Kitu Komya (UID: 404-491-375)
Statistics 101A (Discussion 3A)
Homework 1 (due: 9/30/16)

Note: I'm concurrently taking Stats 20, so I'm still familiarizing myself with R/R Markdown

---

1a. The formula for finding a confidence interval is the following: $\bar{x} \pm z * \frac{s}{\sqrt{n}}$, where

$\bar{x}$ is the sample mean = $23727
z is the z-score = 1.96 (found for a 95% confidence level from a standard z-table)
s is the sample standard deviation = $23624
n is the sample size = 30

$z * \frac{s}{\sqrt{n}}$ is also known as the margin of error and $\frac{s}{\sqrt{n}}$ is known as the standard error. Plugging in these values into the equation, our margin of error is $8453.7.

Thus, our lower bound is $15273.3 and our upper bound is $32180.7, giving us a 95% confidence interval of **($15273.3, $32180.7)** for the mean income of the population.

1b. In order to create an accurate confidence interval, our assumptions revolve around the method of data collection and distribution.

- We assume that the **data has been collected using simple random sampling** which eliminates bias that may result from selection.
- We also assume that our **sampling distribution is normally distributed**, following the Central Limit Theorem.
- We do not know the standard deviation of the entire population, since we are only given the sample standard deviation, *s*, and thus must assume that the **standard deviation of the entire population is similar to the sample standard deviation**.

1c. **No**, the same confidence interval cannot be used to assess a new sample. By definition, a 95% confidence level means that if the same (**the second procedure will likely have a different sample standard deviation**!) procedure for computing a 95% confidence level is used repeatedly on the same population, 95% of the time that particular confidence interval will contain the true population (**not sample**!) parameter value.

1d. We establish the following null and alternate hypotheses, realizing a two-tailed test.

H_0: μ = $25000
H_A: $\mu \neq$ $25000
Since we don't know the population standard deviation, we'll use the 1-mean Student's t-test

$$t = \frac{(\bar{x} - \mu)}{(s/\sqrt{n})}$$, where

$\bar{x}$ is the sample mean = $23727
μ is the hypothesized population mean = $25000
s is the sample standard deviation = $23624
n is the sample size = 30

Plugging in these values, we reach t-value = -0.295
Our degrees of freedom is n – 1 = 30 – 1 = 29

At a degrees of freedom of 29, a significance level at 5% (0.05), for a two-tailed test, we get:
t-critical value = ± 2.045 (using online calculator)
p-value = 0.7701 (using online calculator)

Since t-critical value > |t-value| (2.045 > 0.295) and the p-value of 0.7701 > 0.05 (significance level), we fail to reject the null hypothesis. There is not statistically significant evidence to claim that the hypothesized population mean is not $25000. In other words, yes, the politician's claim that the mean income of US residence in 2000 was $25000 **can be confirmed using this data**.

Be a little careful. We can't confirm his claim, we can only fail to reject it. I won't penalize you for this though since you said that as well

1e. Since the p-value is 0.7701, the smallest significance level would be
1 – 0.7701 = 0.2299 ≈ 0.23, so
at a significance level of 23% we would have been able to reject the null hypothesis.

2a. Excluding all electric vehicles, there are **37723 vehicles** in the data set. This number comes from excluding any car that contained "Electric" in its fuelType1 category.

R code:
```
vehicles2.df <- subset(vehicles, fuelType1 != "Electricity")
vehicles2.df
```

2b. The population mean of highway mileage all cars except for electric is **23.9 MPG**.

R code:
```
mean(vehicles2.df$highway08)
```

2c. R code:
```
set.seed(123)
rand_samp <- sample(vehicles$highway08, 10, replace = TRUE)
rand_samp
mean(rand_samp)
sd(rand_samp)
```

R output: 24 22 18 23 26 23 23 35 30 16

As seen from 1a, the formula for confidence interval is the following: $\bar{x} \pm z * \frac{s}{\sqrt{n}}$, where
$\bar{x}$ is the sample mean = 24 MPG (found via R code)
z is the z-score = 1.96 (found for a 95% confidence level from a standard z-table)
s is the sample standard deviation = 5.5 MPG (found via R code)
n is the sample size = 10

After we plug in the values, we get our lower bound to be 20.6 mpg and our upper bound to be 27.4 mpg, giving us a 95% confidence interval of **(20.6 mpg, 27.4 mpg)** for the mean mpg of the population. **Yes, this interval includes the population mean** of 24 mpg, which is expected.

2d. As seen from the formula, with larger sample sizes, the margin of error will be smaller, since **the relationship is inverse because the "n" is in the denominator of the equation**. Thus, a 95% confidence interval with a sample size of 100 cars **will have a smaller width**. Specifically, using the dependence of n in the equation, the width of the interval will decrease by $\sqrt{\frac{100}{10}} \approx 3.16$ in the larger sample size.

2e. **A confidence level measures the likelihood of containing the true population parameter value in an interval** if the same procedure for computing the confidence level is used repeatedly on the same population.

2f. We establish the following null and alternate hypotheses, realizing a right (one) -tailed test.

H_0: $\mu = 20$ MPG
H_A: $\mu > 20$ MPG

Since we do not know the population standard deviation (we are only using data from the sample), we will use the 1-mean Student's t-test

$$t = \frac{(\bar{x} - \mu)}{(s/\sqrt{n})}, \text{ where}$$

$\bar{x}$ is the sample mean = 24.3 mpg (found via R code)
μ is the hypothesized population mean = 20 mpg
s is the sample standard deviation = 6.06 mpg (found via R code)
n is the sample size = 100

Plugging in these values, we reach t-value = 7.10
Our degrees of freedom is n – 1 = 100 – 1 = 99

At a degrees of freedom of 99, a significance level at 5% (0.05), for a one-tailed test, we get:
t-critical value = 1.66 (using online calculator)
p-value < 0.00001 (using online calculator)

Since t-critical value < t-value (1.66 < 7.10) and the p-value of 0.00001 < 0.05 (significance level), we reject the null hypothesis. In other words**, there is statistically significant evidence to claim that the population mean highway mileage is greater than 20 mpg.**

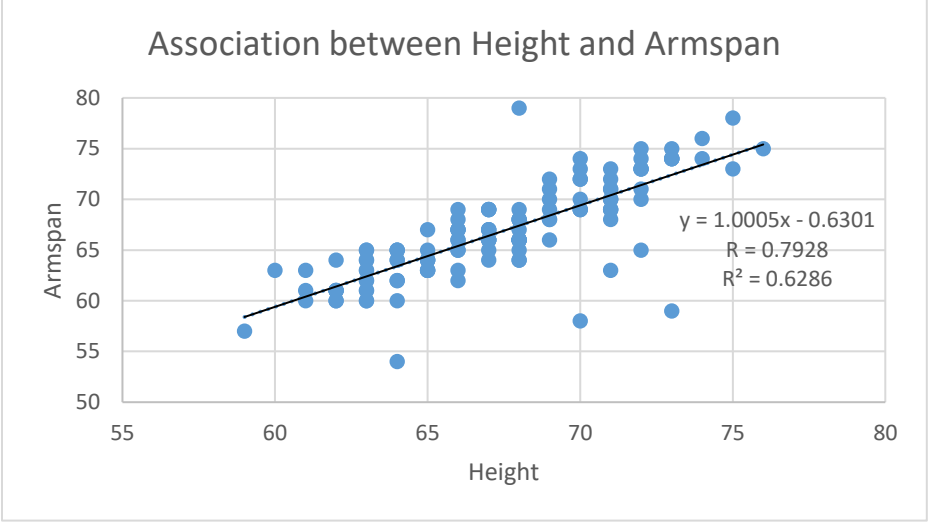
3. We establish the following null and alternate hypotheses, realizing a two-tailed test.

H_0: $\mu_{height} = \mu_{armspan}$
H_A: $\mu_{height} \neq \mu_{armspan}$

After cleaning the data to rid values that were clearly expressed in units of feet, not inches, and also those values that were evidently typos, we are left with 129 observations.

R code: armspan1.df <- subset(armspan, armspan > 30)

By inference of common sense, it seems that height and armspan will be paired, dependent data. To confirm, we create a scatterplot:



The scatterplot makes it evident that there is an association between height and armspan, since the correlation co-efficient R is 0.7928, which describes a moderately strong, positive, linear correlation. Thus, we can proceed with the Welch's Two Sample t-test.

R code:
t.test(x = armspan1.df$height, y = armspan1.df$armspan, mu = 0, alternative = "two.sided")

R Output:

t = 1.1409
p-value = 0.255
alternative hypothesis: true difference in means is not equal to 0

Since the p-value 0.255 > 0.05 (significance level), we fail to reject the null hypothesis. In other words**, there is statistically significant evidence to claim that a person's height is approximately the same as her armspan. Thus, Vitrivus was correct.**

Good!

Note: when we apply the formula for confidence interval, we get the following results

$$\bar{x} \pm z * \frac{s}{\sqrt{n}},$$

height:

$\bar{x} = 67.35$
z = 1.96
s = 3.690
n = 129

95% confidence interval is (66.71, 67.99)

armspan:

$\bar{x} = 66.75$
z = 1.96
s = 4.657
n = 129

95% confidence interval is (65.95, 67.55)

which make it seem as if the overlap of the intervals indicates that the mean values are the same. Using the Welch's Two Sample t-test confirms mathematically that the two means are statistically similar enough to certify our claim, coming again to the conclusion that Vitrivus was correct.