Answers to Practice Final Exam

1. A social worker has been investigating the effect of residence on the lives of her clients. She randomly selected 350 clients who live in public housing and 350 clients who live in standard housing. She found that those who live in public housing have been working in their current job an average of 1.3 years with a standard deviation of 4.3 years. Those living in standard housing have been working in their current job an average of 2.4 years with a standard deviation of 3.6 years.

a. Using 95% confidence level, test whether these two groups differ significantly in their average job tenure. Make sure to state your null and research hypotheses in words as well as using formal notation. After finishing the test, state your formal conclusion with regard to the null hypothesis as well as your substantive answer.
b. Evaluate the practical significance of the difference in average job tenure between these two groups.
c. Evaluate the chances that you are making Type I and Type II errors.

a. Test:

1. H0: $\mu 1 = \mu 2$
   H1: $\mu 1 \neq \mu 2$

In words: H0: Those in public housing and in standard housing have the same average job tenure. H1: Those in public housing and in standard housing have different average job tenure length.

2. α=.05
3. two-tailed test, independent samples

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\left[\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}\right]\left[\frac{n_1+n_2}{n_1 n_2}\right]}}$$

4. numerator=1.9-2.4=-0.5
   denominator= (sqrt(((350-1)*4.3*4.3 + (350-1)*3.6*3.6)/(350+350-2)*(350+350)/(350*350))) = .3
   test statistic t=-0.5/.3=-1.67
5. Get the critical value: df=350+350-2=698; alpha =.05, two-tailed; find critical value in Table B2: t=1.96
6. Compare: test statistic is smaller than the critical value (1.16<1.96)
7. Fail to reject the null hypothesis
8. Conclusion: Based on a sample of 700 individuals, we do not have the evidence to conclude with 95% confidence that there is a difference in the population between these those who live in standard vs public housing in terms of their average job tenure.

b. Practical significance: Let's calculate effect size: (1.9-2.4)/sqrt((4.3^2+3.6^2)/2)=-.126 We interpret the absolute value, .126, which is a very small effect – so it is both not statistically significant and not practically significant.
c. We failed to reject the null → Probability of Type I error is 0, probability of Type II error is small – the sample size is 700, so it is a large sample.

2. You would like to know whether men with more years of formal education work longer hours. You collect data from 20 randomly selected men on the number of years of formal education they obtained as well as the number of hours they work per week and get the following results:
b = 0.453, with standard error of 0.210.

a. Write a sentence summarizing how exactly changes in one of these variables are linked to changes in the other.
b. Evaluate the practical significance of this effect.
c. Using significance level of .1, test whether education leads to longer work hours. Make sure to state your null and research hypotheses in words as well as using formal notation. After finishing the test, state your formal conclusion with regard to the null hypothesis as well as your substantive answer.
d. Evaluate the chances that you are making Type I and Type II error.

a. In the sample, as education increases by 1 year, hours spent at work increase by 0.453 of an hour.
b. Let's see what happens if education increases by 4 (equivalent of a bachelor's degree): 4*.453=1.812. So if education increases by 1 year, hours spent at work increase by 1.8 – that is not huge but substantial, so this effect is practically significant.
c. Test:

1. State hypotheses:
H0: $\beta = 0$ In words: Years of education have no effect on work hours.
H1: $\beta > 0$ In words: More years of education are linked to longer work hours.
2. Select alpha: 0.1
3. Test statistic: Student's t for regression coefficient
4. $t = b/s_b = 0.453/0.210=2.16$
5. Use the table to find critical value: Table B2 (df=n-2 = 20-2=18, alpha = .1, one-tailed) → 1.331
6. Compare computed & critical value: 2.16 > 1.331
7. State your decision: We reject H0 in favor of H1.
β =0.453, p<.10
8. Conclusion: Based on a sample of 20 men, we can be 90% sure that, in the population, higher levels of education are associated with longer hours of work among men. [This relationship is statistically significant at .10 level.]

d. We rejected the null → probability of Type I error is less than .1; probability of Type II error is 0.

3. You would like to know whether there are differences in employment tenure by industry. You sample 4 companies each from manufacturing, retail, and service industries, and get the following data for the average number of years employees stay on the job in each company:
Manufacturing: 10, 7, 8, 7
Retail: 2, 3, 5, 2
Service sector: 2, 1, 3, 2

a. Using 99% confidence level, test whether there are statistically significant differences in employment tenure by industry. Make sure to state your null and research hypotheses in words as well as using formal notation. After finishing the test, state your formal conclusion with regard to the null hypothesis as well as your substantive answer to the research question.
b. Evaluate the chances that you are making Type I and Type II error.

a. Test:
1. Hypotheses:
H0: $\mu_1=\mu_2=\mu_3$ H1: $\mu_1\neq\mu_2\neq\mu_3$
In words: H0: Employment tenure does not differ by industry.
H1: At least one of the three industries has distinct employment tenure.
We use one-tailed test (always for ANOVA)
2. α= .01
3. F statistic
4. Let's calculate all the necessary components, and then construct the variance decomposition table. Means:
Manufacturing: (10+7+8+7)/4=8
Retail: (2+3+5+2)/4=3

Service sector: (2+1+3+2)/4=2
Grand: 4.3

Group	X	$\bar{X}_{group}$	$\bar{X}_{grand}$	$X-\bar{X}_{group}$	$(X-\bar{X}_{group})^2$	$X-\bar{X}_{grand}$	$(X-\bar{X}_{grand})^2$	$\bar{X}_{group}-\bar{X}_{grand}$	$(\bar{X}_{group}-\bar{X}_{grand})^2$
1	10	7	4.3	2	4	5.7	32.49	3.7	13.69
1	7	7	4.3	-1	1	2.7	7.29	3.7	13.69
1	8	7	4.3	0	0	3.7	13.69	3.7	13.69
1	7	7	4.3	-1	1	2.7	7.29	3.7	13.69
2	2	3	4.3	-1	1	-2.3	5.29	-1.3	1.69
2	3	3	4.3	0	0	-1.3	1.69	-1.3	1.69
2	5	3	4.3	2	4	0.7	0.49	-1.3	1.69
2	2	3	4.3	-1	1	-2.3	5.29	-1.3	1.69
3	2	2	4.3	0	0	-2.3	5.29	-2.3	5.29
3	1	2	4.3	-1	1	-3.3	10.89	-2.3	5.29
3	3	2	4.3	1	1	-1.3	1.69	-2.3	5.29
3	2	2	4.3	0	0	-2.3	5.29	-2.3	5.29
Σ				0	14	0	96.68	0	82.68

ANOVA table:

Source	SS	df	Mean SS	F
Between groups	82.68	2	41.34	26.57
Within groups	14	9	1.556	
Total	96.68	11		

Test statistic F = 26.57

5. Critical value: df for numerator=2, df for denominator =9. Our critical value will be the value for these df and α=.01. Therefore, critical value = 8.02.
6. Test statistic > critical value (26.57>8.02)
7. Therefore, we can reject the null hypothesis.
8. Conclusion: Based on the data from a sample of 12 companies, we can be 99% sure that there are differences in average employment tenure by industry in the population.

b. Probability of Type I error is less than .01, probability of Type II error is 0.

4. You would like to test whether Americans with higher income watch TV less. You collect data from a random sample of 20 individuals in the U.S., and obtain r_{xy} =- 0.305.

a. What can you say about the strength and direction of the relationship between income and watching TV in the sample?

b. Using significance level .05, test whether there is in fact a relationship between income and watching TV in the population. Make sure to state your null and research hypotheses in words as well as using formal notation. After finishing the test, state your formal conclusion with regard to the null hypothesis as well as your substantive answer.

c. Evaluate the chances that you are making Type I and Type II error

a. Size and direction: there is a weak negative relationship in the sample (higher income is associated with fewer hours watching TV in the sample).

b. Test:

1. H0:$\rho = 0$ H1: :$\rho < 0$

In words: H0: There is no link between income and hours spent watching TV.

H1: Higher incomes are linked to fewer hours of watching TV.
one-tailed test

2. $\alpha = .05$
3. Correlation coefficient
4. Test statistic r=-0.305
5. Finding critical value: df=n-2=20-2=18, in Table B4, we find critical value = 0.3783
6. Test statistic<critical value (0.305<0.3783)
7. We fail to reject the null hypothesis.
8. Based on this sample of 20 individuals, we do not have the evidence we would need to conclude with 95% confidence that there is a relationship between income and the number of hours watching TV in the larger population.

c. Probability of Type I error is 0, probability of Type II error is high because the sample size is low, only 20.

5. In a survey, women and men were asked whether they feel that (a) friends/social life, (b) job/primary work activity, or (c) health/physical condition contributes most to their general happiness. The results show that among women, 20 persons responded "friends," 20 people responded "job," and 40 people responded "health." Among men, 20 people responded "friends," 60 people responded "job," and 40 people responded "health."

a. Describe the relationship between gender and the opinion on what contributes most to happiness in this sample.
b. Use 99% confidence level to test whether gender has an effect on people's opinions about the main determinant of happiness in the larger population. Make sure to state your null and research hypotheses in words as well as using formal notation. After finishing the test, state your formal conclusion with regard to the null hypothesis as well as your substantive answer.
c. Evaluate the chances that you are making Type I and Type II error.
d. If you conclude that the overall relationship exists in the population, describe the pattern of differences in the population.

.

a. The pattern in the sample: Among women, 25% emphasize friends, but among men, only approximately 17% do. A larger percentage of women than men highlights the importance of health (50% of women vs 33% of men). In contrast, compared to women, men are particularly likely to emphasize jobs as their key determinant (50% of men vs 25% of women do).

b. Test:

	Gender		
	Women	**Men**	**Row Totals**
Friends	20 25.00%	20 16.67%	40 20.00%
Job	20 25.00%	60 50.00%	80 40.00%
Health	40 50.00%	40 33.33%	80 40.00%
Column Totals	80 100.00%	120 100.00%	200 100.00%

1. H0: $O_i = E_i$
H1: $O_i \neq E_i$

H0: the main determinant of happiness does not depend on gender (the two variables are independent)
H1: the main determinant of happiness depends on gender

2. Alpha=.01
3. Test statistic: χ^2
4. Calculate:

Cell	$E= C_t \times R_t/T$	O-E	$(O-E)^2$	$(O-E)^2/E$
C_1R_1	80 x 40/200 = 80 x .2 = 16	20 - 16 = 4	16	16/16 = 1
C_2R_1	120 x 40/200 = 120 x .2 = 24	20 - 24 = -4	16	16/24 = .67
C_1R_2	80 x 80/200 = 80 x .4 = 32	20 - 32 = -12	144	144/32 = 4.5
C_2R_2	120 x 80/200 = 120 x .4 = 48	60 - 48 = 12	144	144/48 = 3
C_1R_3	80 x 80/200 = 80 x .4 = 32	40 - 32 = 8	64	64/32 = 2
C_2R_3	120 x 80/200 = 120 x .4 = 48	40 - 48 = -8	64	64/48 = 1.33
Σ	200	0		χ^2 = 12.5

$\chi^2 = \Sigma((O-E)^2/E) = 12.5$

5. Critical value: alpha=.01, df=(C-1)*(R-1)=1*2=2 → 9.21
6. Critical value is smaller than our chi-square statistic.
7. We reject the null hypothesis of independence.
8. Based on this sample of 200 individuals, we can be 99% confident that the views on the most important determinant of happiness are related to gender in the larger population.

c. Type II error probability is 0, Type I is less than .01.
d. Two residuals would exceed 2.576 cutoff – C1R2 and C2R2 – therefore, the gender difference exists with regard to the importance placed on jobs (men are more likely to rate jobs as most important determinants of happiness than women).

6. A researcher wants to find out whether the use of a new medical device reduces hospital stay for patients with a certain condition. Prior to its introduction, the average length of stay for these patients was 4.2 days. For the 24 patients that participated in the study and used the new device, the average length of stay was 3.7 days, with a standard deviation of 1.1 days.

a. Can we conclude with 95% confidence that the new device shortens the length of stay? Make sure to state your null and research hypotheses in words as well as using formal notation. After finishing the test, state your formal conclusion with regard to the null hypothesis as well as your substantive answer.
b. Evaluate the chances that you are making Type I and Type II errors.

a. Test:
1. H_0: $\mu = 4.2$
H_1: $\mu < 4.2$ -- since we expect a shorter hospital stay among the users of the new device → use a directional research hypothesis → one-tailed test
In words: The null hypothesis is that the hospital stay for those using the new medical stay is the same as it was prior to its introduction, 4.2 days. The research hypothesis is that the hospital stay is shorter for those using the new medical device.
2. We want to use 95% confidence level → we need to use significance level alpha = 0.05
3. Test statistic – Student's t

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}}$$

4. Compute using the formula:
t = (3.7-4.2)/(1.1/sqrt(24))=-2.23
5. Use Table B2 to find the critical value: df=n-1=24-1=23, alpha=0.05, one-tailed test → t_{crit} = 1.714

6. Does computed statistic exceed critical value? We take the absolute value of the computed statistic, without the sign, and we compare: 2.23 is larger than 1.714 so it exceeds the critical value.
7. Conclusion: We reject the null hypothesis H_0: $\mu = 4.2$.
8. Based on this sample of 24 patients, we are 95% confident that the new medical device reduces hospital stays for the patients with this condition in the population.

b. We rejected the null → Probability of Type I error is less than .05, probability of Type II error is 0.

7. Using Stata and the gss2012.dta dataset, calculate the correlation coefficient for the variables *maeduc* and *educ*, and then calculate coefficients of determination and alienation. Explain in words what each of these coefficients tells you about the relationship between one's own education and one's mother's education in the sample.

a. Correlation coefficient:
b. Coefficient of determination:
c. Coefficient of alienation:
d. What can you conclude about the relationship between these two variables in the population? Make sure to state your null and research hypotheses in words as well as using formal notation. After finishing the test, state your formal conclusion with regard to the null hypothesis as well as your substantive answer.
e. Evaluate the chances that you are making Type I and Type II error

a. Correlation coefficient=.447.; there is a moderate positive correlation between respondents' own level of education and their mothers' level of education in the sample.
b. Coefficient of determination=.447*.447= .2; About 20% of variance in respondents' own level of education can be explained by their mothers' level of education (or we can say that about 20% of variance in individuals' own education and their mothers' education is shared).
c. Coefficient of alienation: 1-.2=.8; About 80% of variance in respondents' own level of education cannot be explained by their mothers' level of education (or we can say that about 80% of variance in individuals' own education and their mothers' education is unique).
d. H0: $\rho=0$ There is no relationship between one's own education and one's mother's education in the population.
H1: $\rho\neq0$ There is a relationship between one's own education and one's mother's education in the population.
We select alpha=.05, calculate correlation coefficient in Stata. P value: p<.0001. This p-value is smaller than alpha. We reject the null hypothesis of no relationship. Conclusion: Based on GSS 2012 national sample data, we are 99.9% confident that there is a relationship between one's own education and one's mother's education in the population.

e. After we conducted the test and rejected the null, we can report that Type II error probability is zero, while Type I error probability is less than .0001.

Output:

```
. pwcorr maeduc educ, sig

             |   maeduc     educ
-------------+------------------
      maeduc |   1.0000
             |
             |
        educ |   0.4467   1.0000
             |   0.0000
             |
```

8. Using Stata and the gss2012.dta dataset, regress *realinc* (family income) on *educ*. Create a scatterplot for these two variables and display the regression line as well as lowess line on the scatterplot.

a. Interpret the slope in the sample (both direction and number) in words.
b. In words, what does the intercept mean?
c. What can you conclude about the relationship between these two variables in the population?
d. Evaluate the chances that you are making Type I and Type II error.
e. If the effect exists in the population, describe the slope in words using the 95% confidence interval.
f. Discuss the practical significance of this effect.
g. Write out the regression equation and calculate the predicted value of income for someone with 10 years of education.
h. Calculate the error of estimate for someone with 10 years of education and income of $30,000.
i. Write a sentence describing the R-squared value.
j. Write a sentence discussing whether the relationship is linear.

a. The two variables are positively related: If education increases by one year, income increases by $5168.324.
b. The intercept is the value of income when someone's education is 0 years. Here, that value is negative, -$36,133 would be the predicted income for someone with 0 years of education, which is not a very realistic value to say the least (you cannot have negative income).
c. H0: $\beta=0$ Education level has no effect on family income.
   H1: $\beta\neq0$ Education level has an effect on family income.
   Alpha =.05. P-value in Stata: $p<.001$, which is smaller than alpha.
   We can reject the null hypothesis. Conclusion: Based on GSS 2012 sample data, we can conclude with 99.9% confidence that increases in education are linked to increases in family income in the U.S. population.
d. Before we conducted the test, we would estimate that the Type II error probability is very low because the sample size is large, and Type I error probability is our alpha, .05. After we conducted the test and reject the null, we can report that Type II error probability is zero, while Type I error probability is less than .001.
e. We are 95% sure that, in the population of all Americans, as education increases by on year, income increases by between $4,602 and $5,734.
f. This is a very sizable impact – if one year translates into a something like a $5K increase in income, then a bachelor's degree (4 years) would add app. $20K to one's family income, on average. That is quite large.
g. Y'=-36133+5168*X

If X=10, then Y'=-36133+5168*10=15547

The predicted income for someone with 10 years of education is $15,547

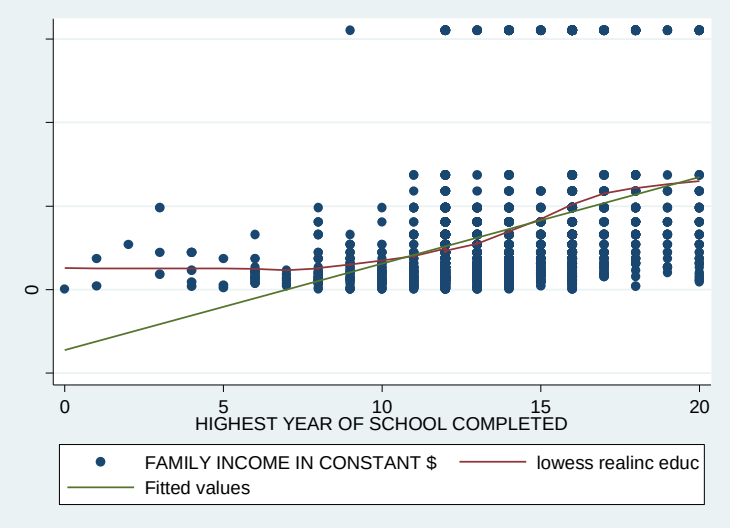
h. Y-Y'=30000-15547=14453 is the error of estimate.
i. R-squared value shows that approximately 15.44% of variance in family income can be explained by respondents' education levels.
j. Overall, the relationship is not linear. It is, however, quite close to linear starting at 10 years of education and up, so regression line describes most of the data accurately. There are fewer people with less than 10 years of education, and for these people, the slope is much flatter – pretty much no relationship up to 8 years of education, and slight increase from 8 to 10.

Output:

```
. reg realinc educ

      Source |       SS       df       MS              Number of obs =    1758
-------------+------------------------------           F(  1,  1756) =  320.69
       Model |  4.2795e+11     1  4.2795e+11           Prob > F      =  0.0000
    Residual |  2.3433e+12  1756  1.3345e+09           R-squared     =  0.1544
-------------+------------------------------           Adj R-squared =  0.1539
       Total |  2.7713e+12  1757  1.5773e+09           Root MSE      =   36530

------------------------------------------------------------------------------
     realinc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   5168.324   288.6075    17.91   0.000     4602.273    5734.374
       _cons |  -36133.21   4028.162    -8.97   0.000    -44033.71   -28232.72
------------------------------------------------------------------------------
. graph twoway (scatter realinc educ) (lowess realinc educ) (lfit realinc educ)
```



9. In Stata, use gss2012.dta dataset and for each of the following variable pairs, do the following:
   a. Examine frequency distributions for the two variables and write down variable type for each (nominal, ordinal, interval/ratio).
   b. Based on your variable examination, write a research question about group differences or a relationship between variables that can be answered using these two variables.
   c. State your null and research hypotheses (both in words and using formal notation). Clarify whether your research hypothesis is directional or non-directional and why, and whether your test will be one- or two-tailed.
   d. Determine the technique you will use to test your null hypothesis. Briefly explain your choice and conduct the analysis in Stata.
   e. Make a formal conclusion about your null hypothesis.
   f. Write the substantive conclusion based on the results of your analysis.
   g. If you do find a relationship/difference, make sure to include a substantive description of that relationship/difference.
   h. Regardless of your conclusion with regard to statistical significance, evaluate the practical significance of differences/effects.

Variable pairs [you would only get one pair on the exam; this is for practice]:

A. race and sibs

B. sibs and educ

C. sex and sibs
D. paeduc and educ

A. race and sibs

a. race: nominal, sibs: ratio
b. Does the average number of siblings differ by race in the U.S.?
c. Null: The average number of siblings does not differ by race.
   Research: The average number of siblings differs by race. (always non-directional for ANOVA, plus we have no specific expectations about how race could be shaping number of siblings).
   H0: $\mu1 = \mu2 = \mu3$
   H1: $\mu1 \neq \mu2 \neq \mu3$
d. We use ANOVA because sibs is ratio and race is nominal with three groups.
e. Using alpha level of .05, we reject the null hypothesis in favor of our research hypothesis. F=84.091, p<.001.
f. Based on GSS 2012 sample data, we can be 99.9% confident that the average number of siblings that people have differ by race in the U.S. population. [This difference is statistically significant at .001 level.]
g. Based on post-hoc pairwise comparison results, we can say that all three racial groups (White, Black, and Other) differ in the average number of siblings they have. Specifically, Blacks have the largest number of siblings (5.4 on average), followed by the Other group (4.7 on average), and finally followed by Whites (3.16 on average).
h. Let's calculate effect sizes for the three comparisons:

White vs Black: ES1= (5.39-3.17)/sqrt((2.48^2+4.42^2)/2)=.62
Black vs Other: ES2=(5.39-4.71)/ sqrt((3.44^2+4.42^2)/2)=.17
White vs Other: ES3=(4.71-3.17)/sqrt((2.48^2+3.44^2)/2)=.51

{This was based on means and SDs from this output:

```
      white |   3.1667797   2.4768908        1475
      black |   5.3866667   4.4282261         300
      other |   4.7142857   3.4403339         196
```

]

White the differences between Whites and the other two groups are medium size and therefore practically significant, the difference between Black and Other groups is very small and not practically significant despite being statistically significant.

Output:

```
. codebook sibs race
-------------------------------------------------------------------------------
sibs                                              NUMBER OF BROTHERS AND SISTERS
-------------------------------------------------------------------------------
                  type:  numeric (byte)
                 label:  LABAD, but 23 nonmissing values are not labeled

                 range:  [0,30]                       units:  1
         unique values:  23                       missing .:  3/1,974

              examples:  1
                         2
                         3
                         6
-------------------------------------------------------------------------------
race                                                        RACE OF RESPONDENT
-------------------------------------------------------------------------------
                  type:  numeric (byte)
                 label:  RACE
```

```
                range:  [1,3]                       units:  1
        unique values:  3                        missing .:  0/1,974

           tabulation:  Freq.   Numeric  Label
                        1,477         1  white
                          301         2  black
                          196         3  other
```

```
. oneway sibs race, means st obs bonferroni

           |  Summary of NUMBER OF BROTHERS AND
    RACE OF |               SISTERS
 RESPONDENT |        Mean   Std. Dev.        Obs.
------------+------------------------------------
      white |   3.1667797   2.4768908        1475
      black |   5.3866667   4.4282261         300
      other |   4.7142857   3.4403339         196
------------+------------------------------------
      Total |    3.658549   3.0797524        1971

                       Analysis of Variance
    Source             SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      1471.08458      2    735.54229     84.09     0.0000
 Within groups      17214.1189   1968   8.74701162
------------------------------------------------------------------------
    Total           18685.2035   1970   9.48487485

Bartlett's test for equal variances:  chi2(2) = 222.1623  Prob>chi2 = 0.000

      Comparison of NUMBER OF BROTHERS AND SISTERS by RACE OF RESPONDENT
                                (Bonferroni)
Row Mean-|
Col Mean |      white      black
---------+----------------------
   black |    2.21989
         |      0.000
         |
   other |    1.54751   -.672381
         |      0.000      0.040
```

B. sibs and educ

a. sibs: ratio; educ: ratio.

b. Does the number of siblings one has reduce the amount of education one gets? (You could also do a more correlation-like research question here – is there a relationship between one's level of education and their number of siblings?)

c. Null: The number of siblings one has doesn't affect the number of years of education.

Research: The number of siblings one has reduces the number of years of education. (directional, one-tailed test; but you can also write a non-directional if you wish)

H0: $\beta=0$

H1: $\beta<0$

d. Regression, because both variables are interval/ratio.

e. Using alpha level of .05, we reject the null hypothesis in favor of our research hypothesis. b=-0.299, p<.001

f. Based on GSS 2012 national sample, we are 99.9% confident that, in the U.S. population. the number of siblings one has reduces the number of years of education one gets.

g. More specifically, with each additional sibling, one's education is reduced by 0.299 of a year.

h. Three extra siblings would reduce someone's education almost by one year (by .9 of a year). That is not a very large but noticeable impact, so the effect is somewhat practically significant.

Output:

```
. codebook educ sibs
-------------------------------------------------------------------------------
educ                                                HIGHEST YEAR OF SCHOOL COMPLETED
-------------------------------------------------------------------------------
                  type:  numeric (byte)
                 label:  LABAB, but 21 nonmissing values are not labeled

                 range:  [0,20]                       units:  1
          unique values:  21                      missing .:  2/1,974

              examples:  12
                         12
                         14
                         16
-------------------------------------------------------------------------------
sibs                                                 NUMBER OF BROTHERS AND SISTERS
-------------------------------------------------------------------------------
                  type:  numeric (byte)
                 label:  LABAD, but 23 nonmissing values are not labeled

                 range:  [0,30]                       units:  1
          unique values:  23                      missing .:  3/1,974

              examples:  1
                         2
                         3
                         6

. reg educ sibs
      Source |       SS       df       MS              Number of obs =    1969
-------------+------------------------------           F(  1,  1967) =  187.03
       Model |  1667.84003     1  1667.84003           Prob > F      =  0.0000
    Residual |  17540.3621  1967  8.91731678           R-squared     =  0.0868
-------------+------------------------------           Adj R-squared =  0.0864
       Total |  19208.2021  1968  9.76026531           Root MSE      =  2.9862
------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        sibs |  -.2988215     .02185   -13.68   0.000    -.3416731   -.2559699
       _cons |   14.62585   .1045144   139.94   0.000     14.42088    14.83083
------------------------------------------------------------------------------
```

C. sex and sibs

a. sibs: interval/ratio; sex: nominal.
b. Do men and women in the U.S. have a different number of siblings on average?
c. Null: Men and women in the U.S. have a similar number of siblings on average.

Research: Men and women in the U.S. have a different number of siblings on average. (non-directional, two-tailed test, because we do not have a reason to expect either gender to have more siblings)

H0: $\mu 1 = \mu 2$
H1: $\mu 1 \neq \mu 2$

d. Two independent samples test of mean difference, because sibs is ratio and sex is nominal with 2 categories.
e. Using alpha level of .05, we fail to reject the null hypothesis. t=-1.184, p=0.237
f. Based on the available evidence from the GSS 2012 national sample, we cannot be 95% sure that women and men in the U.S. have different number of siblings on average.
g. Not applicable.
h. Let's calculate effect size: (3.73-3.57)/sqrt((3.07^2+3.09^2)/2)=.05

That is a tiny effect so it is not practically significant.

Output:

```
. codebook sex sibs
-------------------------------------------------------------------------------
sex                                                             RESPONDENTS SEX
-------------------------------------------------------------------------------
                  type:  numeric (byte)
                 label:  SEX

                 range:  [1,2]                        units:  1
         unique values:  2                        missing .:  0/1,974

            tabulation:  Freq.   Numeric  Label
                           886         1  male
                         1,088         2  female
-------------------------------------------------------------------------------
sibs                                             NUMBER OF BROTHERS AND SISTERS
-------------------------------------------------------------------------------
                  type:  numeric (byte)
                 label:  LABAD, but 23 nonmissing values are not labeled

                 range:  [0,30]                       units:  1
         unique values:  23                       missing .:  3/1,974

              examples:  1
                         2
                         3
                         6

. ttest sibs, by(sex)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    male |     886     3.56772    .1032029    3.071913    3.365169    3.770271
  female |    1085    3.732719    .0936743    3.085571    3.548915    3.916522
---------+--------------------------------------------------------------------
combined |    1971    3.658549    .0693701    3.079752    3.522502    3.794595
---------+--------------------------------------------------------------------
    diff |           -.1649988    .1394387               -.4384617    .1084641
------------------------------------------------------------------------------
    diff = mean(male) - mean(female)                              t =  -1.1833
Ho: diff = 0                                     degrees of freedom =     1969

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.1184         Pr(|T| > |t|) = 0.2368          Pr(T > t) = 0.8816
```

D. sex and arrest

a. sex: nominal; arrest: nominal.
b. Does gender affect the likelihood of being picked up or charged by police in the U.S. population?
c. Null: Gender does not affect the likelihood of being picked up or charged by police in the U.S. population.
   Research: Gender affects the likelihood of being picked up or charged by police in the U.S. population. (Non-directional because we don't have clear prior evidence one way or another.)
   H0: $O_i = E_i$
   H0: $O_i \neq E_i$
d. Chi-square, because both variables are nominal.
e. Using alpha level of .05, we reject the null hypothesis in favor of our research hypothesis. But we have even more confidence in our conclusion: χ^2=72.176, p<.001
f. Based on GSS 2012 national sample data, we are 99.9% confident that gender affects the likelihood of being picked up or charged by police among U.S. adults.

g. Men are much more likely to report an experience of being picked up or charged by police. Specifically, 30.5% of men but only 13.8% of women report such an experience. (All four cells have large residuals.)
h. That difference is practically significant – the percentage of those picked up or charged by police among men is more than double than that among women.

Output:

```
. codebook sex arrest
-------------------------------------------------------------------------------
sex                                                            RESPONDENTS SEX
-------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  SEX

                 range:  [1,2]                        units:  1
         unique values:  2                        missing .:  0/1,974

            tabulation:  Freq.   Numeric  Label
                           886         1  male
                         1,088         2  female
-------------------------------------------------------------------------------
arrest                                       EVER PICKED UP OR CHARGED BY POLICE
-------------------------------------------------------------------------------
                  type:  numeric (byte)
                 label:  LABH

                 range:  [1,2]                        units:  1
         unique values:  2                        missing .:  226/1,974

            tabulation:  Freq.   Numeric  Label
                           373         1  yes
                         1,375         2  no
                           226         .

. tab arrest sex, chi col
+--------------------+
| Key                |
|--------------------|
|     frequency      |
| column percentage  |
+--------------------+
      EVER |
 PICKED UP |
OR CHARGED |    RESPONDENTS SEX
 BY POLICE |      male     female |     Total
-----------+----------------------+----------
       yes |       241        132 |       373
           |     30.51      13.78 |     21.34
-----------+----------------------+----------
        no |       549        826 |     1,375
           |     69.49      86.22 |     78.66
-----------+----------------------+----------
     Total |       790        958 |     1,748
           |    100.00     100.00 |    100.00

          Pearson chi2(1) =  72.1757   Pr = 0.000

. tabchi arrest sex, adj
          observed frequency
          expected frequency
          adjusted residual
----------------------------
EVER      |
PICKED UP |
OR        |
```

```
CHARGED   | RESPONDENTS SEX
BY POLICE |    male   female
----------+----------------
      yes |     241      132
          | 168.576  204.424
          |   8.496   -8.496
          |
       no |     549      826
          | 621.424  753.576
          |  -8.496    8.496
---------------------------
          Pearson chi2(1) =  72.1757   Pr = 0.000
 likelihood-ratio chi2(1) =  72.3373   Pr = 0.000
```

Multiple Choice Questions

1. In hypothesis testing, if we increase the confidence level from 95% to 99% and the sample size stays the same:
   a. the probability of Type II error increases
   b. the probability of Type II decreases
   c. the probability of Type I error increases
   d. the probability of Type I error remains the same
   e. the probability of Type II error remains the same

2. A one-tailed test:
   a. is always performed when we don't know the direction of the difference between means
   b. is always performed when one sample mean is larger than another
   c. requires a larger t-value for the null to be rejected than a two-tailed test
   d. requires a smaller t-value for the null to be rejected than a two-tailed test
   e. requires one t-value while a two-tailed test requires two

3. "Within groups Sum of Squares" in ANOVA table represents:
   a. the variance that is due to the differences among groups
   b. the variance that is due to the differences among observations within each group
   c. the variance that can be explained by the grouping variable
   d. the variance of group means around the grand mean
   e. the variance of the test variable around the grouping variable

4. If the correlation coefficient between variables A and B is equal to .3:
   a. 9% of variance in A is explained by B
   b. 91% of variance in A is explained by B
   c. 3% of variance in A is explained by B
   d. 30% of variance in A is explained by B
   e. 70% of variance in A is explained by B

5. If the difference between two means is statistically significant:
   a. we need to assess separately whether it exists in the population
   b. we need to assess separately whether it exists in the sample
   c. we need to assess separately if it is larger than zero
   d. it can be small and not practically significant
   e. it is always practically significant

6. Expected frequencies in the chi-square test represent:
   a. the distribution that would exist if the two variables were perfectly correlated
   b. the distribution that we expect to observe in the data
   c. the distribution that would exist if the variables were independent of each other
   d. the distribution that would exist if the variables were dependent on each other
   e. the distribution that we would expect to see if we collected the data multiple times

7. To obtain the post-hoc comparisons with a Bonferroni correction after ANOVA by hand, we would have to:
   a. multiply our alpha level by p-value
   b. divide our alpha level by p-value
   c. divide our alpha level by the total number of comparisons
   d. multiply our alpha level by the total number of comparisons
   e. divide our alpha level by the number of groups

8. If Y is regressed on X and the slope of a regression line equals 4:
   a. when X increases by 4, Y will increase by 4
   b. when X decreases by 4, Y will decrease by 4
   c. when X increases by 1, Y will increase by 4
   d. when Y increases by 1, X will increase by 4
   e. when X increases by 1, Y will decrease by 4

9. Coefficient of determination:
   a. ranges from -1 to 1
   b. cannot be zero
   c. shows the proportion of unique variance in X
   d. cannot exceed 1
   e. is calculated by squaring the coefficient of alienation

10. The statistical power of a hypothesis test can be used to calculate:
   a. the confidence level
   b. the sample size
   c. the probability that we will erroneously reject the null hypothesis
   d. the probability that we are making Type I error
   e. the probability that we are making Type II error

11. For a chi-square test, if the critical value is higher than the test statistic, that means:
   a. our Type II error is zero
   b. we have established statistical significance and need to assess practical significance
   c. we reject the null hypothesis
   d. the p-value is larger than alpha
   e. we should do a two-tailed test

12. If a relationship between two interval/ratio variables, X and Y, is not linear, we can use:
   a. a lowess plot to describe that relationship
   b. a scatterplot with a regression line to describe that relationship
   c. correlation analysis to assess the strength of that relationship
   d. regression analysis to assess the strength of that relationship
   e. ANOVA test to assess the strength of that relationship

13. A study reported that the means of Y differ significantly depending on X (F=21.3, p<.001). Assuming that this study used a correct methodology, what are the levels of measurement of X and Y?

a. Both X and Y are interval/ratio variables
b. X is interval/ratio, Y is nominal or ordinal
c. Y is interval/ratio, X is nominal or ordinal
d. Both X and Y are nominal/ordinal
e. We cannot determine the type of variables based on this study report

14. A chi-square test of independence assesses:

a. whether the percentages across the columns are different in the sample
b. whether the differences in observed percentages across the columns are due to the chance
c. whether the differences in expected percentages across the columns are due to the chance
d. whether observed frequencies are larger than expected frequencies
e. whether expected frequencies are larger than observed frequencies

15. If Y is interval/ratio and X is nominal with 2 categories, we can use the following method to assess the relationship between them:

a. Chi-square test
b. Correlation coefficient
c. ANOVA
d. Independent samples t-test
e. Paired samples t-test

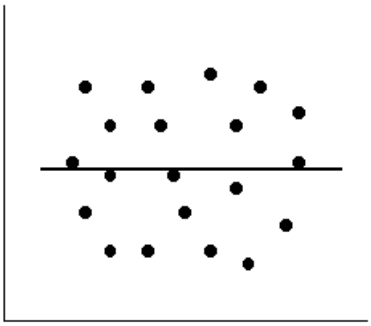
16. When X is the independent variable and Y is the dependent variable, interpolation:

a. is risky since we do not have the data for X in that range
b. is risky since we do not have the data for Y in that range
c. involves predicting a value of X for a value of Y within the range used to do regression
d. involves predicting a value of Y for a value of X within the range used to do regression
e. involves predicting a value of Y for a value of X outside the range used to do regression

17. The analytic technique that can best help us rule out spurious relationships is:

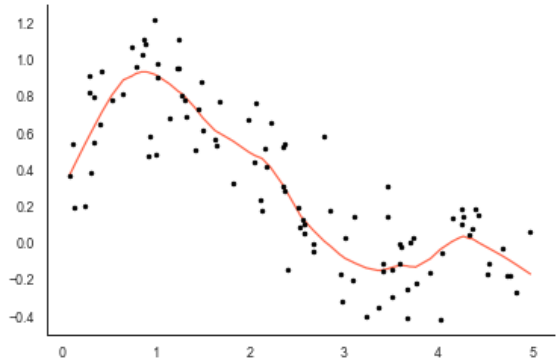
a. t-test for two means
b. ANOVA
c. Bonferroni correction
d. correlation
e. regression

18. The following plot suggests that the corresponding regression line would have a:



a. negative slope, b<0
b. positive slope, b>0
c. zero intercept, a= 0
d. zero slope, b=0
e. negative intercept, a<0

19. This plot suggests that the relationship between two variables:



a. is linear and best described with a Pearson's R
b. is not linear and best described with a Pearson's R
c. is linear and best described with OLS regression
d. is not linear and best described with OLS regression
e. is not linear and best described graphically

20. Bonferroni correction in post-hoc analyses helps us:
a. avoid inflated alpha
b. rule out spurious relationships
c. deal with skewed distributions
d. increase statistical power
e. prevent extrapolation