Stat21 HW2

Dohyun Lee

Feb 17, 2020

1. Q-Q Plots (7 points)

(a) The following are quantile-quantile plots of GRE General Test Verbal scores for students intending graduate study in psychology, classics, and economics. Briefly describe the pattern in each of the Q-Q plots. (One sentence for each plot should suffice.)

Classics students score higher than psychology students in the middle percentile, but score the same at the lowest and highest percentiles

Economics students generally score higher than psychology students at the higher percentiles, but score lower in the lower percentiles. They score about the same at the lowest and highest percentiles.

(b) Suppose we have the SAT verbal and quantitative scores for all current Swarthmore students. How do the following two plots differ: (1) a Q-Q plot of the verbal and quantitative score distributions, and (2) a scatterplot plotting the verbal and quantitative scores for each student? What question is answered by each plot?

(1) A Q-Q plot for the SAT verbal and quantitative scores would show how both groups compare at different percentiles. This would evaluate the distributions of both groups.

(2) A scatterplot would show the relationship between the verbal and quantitative scores. It would show what a student's verbal/quantitative score would be given the other score. This would show us the variability of the two test scores

2. Anova Table (9 points)

The following is part of an ANOVA table from an ANOVA with 50 observations total and 2 groups.

(a) Fill in the seven blanks in the table above.

```
anova_table <- data.frame("Source" = c("Between","Within","Total"),
                          "degrees of freedom" = c(1,48,49),
                          "Sum of Squares" = c(7274.2,5188.8,12463),
                          "Mean Square" = c(7274.2,108.1," "),
                          "F-ratio" = c(67.3," "," "))
anova_table
```

```
##    Source degrees.of.freedom Sum.of.Squares Mean.Square F.ratio
## 1 Between                  1         7274.2      7274.2    67.3
## 2  Within                 48         5188.8       108.1
## 3   Total                 49        12463.0
```

(b) Calculate R^2 for these data.

$R^2 = \frac{SSB}{SST} = \frac{7274.2}{12463}$

```
7274.2/12463
```

```
## [1] 0.5836636
```

$R^2 = 0.58$

3. Elephant Ivory (15 points)

Elephants have declined substantially in population, with losses estimated at 50–75% over the last half-century. A major reason for this decline is the killing of elephants for their ivory (another is habitat loss). Conservation officials would like to be able to trace the source of ivory in order to determine if it was legally or illegally obtained, and to determine where poaching might be taking place. However, this is difficult to do solely by visual inspection of ivory samples. A dataset on stable isotope ratios of five elements was collected on ivory samples from Asia, West Africa, Central Africa, East Africa, and Southern Africa. The goal is to see if these isotope ratios differ in ivory from different regions; if so, then this information may be useful in locating the origin of the ivory. The dataset contains information on the country and region that 495 samples of ivory were obtained from; each sample's ratios of carbon, nitrogen, oxygen, hydrogen, and sulfur; and the latitude and longitude where the ivory was obtained. Source: S. Ziegler, S. Merker, B. Streit, M. Boner, D.E. Jacob (2016): Towards understanding isotope variability in elephant ivory to establish isotopic profiling and source-area estimation. Biological Conservation 196, pp. 154-163.

(a) Download the dataset ivory.csv from Moodle. Then read the dataset into R using the following command:

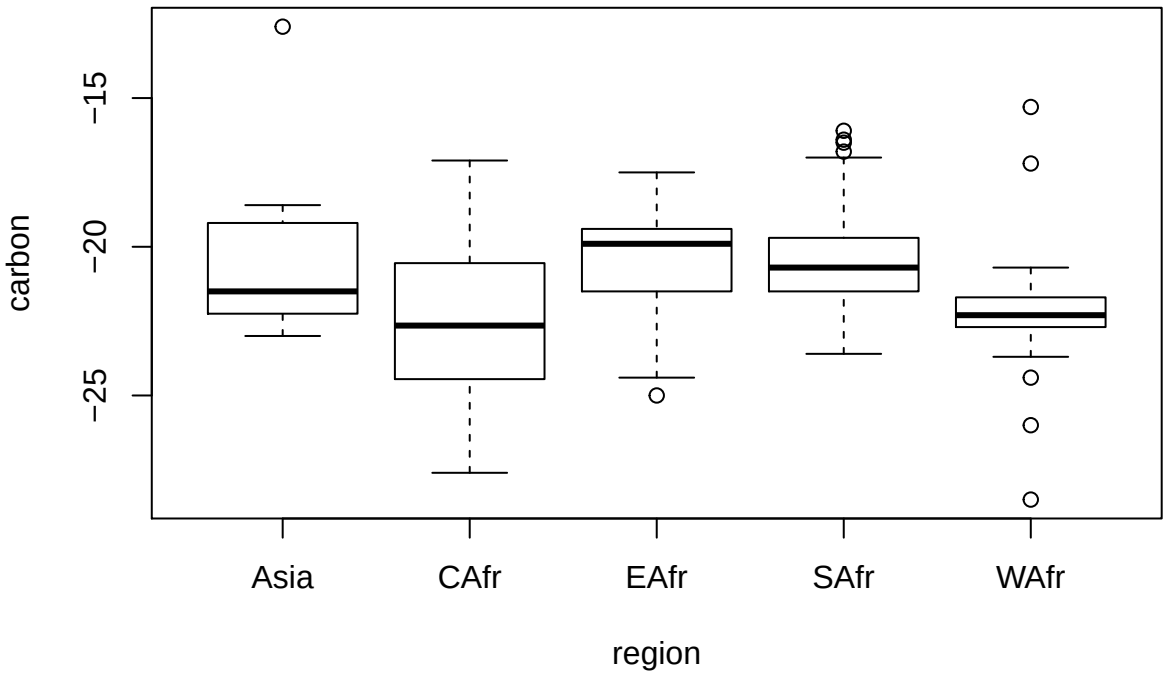
```
data <- read.csv("~/Desktop/school stuff/year2/spring20/stat21/ivory.csv", header=T)
```

(You may have to change the path to the file, depending on where you saved the downloaded file.) Verify that your dataset has 495 rows and 9 columns. In this assignment, we will look only at the region of origin and carbon-13 isotope ratio. Define variables for region and origin as follows:

```
region <- data[,"Region"]
carbon <- data[,"delta13C"]
```

(b) We want to explore whether carbon ratios differ by region. Make boxplots of carbon by region using the following command:

```
boxplot(carbon ~ region)
```



Does ivory from different regions appear to vary in its carbon ratio? Do the assumptions of normality and equal variances for ANOVA appear to be violated?

Ivory carbon ratio does seem to vary in different regions. Just from looking at the boxplots, it's not clear whether they're normal. They do not meet equal variance because the spreads of each boxplot are not the same

(c) Fit an ANOVA model to the data using the following command:

```
fit1 <- lm(carbon ~ region)
anova(fit1)
```

```
## Analysis of Variance Table
##
## Response: carbon
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## region      4  413.65 103.412  31.651 < 2.2e-16 ***
## Residuals 490 1600.93   3.267
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Is there a significant difference in carbon ratios by region?

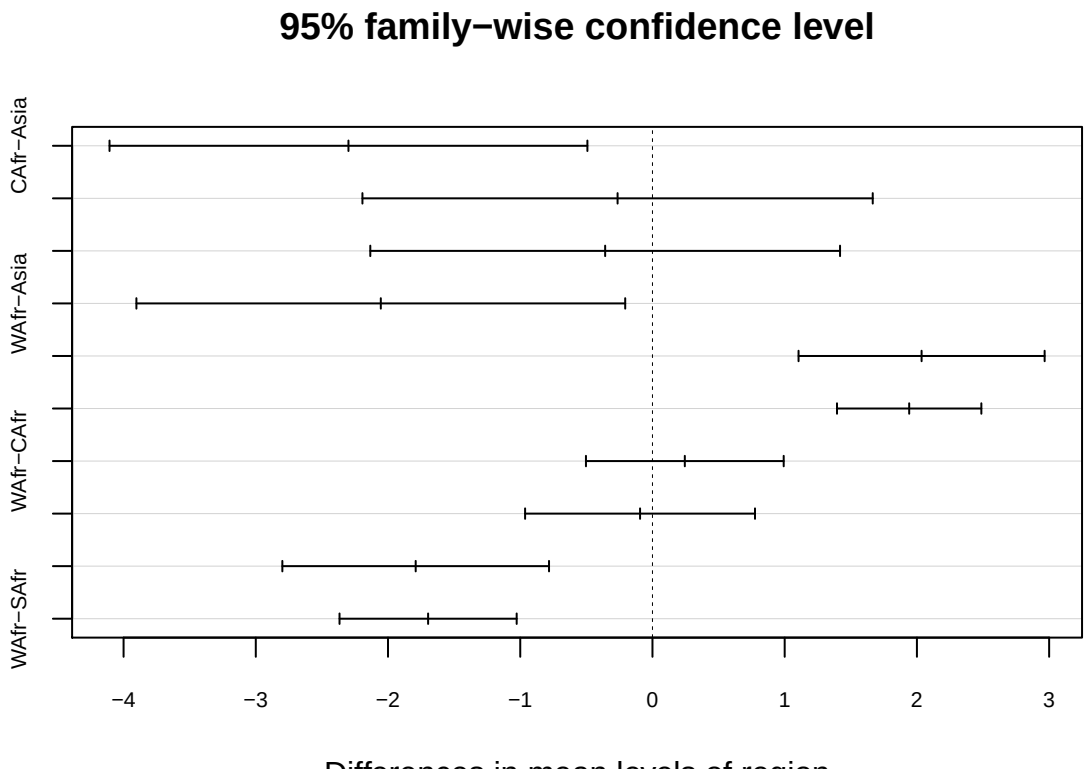
Our p-value 2e-16 is less than 0.05, so we reject H_0. So, there is a difference between the means of the carbon ratios by region.

(d) If there is a significant different among regions, which pairs of regions are significantly different? Use the following commands to carry out a post-hoc test of means between pairs of regions:

```
posthoc <- TukeyHSD(aov(fit1))
posthoc
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = fit1)
##
## $region
##                  diff        lwr        upr     p adj
## CAfr-Asia -2.29916667 -4.1062811 -0.4920522 0.0048776
## EAfr-Asia -0.26385135 -2.1934931  1.6657904 0.9958149
## SAfr-Asia -0.35771073 -2.1340552  1.4186337 0.9817489
## WAfr-Asia -2.05452899 -3.9029126 -0.2061454 0.0207243
## EAfr-CAfr  2.03531532  1.1046905  2.9659401 0.0000000
## SAfr-CAfr  1.94145594  1.3956127  2.4872991 0.0000000
## WAfr-CAfr  0.24463768 -0.5030703  0.9923457 0.8984088
## SAfr-EAfr -0.09385938 -0.9632267  0.7755079 0.9983329
## WAfr-EAfr -1.79067763 -2.7991031 -0.7822521 0.0000155
## WAfr-SAfr -1.69681826 -2.3667468 -1.0268898 0.0000000
```

```
plot(posthoc, cex.axis = .7)
```



You may have to re-size the plot window to see more of the y-axis labels, or try adding the option cex.axis=.7 to the plot(posthoc) command. Which regions are significantly different?

The regions that are significantly different are Central Africa-Asia, West Africa-Asia, East Africa-Central Africa, South Africa-Central Africa, West Africa-East Africa, West Africa-South Africa. There are significant differences in each of these regions because their confidence intervals don't include 0 in them, which would imply that there is no difference.

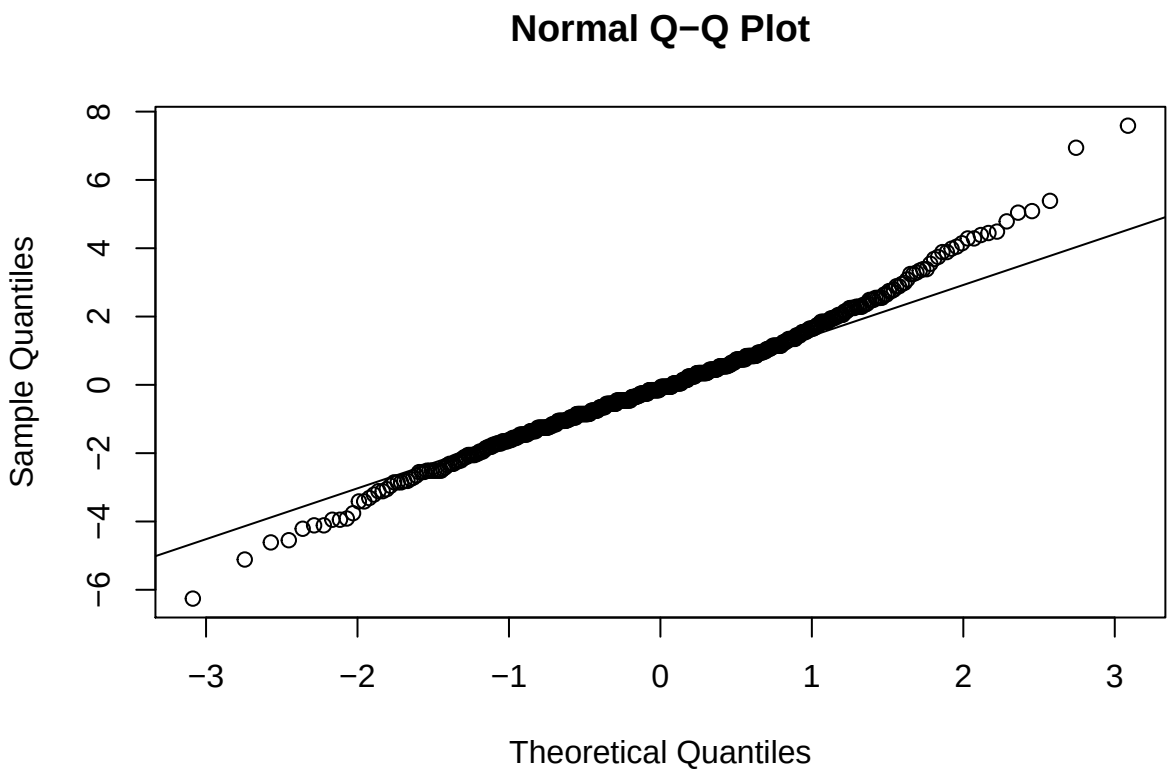
(e) In part (b) we assessed the assumption of normality by eyeballing the symmetry of boxplots. Now let's use an NPP to accomplish the same goal. A strength of the NPP is that it can reveal subtle departures

from normality, so perhaps we will be able to catch something that is not easily seen from the boxplots alone.

Rather than make separate NPPs for each region, it is typical to examine an NPP of the residuals of the entire dataset. Recall from your introductory stat course that a residual is defined as (actual value – predicted value) for an observation. Here the predicted value of an ivory sample is just the mean of the region corresponding to that sample. Subtracting off the mean does not change the distribution of the ivory samples; it merely shifts the distribution to the left or the right. Thus, if the data in each region are normally distributed, then the residuals will be as well. Furthermore, we can aggregate the regions together because if each region is normally distributed with the same variance, then combining the regions will result in a normal distribution with that variance.

To extract the residuals from an lm model object, use the residuals() command. Then make an NPP of the residuals and add the reference line. Do the residuals appear normally distributed?

```
ivory_resid <- residuals(fit1)
qqnorm(ivory_resid)
qqline(ivory_resid)
```



The residuals aren't normal because the tails of the plot are high at the high end and low at the low end and don't align closely with the line.

(f) Calculate R-squared. Would you say this is a relatively high value, or relatively low?

$R^2 = \frac{SSB}{SST} = \frac{414}{414+1601} =$

```
414/(414+1601)
```

```
## [1] 0.2054591
```

$R^2 = 0.21$

This is a relatively low value

(g) Based on (f) and on your boxplots in (b), would you say that an ivory sample's carbon ratio can be used to determine its region of origin? How is this question related to the difference between statistical significance and effect size that we explored in HW 1?

An ivory sample's carbon ratio shouldn't be used to determine the region of origin because of the low R^2 value, which tells us that there big differences between the observed and fitted values. Since the p-value is so low, we can be sure that there is a statistically significant difference between the means. However, this doesn't mean that statistical significance directly leads to a large effect size, which is the magnitude of the difference in means.

(h) Identify a simple way in which this analysis might be modified to improve its predictive power.

We could conduct tests with the other variables like oxygen, nitrogen, and/or hydrogen ratios. We could add a transformation to the data, like squaring or logging the response variable to improve the model and its predictive power. ## 4. F and R^2 (6 points)

In a simple linear regression, the F-statistic (the F-ratio) and R2 both measure the strength of the relationship between Y and X. It should come as no surprise that the two quantities are related. For the case of k = 2 groups, demonstrate this relationship by showing that:

$R^2 = \frac{F}{F+(N-k)}$

(Hint: This is a straightforward proof and can be done in about five steps. Start with the right-hand side of the equation and substitute the definition of F, then simplify.)

$F = \frac{MSB}{MSW} = \frac{\frac{SSB}{n-1}}{\frac{SSW}{N-k}} = \frac{SSB}{n-1} \cdot \frac{N-k}{SSW}$

$\frac{\frac{SSB(N-k)}{SSW(k-1)}}{\frac{SSB(N-k)}{SSW(k-1)}+(N-k)}$

$= \frac{\frac{SSB(N-k)}{SSW(k-1)}}{\frac{SSB(N-k)}{SSW(k-1)} + \frac{SSW(k-1)(N-k)}{SSW(k-1)}}$

$= \frac{SSB(N-k)}{SSW(k-1)} \cdot \frac{SSW(k-1)}{SSB(N-k)+SSW(k-1)(N-k)}$

$= \frac{SSB(N-k)}{SSB(N-k)+SSW(k-1)(N-k)}$

$= \frac{SSB(N-k)}{(N-k)(SSB+SSW(k-1))}$

$= \frac{SSB}{SSB+SSW(k-1)}$

$k = 2$, so:

$\frac{SSB}{SSB+SSW}$

$= \frac{SSB}{SST} = R^2$

5. Is Science Broken? (9 points)

Read the following article: https://fivethirtyeight.com/features/science-isnt-broken/ and watch this John Oliver video: https://youtu.be/0Rnq1NpHdmw (warning: NSFW) If the youtube link doesn't work, try https://www.facebook.com/LastWeekTonight/videos/896755337120143/ or https://www.vox.com/2016/5/9/11638808/john-oliver-science-studies-last-week-tonight

(a) Briefly define p-hacking and researcher degrees of freedom.

P-hacking is a form of bias and data manipulation that researchers implement by using many degrees of freedom to produce the results that they want, which creates a lot of false positives and "statistically significant" results. Researcher degrees of freedom are the decisions that researchers make when they conduct a study, which include which observations to record, which factors to control, and which observations and factors to compare.

(b) Why should you not believe a finding from any single scientific study? When should you consider a finding to be reliable?

You should not believe any single scientific study because there is a good chance that the data from a scientific study has been manipulated to show us results that aren't actually statistically significant. If the same one study is reconducted multiple times using legitimate methods, then you could consider it reliable.

(c) What are replication studies? Why are they rare?

Replication studies are studies that are re-tested multiple times on some study to ensure the findings are legitimate and valid. They are rare because publishers prefer novelty in their journals. Publishers value replication studies less than original research, so that creates less incentive for researchers to do them.