Participation

Appendix A

Python File

```python
'''
Name: Lucas Hasting
Course: DA 380
Instructor: Dr. Michael Floren
Date: 2/24/2026
Description: Get inferential stats from brfss1_cleaned.cvs
'''

#import needed libraries
import numpy as np
import pandas as pd
from scipy import stats
import statsmodels.api as sm
import statsmodels.formula.api as smf

#load the data
df = pd.read_csv("brfss1_cleaned.csv")

#is general health different between people who exercise vs. not? - T-TEST
print("GENHLTH vs. EXERANY2:",end="\n\n")
gen_hlth_exer = df.loc[df["EXERANY2"] == 1, "GENHLTH"].dropna()
gen_hlth_not_exer = df.loc[df["EXERANY2"] == 2, "GENHLTH"].dropna()

print(stats.ttest_ind(gen_hlth_not_exer, gen_hlth_exer, equal_var=False), end="\n\n")
print("Exercise:\n",gen_hlth_exer.agg(["mean", "std"]),end="\n\n",sep="")
print("No Exercise:\n",gen_hlth_not_exer.agg(["mean", "std"]),end="\n\n", sep="")

#general health across income groups - are they different? - ANOVA
```

```python
print("INCOME3 vs. GENHLTH:",end="\n\n")
df["INCOME3_no_na"] = df["INCOME3"].dropna()
df["GENHLTH_no_na"] = df["GENHLTH"].dropna()
model_genhlth = smf.ols("GENHLTH_no_na ~ C(INCOME3_no_na)", data=df).fit()
print(sm.stats.anova_lm(model_genhlth))
print(df.groupby("INCOME3_no_na")["GENHLTH_no_na"].agg(["mean", "std"]), end="\n\n")

#association between exercise participation and median annual household income? - Ch^2

print("INCOME3 vs. GENHLTH:",end="\n\n")
df["EXERANY2_no_na"] = df["EXERANY2"].dropna()
df["income_bin_no_na"] = df["income_bin"].dropna()
print(df.groupby("EXERANY2_no_na")["income_bin_no_na"].value_counts())
print(stats.chi2_contingency(pd.crosstab(df["income_bin_no_na"], df["EXERANY2_no_na"])),end="\n\n")

#BMI linearily correlated with disabled and exercise participation? - OLS Regression

print("BMI vs. EXERANY2 and disabled:",end="\n\n")
df["BMI_no_na"] = df["bmi"].dropna()
df["disabled_no_na"] = df["disabled"].dropna()
model_bmi = smf.ols("BMI_no_na ~ C(EXERANY2_no_na) + C(disabled_no_na)", data=df).fit()
print(sm.stats.anova_lm(model_bmi))
print(model_bmi.summary())
```

Appendix B

Python File Output

```
GENHLTH vs. EXERANY2:

TtestResult(statistic=np.float64(10.469372106336646), pvalue=np.float64(1.2428555292489605e-23), df=np.float64(586.6549314911825))

Exercise:
mean    2.458296
std     0.955383
Name: GENHLTH, dtype: float64

No Exercise:
mean    3.114058
std     1.081881
Name: GENHLTH, dtype: float64

INCOME3 vs. GENHLTH:

                      df       sum_sq    mean_sq          F        PR(>F)
C(INCOME3_no_na)    10.0   187.317149  18.731715  20.000458  1.892123e-34
Residual          1170.0  1095.780226   0.936564        NaN           NaN
                   mean       std
INCOME3_no_na
1.0            3.576923  1.238485
2.0            3.236842  1.101208
3.0            3.116279  1.179372
4.0            3.313725  1.140003
5.0            3.006579  1.026127
6.0            2.692308  1.010381
7.0            2.613953  1.002236
8.0            2.439024  0.914740
```

```
9.0            2.179348  0.859157
10.0           2.173469  0.837590
11.0           2.000000  0.746299

INCOME3 vs. GENHLTH:

EXERANY2_no_na  income_bin_no_na
1.0             1.0                 457
                0.0                 420
2.0             0.0                 235
                1.0                  68
Name: count, dtype: int64
Chi2ContingencyResult(statistic=np.float64(79.05919206447463), pvalue=np.float64(6.027662302195648e-19), dof=1, expected_freq=array([[486.80932203, 168.19067797],
       [390.19067797, 134.80932203]]))

BMI vs. EXERANY2 and disabled:

                       df        sum_sq      mean_sq          F        PR(>F)
C(EXERANY2_no_na)     1.0   1289.123160  1289.123160  28.073209  1.363661e-07
C(disabled_no_na)     1.0    361.244477   361.244477   7.866814  5.107537e-03
Residual           1344.0  61716.547295    45.920050        NaN           NaN
                            OLS Regression Results
==============================================================================

Dep. Variable:              BMI_no_na   R-squared:                       0.026
Model:                            OLS   Adj. R-squared:
    0.025
Method:                  Least Squares   F-statistic:
    17.97
Date:                 Sun, 22 Feb 2026   Prob (F-statistic):           1.99
    e-08
Time:                         17:43:55   Log-Likelihood:
    -4487.2
No. Observations:                 1347   AIC:
    8980.
Df Residuals:                     1344   BIC:
    8996.
Df Model:                            2
Covariance Type:             nonrobust
=========================================================================================

                               coef    std err          t      P>|t|
    [0.025      0.975]
-----------------------------------------------------------------------------------------

Intercept                   27.8740      0.238    117.343      0.000
    27.408      28.340
C(EXERANY2_no_na)[T.2.0]     2.0290      0.439      4.625      0.000
     1.168       2.890
C(disabled_no_na)[T.1]       1.1377      0.406      2.805      0.005
     0.342       1.933
==============================================================================

Omnibus:                       740.497   Durbin-Watson:
    1.990
Prob(Omnibus):                   0.000   Jarque-Bera (JB):
    14851.974
Skew:                            2.113   Prob(JB):
    0.00
```

```
Kurtosis:                      18.709   Cond. No.                         2.76
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```