# ANOVA: First Examples, Model Statement, t tests

Sleuth3 Sections 6.2 and 5.2

## Examples, First Look at Data

### Example 1: Sepal Width of Iris Flowers

#### Study Overview

We have measurements of the characteristics of 150 iris flowers, 50 each of three different species:

- Iris setosa is found in the arctic, including Alaska and Maine in the United States, Canada, Russia, northern China, Korea and other northern countries.
- Iris versicolor is found in the eastern United States and eastern Canada.
- Iris virginica is found in the eastern United States

It's not clear how the flowers were selected for the sample; probably not as a representative sample though. The original purpose of this study was to develop methods for identifying a flower's species based on physical measurements of its characteristics. One of the characteristics that was measured for each flower in our sample was the width of the flower's sepal; the sepal is the part of the plant that sits just below the petals and supports them.

We will investigate differences in the widths of the flowers' sepals for the different species.

#### Look at the Data:

```
library(readr)
library(dplyr)
library(ggplot2)
library(gridExtra)
library(mosaic)
```

```
## Warning: package 'mosaic' was built under R version 3.5.2
```

```
## Warning: package 'ggformula' was built under R version 3.5.2
```

```
options("pillar.sigfig" = 10) # print 10 significant digits in summarize output

head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
dim(iris)
```

```
## [1] 150   5
```

```
iris %>%
  count(Species)
```

```
## # A tibble: 3 x 2
##   Species        n
##   <fct>      <int>
## 1 setosa        50
## 2 versicolor    50
## 3 virginica     50
```

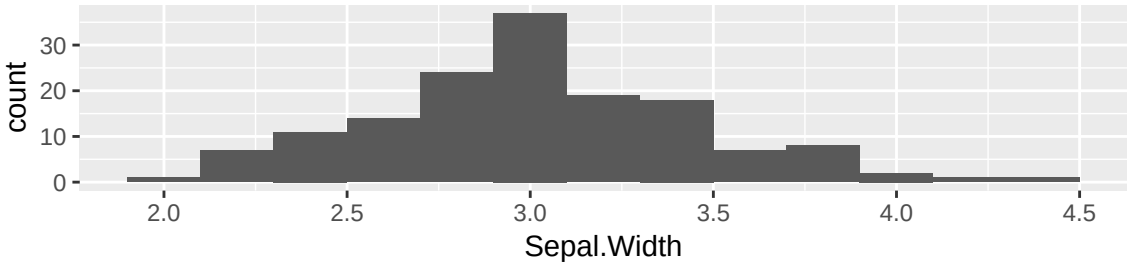
General template of code to make plots with ggplot2:

```
ggplot(data = <name of data frame>,
  mapping = aes(x = <variable for x axis>,
    y = <variable for y axis>,
    color = <variable for color lines>,
    fill = <variable for color area>,
    )) +
  geom_<geometry type>() +
  <optional other things like facetting, axis labels, ...>
```

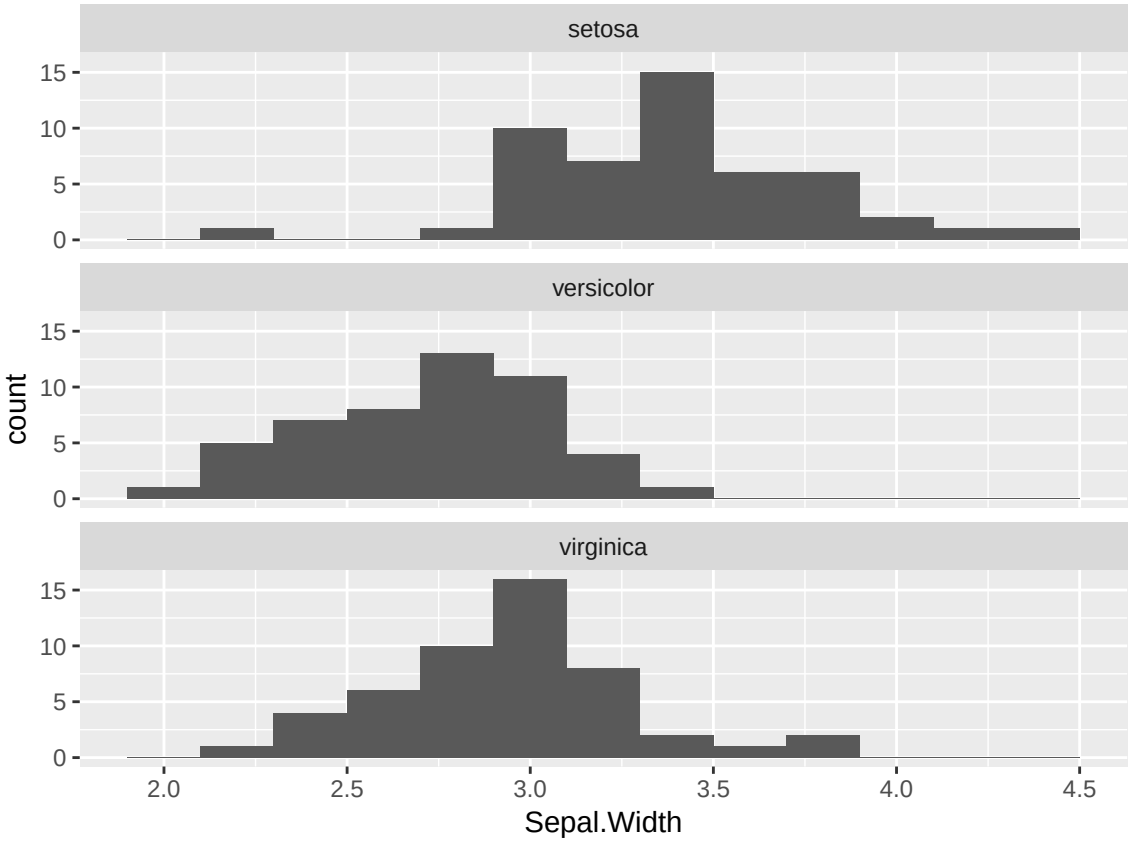
One pair of figures to sum up ANOVA:

- **Key idea** Less variability within each group than across all flowers

```
ggplot(data = iris, mapping = aes(x = Sepal.Width)) +
  geom_histogram(binwidth = 0.2)
```

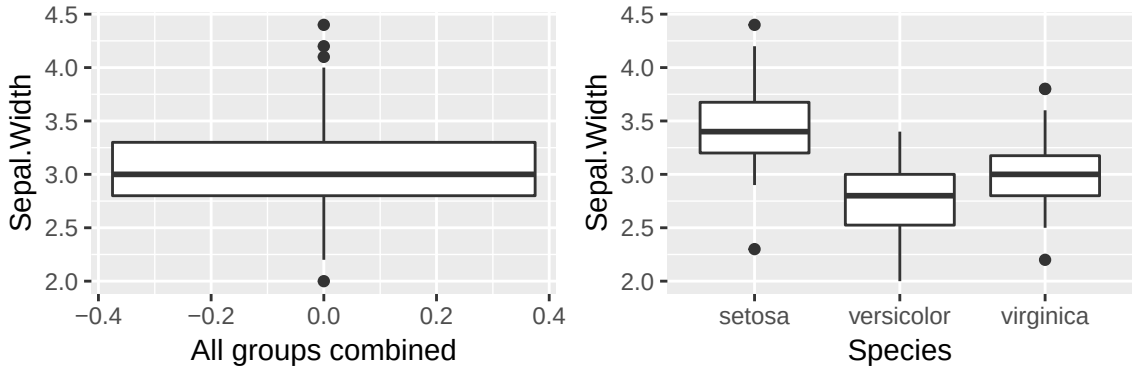



```
ggplot(data = iris, mapping = aes(x = Sepal.Width)) +
  geom_histogram(binwidth = 0.2) +
  facet_wrap( ~ Species, ncol = 1)
```



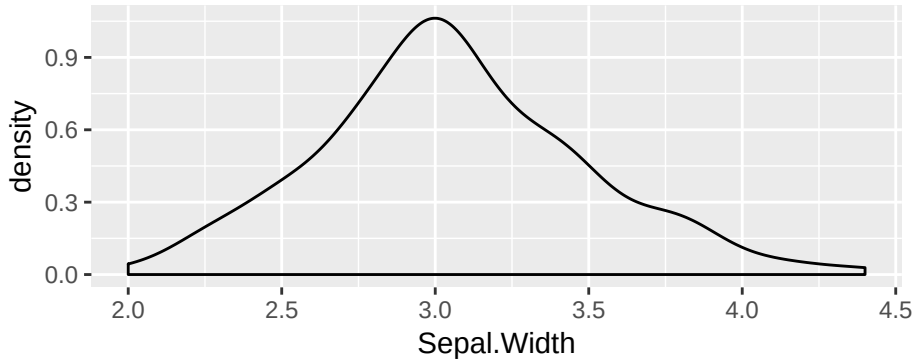
Box plots:

```
plot_combined <- ggplot(data = iris, mapping = aes(y = Sepal.Width)) +
  geom_boxplot() +
  xlab("All groups combined")

plot_bygroup <- ggplot(data = iris, mapping = aes(y = Sepal.Width, x = Species)) +
  geom_boxplot()

grid.arrange(plot_combined, plot_bygroup, ncol = 2)
```

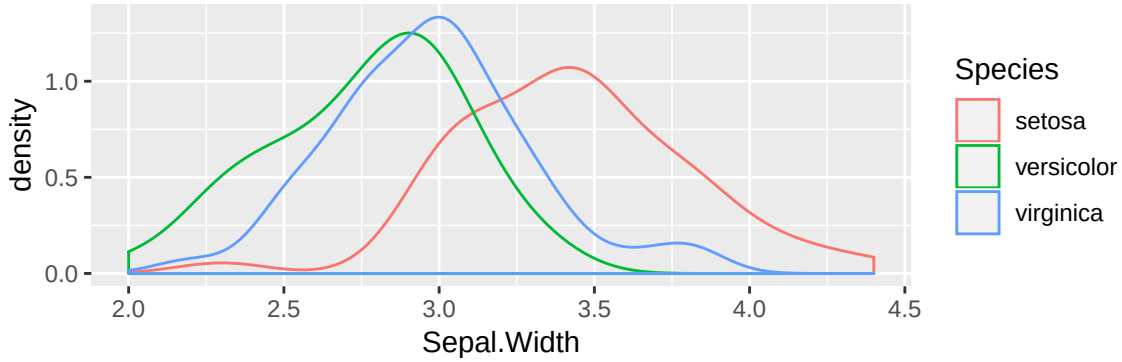



Same idea, but with density plots instead of histograms:

```
ggplot(data = iris, mapping = aes(x = Sepal.Width)) +
  geom_density()
```




```
ggplot(data = iris, mapping = aes(x = Sepal.Width, color = Species)) +
  geom_density()
```



```r
# Calculate overall sample mean and standard deviation
iris %>%
  summarize(
    mean = mean(Sepal.Width),
    sd = sd(Sepal.Width)
  )
```

```
##       mean        sd
## 1 3.057333 0.4358663
```

```r
# Calculate sample means and standard deviations separately for each species
iris %>%
  group_by(Species) %>%
  summarize(
    mean = mean(Sepal.Width),
    sd = sd(Sepal.Width)
  )
```

```
## # A tibble: 3 x 3
##   Species     mean    sd
##   <fct>      <dbl> <dbl>
## 1 setosa      3.43 0.379
## 2 versicolor  2.77 0.314
## 3 virginica   2.97 0.322
```

**Parameters:**

$\mu_{setosa}$ = Average sepal width among all setosa flowers (in the region where the flowers in the sample were found?)

$\mu_{versicolor}$ = Average sepal width among all setosa flowers (in the region where the flowers in the sample were found?)

$\mu_{virginica}$ = Average sepal width among all setosa flowers (in the region where the flowers in the sample were found?)

**Some questions we might ask:**

- Overall, is there a difference in mean post-test scores for the three groups? (F test)
  - $H_0 : \mu_1 = \mu_2 = \mu_3$
- Is the mean for setosa flowers different from the mean for versicolor flowers? (t test)
  - $H_0 : \mu_1 = \mu_2$, or $\mu_1 - \mu_2 = 0$
- Is the mean for setosa flowers different from the mean for virginica flowers? (t test)
  - $H_0 : \mu_1 = \mu_3$, or $\mu_1 - \mu_3 = 0$
- Is the mean for versicolor flowers different from the mean for virginica flowers? (t test)
  - $H_0 : \mu_2 = \mu_3$, or $\mu_2 - \mu_3 = 0$
- Is the mean for the setosa flower, found in the arctic, different from the mean for non-arctic flowers? (t test)
  - $H_0 : \frac{1}{2}(\mu_2 + \mu_3) = \mu_1$, or $\frac{1}{2}(\mu_2 + \mu_3) - \mu_1 = 0$

## Example 2: Women underrepresented on juries?

Quote from our book:

"In 1968, Dr. Benjamin Spock was tried in Boston on charges of conspiring to violate the Selective Service Act by encouraging young men to resist being drafted into military service for Vietnam. The defence in the case challenged the method of jury selection claiming that women were underrepresented. Boston juries are selected in three stages. First 300 names are selected at random from the City Directory, then a venire of 30 or more jurors is selected from the initial list of 300 and finally, an actual jury is selected from the venire in a nonrandom process allowing each side to exclude certain jurors. There was one woman on the venire and no women on the final list. The defence argued that the judge in the trial had a history of venires in which women were systematically underrepresented and compared the judge's recent venires with the venires of six other Boston area district judges."

```
library(readr)
library(ggplot2)
library(dplyr)

juries <- read_csv("http://www.evanlray.com/data/sleuth3/ex0502_women_jurors.csv")
```

```
dim(juries)
```

```
## [1] 46  2
```

```
head(juries)
```

```
## # A tibble: 6 x 2
##   Percent Judge
##     <dbl> <chr>
## 1     6.4 Spock's
## 2     8.7 Spock's
## 3    13.3 Spock's
## 4    13.6 Spock's
## 5    15   Spock's
## 6    15.2 Spock's
```

```
juries %>% count(Judge)
```

```
## # A tibble: 7 x 2
##   Judge       n
##   <chr>   <int>
## 1 A           5
## 2 B           6
## 3 C           9
## 4 D           2
## 5 E           6
## 6 F           9
## 7 Spock's     9
```

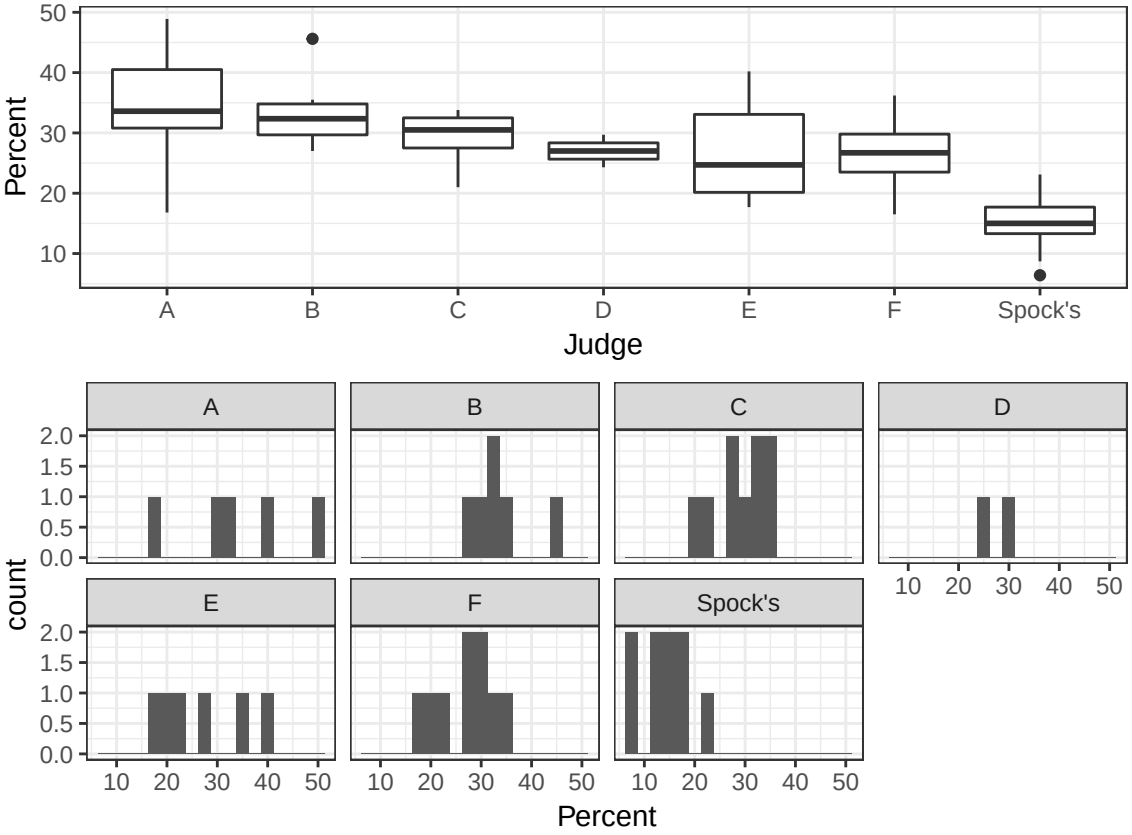
Each observational unit is a venire (jury pool) assembled by one of 7 judges in Boston at this time. We have information about a total of 46 venires across the 7 judges. For each venire, we have recorded:

- The percent of potential jurors in the venire who were women
- Which judge assembled that venire

**Initial Plots**

The GitHub repository URL for this lab is: `https://github.com/mhc-stat242-s2019/Lab2.git`

In R, recreate at least one of the plots below. Also calculate the group means and standard deviations.

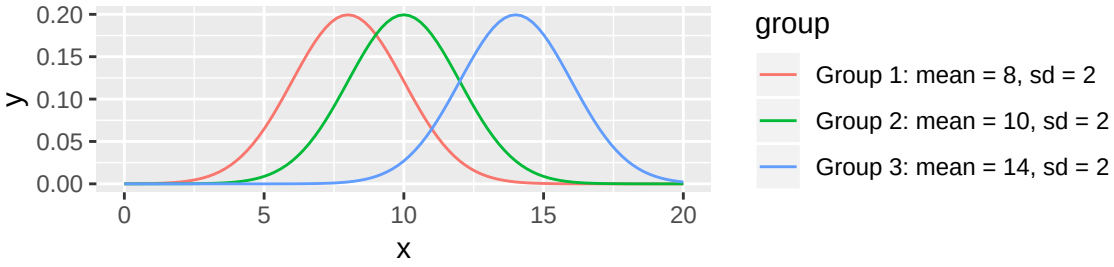


# ANOVA Model, Hypothesis Tests, and Confidence Intervals

## The ANOVA Model

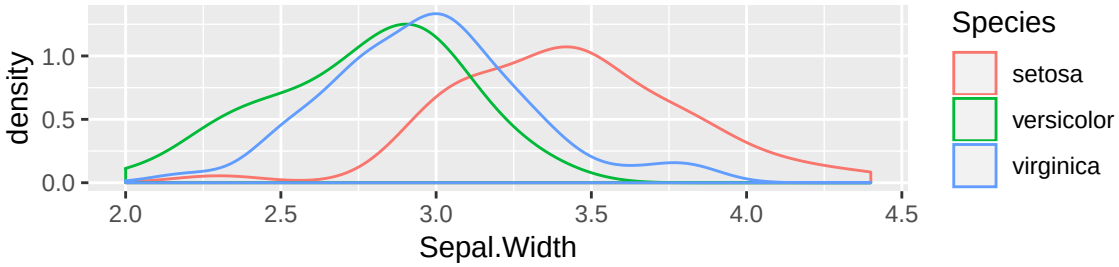
- We have $I$ groups ($I = 3$ for iris example)
- Sample size of $n_i$ for group $i$, total sample size $n = n_1 + n_2 + \cdots + n_I$
- Observations in group $i$ follow a Normal($\mu_i$, $\sigma^2$) distribution
  - (Potentially) different mean for each group
  - Same variance across all groups
- All observations are independent of each other: knowing that one is above its group mean doesn't tell you whether or not another is above its group mean.

**Theoretical model:**




**Compare to the plot for our iris data:**

```
ggplot(data = iris, mapping = aes(x = Sepal.Width, color = Species)) +
  geom_density()
```




## Hypotheses

**General case:**

$H_0 : C_1\mu_1 + C_2\mu_2 + \cdots + C_I\mu_I = 0$

$H_A : C_1\mu_1 + C_2\mu_2 + \cdots + C_I\mu_I \neq 0$

- Notation: Book defines $\gamma$ ("gamma") to be the linear combination we are testing:

$\gamma = C_1\mu_1 + C_2\mu_2 + \cdots + C_I\mu_I$

**Iris example:**

State the hypotheses and the values of the constants $C_1$, $C_2$, and $C_3$ for a test of whether the mean for setosa flowers is different from the mean for versicolor flowers.

## Doing the test and finding a 95% confidence interval

Here's the R code to do this test

```
library(gmodels)

model_fit <- lm(Sepal.Width ~ Species, data = iris)
fit.contrast(model_fit, "Species", c(1, -1, 0), conf = 0.95)
```

```
##                       Estimate Std. Error  t value     Pr(>|t|)  lower CI
## Species c=( 1 -1 0 )     0.658 0.06793755 9.685366 1.832489e-17 0.5237396
##                        upper CI
## Species c=( 1 -1 0 ) 0.7922604
## attr(,"class")
## [1] "fit_contrast"
```

**What is the result of the hypothesis test?**

**What is the interpretation of the confidence interval?**

**If you wanted to conduct a hypothesis test of whether the mean for the setosa flower, found in the arctic, is different from the mean across both non-arctic flowers. What constants $C_1$, $C_2$, and $C_3$ would you use?**

The null hypothesis is $H_0 : \frac{1}{2}(\mu_2 + \mu_3) - \mu_1 = 0$

## Example 2

**(a) Conduct a hypothesis test of the claim that the mean percent of potential jurors who are women in venires assembled by Spock's judge is the same as the mean percent of potential jurors who are women in venires assembled by judge A. Also find and report a 95% confidence interval for the difference in means for those two judges. State your null and alternative hypotheses in terms of equations and written sentences. What are the constants $C_1, \ldots, C_I$ to use for this procedure?**

**(b) Conduct a hypothesis test of the claim that the mean percent of potential jurors who are women in venires assembled by Spock's judge is the same as the mean percent of potential jurors who are women across all 6 other judges. Also find and report a 95% confidence interval for the difference in means between Spock's judge and the average across all 6 other judges. State your null and alternative hypotheses in terms of equations and written sentences. What are the constants $C_1, \ldots, C_I$ to use for this procedure?**

## What's going on behind the scenes?

A bunch of what happens below is more detailed that you really need to know. I will highlight the things that I will ask you to know at the end.

## Sample-Based Estimate of $\gamma$

### General case:

We estimate $\gamma$ with the equivalent combination of group-based sample means:

$g = \hat{\gamma} = C_1\bar{Y}_1 + C_2\bar{Y}_2 + \cdots + C_I\bar{Y}_I$

### Iris example:

```
iris %>%
  group_by(Species) %>%
  summarize(
    group_mean = mean(Sepal.Width)
  )
```

```
## # A tibble: 3 x 2
##   Species    group_mean
##   <fct>           <dbl>
## 1 setosa           3.43
## 2 versicolor       2.77
## 3 virginica        2.97
```

```
3.428 - 2.770
```

```
## [1] 0.658
```

## SD(g) and SE(g)

### SD(g)

- It can be shown (see office hours or Math 342) that

$SD(g) = \sigma\sqrt{\frac{C_1^2}{n_1} + \frac{C_2^2}{n_2} + \cdots + \frac{C_I^2}{n_I}}$

- The population standard deviation $\sigma$ is unknown!

### Estimating $\sigma$

- We have $I$ different estimates $s_i$ of $\sigma$ from the $I$ different groups, each with degrees of freedom $n - 1$:

```
iris %>%
  group_by(Species) %>%
  summarize(
    group_sd = sd(Sepal.Width),
    group_var = var(Sepal.Width)
  )
```

```
## # A tibble: 3 x 3
##   Species    group_sd group_var
##   <fct>         <dbl>     <dbl>
## 1 setosa        0.379    0.144
## 2 versicolor    0.314    0.0985
## 3 virginica     0.322    0.104
```

- To use all the data as effectively as possible, define the *pooled estimate of variance* as a weighted average of these estimates, weighted by each group's degrees of freedom.

$s^2_{pooled} = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2+\cdots+(n_I-1)s_I^2}{(n_1-1)+(n_2-1)+\cdots+(n_I-1)}$

- This is an estimate of $\sigma^2$; estimate $\sigma$ with $s_{pooled} = \sqrt{s^2_{pooled}}$
- The degrees of freedom is the sum of the degrees of freedom from the individual groups:

$df = (n_1 - 1) + (n_2 - 1) + \cdots + (n_I - 1) = n - I$

### SE(g), General case

$SE(g) = s_{pooled}\sqrt{\frac{C_1^2}{n_1} + \frac{C_2^2}{n_2} + \cdots + \frac{C_I^2}{n_I}}$

### SE(g), Iris example

First, find pooled standard deviation:

```
s2_pooled <- ((50 - 1)*0.144 + (50 - 1)*0.098 + (50 - 1)*0.104)/(150 - 3)
s_pooled <- sqrt(s2_pooled)
s2_pooled
```

```
## [1] 0.1153333
```

```
s_pooled
```

```
## [1] 0.3396076
```

Then multiply by $\sqrt{\frac{C_1^2}{n_1} + \frac{C_2^2}{n_2} + \cdots + \frac{C_I^2}{n_I}}$

```
SE_g <- s_pooled * sqrt(1/50 + 1/50)
SE_g
```

```
## [1] 0.06792152
```

## t statistic

### General case:

- Our $t$ statistic will be

$t = \frac{g-\gamma^{null}}{SE(g)} = \frac{g}{SE(g)}$

If the null hypothesis is true, $t \sim t_{n-I}$

- Its degrees of freedom is $n - I$

### Iris example:

Calculating the $t$ statistic:

```
g <- 3.428 - 2.770
g/SE_g
```

```
## [1] 9.687651
```

Degrees of freedom: 150 - 3 = 147

Calculate p-value based on $t$ distribution

```
2 * pt(9.688, df = 147, lower.tail = FALSE)
```

```
## [1] 1.803882e-17
```

## Confidence Interval

### General case:

$g \pm t^* SE(g)$

For a 95% confidence interval, $t^*$ is the 97.5th percentile of a $t_{n-I}$ distribution.

### Iris example:

```
qt(0.975, df = 147)
```

```
## [1] 1.976233
```

```
g - 1.976 * SE_g
```

```
## [1] 0.5237871
```

```
g + 1.976 * SE_g
```

```
## [1] 0.7922129
```

### What out of this section do you actually need to know?

- The estimated difference in group means is the difference in sample means for those groups.
- The pooled estimate of variance combines estimates from different groups; this only makes sense if the variance is approximately the same for all groups.
- The $t$ statistic is

$t = \frac{g - \gamma^{null}}{SE(g)} = \frac{g}{SE(g)}$

- Its degrees of freedom is $n - I$
- A $t$-based confidence interval for $\gamma$ is therefore

$g \pm t^* SE(g)$

where $t^*$ is the 97.5th percentile of a $t_{n-I}$ distribution for a 95% confidence interval.