STAT 515 fa 2021 Final Exam

Karl B. Gregory

- Do not open this exam until told to do so.
- You may have 3 handwritten sheet of notes out during the exam.
- You have 2.5 hours to work on this exam.
- You may use a simple calculator.
- If you are unsure of what a question is asking for, do not hesitate to ask me for clarification.
- *Good luck, and may the odds be ever in your favor!*

Some upper quantiles of some chi-squared distributions:

α	0.10	0.05	0.025	0.01	0.005
$\chi^2_{1,\alpha}$	2.71	3.84	5.02	6.63	7.88
$\chi^2_{2,\alpha}$	4.61	5.99	7.38	9.21	10.60
$\chi^2_{3,\alpha}$	6.25	7.81	9.35	11.34	12.84
$\chi^2_{4,\alpha}$	7.78	9.49	11.14	13.28	14.86

$$S^2_{\text{pooled}} = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}, \quad T_{\text{test}} = \frac{\bar{X}_1 - \bar{X}_2}{S_{\text{pooled}}\sqrt{1/n_1+1/n_2}}, \quad \bar{X}_1 - \bar{X}_2 \pm t_{n_1+n_2-2} S_{\text{pooled}}\sqrt{1/n_1+1/n_2}$$

$$\text{For } X \sim \text{Binomial}(n,p) \quad \text{we have} \quad P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

1. In the USA, approximately 1 out of 700 babies is born with Trisomy 21 (Down syndrome). About 1 in 4 babies born with Trisomy 21 have a heart defect called atrioventricular septal defect. Across the entire population, roughly 1 in 1,860 babies are born with this heart defect. These figures are taken from the following sources:

   - https://www.cdc.gov/ncbddd/heartdefects/avsd.html
   - https://www.cdc.gov/ncbddd/birthdefects/downsyndrome.html
   - https://www.mottchildren.org/conditions-treatments/ped-heart/conditions/atrioventricular-septal-defect

   Consider a randomly selected newborn:

   (a) Given the presence of the atrioventricular septal defect, give the probability of Trisomy 21.

   Let T represent the event that the newborn has Trisomy 21 and let H represent the event that the newborn has the heart defect. We have

   $$P(T|H) = \frac{P(T \cap H)}{P(H)} = \frac{P(H|T)P(T)}{P(H)} = \frac{1/4(1/700)}{1/1860} = 0.6642857.$$

   (b) Given the absence of Trisomy 21, give the probability of the atrioventricular septal defect.

   We have

   $$P(H|T^c) = \frac{P(T^c|H)P(H)}{P(T^c)} = \frac{(1 - 0.6642857)(1/1860)}{799/800} = 0.0001807175,$$

   where we have used $P(T^c|H) = 1 - P(T|H) = 1 - 0.6642857$.

2. A survey of the marital and employment statuses of a sample of males aged 35–44 resulted in the table:

	Married	Widowed, divorced, or separated	Never married
Employed	638	133	102
Unemployed	27	8	6
Not in labor force	35	12	20

It is of interest whether there is an association between marital and employment statuses (See pg. 490 of [2]).

(a) Give the bottom row of the table of expected values under the null hypothesis of no association.

The entire table of expected counts is

	Married	Widowed, divorced, or separated	Never married	Total
Employed	622.94	136.16	113.91	873
Unemployed	29.26	6.39	5.35	41
Not in labor force	47.81	10.45	8.74	67
Total	700	153	128	981

(b) Which entry in the bottom row of the observed table will make the greatest contribution to the value of the test statistic $W_{\text{test}} = \sum_{j=1}^{J} \sum_{k=1}^{K} (O_{jk} - E_{jk})^2 / E_{jk}$?

The largest contribution to the test statistic comes from $(20 - 8.74)^2/8.74 = 14.50659$. A higher proportion of the never-marrieds are not in the labor force as compared with the other groups.

(c) Give the $\alpha = 0.05$ critical value for rejecting the null hypothesis (use table on front of exam).

The critical value is $\chi^2_{4,0.05} = 9.49$, where the degrees of freedom are 4 because the table has dimension 3×3.

(d) Determine whether or not to reject the null hypothesis at the $\alpha = 0.05$ significance level based on the test statistic $W_{\text{test}} = \sum_{j=1}^{J} \sum_{k=1}^{K} (O_{jk} - E_{jk})^2 / E_{jk}$.

We reject the null hypothesis of no association; we can determine this solely from the contribution to the value of the test statistic of the observed count in the third row and third column, which is $(20 - 8.74)^2/8.74 = 14.50659$. This value exceeds the $\alpha = 0.05$ critical value 9.49.

3. An experiment conducted in Nigeria compared the chlorophyll content (an indexed reading from -9.9 to 199.9 from a device clamped onto a leaf for two seconds) in the leaves of 145 maize seedlings, of which by random assignment 36, 27, 39, and 43 were treated with 0, 5, 10, and 20 grams, respectively, of an N-P-K (nitrogen-phosphorus-potassium) fertilizer [1]. The experiment resulted in the values $SS_{Total} = 21{,}815.2$ and $SS_{Treatment} = 11{,}373.06$.

(a) Write down the hypotheses of interest to the researchers.

Letting μ_0, μ_5, μ_{10}, and μ_{20} represent the mean chlorophyll content of the treatment groups

(b) Give the values (i)–(vi) that are missing from the ANOVA table:

Source	Sum of Sq	df	Mean Sq	F	p-value
Treatment	11373.06	(i)	(ii)	(vi)	
Error	(iii)	(iv)	(v)		
Total	21815.2	144			

i. 3

ii. $11373.06/3 = 3791.02$

iii. $21815.2 - 11373.06 = 10442.14$

iv. $145 - 4 = 141$

v. $10442.14/141 = 74.058$

vi. $3791.02/74.058 = 51.19$

(c) The table below gives some upper quantiles of the F-disribution relevant to this experiment:

α	0.10	0.05	0.025	0.01	0.005
$F_{\text{(i)},\text{(iv)},\alpha}$	2.12	2.67	3.21	3.92	4.46

State whether you think the null hypothesis should be rejected and explain why.

The value of the F-statistic is 51.19, which by far exceeds even the largest of these critical values. Therefore the p-value is very small—much smaller than 0.005. These are strong grounds for rejecting H_0 and claiming that the amount of fertilizer makes a difference in the chlorophyll content of the leaves.

(d) If you had access to the entire data set:

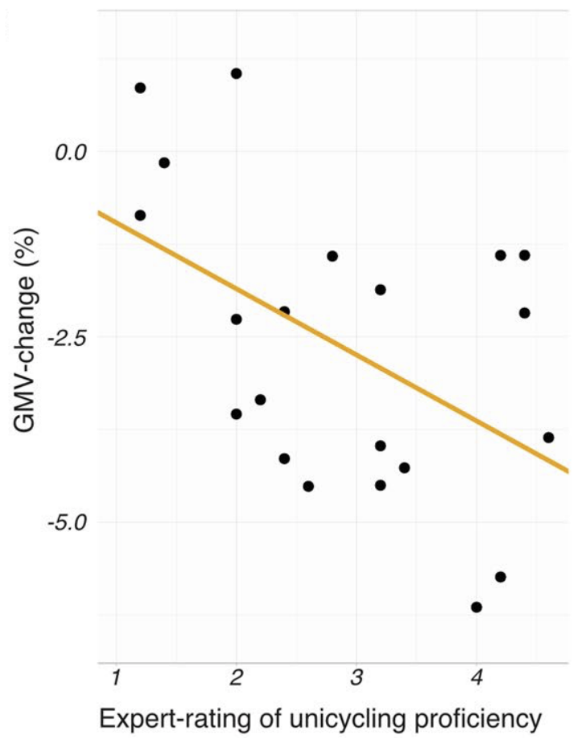
i. How could you check the assumption of equal variances among the treatment groups?

We would plot the residuals against the fitted values (which are the observed treatment means) and check whether the vertical spread of the points is equal across the treatment groups. Alternatively, we could make boxplots of the response values in each treatment group and plot these side by side, inspecting them to see if the spread of the responses is the roughly same in each treatment group. Since there are a large number of experimental units in each treatment group, nice boxplots could be made for each treatment group (which would not be the case if the treatment groups were small), so this would be a nice way to check the assumption of equal variances.

ii. How could you check the assumption that the responses are Normally distributed around the treatment means?

We would obtain the residuals (the observed response values minus the corresponding treatment means) and make a Normal quantile-quantile plot of them. We would check to see if the points fall close to a straight line, which would suggest Normality.

4. In a study of neuroplasticity—the ability of the brain to reorganize itself in response to environmental demands—several female subjects with no prior unicycling experience were trained by semi-professional unicyclists for three weeks, before and after which their brains were imaged [3]. The figure below plots the percent changes in gray matter volume at the right superior temporal gyrus (a region of the brain) from before to after the training against expert ratings of post-training unicycling proficiency for the subjects in the study.



The authors reported a p-value equal to 0.030 for the significance of the relationship between percent change in gray matter volume and the expert rating of unicycling proficiency.

(a) Write down a simple linear regression model for the relationship between the percent change in gray matter volume and the expert rating of post-training unicycling proficiency. Define your variables and your parameters.

> Letting Y_i be the percent change in gray matter volume and x_i be the expert-rating of post-training unicycling proficiency of subject i, for $i = 1, \dots, 21$, we assume the model
>
> $$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \text{for } i = 1, \dots, 21,$$
>
> where β_0 is the intercept parameter and β_1 is the slope parameter and $\varepsilon_1, \dots, \varepsilon_{21} \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma^2)$ are error terms.

(b) Do your best to ascertain (approximately) the values of the least-squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of the regression coefficients by studying the plot.

From the plot, it looks like $\hat{\beta}_1 \approx -1.1$ and $\hat{\beta}_0 \approx 0.1$

(c) What difference in the percent change of gray matter volume is associated with an increase of 1 point in a subject's unicycling proficiency rating?

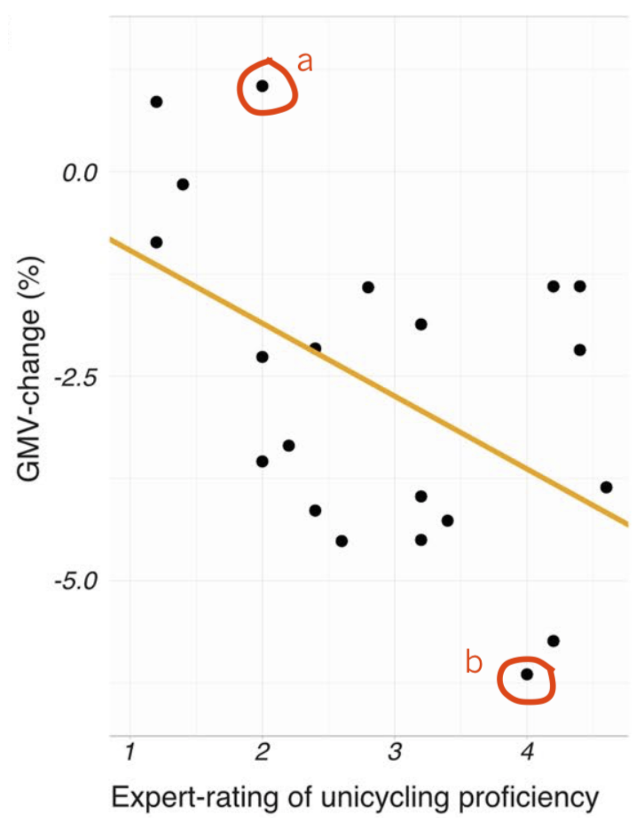
This is the interpretation of the value of the slope parameter. It looks to be something like -1.1 from the plot.

(d) Write down the null and alternate hypotheses to which the reported p-value of 0.030 corresponds.

The authors are testing

$$H_0\colon\ \beta_1 = 0 \text{ versus } H_1\colon\ \beta_1 \neq 0.$$

(e) Identify two points on the scatter plot (label them) which have a strong influence on the least-squares regression line. Explain why you selected these two points (there is more than one correct choice of two points).



The points circled and labeled a and b exert a strong influence over the least-squares line because they are both (i) far from the line and (ii) somewhat far from the mean of the x values. If either point were removed from the data set, the line would probably change noticeably. Other points could be chosen and discussed.

(f) The authors reported one of the following values for the correlation between the percent change in gray matter volume and the expert rating of post-training unicycling proficiency. Which one is it?

A. $r_{xY} = -0.030$

B. $r_{xY} = -1.101$

C. $r_{xY} = -0.876$

D. $r_{xY} = -0.473$

E. $r_{xY} = 0.473$

F. $r_{xY} = 1.101$

G. $r_{xY} = 0.030$

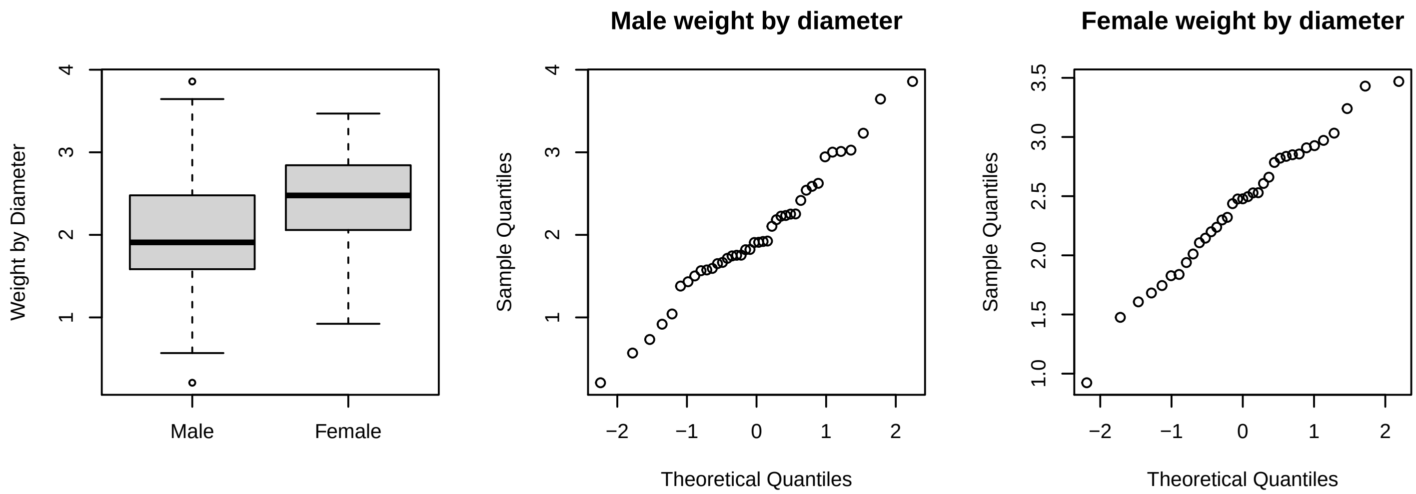
(g) What percent change of gray matter volume would you expect for subjects with an expert rating of unicycling proficiency equal to 6?

> To use the fitted linear regression line to answer this question would be extrapolation, since a rating of 6 falls outside the range of observed data. In fact, experts were instructed to assign ratings between 1 and 5.

5. From large populations of male and female abalones, 40 males and 35 females were sampled. The whole weight divided by the diameter of each of these was recorded. The means and standard deviations of the weight-by-diameter values for the two samples were

$$\bar{X}_M = 2.006, \quad \bar{X}_F = 2.420, \quad S_M = 0.787, \quad \text{and} \quad S_F = 0.572,$$

where M and F denote male and female, respectively. Below are boxplots and Normal quantile-quantile plots of the samples.



Suppose it if of interest whether there is a difference in the weight-by-diameter ratio of male and female abalones.

(a) Write down the hypotheses of interest (male minus female).

We are interested in testing

$$H_0\colon\ \mu_M - \mu_F = 0 \text{ versus } H_1\colon\ \mu_M - \mu_F \neq 0.$$

(b) Give the value of the test statistic for testing the hypotheses of interest; assume equal variances.

First we compute

$$S^2_{\text{pooled}} = \frac{(40-1)(0.787)^2 + (35-1)(0.572)^2}{40+35-2} = 0.483.$$

Then the test statistic is given by

$$T_{\text{test}} = \frac{2.006 - 2.420}{\sqrt{0.483}\sqrt{1/40 + 1/35}} = -2.571.$$

(c) Use the t-table to determine a range within which the p-value lies.

The p-value is twice the area under the probability density function of the t_{73} distribution to the left of -2.571, where 73 comes from $40+35-2$. Though the t-table does not show quantiles of the t_{73} distribution, we can see that the value 2.571 falls between the upper 0.010 and 0.005 quantiles. Therefore the p-value must lie in the interval $(0.01, 0.02)$.

(d) Give a 95% confidence interval for the true difference in weight-by-diameter ratio means. If you cannot find the row of the t-table you need, just use the row above it.

We have

$$2.006 - 2.420 \pm \underbrace{t_{73,0.025}}_{\approx 2.0003} \sqrt{0.483}\sqrt{1/40 + 1/35} = (-0.735, -0.092).$$

(e) State whether you should conclude, based on these data, that the weight-by-diameter means differ.

Since the p-value is quite small (smaller than 0.05), we would reject the null hypothesis of equal means and conclude that the means differ; moreover, we would conclude that the mean weight-by-diameter ratio of females is greater than that of males. We see this also in that our 95% confidence interval for the difference in means contains only negative values.

6. You are hosting 8 friends for an afternoon gathering and you are wondering how many cups of coffee to brew. For each guest, the probability that he or she will want coffee is 3/5. Find the probability of the events (evaluate your expressions):

(a) All 8 guests want coffee.

If X is the number of guests who want coffee, then $X \sim \text{Binomial}(8, 3/5)$.

$$P(X = 8) = (3/5)^8 = 0.01679616.$$

(b) At least one guest wants coffee.

We have

$$P(X \geq 1) = 1 - P(X = 0) = 1 - (1 - 3/5)^8 = 0.9993446.$$

(c) Exactly 6 guests want coffee.

We have

$$P(X = 6) = \binom{8}{6}(3/5)^6(1 - 3/5)^2 = \frac{8 \cdot 7}{2}(3/5)^6(1 - 3/5)^2 = 0.2090189$$

(d) No more than 6 guests want coffee.

We have

$$\begin{aligned} P(X \le 6) &= 1 - P(X \ge 7) \\ &= 1 - (P(X=7) + P(X=8)) \\ &= 1 - \binom{8}{7}(3/5)^7(1-3/5)^1 - \binom{8}{8}(3/5)^8(1-3/5)^0 \\ &= 1 - 8(3/5)^7(1-3/5) - (3/5)^8 \\ &= 0.8936243. \end{aligned}$$

References

[1] Hussaini Abubakar, Haruna Danyaya Abubakar, and Aminu Salisu. One way anova: Concepts and application in agricultural system.

[2] David Freedman, Robert L. Pisani, Roger Purves, and Ani Adhikari. *Statistics, Second Edition*. W.W. Norton & Company, Inc., 1991.

[3] Ilona Papousek, Bernhard Weber, Karl Koschutnig, Andreas Schwerdtfeger, Christian Rominger, Elisabeth M Weiss, Markus Tilp, and Andreas Fink. Learning unicycling evokes manifold changes in gray and white matter networks related to motor and cognitive functions. *Scientific Reports*, 9(1):4324–4324.