SMML Class 2 Lab

John Kubale

9/3/2024

We will use `Wage` data in R package `ISLR`

```r
data("Wage")
dim(Wage)
```

```
## [1] 3000   11
```

```r
summary(Wage)
```

```
##       year           age                     maritl           race     
##  Min.   :2003   Min.   :18.00   1. Never Married: 648   1. White:2480  
##  1st Qu.:2004   1st Qu.:33.75   2. Married      :2074   2. Black: 293  
##  Median :2006   Median :42.00   3. Widowed      :  19   3. Asian: 190  
##  Mean   :2006   Mean   :42.41   4. Divorced     : 204   4. Other:  37  
##  3rd Qu.:2008   3rd Qu.:51.00   5. Separated    :  55                  
##  Max.   :2009   Max.   :80.00                                          
##                                                                        
##               education                     region               jobclass   
##  1. < HS Grad      :268   2. Middle Atlantic   :3000   1. Industrial :1544  
##  2. HS Grad        :971   1. New England       :   0   2. Information:1456  
##  3. Some College   :650   3. East North Central:   0                        
##  4. College Grad   :685   4. West North Central:   0                        
##  5. Advanced Degree:426   5. South Atlantic    :   0                        
##                           6. East South Central:   0                        
##                           (Other)              :   0                        
##             health      health_ins      logwage           wage       
##  1. <=Good     : 858   1. Yes:2083   Min.   :3.000   Min.   : 20.09  
##  2. >=Very Good:2142   2. No : 917   1st Qu.:4.447   1st Qu.: 85.38  
##                                      Median :4.653   Median :104.92  
##                                      Mean   :4.654   Mean   :111.70  
##                                      3rd Qu.:4.857   3rd Qu.:128.68  
##                                      Max.   :5.763   Max.   :318.34  
## 
```

1. Focus on the variable, `wage`

A. Mean and variance of `wage`

```r
mean(Wage$wage)
```

```
## [1] 111.7036
```

```r
var(Wage$wage)
```

```
## [1] 1741.276
```

Does the var() function give you the population or sample variance (hint: ?var)? - Sample variance, because it comes from the sample.

B. Manually calculate the population and sample variance for wage. The code for calculating the population variance is provided. You will have to tweak it to calculate the sample variance.

```r
library(dplyr)

n <- dim(Wage)[[1]]


# calculate population variance
# Wage %>%
#   mutate(dif2 = (wage - mean(wage))^2
#          ) %>%
#   summarise(pop_var = (sum(dif2)/3000))

Wage |>
  summarise(sam_var = sum((wage - mean(wage))^2) / (n() - 1),
            pop_var <- sum((wage - mean(wage))^2) / (n() ))
```

```
##    sam_var pop_var <- sum((wage - mean(wage))^2)/(n())
## 1 1741.276                                    1740.695
```

```r
# calculate sample variance
```

Which matches what you got using var()? Which is larger and why do you think that is? - The Var matches the sample variance.

B. Using the sample variance of the estimated mean you've already calculated, estimate the 95% confidence interval of the true mean?

```r
# calculate sample size of Wage and save sample variance of Wage$wage as object called
n <- dim(Wage)[[1]]
sampvar <- var(Wage$wage)

# calculate the appropriate t statistic to calculate a 95% CI for mean of wage
```

```r
t.score<-qt(p=.05/2, df=n-1, lower.tail=F)
t.score
```

```
## [1] 1.960755
```

```r
# calculate 95% CI for estimated mean of wage
lowCI <- mean(Wage$wage)-t.score*sqrt(sampvar)/sqrt(n)
upCI <- mean(Wage$wage)+t.score*sqrt(sampvar)/sqrt(n)
print(c(lowCI,upCI))
```

```
## [1] 110.2098 113.1974
```

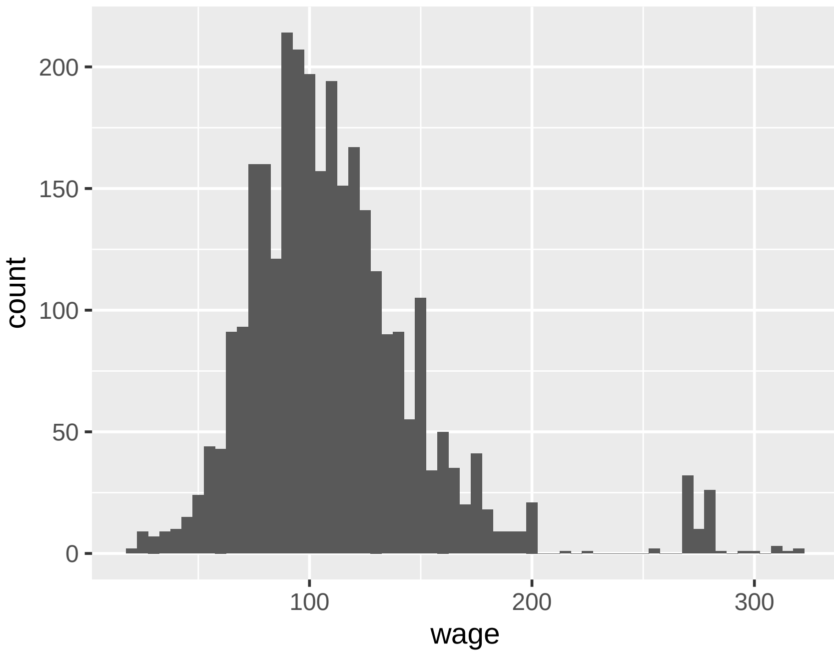
```r
# Conduct single sample t-test of Wage$wage with 95% CI.
t.test(Wage$wage, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  Wage$wage
## t = 146.62, df = 2999, p-value < 0.0000000000000022
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  110.2098 113.1974
## sample estimates:
## mean of x
##  111.7036
```

- How would you interpret the 95% confidence interval of the mean?
- If we repeat this study sampling from the mean with N sample size, we would expect this interval to contain the true mean 95% of the time.
- What are the null and alternative hypotheses associated with the t-test you ran above?
- Null hypothesis is associated with no difference in the means, while the alternative hypothesis is associated with a two-sided test meaning that the difference in the sample means are probably different. Nully hypothesis is equal to 0.
- How does the 95% CI from t.test() compare to the interval you calculated by hand?
- The confidence intervals calculated in both seem to account for a two sided test.

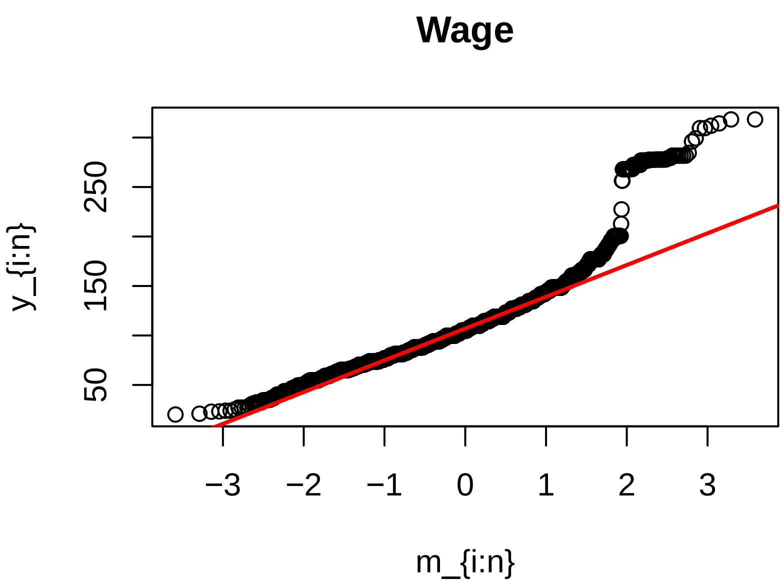
C. Does `wage` follow a normal distribution?

```r
ggplot(Wage, aes(x=wage)) + geom_histogram(binwidth=5)
```




```r
qqnorm(Wage$wage, main="Wage", ylab="y_{i:n}", xlab="m_{i:n}") +
  qqline(Wage$wage, col="red",lwd=2)
```

```
## Error in qqnorm(Wage$wage, main = "Wage", ylab = "y_{i:n}", xlab = "m_{i:n}") + : non
```



* Based on the figures would you say wage is normally distributed? - It appears that Wages is positive skewed.

```r
shapiro.test(Wage$wage)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Wage$wage
## W = 0.87957, p-value < 0.00000000000000022
```

- How would you interpret the results of the Shapiro-Wilk Normality test?

- Based on the Shapiro-Wilk test the Wage follows a normal distribution. Although, this could be due to the large sample size.

D. What are the steps to take to compare `wage` of those without vs. with college or higher education? (Hints: What do you need to assess before using a two sample t-test?)

$$H_0 : \mu_{\geq CollEduc} = \mu_{<CollEduc} vs. H_A : \mu_{\geq CollEduc} \neq \mu_{<CollEduc}$$

- Step 1) Recode `education`
- Step 2) Check means and variances by recoded education
- Step 3) ?
- Step 4) Conduct proper testing

Step 1) Recode `education`

```
# library(tidyverse) -- this will load dplyr and a number of other packages
table(Wage$education)
```

```
##
##       1. < HS Grad          2. HS Grad    3. Some College    4. College Grad
##                268                 971                650                685
## 5. Advanced Degree
##                426
```

```
Wage <- Wage%>%
  mutate(CollEduc=ifelse(education=="4. College Grad"|
                         education=="5. Advanced Degree",1,0))
```

- Look at the help page for the ifelse() function. What is the code above doing?
- ifelse function is is recoding the education variable if 4 or 5 as = 1, all else as 0.

Step 2) Check means and variances by recoded education

```
Wage %>%
  group_by(CollEduc)%>%
  summarize(m=mean(wage),
            var=var(wage))
```

```
## # A tibble: 2 x 3
##   CollEduc     m   var
##      <dbl> <dbl> <dbl>
## 1        0  98.2  910.
## 2        1 135.  2324.
```

- $\hat{\mu}_{<CollEduc} = \hat{\bar{y}}_{<CollEduc} = 98.2$ and $\hat{\sigma}^2_{<CollEduc} = s^2_{<CollEduc} = 910$
- $\hat{\mu}_{\geq CollEduc} = \hat{\bar{y}}_{\geq CollEduc} = 135$ and $\hat{\sigma}^2_{\geq CollEduc} = s^2_{\geq CollEduc} = 2324$

Step 3) Test equal variance * Corresponding hypothesis: $H_0 : \sigma^2_{<CollEduc} = \sigma^2_{\geq CollEduc}$ vs. $H_A : \sigma^2_{<CollEduc} \neq \sigma^2_{\geq CollEduc}$

```r
var.test(wage ~ CollEduc, Wage, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  wage by CollEduc
## F = 0.39169, num df = 1888, denom df = 1110, p-value <
## 0.00000000000000022
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3524383 0.4346930
## sample estimates:
## ratio of variances
##          0.3916941
```

- What are the null and alternative hypotheses associated with the var.test() above?
- The F-test is testing the Null of whether the variances are not equal, and here the alternative hypothesis is that the variances are equal in our two samples.
- How do you interpret the results?

- Since, our p-value is less than .05, we cannot reject the Null, and we conclude that the variances are not equal.

Step 4) Conduct proper testing based on what you found in the previous step. * Corresponding hypothesis: $H_0 : \mu_{<CollEduc} = \mu_{\geq CollEduc}$ vs. $H_A : \mu_{<CollEduc} \neq \mu_{\geq CollEduc}$

```r
higher_ed <- Wage |> filter(CollEduc == 1) |> select(wage)
lower_ed <- Wage |> filter(CollEduc == 0) |> select(wage)

t_test <- t.test(higher_ed, lower_ed, var.equal = FALSE)
t_test
```

```
##
##  Welch Two Sample t-test
##
## data:  higher_ed and lower_ed
## t = 22.651, df = 1629.5, p-value < 0.00000000000000022
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  33.19245 39.48578
## sample estimates:
## mean of x mean of y
## 134.58514  98.24602
```

- What are the null/alternative hypotheses associated with this t-test?

- The Null in this Welch 2 sample t-test is that the means are the same, while the alternative hypothesis is that these means are not the same.
- How do you interpret the results?
- The mean wage for college education respondents are statistically higher than those without a higher education level.

- How would you conduct this test if you came to the opposite conclusion (regarding the two sample variances) in the previous step?

- By using the argument var.equal = FALSE.

E. What is the correlation coefficient between `wage` and `age`?

```r
cor(Wage$wage, Wage$age)
```

```
## [1] 0.1956372
```

```r
cor.test(Wage$wage, Wage$age)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Wage$wage and Wage$age
## t = 10.923, df = 2998, p-value < 0.00000000000000022
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1609777 0.2298147
## sample estimates:
##       cor
## 0.1956372
```

What are the conclusions from F? Can we use conclusions from above? - Wage and age has a small relationship based on our Pearson's correlation analysis.

F. If we're concerned that the wage distribution in the groups we want to compare (here it is those with/without college education) we should consider using a non-parametric test for comparing means like the Wilcoxon test.

```r
# stratify by college education status and look at each distribution as before (i.e.,
coll_edu <- filter(Wage, CollEduc==1) ## Subset Wage data to only include those with co
nocoll_edu <- filter(Wage, CollEduc!=1) ## Subset Wage data to only include those with
```

How could we compare the two wage distributions since? Conduct a Wilcoxon rank sum (AKA Mann-Whitney U test) test comparing the mean wage of those without vs. with college or higher education.

```r
# conduct non-parametric test
wilcox.test(wage ~ CollEduc, data = Wage,
                    exact = FALSE, conf.int=0.95)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  wage by CollEduc
## W = 496842, p-value < 0.00000000000000022
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -33.95447 -29.21830
## sample estimates:
## difference in location
##              -31.56618
```

- How do the results compare to the t-test you conducted earlier?

- The results tell us the same story, yet the 95% CI are different.

2. Focus on the variable, `logwage`

A. What is the estimated mean and variance of the sample?

```
Wage |>
  summarise(mean(logwage), var(logwage))
```

```
##   mean(logwage) var(logwage)
## 1      4.653905    0.1237299
```

B. What is the sample standard error for logwage? Use it to calculate 95% confidence interval of the true mean.

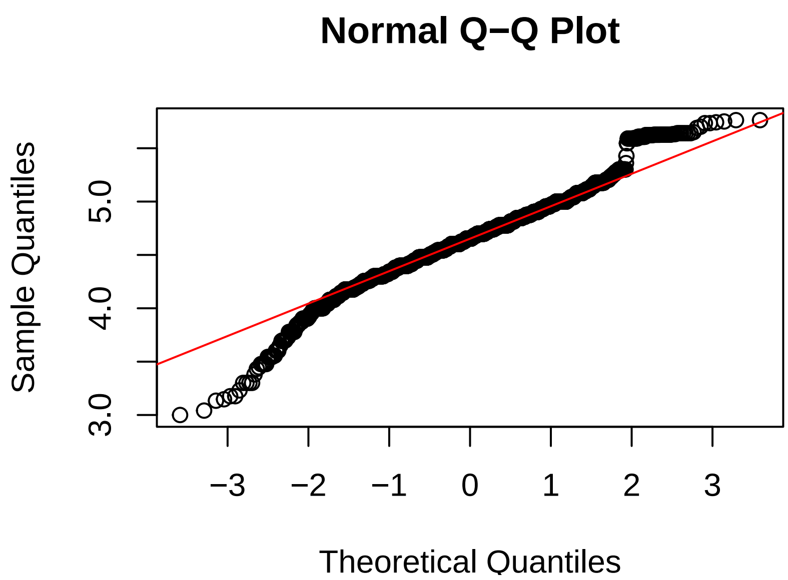
```
Wage |>
  summarise(se = sd(logwage) / sqrt(n()),
            # calculate 95% CI
            mean(logwage),
            low = mean(logwage) - 1.96*se,
            upper = mean(logwage) + 1.96*se)
```

```
##            se mean(logwage)      low    upper
## 1 0.006422094      4.653905 4.641318 4.666492
```

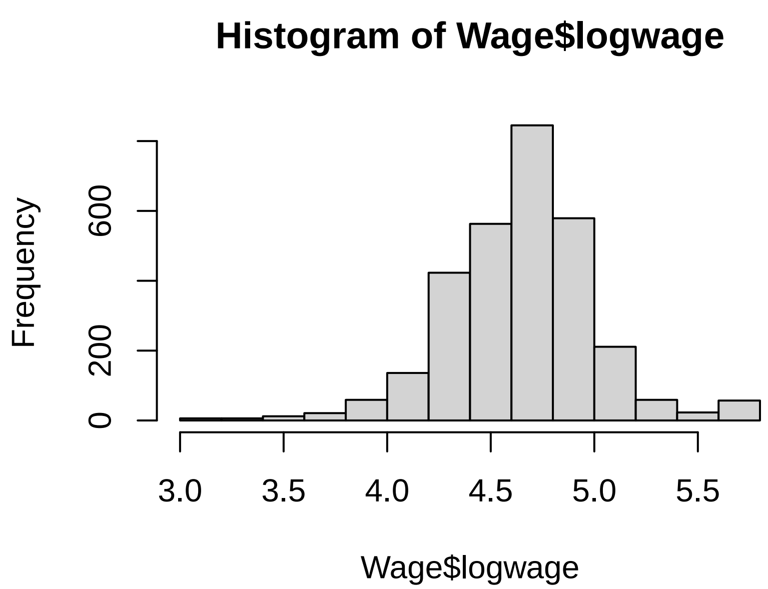
- The sample standard error is .0064.
- The 95% confidence interval of $\mu_{logwage}$ is $[4.64, 4.67]$

C. Does `logwage` follow a normal distribution? Evaluate its distribution both graphically and statistically. - There is some negative skewness in the log of wages data.

```
qqnorm(Wage$logwage)
qqline(Wage$logwage, col="red")
```




```r
hist(Wage$logwage)
```



- Based on the Shapiro-Wilk test the log of Wage follows a normal distribution. Although, this could be due to the large sample size.

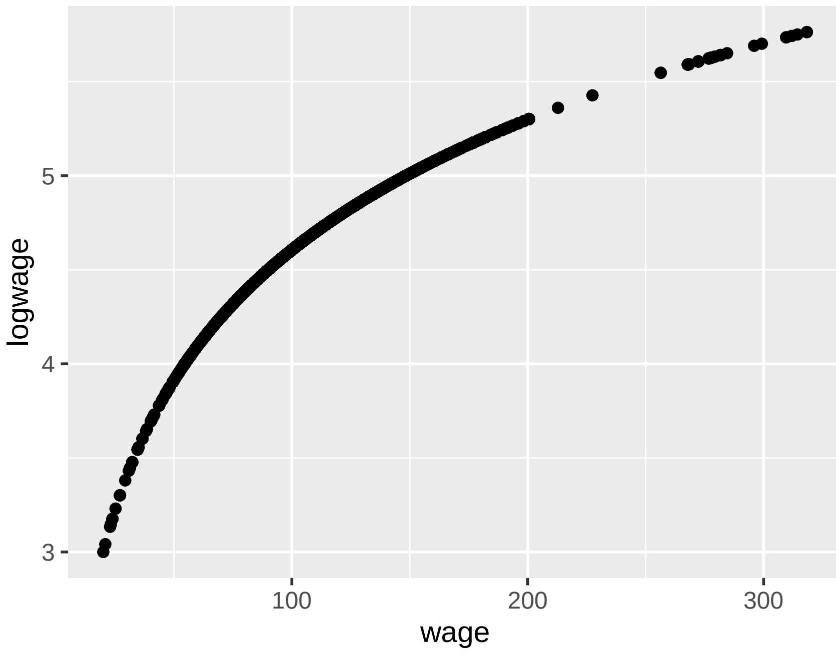
```r
shapiro.test(Wage$logwage)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Wage$logwage
## W = 0.97696, p-value < 0.00000000000000022
```

- While, compared to `wage`, `logwage` appears more normally distributed, it still fails to meet the normal distribution requirements.

D. Assess the relationship between `wage` and `logwage`.

```
ggplot(Wage, aes(x=wage,y=logwage)) + geom_point()
```



- How would you describe the relationship between Wage and logwage?

- The relationship between wage and logwage follows a curvlinear relationship.

```
# There are often multiple ways in R to achieve the same result.
mean(log(Wage$wage))
```

```
## [1] 4.653905
```

```
mean(Wage$logwage)
```

```
## [1] 4.653905
```