# Section 7.5
# Comparing several means: ANOVA

# Comparing many means with ANOVA

Example: Comparing Exercise and Weight loss

| *Treatment* | *Sample Size* | *Sample Mean* | *Sample St. Dev* |
|---|---|---|---|
| *Long exercise periods* | 37 | 10.6 | 4.6 |
| *Short exercise periods* | 36 | 11.2 | 4.2 |
| *Short periods with equipment* | 42 | 9.3 | 4.7 |

# Questions

- Is there a significant difference in the sample means?
- Is the difference reasonably explained by chance?

# Three 95% confidence intervals

We can build a 95% confidence interval for each sample using

$$\overline{x} \pm t^* s/\sqrt{n}$$

- Group 1: (9.1 to 12.1)
- Group 2: (9.8 to 12.6)
- Group 3: (7.8 to 10.8)

My hunch: These confidence intervals all overlap, which suggests that there is NOT convincing evidence that at least one of the population means is different than the others, but can a single test establish this?

# ANOVA Can

The Idea of Analysis of Variance (ANOVA)

- The sample mean from a random sample with a small standard deviation is more likely to be close to the population mean than a sample mean from a sample having a large standard deviation.
- The test for a comparison between more than two means must measure how far apart the sample means are relative to how much difference there is between the individual data items.

# The F test statistic, informally

- Informally, the test statistic we calculate from our data, which we call $F$, gives a positive number quantifying how far apart the sample means are relative to the variability of the individual observations.
- $F$ is computed under the assumption that all population means in the study are equal (this assumption is the null hypothesis).

$$F = \frac{\text{variation among the sample means}}{\text{variation among individuals in the same sample}}$$

# The $F$ distribution

- The $F$ distribution is actually a family of distributions, like the $t$-distributions and the chi-square distributions.
- An $F$ distribution is described by two parameters. Let's assume we're comparing $k$ populations, and we have $N$ total data values. Then we have:
  - Numerator degrees of freedom $= k - 1$
  - Denominator degrees of freedom $= N - k$
- $F$ distributions are skewed right, taking only non-negative values, and the shape can vary depending on the numerator and denominator df.

# The $F$ Statistic (more formally)

Suppose we are comparing $k$ populations

- We draw a random sample from each population
- Let $n_i$ = the sample size from population $i$
- Let $\overline{x}_i$ = the sample mean from population $i$
- Let $s_i$ = the sample standard deviation from population $i$
- Let $N = n_1 + n_2 + \cdots + n_k$ denote overall sample size
- The overall sample mean is

$$\overline{x} = \frac{n_1\overline{x}_1 + n_2\overline{x}_2 + \cdots + n_k\overline{k}_k}{N}$$

# The $F$ Statistic (for more than two means)

$$F = \frac{\text{MSG}}{\text{MSE}}$$

where MSG is "mean square for groups":

$$\text{MSG} = \frac{n_1(\overline{x}_1 - \overline{x})^2 + n_2(\overline{x}_2 - \overline{x})^2 + \cdots + n_k(\overline{x}_k - \overline{x})^2}{k - 1}$$

and MSE is "mean square for error":

$$\text{MSE} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2}{N - k}.$$

## The $F$ Statistic (for more than two means)

This formula is computationally involved, so use R to do the calculation, but it is worthwhile to see what the formula is really capturing.

- The numerator, MSG term, adds up terms that look like this: $(\overline{x}_i - \overline{x})^2$. That is, MSG computes variation **among** the different groups, and is larger when the sample mean for one group is dramatically different than the overall sample mean.
- The denominator, MSE term, adds up the sample variances for all the groups, weighting each one by the size of the sample for that group. MSE computes variation in values that is happening **within** the groups.
- $F$ is the ratio of variation of among groups to variation within groups. So a large value for $F$ means variation in values from group to group outweighs the variation in values that is happening within the groups.
- The larger $F$ gets, the stronger the evidence is that the population means in the different groups are not all equal.

# Assumptions for using ANOVA

- As usual, we assume independent samples
- In theory, each population has a normal distribution
- In theory, all populations have the same standard deviation.
- In practice, the test can still be reliable if these last two conditions aren't exactly met.
- **Rule of Thumb**: The largest sample standard deviation should not be more than twice as big as the smallest sample standard deviation.
- **Suggestion**: When designing a study with the expectation of using ANOVA, try to take samples of the same size from all groups you want to compare.

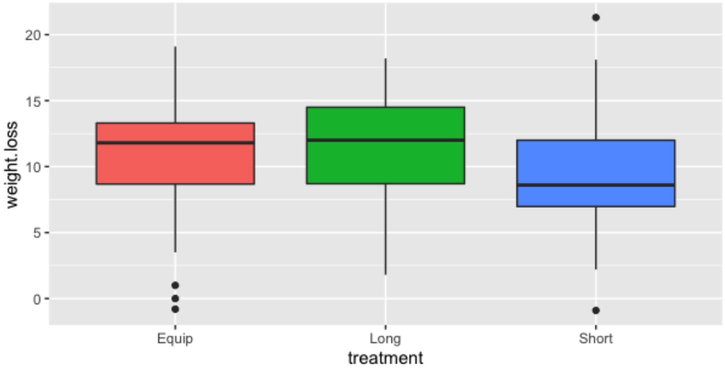
# The Weight loss Example in R

Data entered into R as a data frame with two variables: `treatment` and `weight.loss`

| | treatment | weight.loss |
|---|---|---|
| 1 | Long | 13.5 |
| 2 | Long | 13.9 |
| 3 | Long | 14.9 |
| 4 | Long | 11.4 |
| ⋮ | ⋮ | ⋮ |
| 114 | Equip | 9.9 |
| 115 | Equip | 8.6 |

# The Weight loss Example in R

Side by side boxplots:



# The Weight loss Example in R

Recall the summary statistics

| *Treatment* | *Sample Size* | *Sample Mean* | *Sample St. Dev* |
|---|---|---|---|
| *Long exercise periods* | 37 | 10.6 | 4.6 |
| *Short exercise periods* | 36 | 11.2 | 4.2 |
| *Short periods with equipment* | 42 | 9.3 | 4.7 |

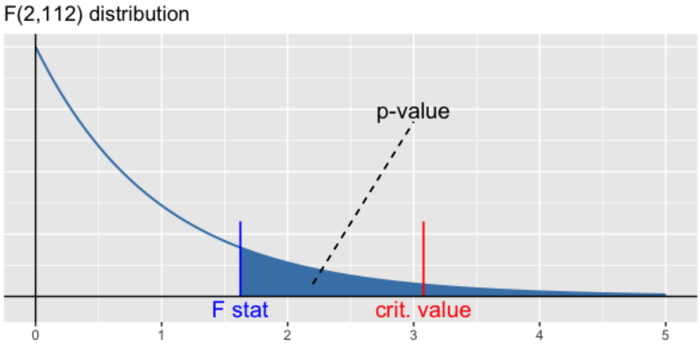
# The Weight loss Example in R

Conducting ANOVA in R right from the raw data (the data frame is called `df`):

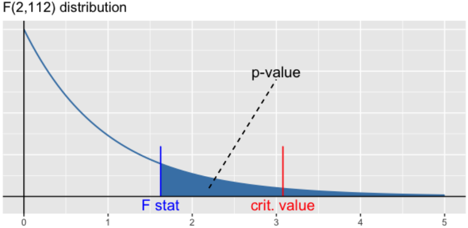
```
anova(lm(weight.loss~treatment,df))
```

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| treatment | 2 | 66.11 | 33.06 | 1.62 | 0.2017 |
| Residuals | 112 | 2279.92 | 20.36 | | |

# The Weight loss Example in R



# The Weight loss Example in R




**Conclusion**: We do not reject the null hypothesis since the p-value is greater than $\alpha = .05$. We do not have significant evidence that at least one of the population means is different than the others. If, the mean weight loss for all groups is equal in all populations (that is, for all exercise groups), we would still have about a 20% chance of obtaining data with such different sample means as what we had in our study.