Inference for Numerical Data

IS381 - Statistics and Probability with R

Jason Bryer, Ph.D.

April 20, 2026

One Minute Paper Results

What was the most important thing you learned during this class?

```
## NULL
```

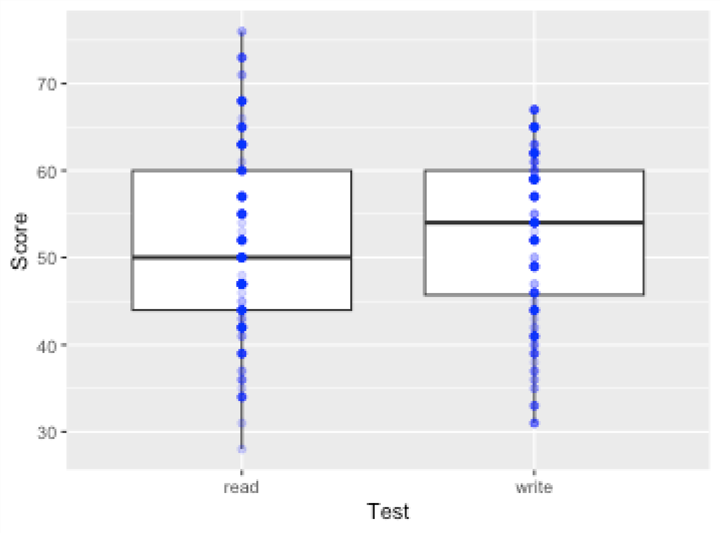
What important question remains unanswered for you?

```
## NULL
```

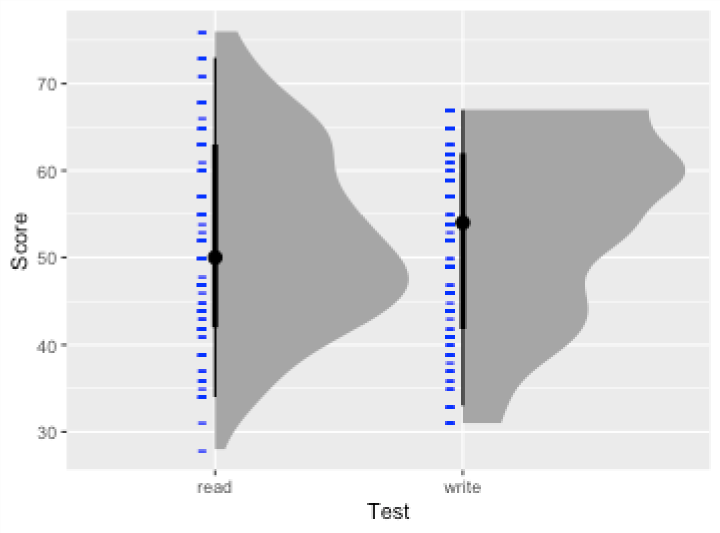
High School & Beyond Survey

200 randomly selected students completed the reading and writing test of the High School and Beyond survey. The results appear to the right. Does there appear to be a difference?

```
data(hsb2, package = 'openintro') # in openintro packag
hsb2.melt <- melt(hsb2[,c('id','read', 'write')], id='
ggplot(hsb2.melt, aes(x=variable, y=value)) +      geom_
    geom_point(alpha=0.2, color='blue') + xlab('Test')
```




```
ggplot(hsb2.melt, aes(x=variable, y=value)) +
    ggdist::stat_halfeye() +
    geom_point(color='blue', position = position_nudge(
    xlab('Test') + ylab('Score')
```



High School & Beyond Survey

```
head(hsb2)
```

```
## # A tibble: 6 × 11
##      id gender race  ses    schtyp prog        read write  math science socst
##   <int> <chr>  <chr> <fct>  <fct>  <fct>      <int> <int> <int>   <int> <int>
## 1    70 male   white low    public general       57    52    41      47    57
## 2   121 female white middle public vocational    68    59    53      63    61
## 3    86 male   white high   public general       44    33    54      58    31
## 4   141 male   white high   public vocational    63    44    47      53    56
## 5   172 male   white middle public academic      47    52    57      53    61
## 6   113 male   white middle public academic      44    52    51      63    61
```

Are the reading and writing scores of each student independent of each other?

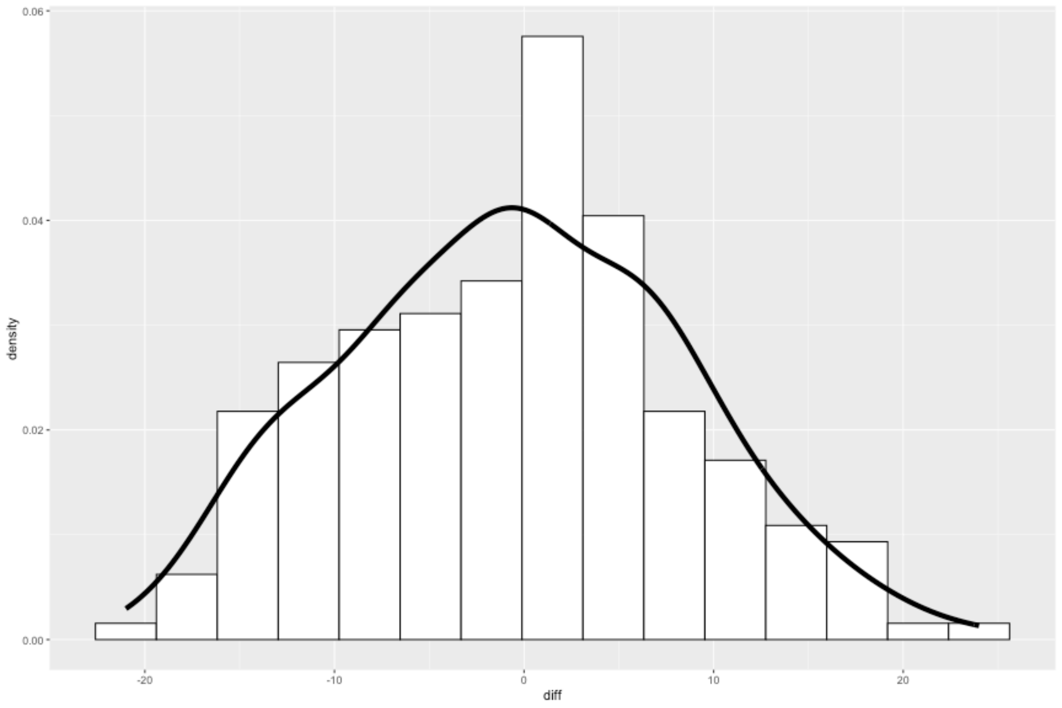
Analyzing Paired Data

- When two sets of observations are not independent, they are said to be paired.
- To analyze these type of data, we often look at the difference.

```
hsb2$diff <- hsb2$read - hsb2$write
head(hsb2$diff)
```

```
## [1]  5  9 11 19 -5 -8
```

```
ggplot(hsb2, aes(x = diff)) +
    geom_histogram(aes(y = ..density..), bins = 15, col
    geom_density(size = 2)
```



Setting the Hypothesis

What are the hypothesis for testing if there is a difference between the average reading and writing scores?

H_0: There is no difference between the average reading and writing scores.

$$\mu_{diff} = 0$$

H_A: There is a difference between the average reading and writing score.

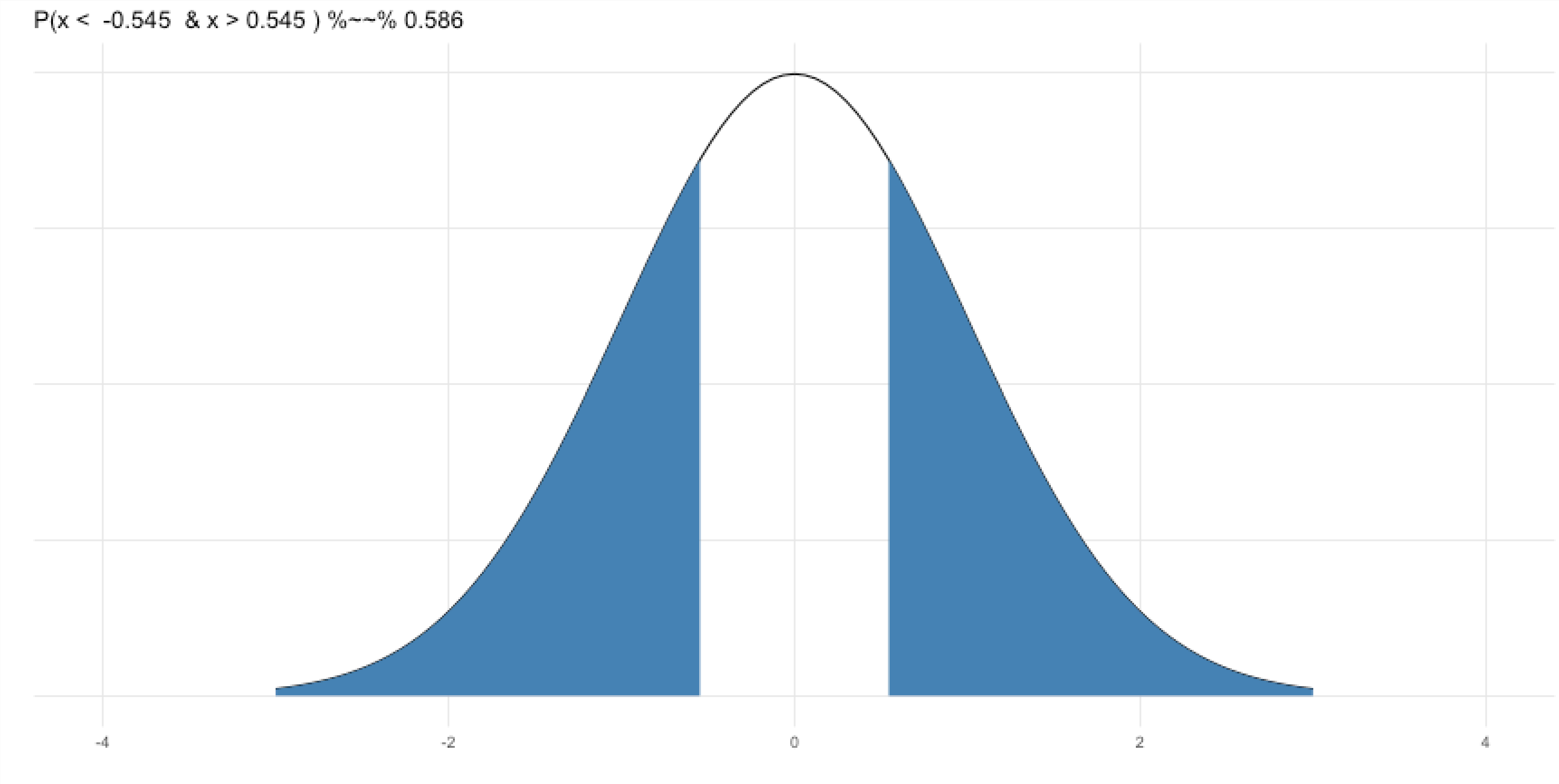
$$\mu_{diff} \neq 0$$

Nothing new here...

- The analysis is no different that what we have done before.
- We have data from one sample: differences.
- We are testing to see if the average difference is different that 0.

Calculating the test-statistic and the p-value

The observed average difference between the two scores is -0.545 points and the standard deviation of the difference is 8.887 points. Do these data provide convincing evidence of a difference between the average scores on the two exams (use $\alpha = 0.05$)?



Calculating the test-statistic and the p-value

$$Z = \frac{-0.545 - 0}{\frac{8.887}{\sqrt{200}}} = \frac{-0.545}{0.628} = -0.87$$

$$p - value = 0.1949 \times 2 = 0.3898$$

Since p-value > 0.05, we **fail to reject the null hypothesis**. That is, the data do not provide evidence that there is a statistically significant difference between the average reading and writing scores.

```
2 * pnorm(mean(hsb2$diff), mean=0, sd=sd(hsb2$diff)/sqrt(nrow(hsb2)))
```

```
## [1] 0.3857741
```

Evaluating the null hypothesis

Interpretation of the p-value

The probability of obtaining a random sample of 200 students where the average difference between the reading and writing scores is at least 0.545 (in either direction), if in fact the true average difference between the score is 0, is 38%.

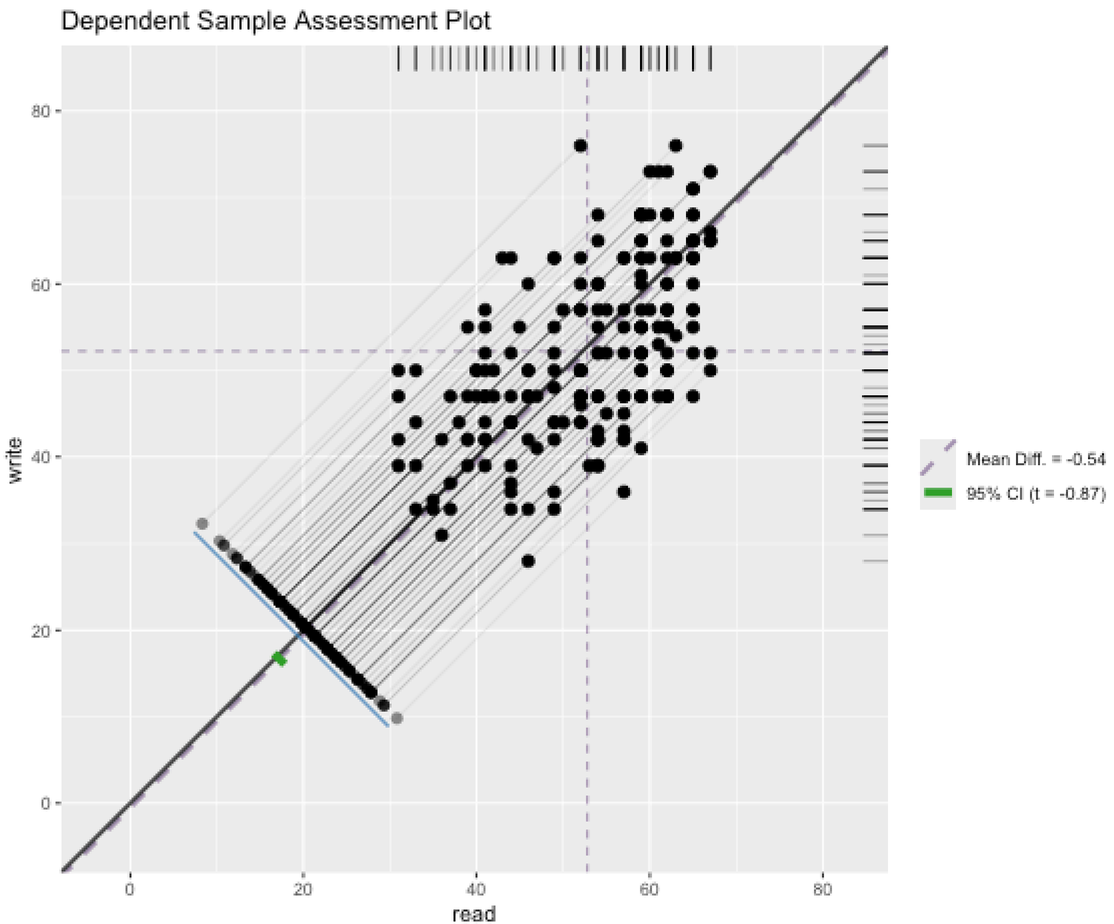
Calculating 95% Confidence Interval

$$-0.545 \pm 1.96 \frac{8.887}{\sqrt{200}} = -0.545 \pm 1.96 \times 0.628 = (-1.775, 0.685)$$

Note that the confidence interval spans zero!

Visualizing Dependent Sample Tests

```r
# remotes::install_github('briandk/granovaGG')
library(granovaGG)
granovagg.ds(as.data.frame(hsb2[,c('read', 'write')]))
```



SAT Scores by Sex

```r
data(sat, package = 'DATA606')
head(sat)
```

```
##   Verbal.SAT Math.SAT Sex
## 1        450      450   F
## 2        640      540   F
## 3        590      570   M
## 4        400      400   M
## 5        600      590   M
## 6        610      610   M
```

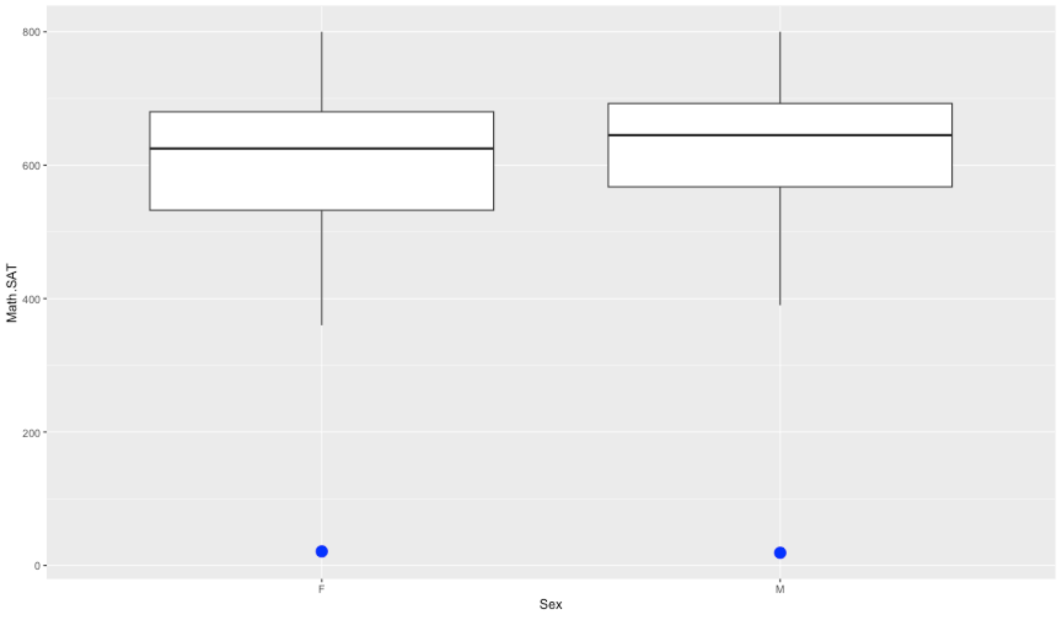
Is there a difference in math scores between males and females?

SAT Scores by Sex

```
tab <- describeBy(sat$Math.SAT,
           group=sat$Sex,
           mat=TRUE, skew=FALSE)
tab[,c(2,4:7)]
```

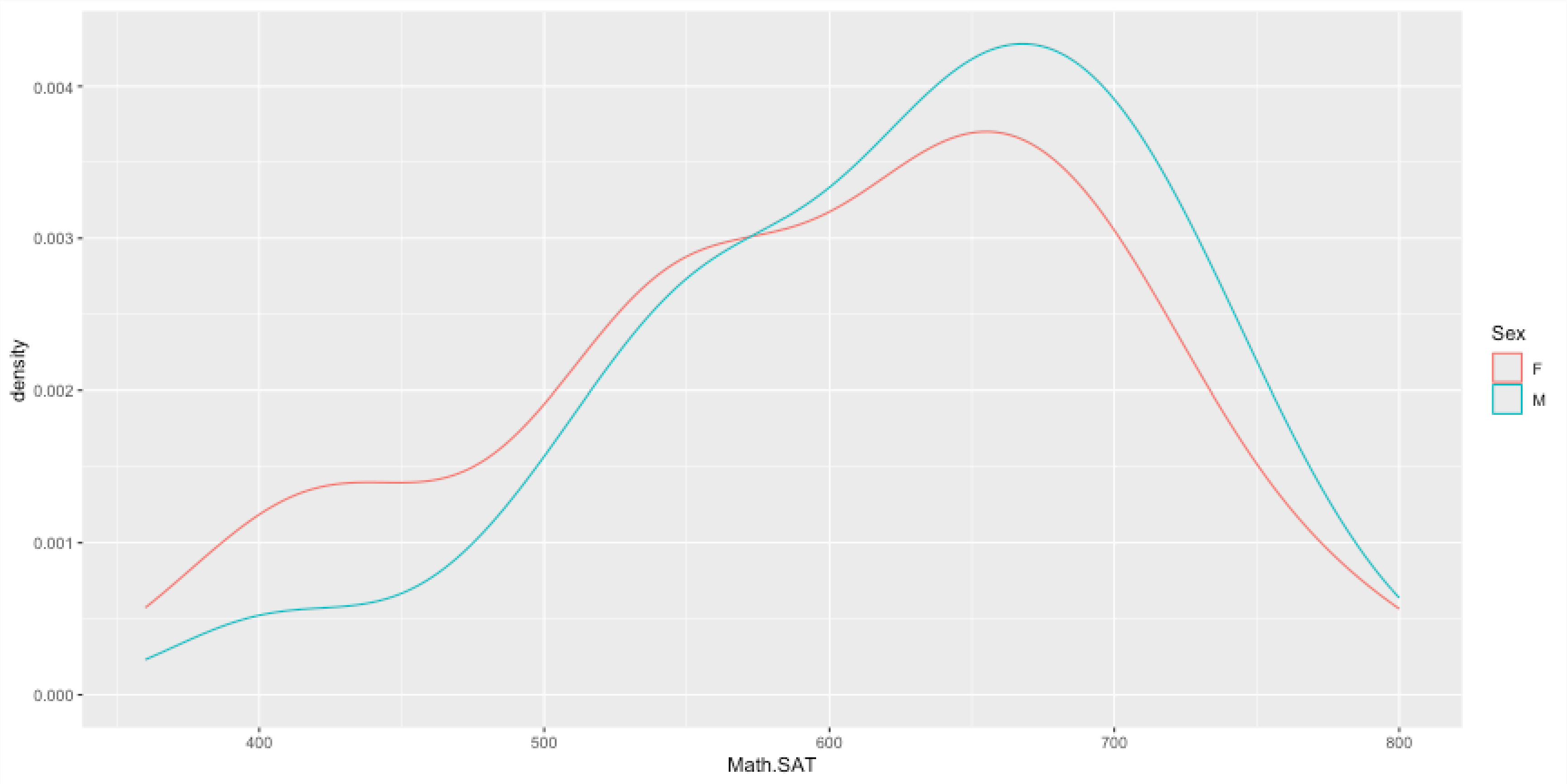
```
##      group1  n    mean       sd median
## X1*1      F 82 20.9878 9.571684   23.5
## X1*2      M 80 18.9250 8.185778   20.5
```

```
ggplot(sat, aes(x=Sex, y=Math.SAT)) +
    geom_boxplot() +
    geom_point(data = tab, aes(x=group1, y=mean),
               color='blue', size=4)
```



Distributions

```
ggplot(sat, aes(x=Math.SAT, color = Sex)) + geom_density()
```



95% Confidence Interval

We wish to calculate a 95% confidence interval for the average difference between SAT scores for males and females.

Assumptions:

1. Independence within groups.
2. Independence between groups.
3. Sample size/skew

Confidence Interval for Difference Between Two Means

- All confidence intervals have the same form: point estimate ± ME
- And all ME = critical value * SE of point estimate
- In this case the point estimate is $\bar{x}_1 - \bar{x}_2$ Since the sample sizes are large enough, the critical value is z* So the only new concept is the standard error of the difference between two means...

Standard error for difference in SAT scores

$$SE_{(\bar{x}_M - \bar{x}_F)} = \sqrt{\frac{s_M^2}{n_M} + \frac{s_F^2}{n_F}}$$

$$SE_{(\bar{x}_M - \bar{x}_F)} = \sqrt{\frac{90.4}{80} + \frac{103.7}{82}} = 1.55$$

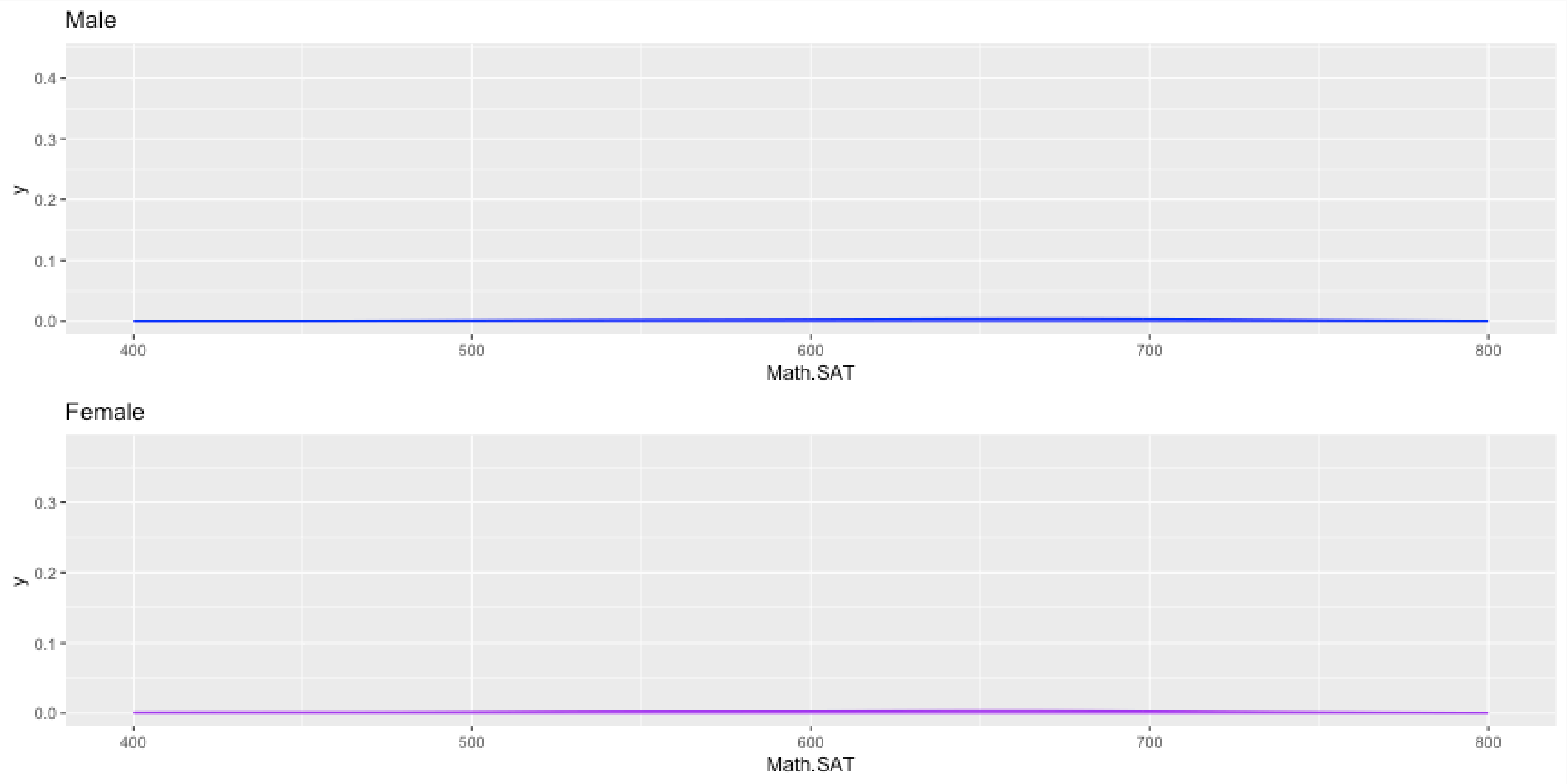
Calculate the 95% confidence interval:

$$(\bar{x}_M - \bar{x}_F) \pm 1.96 SE_{(\bar{x}_M - \bar{x}_F)}$$

$$(626.9 - 597.7) \pm 1.96 \times 1.55$$

$$29.2 \pm 3.038 = (26.162, 32.238)$$

Visualizing independent sample tests



What about smaller sample sizes?

What if you want to compare the quality of one batch of Guinness beer to the next?

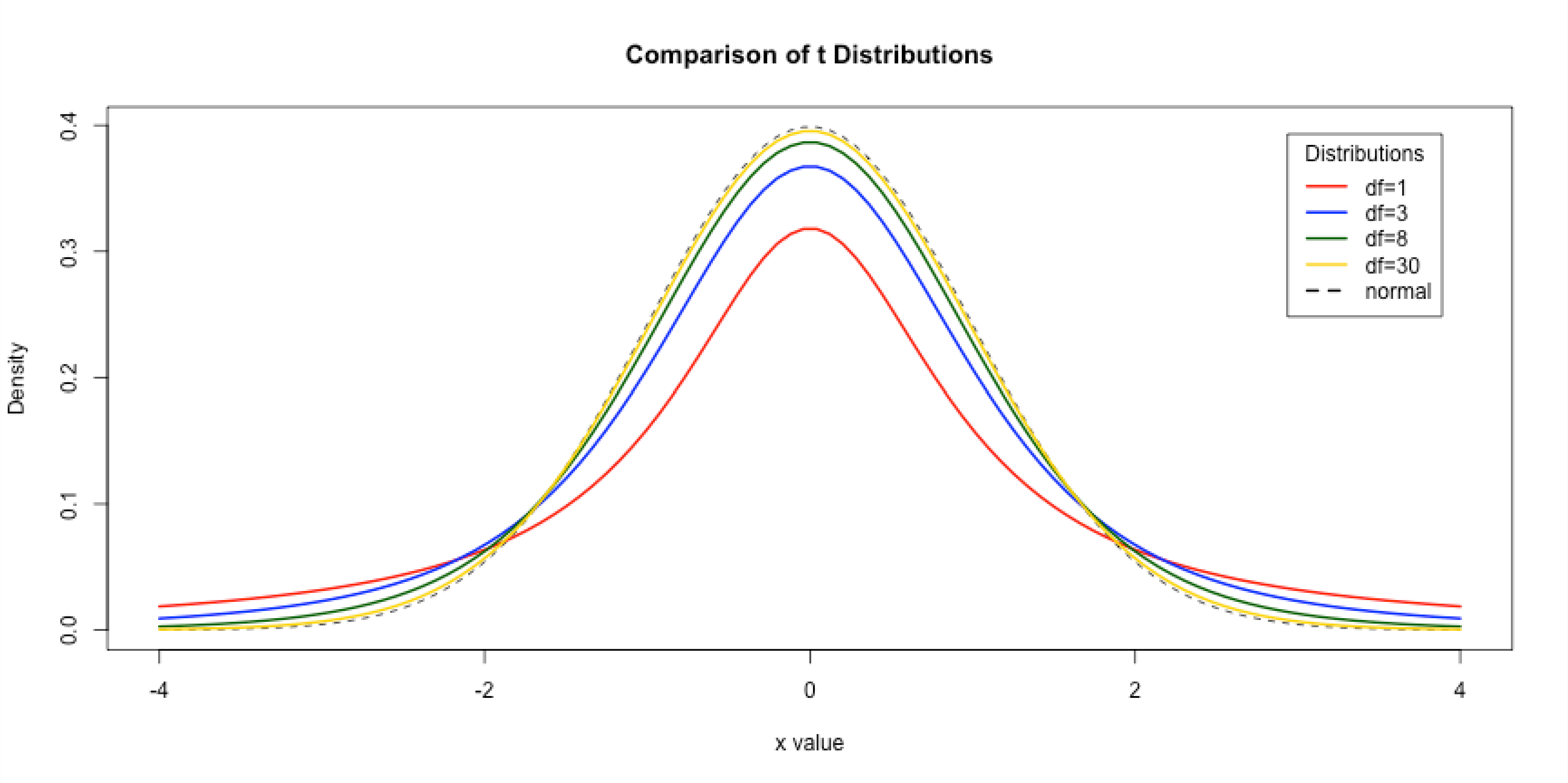
- Sample sizes necessarily need to be small.
- The CLT states that the sampling distribution approximates normal as n -> Infinity
- Need an alternative to the normal distribution.
- The *t* distribution was developed by William Gosset (under the pseudonym *student*) to estimate means when the sample size is small.

Confidence interval is estimated using

$$\overline{x} \pm t_{df}^{*} SE$$

Where *df* is the degrees of freedom (*df* = *n* -1)

t-Distributions



t-test in R

The `pt` and `qt` will give you the *p*-value and critical value from the *t*-distribution, respectively.

Critical value for p = 0.05, degrees of freedom = 10

```
qt(0.025, df = 10)
```

```
## [1] -2.228139
```

p-value for a critical value of 2, degrees of freedom = 10

```
pt(2, df=10)
```

```
## [1] 0.963306
```

The `t.test` function will calculate a null hypothesis test using the *t*-distribution.

```
t.test(Math.SAT ~ Sex, data = sat)
```

```
##
##     Welch Two Sample t-test
##
## data:  Math.SAT by Sex
## t = -1.9117, df = 158.01, p-value = 0.05773
## alternative hypothesis: true difference in means bet
## 95 percent confidence interval:
##  -59.3527145   0.9685682
## sample estimates:
## mean in group F mean in group M
##        597.6829        626.8750
```

One Minute Paper

1. What was the most important thing you learned during this class?
2. What important question remains unanswered for you?

https://forms.gle/bz8GvYWfKdMggRv38