Solutions to Statistics 302 Midterm 3

Larget, Spring 2014

1. **5 parts, 6 points each, 30 points total.** A group of middle school students in Michigan (grades 4 to 6) were asked whether grades, sports, or popularity was most important to them. The results are shown in the following table.

	Grades	Sports	Popularity
Boy	117	60	50
Girl	130	30	91

(a) The table shows the proportion of boys that select each of grades, sports, and popularity as most important. Complete the table for girls by writing proportions rounded to three digits.

Solution: Divide 130, 30, and 91 by their sum 251 to find values for the second row of numbers.

	Grades	Sports	Popularity
Boy	0.515	0.264	0.220
Girl	0.518	0.120	0.363

Note that the sample proportions of boys and girls that select grades are very close, but that boys are more likely to select sports and girls popularity where the differences look to be potentially large enough to be considered statistically significant.

(b) Treat the students as a representative sample of all middle school students in Michigan. Write null and alternative hypotheses about the proportions of middle school boys and girls in Michigan whom would select each of these three items as most important to them when asked to select one of the three.

Solution: The null hypothesis is that there is no difference between the sexes in population proportions of students that select each of these three items as most important. The alternative is at least one difference between boys and girls.

More formally, if the population proportions for boys and girls that select each of the three items are denoted in the obvious way, the null hypothesis is

$$H_0: \quad p_{\text{boy,grades}} = p_{\text{girl,grades}} \quad \text{and} \quad p_{\text{boy,sports}} = p_{\text{girl,sports}} \quad \text{and} \quad p_{\text{boy,popularity}} = p_{\text{girl,popularity}}$$

versus the alternative that at least one of these equalities does not hold.

(c) Assuming that sex is independent of the proportion in the population that would select each item, find the expected count for each cell of the table. Write expected counts in the table (rounded to one decimal place).

Solution: Each expected count is the row sum times column sum divided by the table sum. The logic is that if independent, the probability of being in a given cell is the row probability times the column probability, so that the expected count is

$$(\text{sample size}) \times (\text{row probability}) \times (\text{column probability})$$

Here, the sample size is $n = 478$ and each row and column probability is estimated (by maximum likelihood) to be its sum over n.

$$\text{Expected count} = n \times \left(\frac{\text{row sum}}{n}\right) \times \left(\frac{\text{column sum}}{n}\right) = \frac{(\text{row sum}) \times (\text{column sum})}{n}$$

	Grades	Sports	Popularity
Boy	117.3	42.7	67.0
Girl	129.7	47.3	74.0

(d) Calculate a test statistic and fill in the blanks or circle the correct option to describe how to find a p-value for the hypothesis test for this problem.

Solution: The test statistic is

$$\begin{aligned} X^2 &= \sum_i \sum_j \frac{((\text{observed}_{ij}) - (\text{expected}_{ij}))^2}{(\text{expected}_{ij})} \\ &= \frac{(117 - 117.3)^2}{117.3} + \cdots + \frac{(91 - 74.0)^2}{74.0} = 21.5 \end{aligned}$$

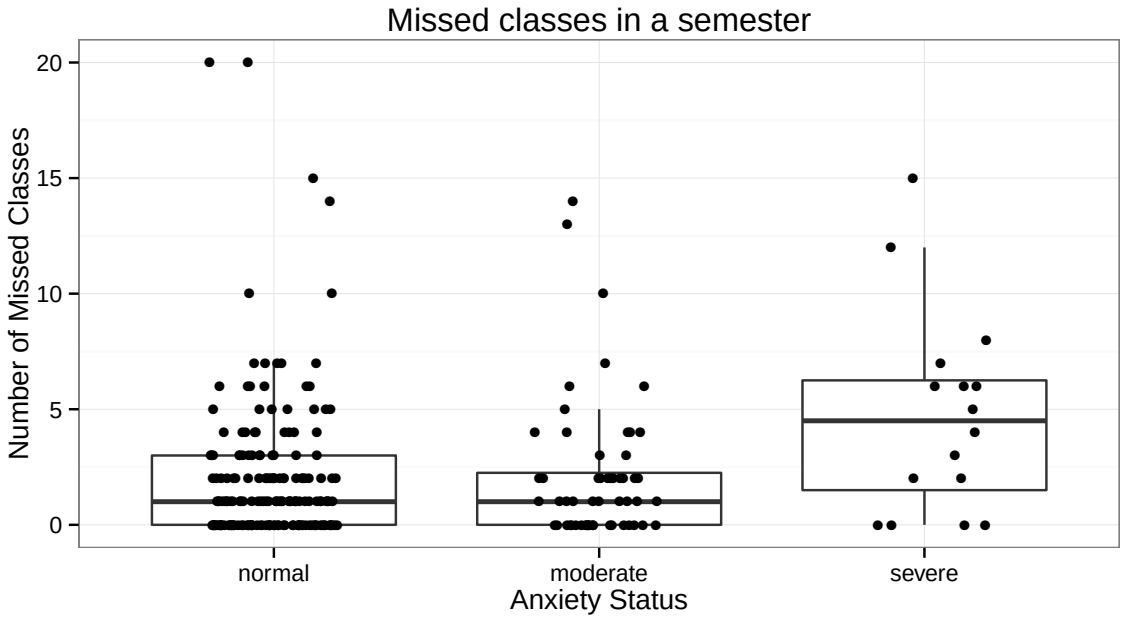
The p-value is the area to the RIGHT of 21.5 under the density curve of a χ^2 distribution with 2 degrees of freedom.

There are 2 rows and 3 columns in the table: $\text{df} = (2-1)(3-1) = 2$.

(e) Assume that the p-value is 0.00002. Interpret the results of the hypthesis test in context.

Solution: There is very strong evidence that middle school boys and girls in Michigan have different proportions who select what is most important from grades, sports, and popularity.

2. **5 parts, 7 points each, 35 points total.** Researchers conducted a study that examined sleep habits of college students and academic performance. Two variables in the study were **anxiety**, which was scored categorically as one of *normal*, *moderate*, or *severe*, and **classes missed** which counted the number of classes that the student missed during the semester. The study includes a sample of 253 subjects whom you may consider to be a representative sample from a population of college students. Raw data is plotted below (points jittered horizontally to lessen overlap).



Summary statistics of the number of classes missed are tabulated here.

Anxiety Status	n	$\bar{x}$	s
Normal	181	2.01	3.11
Moderate	56	2.13	3.05
Severe	16	4.75	4.37

The following data may also be useful:

- The grand mean of all observations is $\bar{y} = 2.21$.
- The sum of squares of individual **classes missed** observations around the grand mean is $\sum_{i=1}^{n}(y_i - \bar{y})^2 = 2645.9$ so that the standard deviation of **classes missed** is $\sqrt{2645.9/252} = 3.24$.
- Ignoring round-off errors in sample standard deviations, the following expression sums the squared deviations of each observation from its sample mean.

$$180 \times (3.11)^2 + 55 \times (3.05)^2 + 15 \times (4.37)^2 \doteq 2535.1$$

(a) Fill in all missing values in this ANOVA table.

Solution:

Source	df	SS	MS	F	p-value
Groups	2	110.8	55.4	5.46	0.0048
Error	250	2535.1	10.14		
Total	252	2645.9			

There are three groups ($k = 3$) and $n = 253$ total observations, so there are $k-1 = 2$ degrees of freedom for groups, $n-k = 250$ degrees of freedom for error, and $n-1 = 252$ degrees of freedom in total.

The total sum of squared deviations around the grand mean is 2645.9 and the error sum of squares is 2535.1: this is also the weighted sum of sample standard deviations, weighted by the degrees of freedom.

The difference is the sum of squares for groups, $2645.9 - 2535.1 = 110.8$. Each sum of squares is divided by the corresponding degrees of freedom to find the mean square. The F statistic is the ratio of mean squares.

(b) State null and alternative hypotheses tested in the ANOVA table. Assuming a valid test, interpret the results of the test in context.

Solution: The null hypothesis is that the population mean number of classes missed is equal in all three populations of individuals determined by anxiety level. In symbols,

$$H_0 : \mu_{\text{normal}} = \mu_{\text{moderate}} = \mu_{\text{severe}}$$

versus the alternative that there is at least one mean different from another.

In context, noting the small p-value of 0.0048:

> There is strong evidence that the mean number of classes missed differs among the populations defined by anxiety levels ($p = 0.00484$, F-test from one-way ANOVA with 2 and 250 df).

(c)

Solution: The pooled estimate of the common population standard deviation is 3.18.

Note that the value 3.25 from the previous page is the standard deviation of the combined sample and ignores separate groups. This is not the pooled estimate of the standard deviation, as it will be inflated if the null hypothesis of common means is false. In general, the estimate of the common sd is the square root of the MSE, $\sqrt{10.14} \doteq 3.18$.

(d) Fill in the missing parts for a 99% confidence interval for the difference in population mean number of classes missed between the severe and normal anxiety groups.

Solution:

$$\underline{2.74} \pm t^* s_{\mathsf{e}} \sqrt{\frac{1}{\boxed{181}} + \frac{1}{\boxed{16}}}$$

where t^* is the $\underline{0.995}$ quantile from a t distribution with $\underline{250}$ degrees of freedom and

s_{e} has numerical value $\underline{3.18}$.

(e) ANOVA assumes independent random samples from each group. (1) What additional *two assumptions* does ANOVA make about the distribution of observations within a group? (2) What about the plot of the data indicates *the most serious evidence* against one of these two assumptions?

Solution: The additional *assumptions* are that the populations are normally distributed and that the population standard deviations are equal to one another. These are stated in the textbook on page 498.

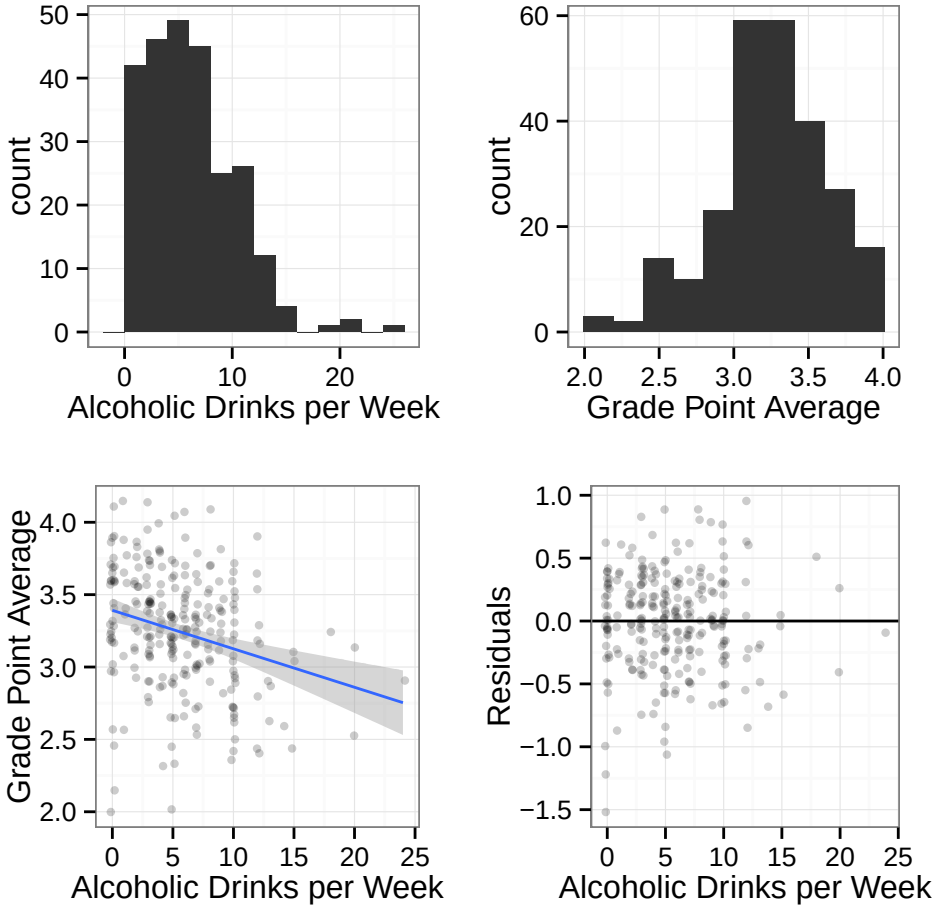
The sample standard deviations are not too different from each other, but the plot shows substantial right skewness in two of the samples, which has the potential to affect reliability.

In addition, the textbook provides criteria for when ANOVA is reliable, which states that either samples are large (using 30 as a cutoff) or approximately normal and that variability is similar. They give a rough rule of thumb that the largest sample sd is not more than twice the smallest sample sd.

There is no fixed sample size that guarantees that the sampling distribution of the F statistic will closely follow the F distribution, regardless the nonnormality in the population (especially skewness). So, I dislike invoking a cutoff like $n \geq 30$ to indicate that a sample is large enough that normality no longer matters. However, large samples greatly lessen the effects of nonnormality in the F test. Examining the robustness of the F test to deviations from normality and equal variance can be tested through simulation.

3. **5 parts, 7 points each, 35 points total.**

The same study as in the previous problem on a sample of 253 college students included the variables **Drinks**, which is the number of alcoholic drinks per week the student consumes and **GPA**, which is the grade point average of the student. A regression analysis seeks to explain how the **Drinks** might explain **GPA**.



(a) Based on the graphs above, *circle the value* of each summary statistic where x is **Drinks**, y is **GPA**, and r is the correlation coefficient between these variables.

Solution:

$\bar{x}$	[5.57]	11.57	13.57	15.57
s_x	0.10	0.60	1.10	[4.10]
$\bar{y}$	2.044	2.644	[3.244]	3.844
s_y	[0.404]	1.404	2.404	3.404
r	-0.84	[-0.27]	0.36	0.79

From the choices given, 5.57 must be the mean number of drinks per week. Each of the other values is so far in the right tail of the observed values that it cannot be the balancing point. Similarly, 3.244 must be the mean of GPA. This distribution has only small skew left, and the mean must be near the middle peak between 3.0 and 3.5. Once the mean is determined, the mean plus or minus one standard deviation should contain more than the majority, but not close to all of the values. For **Drinks**, the

three lowest values are much too small as only a small fraction of individual values would be within an sd of the mean. For **GPA**, the three largest values are much too large as the mean plus or minus an sd would contain all or almost of the data. The correlation coefficient r is negative, but fairly weak, because the points are not tightly clustered around the line.

(b) Find the slope and intercept of the regression line.

Solution: Using the important equations from lecture, $\hat{\beta}_1 = r\frac{s_y}{s_x}$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$. Filling in numbers, the slope is $-0.27 \times 0.404/4.10 \doteq -0.0266$ and the intercept is $3.244 - (-0.0266)(5.57) \doteq 3.39$.

Another way to determine these values is from the table below. Each t value is computed as the estimate over the standard error, so the slope is estimated as -4.43×0.006001 and the intercept is estimated as 81.82×0.041455.

You can also empirically guess from the graphed line.

(c) Here is partial computer output of the fitted regression line.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)              0.041455   81.82  < 2e-16 ***
Drinks                   0.006001   -4.43 1.41e-05 ***
```

Find a 95% confidence interval for the slope of the regression line. You may round t^* to the nearest integer for your calculation.

Solution: Note that the t^* for the confidence interval is the value such that 95% of the area under a t density is between $-t^*$ and t^*. As a t distribution with 251 degrees of freedom is so close to a standard normal distribution, this quantile will be 2, rounded to the nearest integer. Hence,

$$\begin{array}{rcl} -0.0266 & \pm & 2(0.006001) \\ -0.039 & \text{to} & -0.015 \end{array}$$

(d) Call the endpoints of the confidence interval you found in the previous part a and b. Circle the best interpretation.

Solution:

(1) We are 95% confident that the mean GPA in the population of college students that consume a given number of alcoholic drinks per week changes by an amount between a and b points for each additional drink consumed per week.

(2) 95% of all college students consume between a and b drinks per week.

(3) 95% of all college students would see their GPA change by an amount between a and b points if they consumed one more drink per week.

(4) We are 95% confident that drinking alcohol causes a change in GPA from between a and b points on average.

We are confident in a rate of change between the population mean GPA and number of alcoholic drinks per week. The confidence interval is not about a proportion of individuals or the probability that there is a change.

(e) The band around the fitted regression line is a 95% confidence interval. If a 95% prediction interval had been graphed, would it be wider or narrower? Briefly defend your response.

Solution: The prediction interval will be wider. The formula for the interval contains an extra $+1$, so it must be wider. Conceptually, it is wider because it incorporates both uncertainty in the location of the regression line and the estimate of the individual variation associated with a single new observation. A confidence interval is only affected by uncertainty in the location of the regression line.