Name: Sachi Graber

UmichID: sgraber

What dataset are you working with: drinks

List 3 questions that you can ask with your dataset.

Q1: Is the distribution of drinks in Germany different than the distribution of drinks in Austria?

Q2: Are more servings of beer drunk in countries with more than 100 servings of spirits than in countries with less than 100 servings of spirits?

Q3: Is there a correlation between the amount of beer drunk in a country and the amount of wine drunk in that country?

List the associated null hypothesis for each question:

Q1: People in Austria drink things in the same proportion as people in Germany

Q2: People in countries where the average servings of spirits is less than 100 drink the same amount of beer as in countries where the average servings of spirits is more than 100.

Q3: There is no correlation between these two variables.

What statistical test(s) will you use to answer each of the questions:

Q1: Chi-square test for independence will tell us if there is any association between the samples

Q2: two-sample, one-tail unpaired t-test

Q3: linear model/regression

Make a visual plot showing the relationship that you will analyze statistically (e.g. boxplot for t-test or ANOVA; scatterplot for regression; table for chi-square).

Q1:

[,1] [,2]

country "Austria" "Germany"

beer\_servings 279 346

spirit\_servings 75 117

wine\_servings 191 175

total\_litres\_of\_pure\_alcohol 9.7 11.3

Q2: 

Q3: 

Do your data meet the assumptions required for the statistical test you want to run? Please state the assumptions you examined and whether or not your data meet those assumptions:

Q1: Independent observations – I hope so, because they are in different countries

Random sample – I assume so, but I guess I’m not sure

Sample size >30 – I think so, because some of the cells have a value greater than 30

Cell count >5 or 10 – yes

Q2: Continuous data: yes, since # of drinks is a continuous count

Random sample: not sure, but I hope so since I didn’t collect the data

Independent samples: Yes, since different countries represented

Normally distributed: sure, since we have large sample sizes

Equal variance: pretty close. Same order, etc.

> var(low\_spirits$beer\_servings)

[1] 7558.892

> var(high\_spirits$beer\_servings)

[1] 9738.237

Q3: There is a linear relationship between the data – I am not sure about this yet!

Data are homoscedastic – I BROKE THIS ASSUMPTION because the line I get here is not horizontal – skedasticity increases with increasing values. I can also look at the plot above and see that points are closer to each other near the origin, and farther apart as you get farther away.



Errors are statistically independence – I think so, they don’t really follow a path or something

Normal error distribution 

Run the statistical test! Put your results here:

Q1:

X-squared = 9.819, df = 3, p-value = 0.02017

Q2:

data: low\_spirits$beer\_servings and high\_spirits$beer\_servings

t = -6.2399, df = 113.89, p-value = 3.853e-09

alternative hypothesis: true difference in means is less than 0

Q3:

lm(formula = wine\_servings ~ beer\_servings, data = drinks)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.35221 7.09531 0.754 0.452

beer\_servings 0.41539 0.04845 8.574 3.38e-15 \*\*\*

Interpret your results!

Q1: The distribution of drinks drunk in Austria is significantly different than the number of drinks drunk in Germany at alpha=0.05 level.

Q2: There is a difference in the means and we can say that the number of beers drunk in countries that drink more spirits is significantly greater than the beers drunk in countries that drink less spirits. This is significant at the 0.01 level since p<<0.01.

Q3: There is a positive and significant correlation between the amount of beer and wine drunk in a country. If you know the amount of beer drunk increases at the margin, you can assume that the amount of wine also increases at the margin. For each additional serving of beer, an additional 0.4 servings of wine is normally drunk.