Homework 1 - Stat 495

Example Solution

Due Friday, Sept. 15th by midnight

PROBLEMS TO TURN IN: Vis 1, Vis 2, CASI 1.1, CASI 1.4, Portfolio Reflection

Visualization Problems (Adapted from an assignment by Prof. Horton)

Vis 1 - Compelling

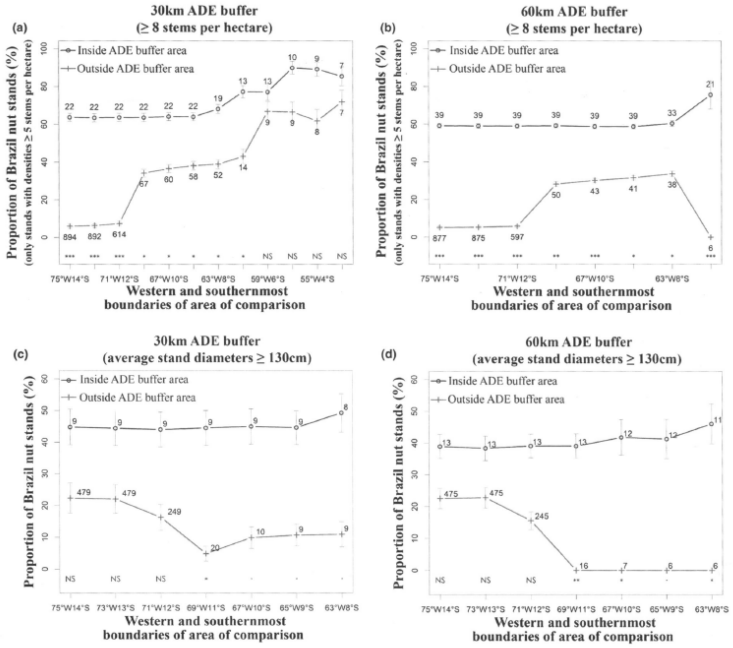


Figure 4 Proportions of Brazil nut stands with densities of 8 trees ha^{-1} or higher (only stands with ≥5 trees ha^{-1}) (a and b) and average diameters of 130 cm or higher (c and d) inside and outside of 30 and 60 km Amazonian Dark Earth (ADE) buffer areas, respectively, along a gradient from south-western to north-eastern Amazonia. Numbers associated with points are sample sizes. Significance levels are given above the *x*-axis (NS, not significant; ·$P < 0.10$; * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$; bootstrap comparisons). Error bars represent standard error the mean.

Figure 1: Figure 4 from Thomas et al. 2015

This article was found using JSTOR.

Citation: Uncovering spatial patterns in the natural and human history of Brazil nut (Bertholletiaexcelsa) across the Amazon Basin. Evert Thomas, Carolina Alcázar Caicedo, Crystal H. McMichael, Ronald Corvera and Judy Loo. Journal of Biogeography, Vol. 42, No. 8 (August 2015), pp. 1367-1382.

Commentary: I felt this visual was compelling because it was able to incorporate a lot of information without being overwhelming. Each scatterplot represents the relationships within a different setting. Then, within each scatterplot, not only do we have the observed percentages and standard errors, but also the sample sizes

and the results of a test comparing the behavior inside and outside the buffer areas. And this is all displayed relatively cleanly and concisely.

Vis 2 - Suboptimal

This article was found using Google Scholar.

Citation: Valsecchi, A., Irurita Olivares, J., & Mesejo, P. (2019). Age estimation in forensic anthropology: methodological considerations about the validation studies of prediction models. International Journal of Legal Medicine, 133(6), 1915-1924.

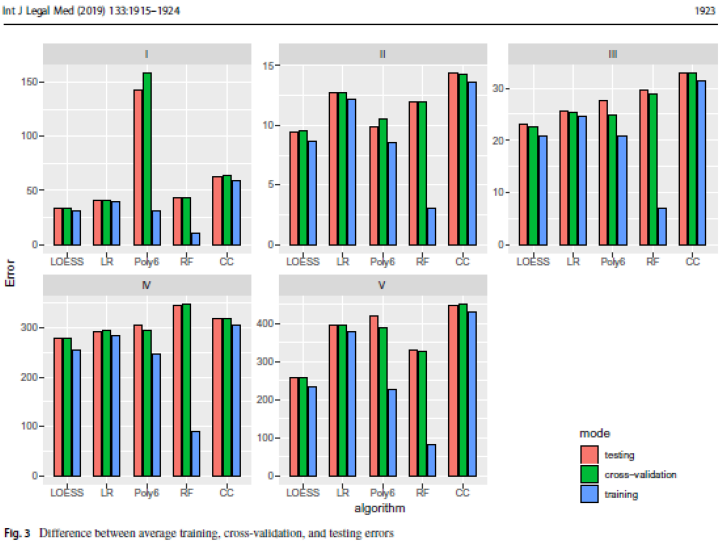


Figure 2: Figure 3 from Valsecchi et al. 2019

Commentary: The concept behind this plot is actually great. You want to compare these different error rates, and it's a major way to compare the methods. However, bar charts are not great for this, and also, the y-axis scale is different for every plot. We also have no sense of error (i.e. spread). I would have personally preferred a table with means and sds from simulations or something similar. Perhaps a way can be found to standardize each graph as well, so the y-axes are made to be equal. This is also not a good color scheme for someone who is color-blind.

CASI 1.1

This problem was chosen to help you remember how to perform regressions and obtain predictions. There are multiple ways to code the solution.

```
kidney <- read.table("http://web.stanford.edu/~hastie/CASI_files/DATA/kidney.txt", header = TRUE)
```

(a) Fit a cubic regression, as a function of age, to the kidney data of Figures 1.1 and 1.2, calculating estimates and standard errors at ages 20, 30, 40, 50, 60, 70, 80.

SOLUTION:

```
# This is one way to fit the model. There are others.
# Be sure you see all three terms in the cubic showing up.
mymod <- lm(tot ~ age + I(age^2) + I(age^3), data = kidney)
age <- c(20, 30, 40, 50, 60, 70, 80)
newdata <- data.frame(age)
kidneycubic <- augment(mymod, kidney, newdata, se_fit = TRUE)
kidneycubic
```

```
## # A tibble: 7 x 3
##     age .fitted .se.fit
##   <dbl>   <dbl>   <dbl>
## 1    20   1.25    0.372
## 2    30   0.509   0.195
## 3    40  -0.243   0.259
## 4    50  -1.01    0.278
## 5    60  -1.82    0.335
## 6    70  -2.66    0.389
## 7    80  -3.56    0.668
```

We fit the cubic regression and use the augment function on a newdata set containing just the age values asked about. The resulting output shows the age along with the predicted response (.fitted) and corresponding standard error.

You can also use the *poly* function to do this. You can see below that it gives identical results.

```
mymod2 <- lm(tot ~ poly(age, 3), data = kidney)
kidneycubicpoly <- augment(mymod2, kidney, newdata, se_fit = TRUE)
kidneycubicpoly
```

```
## # A tibble: 7 x 3
##     age .fitted .se.fit
##   <dbl>   <dbl>   <dbl>
## 1    20   1.25    0.372
## 2    30   0.509   0.195
## 3    40  -0.243   0.259
## 4    50  -1.01    0.278
## 5    60  -1.82    0.335
## 6    70  -2.66    0.389
## 7    80  -3.56    0.668
```

(b) How do the results compare with those in Table 1.1?

SOLUTION:

We look at the fitted values and then the SEs.

```
LRfit <- c(1.29, 0.50, -0.28, -1.07, -1.86, -2.64, -3.43)
Lowessfit <- c(1.66, 0.65, -0.59, -1.27, -1.91, -2.68, -3.50)
```

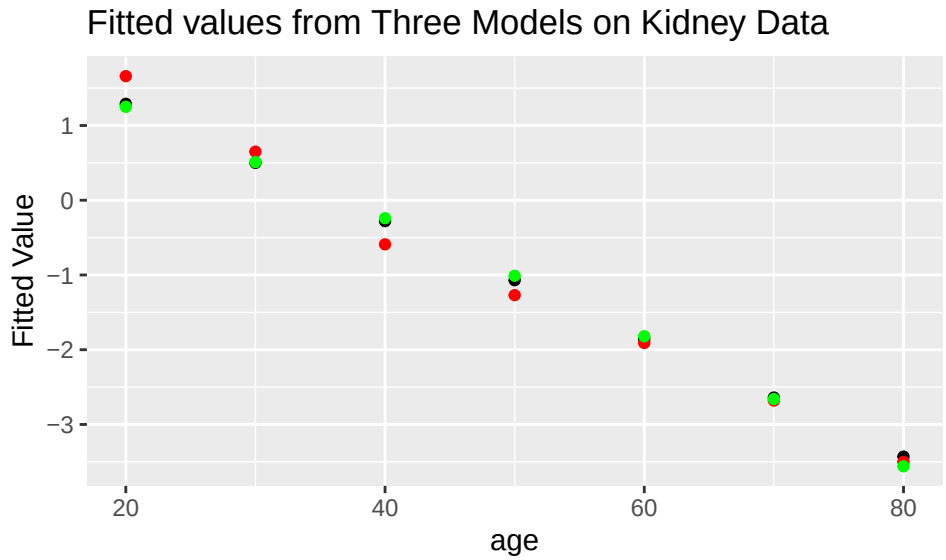
Table 1: Comparisons of Fitted Values from Models on Kidney Data

age	LRfit	Lowessfit	Cubicfit
20	1.29	1.66	1.25
30	0.50	0.65	0.51
40	-0.28	-0.59	-0.24
50	-1.07	-1.27	-1.01
60	-1.86	-1.91	-1.82
70	-2.64	-2.68	-2.66
80	-3.43	-3.50	-3.56

```
Cubicfit <- round(kidneycubic$.fitted, 2)
fitdata <- data.frame(age, LRfit, Lowessfit, Cubicfit)
kable(fitdata, caption = "Comparisons of Fitted Values from Models on Kidney Data")
```

We have pulled the fitted values from the linear and lowess regressions from the textbook. Here, we can see that the cubic fit usually has predictions that are closer to 0 than the other two methods, though this is not true at the higher ages. The cubic fitted values are not always in between the other two fitted values or in any pattern like that, though they do tend to be closer to the fitted values from the linear regression than the lowess fit (age 80 is the exception).

```
gf_point(LRfit ~ age, data = fitdata) %>%
  gf_point(Lowessfit ~ age, data = fitdata, color = "red") %>%
  gf_point(Cubicfit ~ age, data = fitdata, color = "green") %>%
  gf_labs(title = "Fitted values from Three Models on Kidney Data", y = "Fitted Value")
```



Next we examine the SEs.

```
LRse <- c(0.21, 0.15, 0.15, 0.19, 0.26, 0.34, 0.42)
Lowessse <- c(0.71, 0.23, 0.31, 0.32, 0.37, 0.47, 0.70)
Cubicse <- round(kidneycubic$.se.fit, 2)
sedata <- data.frame(age, LRse, Lowessse, Cubicse)
kable(sedata, caption = "Comparisons of SEs from Models on Kidney Data")
```

We have pulled the values of the SEs from the linear and lowess regressions from the textbook. We can

Table 2: Comparisons of SEs from Models on Kidney Data

age	LRse	Lowessse	Cubicse
20	0.21	0.71	0.37
30	0.15	0.23	0.19
40	0.15	0.31	0.26
50	0.19	0.32	0.28
60	0.26	0.37	0.33
70	0.34	0.47	0.39
80	0.42	0.70	0.67

see here that in each case, the cubic standard error for the fitted value is in between the linear and lowess standard errors. A few times, it is closer to the linear SE, and a few times it is closer to the lowess SE. It appears to be a compromise between the two methods presented in the text. You could make a plot again here, but the pattern seems fairly clear from the table.

CASI 1.4 - Slightly Modified

This problem was chosen to help you remember the concepts of the bootstrap and permutation/randomization tests. You may have seen these concepts only briefly before. Both are extremely valuable concepts to have knowledge of in your statistical understanding.

```
# Load and format data
leukemia_big <- read.csv("http://web.stanford.edu/~hastie/CASI_files/DATA/leukemia_big.csv")
# says pictures from row 136
gene136 <- t(leukemia_big[136, ]) #says pictures from row 136
# Need to get ALL and AML tags in
type <- c(rep("ALL", 20), rep("AML", 14), rep("ALL", 27), rep("AML", 11))

# Set up dataset
leukemia <- data.frame(gene136, type)
leukemia <- rename(leukemia, gene136 = X136)
favstats(~ gene136 | type, data = leukemia)
```

```
##   type      min       Q1   median       Q3     max     mean       sd  n missing
## 1  ALL 0.210578 0.560344 0.732703 0.855127 1.63400 0.752479 0.275342 47       0
## 2  AML 0.324703 0.771426 0.967796 1.096250 1.42548 0.949973 0.243019 25       0
```

We want to see if there is a significant difference in mean gene expression for gene 136 for the ALL and AML groups.

(a) Record the means of the ALL and AML groups for the gene 136 data available for reference.

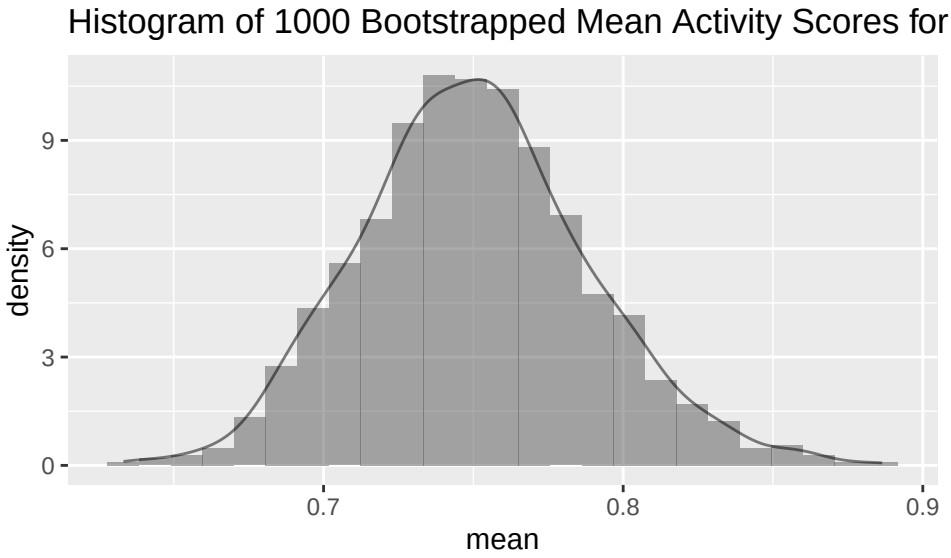
SOLUTION:

The mean for ALL is 0.7525 and the mean for AML is 0.9500.

(b) Perform 1000 nonparametric bootstrap replications for the mean of ALL for gene 136. Describe the distribution of the resulting means. You can perform the bootstrap in any way you see fit (the functions do and resample might prove useful).

SOLUTION:

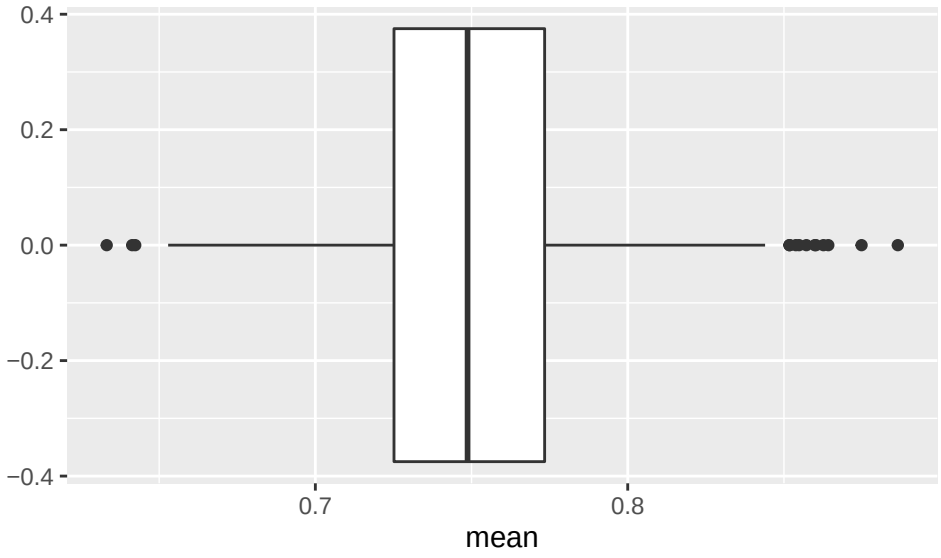
```
set.seed(495)
ALLmeans <- do(1000)*mean(~ gene136, data = resample(filter(leukemia, type == "ALL"), replace = TRUE))
gf_dhistogram(~ mean, data = ALLmeans) %>%
  gf_dens() %>%
  gf_labs(title = "Histogram of 1000 Bootstrapped Mean Activity Scores for ALL")
```




```
favstats(~ mean, data = ALLmeans)
```

```
##       min       Q1   median       Q3      max     mean        sd    n missing
##  0.633203 0.725177 0.748724 0.773395 0.886461 0.750237 0.0382822 1000       0
```

```
gf_boxplot(~ mean, data = ALLmeans)
```



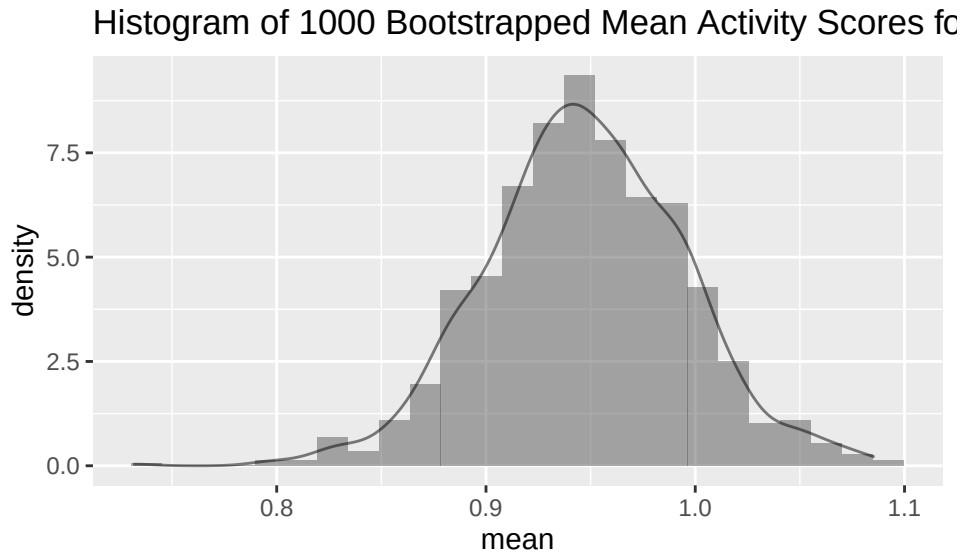
The distribution of means for ALL seems fairly bell-shaped and unimodal, with a mean of 0.75, and a standard deviation of 0.038. The mean is as expected, since that was the sample mean. There are some outliers on both sides of the distribution, more in the upper-tail with my seed. (Different seeds may get different results with that).

(c) Repeat (b) for AML.

SOLUTION:

```
set.seed(495)
AMLmeans <- do(1000)*mean(~ gene136, data = resample(filter(leukemia, type == "AML"), replace = TRUE))
gf_dhistogram(~ mean, data = AMLmeans) %>%
```

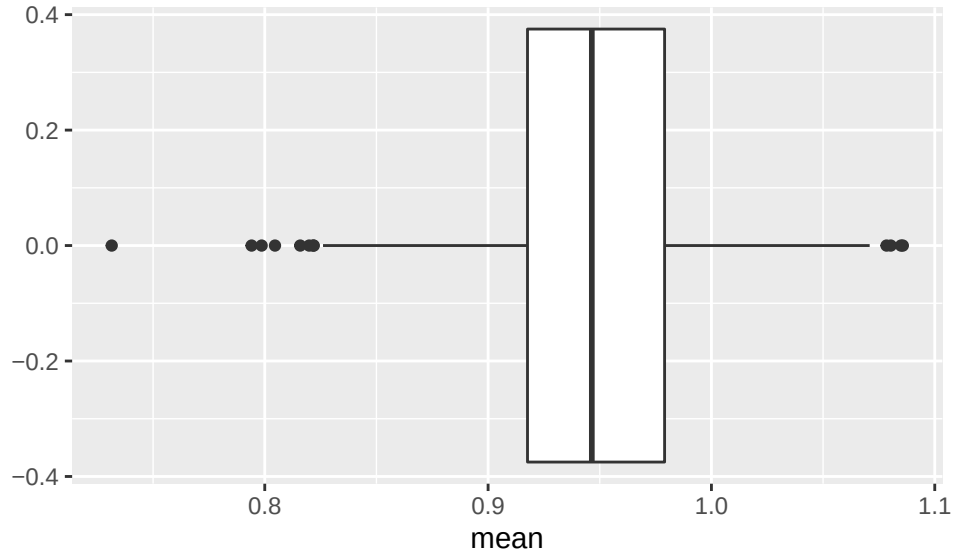
```
  gf_dens() %>%
  gf_labs(title = "Histogram of 1000 Bootstrapped Mean Activity Scores for AML")
```




```
favstats(~ mean, data = AMLmeans)
```

```
##       min      Q1   median       Q3     max    mean       sd    n missing
##  0.731458 0.91765 0.946449 0.979033 1.08571 0.94699 0.047604 1000       0
```

```
gf_boxplot(~ mean, data = AMLmeans)
```

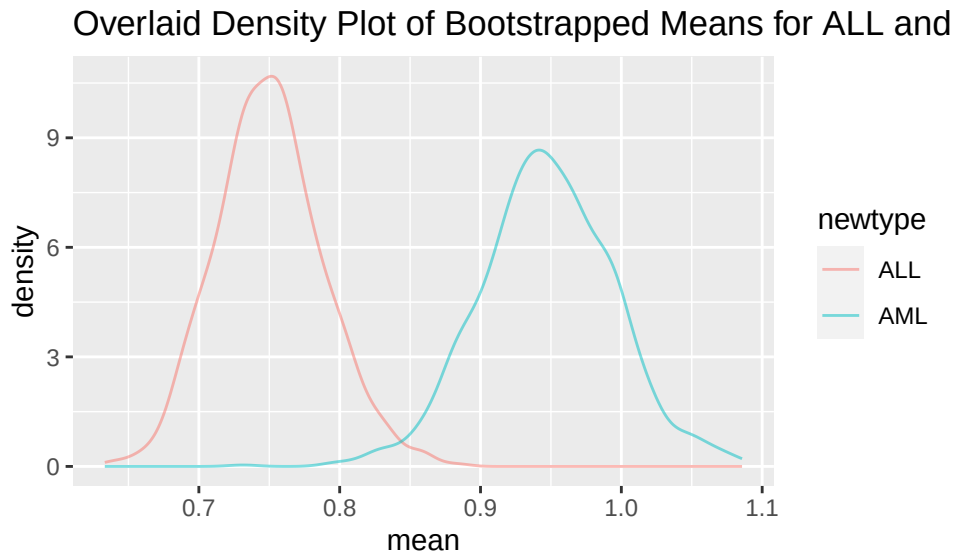


This distribution of means for AML is also fairly bell-shaped with a clear single central peak. The mean is 0.947 for my seed, very similar to the sample mean, and the standard deviation is 0.0476. There are outliers on both sides of the distribution, but one more extreme on the low end for my chosen seed.

(d) Suggest an inference. In other words, what do your results in (b) and (c) suggest about whether there is a difference in means for the ALL and AML groups for gene 136?

SOLUTION:

```
mydata <- rbind(ALLmeans, AMLmeans)
newtype <- c(rep("ALL", 1000), rep("AML", 1000))
combdata <- data.frame(mydata, newtype)
gf_dens(~ mean, data = combdata, color = ~ newtype) %>%
  gf_labs(title = "Overlaid Density Plot of Bootstrapped Means for ALL and AML")
```




```
qdata(~ mean, c(0.025, 0.975), data = ALLmeans)
```

```
##     2.5%    97.5%
## 0.680755 0.830720
```

```
qdata(~ mean, c(0.025, 0.975), data = AMLmeans)
```

```
##     2.5%    97.5%
## 0.853861 1.044959
```

The density plot shows the AML peak is shifted to the right of the ALL peak, and there is very little overlap in the distributions. The estimated 95 percent CIs for each mean are (0.68, 0.83) for ALL and (0.85, 1.04) for AML. These do not overlap. This suggests that there is a difference in means here. Ideally, we'd want to now combine the procedures, and look at differences in mean values in some way.

(e) Brainstorm an alternative way to approach the problem via a randomization/permutation test. Describe what you would do in a way that someone else could code it up. (You do not need to actually code this up, but you can if you want to see what the result is.)

SOLUTION:

Basically, we would compute the t statistic for the difference in means on the original data, and save it. Then, we would randomly permute the type labels (45 ALL and 27 AML), recompute the t statistic, and save it. You would repeat this process of permuting the labels and saving the t statistic many times. (In R, the *shuffle* function can be used for this.) Then, you would look at the distribution of t-statistics obtaining by randomly permuting the labels - that behaves like the distribution of t-statistics you would obtain if the null hypothesis of no difference in means was true. You then compare your observed t-statistic from the original data to that empirically generated null distribution. If the original observed statistic is in the tails, you have evidence of a significant difference. You can obtain an empirical p-value.

Portfolio Reflection

Look at our portfolio review and in-class activities. In a separate word or pdf document, in a few paragraphs, reflect on how the items in your portfolio demonstrate:

- how your statistical analytical skills have developed over time
- how your statistical writing skills have developed over time
- skills you have a solid grasp of (such as R code or visuals or regression)
- skills you would like to improve on

Then, set some goals for what you'd like to work on improving in future statistical reports/work. (Yes, you brainstormed some before, this is asking you to pick some to really focus on!)

Upload this portfolio reflection and goals document for future reports to your portfolio folder in your personal class repo.

Given what you are asked to include above, I expect the document you generate to have at least 3 paragraphs and contain at least 3 goals for future work.