Dana Hellman
USP 634 Data Analysis I
Homework II

1. According to 2016 American Community Survey, the point estimates of average per capita income for Beaverton, Hillsboro, and Portland and the corresponding Margin of Error is listed in the table below. Can you state with 95% confidence level that the average per capita income in Portland is higher than Beaverton and Hillsboro? Hint: ACS uses margin of errors corresponding to 90% confidence level. (15 pts)

Beaverton City	**$34,414**	**+/-2,606**
Hillsboro city	**$32,296**	**+/-2,930**
Portland city	**$37,513**	**+/-989**

Since the given margin of error values correspond to 90%, I worked backwards to determine the Standard Error (SE) value for each city, using the 90% confidence Z value of 1.65; I then multiplied the SE values by 1.96 (the Z value for 95% confidence) to get an updated margin of error for each city:

Beaverton: 2606 = 1.65 * SE SE = 2606/1.65 = 1579.39
Margin of error for 95% confidence = 1.96 * 1579.39 = **$3,096**
Hillsboro: 2930 = 1.65 * SE SE = 2930/1.65 = 1775.76
Margin of error for 95% confidence = 1.96 * 1775.76 = **$3,480**
Portland: 989 = 1.65 * SE SE = 989/1.65 = 599.39
Margin of error for 95% confidence = 1.96 * 599.39 = **$1,175**

Using the newly calculated margin of error values, I then calculated a range for each city at 95% confidence:

Beaverton Average Income Range = $34,414 +/- $3,096 = ($31,318; $37,510)
Hillsboro Average Income Range = $32,296 +/- $3,480 = ($28,816; $35,776)
Portland Average Income Range = $37,513 +/- $1,175 = ($36,338; $38,688)

Based on these results, I can say that Portland's incomes are higher than Hillsboro's, given that there is no overlap between the two cities' income ranges - the lowest value in the Portland range is higher than the highest value in the Hillsboro range. However, I cannot say that Portland's incomes are higher than Beaverton's because there *is* some overlap in the values of those two cities; Beaverton's highest value is approximately $1,000 more than Portland's lowest value, meaning that the average income in Beaverton could conceivably be higher than that in Portland.

2. You have been asked to manage a regional study on attitudes toward regional growth management. You've been given a budget of $75,000 to conduct a regional mail survey of households. Your Board wants to know the proportion of residents in favor of an urban growth boundary within an accuracy of plus or minus two percent and at a confidence level of 90 percent. (Recall that when you don't know the true population proportion, to be conservative you should assume a maximum standard error). What is the minimum number of households you need to randomly sample? Can you conduct the survey within budget given an estimated cost of $25 for administering each survey? (15 pts)

In order to determine the needed number of households (n), we can work backwards from:
$E = Z^* \sqrt{ps(1-ps)/n}$

Use margin of error E=2%, ps = 50% , and Z for 90% = 1.65

$.02 = 1.65 * \sqrt{.5(1-.5)/n}$
$.02 = 1.65 * \sqrt{.25/n}$
$\sqrt{n} = (1.65* .5)/.02 = 41.25$
n = 1,701.56 ---- round up to 1,702

Based on this calculation, you would need to randomly sample at least 1,702 homes. At $25 per survey, the cost = $42,550, which is within budget.
.

3. A random sample of adults living in 60 traditional, mixed-use, pedestrian-friendly neighborhoods and adults living in 55 postwar, auto-oriented neighborhoods, all with comparable household income levels, revealed the following:

Traditional neighborhoods averaged 15.8 daily vehicle miles traveled (VMT) per adult household member, with a standard deviation of 5.3 VMT.

Auto-oriented neighborhoods averaged 18.3 VMT per adult household member and a standard deviation of 7.5 VMT.

Create 90% confidence intervals for VMT for each type of neighborhoods. Do the intervals overlap with each other? Using the hypothesis testing process, test at α = .10 level the hypothesis of New Urbanists that people living in traditional neighborhoods have lower automobile usage. (20 pts)

TRADITIONAL NEIGHBORHOODS

n = 60 $\bar{x}$ = 15.8 s = 5.3 SE = 5.3/ $\sqrt{60}$ = .684 df = 60-1 = 59 t_{59} = 1.67

90% CI = 15.8 +/- 1.67 * .684 = (14.66, 16.94)

AUTO-ORIENTED NEIGHBORHOODS

n = 55 $\bar{x}$ = 18.3 s = 7.5 SE = 7.5/ $\sqrt{55}$ = 1.011 df = 55-1 = 54 t_{54} = 1.67

90% CI = 18.3 +/- 1.67 * 1.011 = (16.61, 19.99)

There is some overlap between these two intervals, in the values ranging from 16.61 to 16.94.

HYPOTHESIS TEST at α = .10 using *difference between two means*:

H_0 = Average VMT for Traditional & Auto-Oriented neighborhoods are not different

H_A = Average VMT for Traditional neighborhoods are less than for Auto-Oriented ones

Point Estimate = 15.8 - 18.3 = -2.5

$SE_{(\text{Traditional VMT - Auto Oriented VMT})} = \sqrt{\ }\ (5.3^2/60) + (7.5^2/55) = 1.22$

T = -2.5/1.22 = -2.05

p value = .02

The p value of .02 is less than the α = .10 threshold, so we can reject the null hypothesis; there is evidence to support the alternative hypothesis.

[p value was calculated using R: pt(-abs(-2.05), df = pmin(60, 59)-1)]

4. Housing values were compared between residences with a view of Mount Hood and otherwise comparable residences (e.g., amenities, neighborhood quality) though without a view. The analysis sought to measure the imputed value of a view on home sales prices. Two factors were also controlled for in the analysis: age of home and distance to downtown Portland. Thus, randomly selected homes that just sold were matched on these two factors, yielding the data shown in the excel spreadsheet view.csv. Test at α = .05 level that homes with a view enjoy significant value premiums, as reflected by differences in housing values. (20 pts)

HYPOTHESIS TEST at $\alpha = .05$ using *paired data*:
H_0 = The cost of homes is not affected by the availability of a view
H_A = Homes with no view cost less than homes with a view
$n_{diff} = 25$ $\bar{x} = 20.68$ $s = 19.36$ $SE = 19.36/\sqrt{25} = 3.87$
$T = 20.68/3.87 = 5.34$
$df = 25-1 = 24$

In this case, the p value is 8.806863e-06 [determined using R: pt (-abs(5.34), df = 24)] -- as this value is well below the threshold of .05, we can reject the null hypothesis; there is evidence to support the alternative hypothesis.

5. Using your own data set and test a hypothesis with t-test. First select (or create by recoding) a nominal variable ("V1") with two categories (for example, male and female, income above or below poverty line, before or after a "treatment"), then choose a continuous variable (or a discrete or even an ordinal variable if there is no continuous variable in your dataset) ("V2"), and finish the following tasks (30 points).

Nominal Variable: PERCENT NON-WHITE
[categories = Under 25% Non-White Population & Over 25% Non-White Population]

Continuous Variable: OBESITY [measured as a percentage of total census tract population]

1.) Calculate confidence intervals for V2, first for all observations and then for each of the two groups (categories) in variable V1;

95% CI for OBESITY (all values)
n = 145
$\bar{x}$ = 25.30
s = 3.14
SE = 3.14/ $\sqrt{145}$ = .26
CI = 25.30 +/- (1.96 * .26) = 25.30 +/- .51 = (24.79, 25.81)

95% CI for OBESITY when NON-WHITE is Under 25%
n = 92
$\bar{x}$ = 23.87
s = 2.39
SE = 2.39/ $\sqrt{92}$ = .25
CI = 23.87 +/- (1.96 * .25) = 23.87 +/- .49 = (23.38, 24.36)

95% CI for OBESITY when NON-WHITE is Over 25%
n = 53
$\bar{x}$ = 27.80
s = 2.71
SE = 2.71/ $\sqrt{53}$ = .37
CI = 27.80 +/- (1.96 * .37) = 27.80 +/- .73 = (27.07, 28.53)

2.) Formulate a hypothesis for the relationship between the two groups (categories), conduct a t-test for your hypothesis, and interpret your hypothesis testing results;

HYPOTHESIS TEST at $\alpha = .05$:
H_0 = Prevalence of obesity is not affected by non-white population
H_A = Populations that are Over 25% Non-White exhibit a higher rate of obesity

Point Estimate = 23.87 - 27.80 = -3.93
$SE_{(\text{Under 25\% - Over 25\%})} = \sqrt{\ } \ (2.39^2/92) + (2.71^2/53) \ = .45$

T = -3.93/.45 = -8.7

In this case, the p value will be extremely low, well under .05. As such, we can reject the null hypothesis; there is evidence to support the alternative hypothesis that a higher non-white population corresponds with a higher prevalence of obesity.

Two Different Outcomes from using different methods in R:

Using code **pt(-abs(-8.7), df = pmin(92, 53)-1)** I get p = 5.070158e-12

Using the Welch two sample t test **t.test(OBESITY ~ PER_NONWHT, data = DataHW2_xls, alternative = 'greater', conf.level = .95)** I get p = 2.768e-14

In either case, the value is well below .05, so the null can be rejected.

3.) Describe and discuss your findings, as appropriate.

These results all indicate that there is a relationship between non-white population and obesity; these two variables are not independent. This is evident from the confidence intervals alone, as the mean obesity ranges for the UNDER 25% and OVER 25% have absolutely no overlap; the stark difference between the two provides evidence that the difference in non-white status has an effect on mean obesity. This conclusion is further supported by a hypothesis test, which yields a p value of nearly zero (much less than the designated threshold value of $\alpha = .05$). This result strongly suggests that we can reject the null hypothesis, which states that non-white population and obesity are not related. There is evidence to support the alternative hypothesis high non-white populations (over 25%) exhibit higher obesity rates than low non-white populations.