Testing Differences in Group Means

Ryan Safner

ECON 480 - Econometrics

Often we want to compare the means between two groups, and see if the difference is statistically significant. As an example, **is there a statistically significant difference in average hourly earnings between men and women**? Let:

- μ_W: mean hourly earnings for female college graduates
- μ_M: mean hourly earnings for male college graduates

We want to run a hypothesis test for the difference (d) in these two population means:

$$\mu_M - \mu_W = d_0$$

Our null hypothesis is that there is *no* statistically significant difference. Let's also have a two-sided alternative hypothesis, simply that there *is* a difference (positive or negative).

- $H_0 : d = 0$
- $H_1 : d \neq 0$

Note a logical one-sided alternative would be $H_2 : d > 0$, i.e. men earn more than women

The Sampling Distribution of d

The *true* population means μ_M, μ_W are unknown, we must estimate them from *samples* of men and women. Let: - $\bar{Y}_M$ the average earnings of a sample of n_M men
- $\bar{Y}_W$ the average earnings of a sample of n_W women

We then estimate $(\mu_M - \mu_W)$ with the sample $(\bar{Y}_M - \bar{Y}_W)$.

We would then run a **t-test** and calculate the **test-statistic** for the difference in means. The formula for the test statistic is:

$$t = \frac{(\bar{Y_M} - \bar{Y_W}) - d_0}{\sqrt{\frac{{s_M}^2}{n_M} + \frac{{s_W}^2}{n_W}}}$$

We then compare t against the critical value t^*, or calculate the p-value $P(T > t)$ as usual to determine if we have sufficient evidence to reject H_0

```r
# Our data comes from WAGE1.dta which you can find in Blackboard under data

# Load WAGE1 as wages
library("foreign") # to load .dta Stata files
wages<-read.dta("../Data/WAGE1.dta")
```

```
## Warning in read.dta("../Data/WAGE1.dta"): cannot read factor labels from
## Stata 5 files
```

```r
# there's a lot of variables in wages, let's only look at wage and female for now
wages<-subset(wages, select=c("wage","female"))

# just get a sense of the data
```

```r
head(wages)
```

```
##   wage female
## 1 3.10      1
## 2 3.24      1
## 3 3.00      0
## 4 6.00      0
## 5 5.30      0
## 6 8.75      0
```

```r
# we now want to look at the data under certain CONDITIONS
# conditionals require subsetting data with square brackets []
# such as: data[df$variable==condition]

# look at average wage for men
summary(wages$wage[wages$female==0])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.500   4.143   6.000   7.099   8.765  24.980
```

```r
sd(wages$wage[wages$female==0]) # get sd
```

```
## [1] 4.160858
```

```r
# look at average wage for women
summary(wages$wage[wages$female==1])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.530   3.000   3.750   4.588   5.510  21.630
```

```r
sd(wages$wage[wages$female==1]) # get sd
```

```
## [1] 2.529363
```

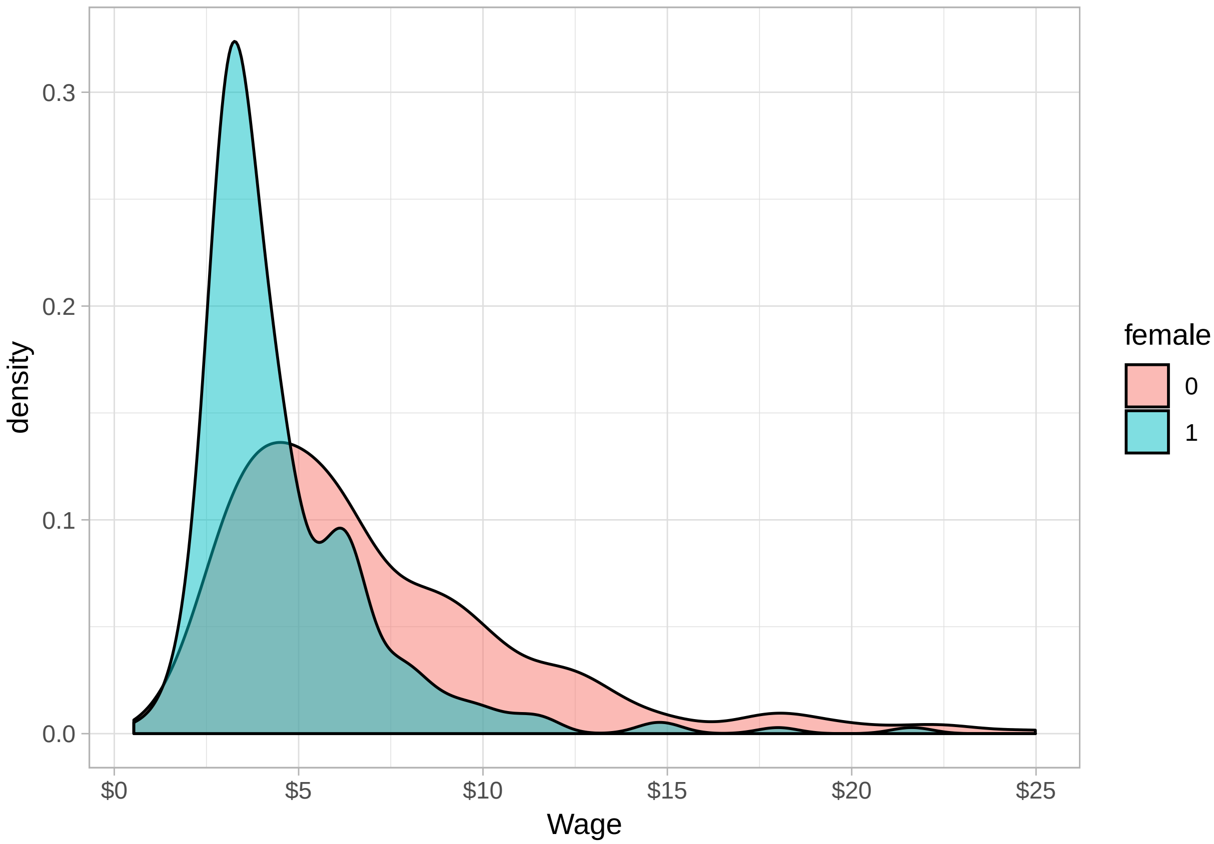
So our data is telling us that male and female average hourly earnings are distributed as such:

$$\begin{aligned}
\bar{Y}_M &\sim N(7.10, 4.16) \\
\bar{Y}_W &\sim N(4.59, 2.53)
\end{aligned}$$

We can plot this to see visually. There is a lot of overlap in the two distributions, but the male average is higher than the female average, and there is also a lot more variation in males than females, noticeably the male distribution skews further to the right.

```r
wages$female<-as.factor(wages$female)

library("ggplot2")
ggplot(data=wages,aes(x=wage,fill=female))+
  geom_density(alpha=0.5)+
  scale_x_continuous(seq(0,25,5),name="Wage",labels=scales::dollar)+
  theme_light()
```



Knowing the distributions of male and female average hourly earnings, we can estimate the **sampling distribution of the difference in group eans** between men and women as:

The mean:

$$
\begin{aligned}
\bar{d} &= \bar{Y}_M - \bar{Y}_W \\
\bar{d} &= 7.10 - 4.59 \\
\bar{d} &= 2.51
\end{aligned}
$$

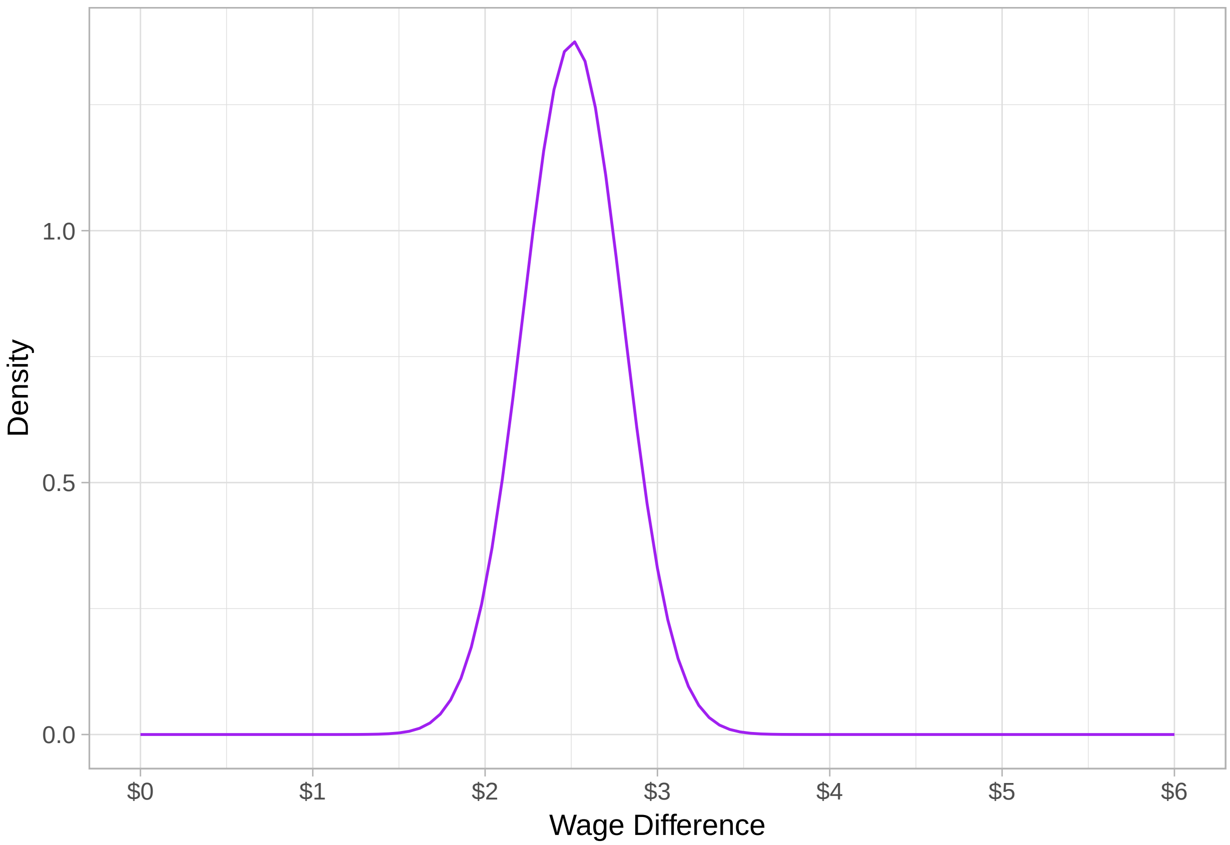
The standard error of the mean:

$$
\begin{aligned}
SE(\bar{d}) &= \sqrt{\frac{{s_M}^2}{n_M} + \frac{{s_W}^2}{n_W}} \\
&= \sqrt{\frac{4.16^2}{274} + \frac{2.33^2}{252}} \\
&\approx 0.29
\end{aligned}
$$

So the sampling distribution of the difference in group means is distributed:

$$\bar{d} \sim N(2.51, 0.29)$$

```r
ggplot(data.frame(x=0:6),aes(x=x))+
  stat_function(fun=dnorm, args=list(mean=2.51, sd=0.29), color="purple")+
  ylab("Density")+
  scale_x_continuous(seq(0,6,1),name="Wage Difference",labels=scales::dollar)+
  theme_light()
```



Now we the ***t*-test** like any other:

$$
\begin{aligned}
t &= \frac{\text{estimate} - \text{null hypothesis}}{\text{standard error of the estimate}} \\
&= \frac{d - 0}{SE(d)} \\
&= \frac{2.51 - 0}{0.29} \\
&= 8.66
\end{aligned}
$$

This is statistically significant. The p-value, $P(t > 8.66) =$ is 0.0000000000000000000410, or basically, 0.

```r
pt(8.66,456.33, lower.tail=FALSE)
```

```
## [1] 4.102729e-17
```

The t-test in R

```
t.test(wage~female, data=wages, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  wage by female
## t = 8.44, df = 456.33, p-value = 4.243e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.926971 3.096690
## sample estimates:
## mean in group 0 mean in group 1
##        7.099489        4.587659
```

```
reg<-lm(wage~female, data=wages)
summary(reg)
```

```
##
## Call:
## lm(formula = wage ~ female, data = wages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5995 -1.8495 -0.9877  1.4260 17.8805
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.0995     0.2100  33.806  < 2e-16 ***
## female1      -2.5118     0.3034  -8.279 1.04e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.476 on 524 degrees of freedom
## Multiple R-squared:  0.1157, Adjusted R-squared:  0.114
## F-statistic: 68.54 on 1 and 524 DF,  p-value: 1.042e-15
```