Agenda

1. The Bootstrap

Warumup: 4.39 – Coffee, depression, and physical activity Caffeine is the world's most widely used stimulant, with approximately 80% consumed in the form of coffee. However, studies that analyze the relationship between coffee / caffeine consumption and depression risk are scarce. Since depression is also known to have an association with physical activity, participants in a study investigating the relationship between coffee consumption and depression were asked to report the number of hours they spent per week on moderate (e.g., brisk walking) and vigorous (e.g., strenuous sports and jogging) exercise. Based on these data the researchers estimated the total hours of metabolic equivalent tasks (MET) per week, a value always greater than 0. The table below gives summary statistics of MET for women in this study based on the amount of coffee consumed.

	Caffeinated coffee consumption					
	$\leq$ 1 cup/week	2-6 cups/week	1 cup/day	2-3 cups/day	$\geq$ 4 cups/day	Total
Mean	18.7	19.6	19.3	18.9	17.5	
SD	21.1	25.5	22.5	22.0	22.0	
n	12,215	6,617	17,234	12,290	2,383	50,739

1. Write the hypotheses for evaluating if the average physical activity level varies among the different levels of coffee consumption.

2. Check conditions and describe any assumptions you must make to proceed with the test.

3. Below is part of the output associated with this test. Fill in the empty cells.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
coffee					0.0003
Residuals		25,564,819			
Total		25,575,327			

4. What is the conclusion of the test?

5. Given that the overall mean is 19, write out the three different forms of the model

The Bootstrap The bootstrap is a powerful computational technique for estimating all kinds of things. [Do not believe the hyperbolic warnings in Section 4.5.3 of the book!] It is particularly useful when our actual data sample is non-normal.

- The bootstrap works in three steps:
  1. Construct a sample of n items from your original data set, sampling *with replacement* (`resample()`)
  2. Compute the statistic of interest on this sample (in our case, the mean (`mean()`))
  3. Repeat this process many, many times and collect the results (`do()`)
- This *bootstrap distribution* is an approximation of the sampling distribution of your statistic
- Big Idea: The middle $P\%$ of the bootstrap distribution makes a $P\%$ confidence interval for the statistic in question, without making many assumptions about the distribution of X!

Wage example Consider the following sample of 534 hourly wages from the Current Population Survey (of 1985):

```
favstats(~wage, data = CPS85)

##  min   Q1 median    Q3  max     mean       sd   n missing
##    1 5.25   7.78 11.25 44.5 9.024064 5.139097 534       0
```

Distributional assumptions

1. Construct a 95% confidence interval for the mean wage in the 1985 CPS, based on this sample. Assume that 5.139 is the true population standard deviation, and that wages are normally distributed.

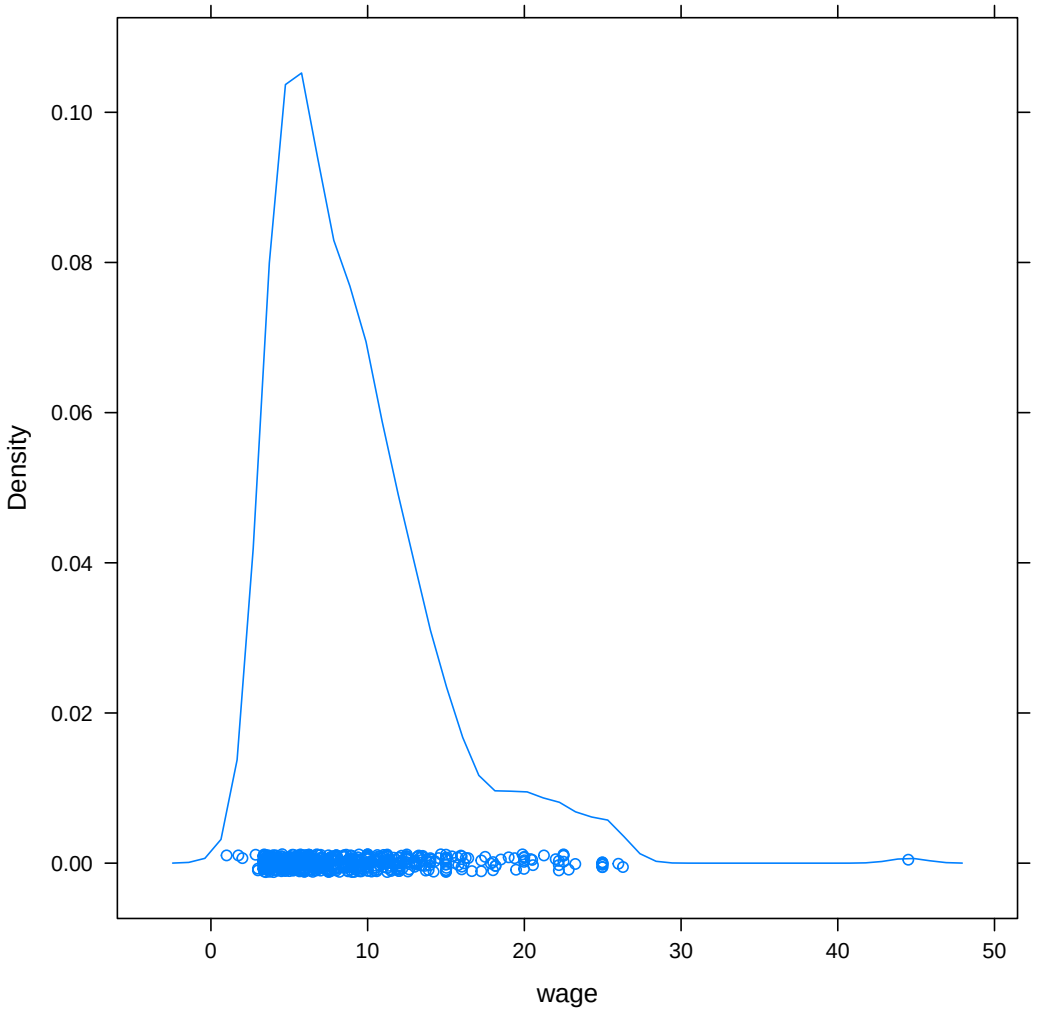
```
x.bar = mean(~wage, data=CPS85)
sd = sd(~wage, data=CPS85)
n = nrow(CPS85)
z.star = qnorm(c(0.025, 0.975))
se = sd / sqrt(n)
x.bar + z.star * se

## [1] 8.588186 9.459941
```

2. Using the t-statistic below, construct a 95% confidence interval for the mean wage that makes no assumption about the population standard deviation, but assumes that wages are normally distributed.

```
qt(0.975, df = n-1)

## [1] 1.964425

t.star = qt(c(0.025, 0.975), df=n-1)
x.bar + t.star * se

## [1] 8.587194 9.460933
```

3. Examine the distribution of *wage*. Is it normally distributed?

```
densityplot(~wage, data=CPS85)
```



The bootstrap Using the bootstrap, construct a 95% confidence interval for the mean wage that does not assume that wages are normally distributed.

```
bstrap <- do(10000) * mean(~wage, data = resample(CPS85))
qdata(~mean, p = c(0.025, 0.975), data = bstrap)

##       quantile     p
## 2.5%  8.596251 0.025
## 97.5% 9.462125 0.975
```

Compare the three confidence intervals you constructed. Do you see any important differences?