STAT 217, Fall 2016 Sep 9 Name:

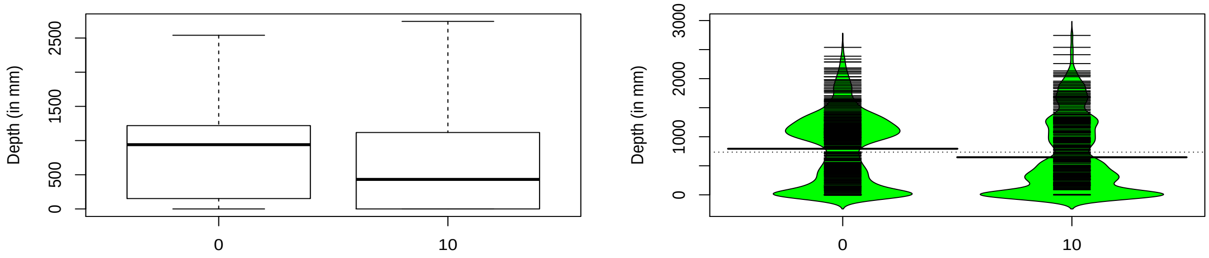
Activity for 2 independent sample means:

Description of snow1 dataset

Karl Wetlaufer, Jordy Hendrikx, and Lucy Marshall wrote a paper for a journal on snow density (the full article is on D2L) and different areas of the West Fork of the Gallatin River Basin in southwest Montana. we will focus on two of the variables the team collected data on: snow depth (`depth`) and forest cover (`cover`). The snow depth was measured in mm and the areas either forested (a 10 was recorded) or unforested (a 0 was recorded). We want to investigate if this data suggest some association between forest cover and snow depth, on average. If there is an association, we would like to know by how much snow depth the areas might differ on average.

Explore the data

```
# TODO: load in the data and call it snow (on D2L it is snow1)

# look at the summary statistics
require(mosaic)
favstats(depth ~ cover, data = snow)

##   cover min  Q1 median   Q3  max     mean       sd   n missing
## 1     0   0 152    940 1219 2540 793.5228 603.4008 614       0
## 2    10   0   0    432 1118 2743 646.8362 636.3934 403       0


# look at the distributions
par(mfrow = c(1, 2))
boxplot(depth ~ cover, data = snow, ylab = "Depth (in mm)")
require(beanplot)
beanplot(depth ~ cover, data = snow, method = "jitter", col = "green",
    ylab = "Depth (in mm)")
```



1. Does it appear from the summary measures or from the beanplots that there may be an association between forest cover and snow depth, on average? Explain.

Hypotheses for the test

2. What is the null hypothesis of the test?

(a) Using proper notation (making sure to define all parameters used):

(b) In words:

(c) Write out the null model making sure to define all parameters used:

3. What is the alternative hypothesis of the test?

(a) Using proper notation (making sure to define all parameters used):

(b) In words:

(c) Write out the null model making sure to define all parameters used:

Assess the validity conditions

4. **Independence of observations:** The 'observations' were the locations that snow depth was measured. The West Fork Basin was split into sampling areas to make the data collection process more feasible. The areas were selected to ensure all of the Basin could be reasonably accounted for. As such, the sampling areas themselves were not random (although they may be representative of the Basin itself). However, within each area a random location was chosen to obtain the snow depth measurements. As a result, there is weak evidence to suggest a clustering of observations or that the observations are related in any way.

5. **Equal variances in the groups:** We see from the beanplots above that the spreads, or variability, of the distributions of snow depth are similar for each group providing little evidence against this assumption.

6.

(a) **Normal distributions of the observations in each group (parametric):** The distributions are not symmetric as the distribution of snow depth for the forested locations is right-skewed and the distribution of snow depth for the unforested locations is bimodal. There are really large sample sizes with 614 snow depth measurements at unforested locations and 403 snow depth measurements at forested locations. With these large sample sizes, due to the Central Limit Theorem, this assumption becomes less problematic.

(b) **Similar distributions of the observations in each group (nonparametric):** The distributions are not similar as the distribution of snow depth for the forested locations is right-skewed and the distribution of snow depth for the unforested locations is bimodal (and left-skewed based on how the mean snow depth compares to the median snow depth).

Find the value of the appropriate test statistic

7. Use `t.test()` to calculate the test statistic. (This is the test statistic for both the parametric and nonparametric approaches.)

```
# calculate the test statistic
Tobs <- t.test(depth ~ cover, data = snow, var.equal = T)$statistic
Tobs
```

```
##        t
## 3.710287
```

(a) What is the value of the test statistic and what type of statistic is it (t-statistic, z-statistic, F-statistic, etc.)?

(b) What was the order of subtraction in order to get the test statistic?

Find the p-value

8. Use `t.test()` to calculate the p-value using a parametric approach

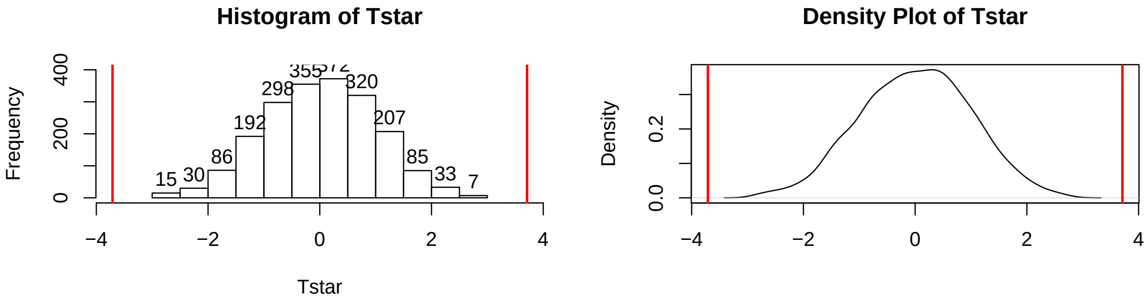
```
# parametric t test
t.test(depth ~ cover, data = snow, var.equal = T, conf.level = 0.95)

##
##  Two Sample t-test
##
## data:  depth by cover
## t = 3.7103, df = 1015, p-value = 0.0002183
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   69.10668 224.26647
## sample estimates:
##  mean in group 0 mean in group 10
##         793.5228         646.8362
```

(a) Under the null hypothesis, what is the sampling distribution of the test statisic?

(b) What is the value of the p-value? Interpret this value in the context of the study.

9. Use permutations to **build** the null distribution (a nonparametric approach):

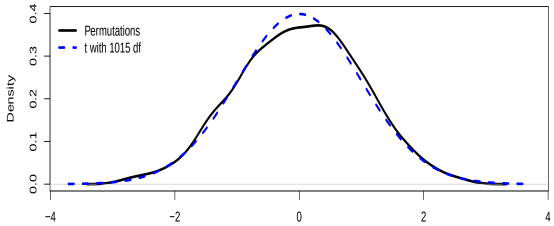
```
B <- 2000  ## number of permutations we will perform
Tstar <- matrix(NA, nrow = B)  ## empty slots for storage
set.seed(3)  ## good practice when dealing with randomness
for (b in 1:B) {
    ## calculate the t statistic obtained from each shuffle
    Tstar[b] <- t.test(depth ~ shuffle(cover), data = snow, var.equal = T)$statistic
}
pdata(abs(Tstar), abs(Tobs), lower.tail = F)  ## obtain a p-value

## t
## 0
```



(a) How many of the permutations resulted in a test statistic as or more extreme than the one observed in the original sample?

(b) What is the value of the p-value? (**Note:** If your p-value= 0 then you should say p-value < 0.001.)

(c) Compare the sampling distributions of the test statistic under the null hypothesis between non-parametric (permutations) and parametric (t_{1015}) using the following plot.



Make a decision

10. Make a decision for the test at the $\alpha = 0.05$ significance level.

Write a conclusion

11. Write a conclusion specific to the problem, including scope of inference discussion.

Estimate the difference in means

12. Calculate the point estimate for the difference in means, including the notations. Make sure to label it correctly (that includes the order of subtraction).

```
## estimate the difference in means
Tobs_d <- diffmean(depth ~ cover, data = snow)
Tobs_d
```

13. Report the 95% confidence interval from problem 8. Was this CI obtained parametrically or nonparametrically?

14. Now find the 95% confidence interval using bootstrapping. Is this CI being obtained parametrically or nonparametrically?

```
## bootstrap
B <- 1000  #we will resample 1000 times
set.seed(5)
Tboot <- matrix(NA, nrow = B)  #empty slots for each statistic
for (b in 1:B) {
    # different from the text but is more in line with 216 methods
    Tboot[b] <- diffmean(depth ~ cover, data = resample(snow, groups = cover))
}

## calculate the 95% CI
qdata(Tboot, c(0.025, 0.975))

##         quantile     p
## 2.5%  -228.37730 0.025
## 97.5%  -71.39043 0.975
```

15. Interpret the confidence interval from 14 in the context of the study.

16. Does your confidence interval from 14 match the decision for the test based on (9b)? Explain.