A Second Semester Statistics Course with R

Mark Greenwood and Katherine Banner

2017-07-17

Contents

Acknowledgments

Chapter 1

Placeholder

Chapter 2

Placeholder

Chapter 3

One-Way ANOVA

3.1 Situation

In Chapter **??**, tools for comparing the means of two groups were considered. More generally, these methods are used for a quantitative response and a categorical explanatory variable (group) which had two and only two levels. The full prisoner rating data set actually contained three groups (Figure 3.1 with *Beautiful*, *Average*, and *Unattractive* rated pictures randomly assigned to the subjects for sentence ratings. In a situation with more than two groups, we have two choices. First, we could rely on our two group comparisons, performing tests for every possible pair (*Beautiful* vs *Average*, *Beautiful* vs *Unattractive*, and *Average* vs *Unattractive*). We spent Chapter **??** doing inferences for differences between *Average* and *Unattractive*. The other two comparisons would lead us to initially end up with three p-values and no direct answer about our initial question of interest – is there some overall difference in the average sentences provided across the groups? In this chapter, we will learn a new method, called ***Analysis of Variance***, or ***One-Way ANOVA*** since there is just one[1] grouping variable. After we perform our One-Way ANOVA test for overall evidence of a difference, we will revisit the comparisons similar to those considered in Chapter **??** to get more details on specific differences among *all* the pairs of groups – what we call ***pair-wise comparisons***. An issue is created when you perform many tests simultaneously and we will augment our previous methods with an adjusted method for pairwise comparisons to make our results valid called ***Tukey's Honest Significant Difference***.

To make this more concrete, we return to the original MockJury data, making side-by-side boxplots and beanplots (Figure 3.1 as well summarizing the sentences for the three groups using `favstats`.

```
require(heplots)
require(mosaic)
data(MockJury)
par(mfrow=c(1,2))
boxplot(Years~Attr,data=MockJury)
require(beanplot)
beanplot(Years~Attr,data=MockJury,log="",col="bisque",method="jitter")

favstats(Years~Attr,data=MockJury)
```

```
##            Attr min Q1 median   Q3 max     mean       sd  n missing
## 1    Beautiful   1  2      3  6.5  15 4.333333 3.405362 39       0
## 2      Average   1  2      3  5.0  12 3.973684 2.823519 38       0
## 3 Unattractive   1  2      5 10.0  15 5.810811 4.364235 37       0
```

There are slight differences in the sample sizes in the three groups with 37 *Unattractive*, 38 *Average* and 39 *Beautiful* group responses, providing a data set has a total sample size of $N = 114$. The *Beautiful* and *Average* groups do not appear to be very different with means of 4.33 and 3.97 years. In Chapter **??**, we found moderate evidence regarding the difference in *Average*and *Unattractive*. It is less clear whether we might find evidence of a difference between *Beautiful* and *Unattractive* groups since we are comparing means of 5.81 and 4.33 years. All the distributions appear to be right skewed with relatively similar shapes. The variability in *Average* and *Unattractive* groups seems like it could be slightly different leading to an overall concern of whether the variability is the same in all the groups.

[1]In Chapter 4, methods are discussed for when there are two categorical explanatory variables that is called the Two-Way ANOVA and related ANOVA tests are used in Chapter 8 for working with extensions of these models.

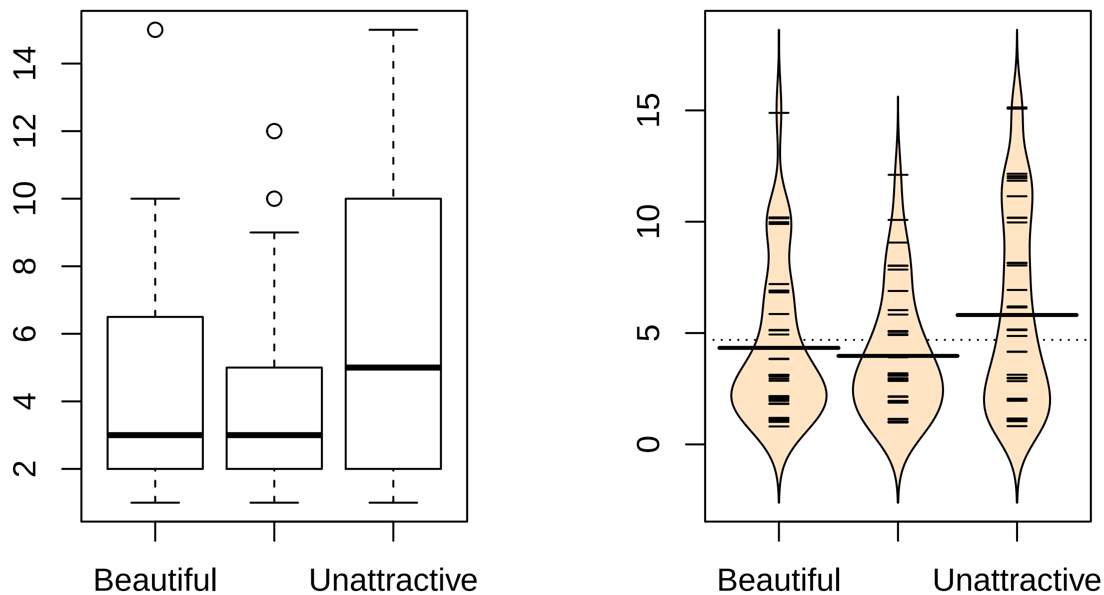


Figure 3.1: Boxplot and beanplot of the sentences (years) for the three treatment groups.

3.2 Linear model for One-Way ANOVA (cell-means and reference-coding)

We introduced the statistical model $y_{ij} = \mu_j + \varepsilon_j$ in Chapter **??** for the situation with $j = 1$ or 2 to denote a situation where there were two groups and, for the model that is consistent with the alternative hypothesis, the means differed. Now we have three groups and the previous model can be extended to this new situation by allowing j to be 1, 2, or 3. Now that we have more than two groups, we need to admit that what we were doing in Chapter **??** was actually fitting what is called a ***linear model***. The linear model assumes that the responses follow a normal distribution with the linear model defining the mean, all observations have the same variance, and the parameters for the mean in the model enter linearly. This last condition is hard to explain at this level of material – it is sufficient to know that there models where the parameters enter the model nonlinearly and that they are beyond the scope of this course. The result of this constraint is that we will be able to use the same general modeling framework for the methods introduced in Chapters 3, 4, 6, 7, and 8.

As in Chapter **??**, we have a null hypothesis that defines a situation (and model) where all the groups have the same mean. Specifically, the ***null hypothesis*** in the general situation with J groups ($J \geq 2$) is to have all the **true** group means equal,

$$H_0 : \mu_1 = \ldots \mu_J.$$

This defines a model where all the groups have the same mean so it can be defined in terms of a single mean, μ, for the i^{th} observation from the j^{th} group as $y_{ij} = \mu + \varepsilon_{ij}$. This is not the model that most researchers want to be the final description of their study as it implies no difference in the groups. There is more caution required to specify the alternative hypothesis with more than two groups. The ***alternative hypothesis*** needs to be the logical negation of this null hypothesis of all groups having equal means; to make the null hypothesis false, we only need one group to differ but more than one group could differ from the others. Essentially, there are many ways to "violate" the null hypothesis so we choose some delicate wording for the alternative hypothesis when there are more than 2 groups. Specifically, we state the alternative as

$$H_A : \text{ Not all } \mu_j \text{ are equal}$$

or, in words, **at least one of the true means differs among the J groups**. You will be attracted to trying to say that all means are different in the alternative but we do not put this strict a requirement in place to reject the null hypothesis. The alternative model allows all the true group means to differ but does require that they differ with

$$\mu_j + \varepsilon_{ij}.$$

This linear model states that the response for the i^{th} observation in the j^{th} group, $\mathbf{y_{ij}}$, is modeled with a group j ($j = 1, \ldots, J$) population mean, μ_j, and a random error for each subject in each group ε_{ij}, that we assume follows a normal distribution and that all the random errors have the same variance, σ^2. We can write the assumption about the random errors, often called the ***normality assumption***, as $\varepsilon_{ij} \sim N(0, \sigma^2)$. There is a second way to write out this model that allows extension to more complex models discussed below, so we need a name for this version of the model. The model written in terms of the μ_j's is called the **cell means model** and is the easier version of this model to understand.

One of the reasons we learned about beanplots is that it helps us visually consider all the aspects of this model. In the right panel of Figure 3.1, we can see the wider, bold horizontal lines that provide the estimated group means. The bigger the differences in the sample means, the more likely we are to find evidence against the null hypothesis. You can also see the null model on the plot that assumes all the groups have the same as displayed in the dashed horizontal line at 4.7 years (the R code below shows the overall mean of *Years* is 4.7). While the hypotheses focus on the means, the model also contains assumptions about the distribution of the responses – specifically that the distributions are normal and that all the groups have the same variability. As discussed previously, it appears that the distributions are right skewed and the variability might not be the same for all the groups. The boxplot provides the information about the skew and variability but since it doesn't display the means it is not directly related to the linear model and hypotheses we are considering.

```r
mean(MockJury$Years)
```

```
## [1] 4.692982
```

There is a second way to write out the One-Way ANOVA model that provides a framework for extensions to more complex models described in Chapter 4 and beyond. The other ***parameterization*** (way of writing out or defining) of the model is called the **reference-coded model** since it writes out the model in terms of a ***baseline group*** and deviations from that baseline or reference level. The reference-coded model for the i^{th} subject in the j^{th} group is $y_{ij} = \boldsymbol{\alpha} + \boldsymbol{\tau_j} + \varepsilon_{ij}$ where $\boldsymbol{\alpha}$ (alpha) is the true mean for the baseline group (first alphabetically) and the $\boldsymbol{\tau_j}$ (tau j) are the deviations from the baseline group for group j. The deviation for the baseline group, $\boldsymbol{\tau_1}$, is always set to 0 so there are really just deviations for groups 2 through J. The equivalence between the two models can be seen by considering the mean for the first, second, and J^{th} groups in both models:

	Cell means:	Reference-coded:
Group 1:	${\color{red}{\mu_1}}$	${\color{purple}{\boldsymbol{\alpha}}}$
Group 2:	${\color{red}{\mu_2}}$	${\color{red}{\boldsymbol{\tau_2}}}$
$\ldots$	$\ldots$	$\ldots$
Group J:	${\color{red}{\mu_J}}$	${\color{purple}{\boldsymbol{\tau_J}}}$

The hypotheses for the reference-coded model are similar to those in the cell-means coding except that they are defined in terms of the deviations, $\boldsymbol{\tau_j}$. The null hypothesis is that there is no deviation from the baseline for any group – that all the $\boldsymbol{\tau_j}$'s $= 0$,

$$\boldsymbol{H_0 : \tau_2 = \ldots = \tau_J = 0}.$$

The alternative hypothesis is that at least one of the deviations is not 0,

$$\boldsymbol{H_A :} \textbf{ Not all } \boldsymbol{\tau_j} \textbf{ equal } \boldsymbol{0}.$$

In this chapter, you are welcome to use either version (unless we instruct you otherwise) but we have to use the reference-coding in subsequent chapters. The next task is to learn how to use R's linear model `lm` function to get estimates of the parameters in each model, but first a quick review of these new ideas:

Cell Means Version

- $H_0 : \mu_1 = \ldots \mu_J$ $\qquad$ $H_A :$ Not all μ_j equal
- Null hypothesis in words: No difference in the true means between the groups.
- Null model $y_{ij} = \mu_j + \varepsilon_{ij}$
- Alternative hypothesis in words: At least one of the true means differs between the groups.
- Alternative model: $y_{ij} = \mu_j + \varepsilon_{ij}$.

Reference-coded Version

- $H_0 : \boldsymbol{\tau_2 \ldots \tau_J = 0}$ $\qquad$ $H_A :$ Not all τ_j equal
- Null hypothesis in words: No deviation of the true mean for any groups from the baseline group.
- Null model: $y_{ij} = \boldsymbol{\alpha} + \tau_j + \varepsilon_{ij}$
- Alternative hypothesis in words: At least one of the true deviations is different from 0 or that at least one group has a different true mean than the baseline group.
- Alternative model: $y_{ij} = \boldsymbol{\alpha + \tau_j} + \varepsilon_{ij}$

In order to estimate the models discussed above, the `lm` function is used. If you look closely in the code for the rest of the book, any model for a quantitative response will use this function, suggesting a common thread in the most commonly used statistical models. The `lm` function continues to use the same format as previous functions, `lm(Y~X, data=datasetname)`. It ends up that this code will give you the reference-coded version of the model by default (R thinks it is that important!). We want to start with the cell-means version of the model, so we have to override the standard technique and add a "`-1`" to the formula interface to tell R that we want to the cell-means coding. Generally, this looks like `lm(Y~X-1 , data=datasetname).` Once we fit a model in R, the `summary` function run on the model provides a useful "summary" of the model coefficients and a suite of other potentially interesting information. When fitting this version of the One-Way ANOVA model, you will find a row of output for each group relating the μ_j's. The output contains columns for an estimate (`Estimate`), standard error (`Std.Error`), t-value (`t value`), and p-value (`Pr(>|t|)`). We'll learn to use all the output in the following material, but for now just focus on the estimates of the parameters that the function provides that we put in bold.

```r
lm1 <- lm(Years ~ Attr-1, data=MockJury)
summary(lm1)
```

```
##
## Call:
## lm(formula = Years ~ Attr - 1, data = MockJury)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8108 -2.8108 -0.9737  2.1892 10.6667
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## AttrBeautiful      4.3333     0.5730   7.563 1.23e-11
## AttrAverage        3.9737     0.5805   6.845 4.41e-10
## AttrUnattractive   5.8108     0.5883   9.878  < 2e-16
##
## Residual standard error: 3.578 on 111 degrees of freedom
## Multiple R-squared:  0.6449, Adjusted R-squared:  0.6353
## F-statistic: 67.21 on 3 and 111 DF,  p-value: < 2.2e-16
```

	Estimate	Std. Error	t value	Pr(>|t|)
AttrBeautiful	**4.33**	0.573	7.56	1.23e-11
AttrAverage	**3.97**	0.58	6.85	4.41e-10
AttrUnattractive	**5.81**	0.588	9.88	6.86e-17

In general, we denote estimated parameters with a hat over the parameter of interest to show that it is an estimate. For the true mean of group j, μ_j, we estimate it with $\hat{\mu}_j$, which is just the sample mean for group j, $\bar{x}_j$. The model suggests an estimate for each observation that we denote as $\hat{y}_{ij}$ that we will also call a ***fitted value*** based on the model being considered. The three estimates are bolded in the previous output, with the same estimate used for all observations in the same group. R tries to help you to sort out which row of output corresponds to which group by appending the group name l with the variable name. Here, the variable name was `Attr` and the first group alphabetically was *Beautiful*, so R provides a row labeled `AttrBeautiful` with an estimate of 4.3333. The sample means from the three groups can be seen to directly match that and the other two results.

```r
mean(Years ~ Attr, data=MockJury)
```

```
##    Beautiful      Average Unattractive
##     4.333333     3.973684     5.810811
```

The reference-coded version of the same model is more complicated but ends up giving the same results once we understand what it is doing. It uses a different parameterization to accomplish this so has different model output. Here is the model summary:

```r
lm2 <- lm(Years ~ Attr, data=MockJury)
summary(lm2)
```

```
##
## Call:
## lm(formula = Years ~ Attr, data = MockJury)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8108 -2.8108 -0.9737  2.1892 10.6667
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        4.3333     0.5730   7.563 1.23e-11
## AttrAverage       -0.3596     0.8157  -0.441   0.6601
## AttrUnattractive   1.4775     0.8212   1.799   0.0747
##
## Residual standard error: 3.578 on 111 degrees of freedom
## Multiple R-squared:  0.04754,    Adjusted R-squared:  0.03038
## F-statistic:  2.77 on 2 and 111 DF,  p-value: 0.067
```

The estimated model coefficients are $\hat{\alpha} = 4.333$ years, $t\hat{au}_2 = -0.3596$ years, $\hat{\tau}_3 = 1.4775$ years where R selected group 1 for *Beautiful*, 2 for *Average*, and 3 for *Unattractive*. The way you can figure out the baseline group (group 1 is *Beautiful* here) is to see which category label is *not present* in the output. **The baseline level is typically the first group label alphabetically**, but you should always check this. Based on these definitions, there are interpretations available for each coefficient. For $\hat{\alpha} = 4.333$ years, this is an estimate of the mean sentencing time for the *Beautiful*group. $\hat{\tau}_2 = -0.3596$ years is the deviation of the *Average* group's mean from the *Beautiful* groups mean (specifically, it is 0.36 years lower). Finally, $\hat{\tau}_3 = 1.4775$ years tells us that the *Unattractive* group mean sentencing time is 1.48 years higher than time. These interpretations lead directly to reconstructing the estimated means for each group by combining the baseline and pertinent deviations as shown in Table 3.2.

We can also visualize the results of our linear models using what are called ***term-plots*** or ***effect-plots*** (from the `effects` package; Fox (2003), (Fox et al., 2016)) as displayed in Figure 3.2. We don't want to use the word "effect" for these model components unless we have random assignment in the study design so we

Table 3.2: Constructing group mean estimates from the reference-coded linear model estimates.

Group	Formula	Estimates
Beautiful	$\hat{\alpha}$	**4.3333** years
Average	$\hat{\alpha}+\hat{\tau}_2$	4.3333 - 0.3596 = **3.974** years
Unattractive	$\hat{\alpha}+\hat{\tau}_3$	4.3333 + 1.4775 = **5.811** years

generically call these ***term-plots*** as they display terms or components from the model in hopefully useful ways to aid in model interpretation even in the presence of complicated model parameterizations. Specifically, these plots take an estimated model and show you its estimates along with 95% confidence intervals generated by the linear model. To make this plot, you need to install and load the `effects` package and then use `plot(allEffects(...))` functions together on the `lm` object called `lm2` that was estimated above. You can find the correspondence between the displayed means and the estimates that were constructed in Table 3.2.

```
require(effects)
plot(allEffects(lm2))
```

In order to assess evidence for having different means for the groups, we will compare either of the previous models (cell-means or reference-coded) to a null model based on the null hypothesis ($H_0 : \mu_1 = \ldots = \mu_J$) which implies a model of $y_{ij} = \mu_j + \varepsilon_{ij}$ in the cell-means version where μ is a common mean for all the observations. We will call this the **mean-only** model since it only has a single mean in it. In the reference-coding version of the model, we have a null hypothesis that $H_0 : \tau_2 = \ldots = \tau_J = 0$, so the "mean-only" model is $y_{ij} = \boldsymbol{\alpha} + \varepsilon_{ij}$ with $\boldsymbol{\alpha}$ having the same definition as μ for the cell means model – it forces a common value for the mean for all the groups. Moving from the *reference-coded* model to the *mean-only* model is also an example of a situation where we move from a "full" model to a "reduced" model by setting some coefficients in the "full" model to 0 and, by doing this, get a simpler or "reduced" model. Simple models can be good as they are easier to groups that suggests no difference in the groups is not a very exciting result in most, but not all, situations[2]. In order for R to provide results for the mean-only model, we remove the grouping variable, `Attr`, from the model formula and just include a "1". The `(Intercept)` row of the output provides the estimate for the mean-only model as a reduced model from either the cell-means or reference-coded models when we assume that the mean is the same for all groups:

```
lm3 <- lm(Years ~ 1, data=MockJury)
summary(lm3)
```

```
## $coefficients
##             Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 4.692982  0.3403532 13.78857 5.765681e-26
```

This model provides an estimate of the common mean for all observations of $4.693 = \hat{\mu} = \hat{\alpha}$ years. This value also is the dashed, horizontal line in the beanplot in Figure 3.1. Some people call this mean-only estimate the grand or overall mean.

3.3 One-Way ANOVA Sums of Squares, Mean Squares, and F-test

The previous discussion showed two ways of parameterizing models for the One-Way ANOVA model and getting estimates from output but still hasn't addressed how to assess evidence related to whether the observed differences in the means among the groups is "real". In this section, we develop what is called the ***ANOVA F-test*** that provides a method of aggregating the differences among the means of 2 or more groups and testing our null hypothesis of no difference in the means vs the alternative. In order to develop the test, some additional notation is needed. The sample size in each group is denoted n_j and the total sample size is $\boldsymbol{N = \Sigma n_j = n_1 + n_2 + \ldots + n_J}$ where Σ (capital sigma) means "add up over whatever follows". An estimated ***residual*** (e_{ij}) is the difference between an observation, y_{ij}, and the model estimate, $\hat{y}_{ij} = \hat{\mu}_j$, for that observation, $y_{ij} - \hat{y}_{ij} = e_{ij}$. It is basically what is left over that the mean part of the model ($\hat{\mu}_j$) does not explain. It is also a window into how "good" the model might be.

Consider the four different fake results for a situation with four groups ($J = 4$) displayed in Figure 3.3. Which of the different results shows the most and least evidence of differences in the means? In trying to answer this,

[2]Suppose we were doing environmental monitoring and were studying asbestos levels in soils. We might be hoping that the mean-only model were reasonable to use if the groups being compared were in remediated areas and in areas known to have never been contaminated.

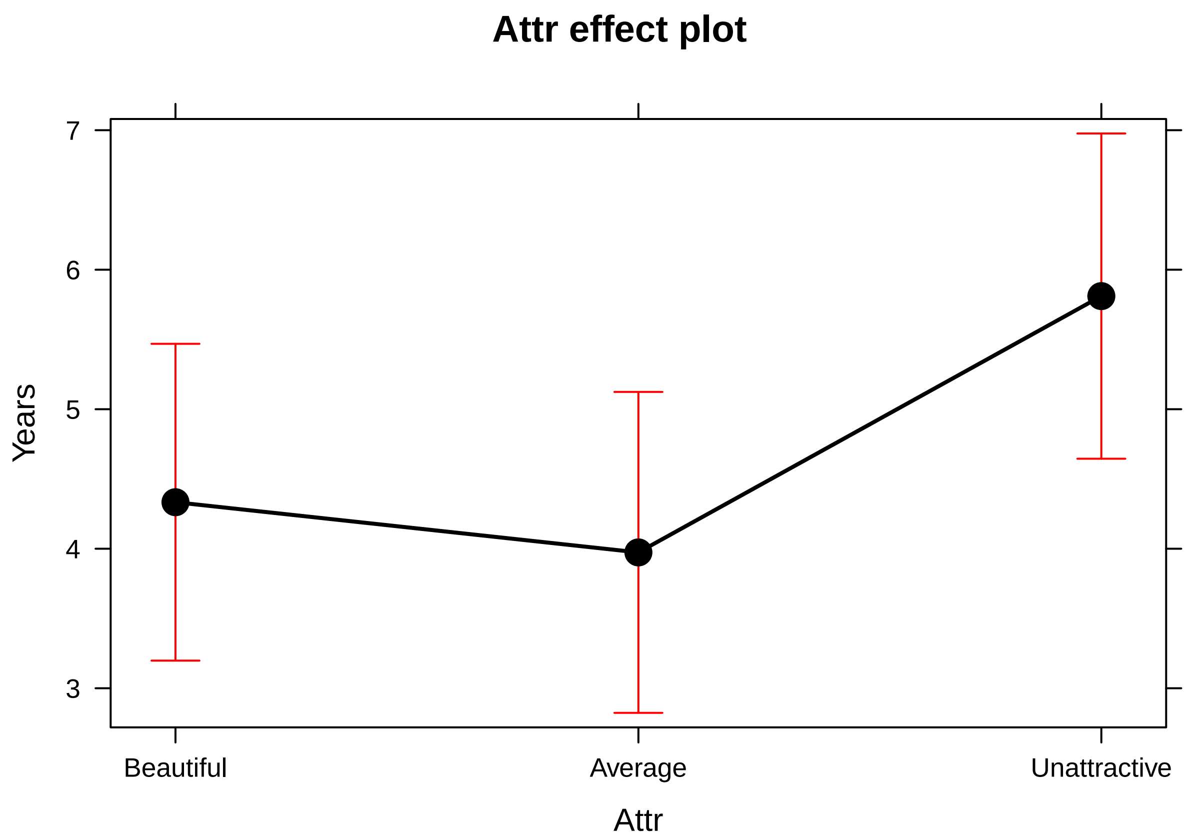


Figure 3.2: Plot of the estimated group mean sentences from the reference-coded model for the MockJury data.

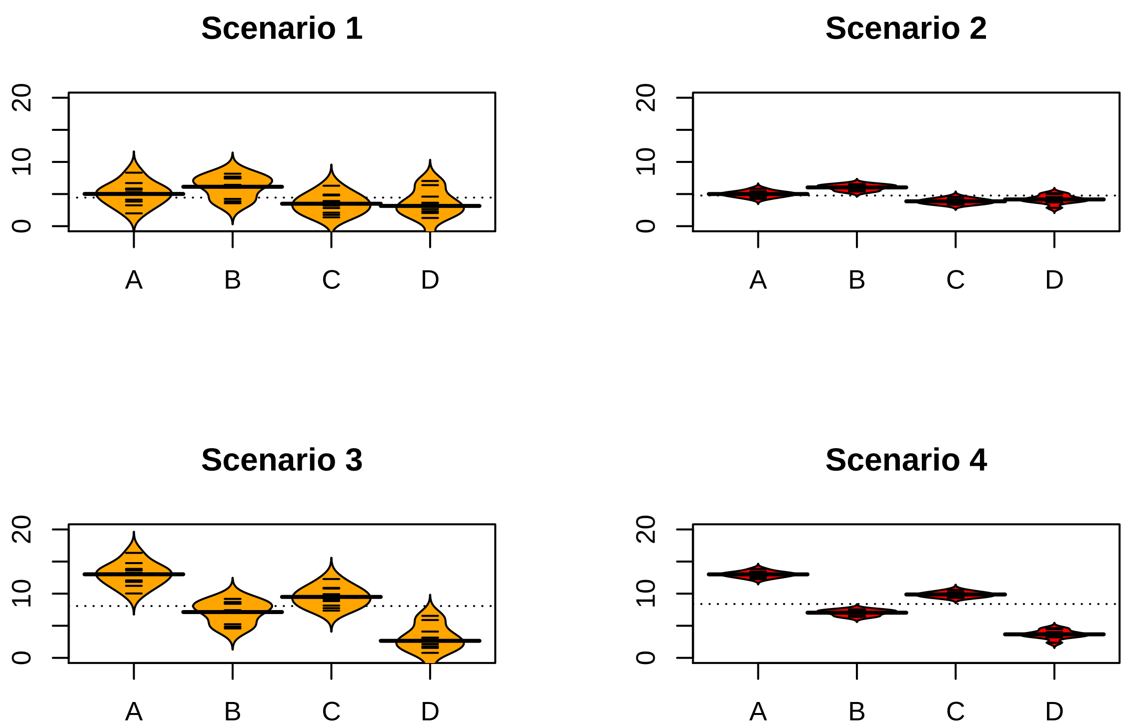


Figure 3.3: Demonstration of different amounts of difference in means relative to variability. Scenarios have same means in rows and same variance around means in columns of plot.

think about both how different the means are (obviously important) and how variable the results are around the mean. These situations were created to have the same means in Scenarios 1 and 2 as well as matching means in Scenarios 3 and 4. The variability around the means matches by shading (lighter or darker). In Scenarios 1 and 2, the differences in the means is smaller than in the other two results. But Scenario 2 should provide more evidence of what little difference in present than Scenario 1 because it has less variability around the means. The best situation for finding group differences here is Scenario 4 since it has the largest difference in the means and the least variability around those means. Our test statistic somehow needs to allow a comparison of the variability in the means to the overall variability to help us get results that reflect that Scenario 4 has the strongest evidence of a difference and Scenario 1 would have the least.

The statistic that allows the comparison of relative amounts of variation is called the ***ANOVA F-statistic*** . It is developed using ***sums of squares*** which are measures of total variation like are used in the numerator of the standard deviation ($\Sigma_1^N(y_i - \bar{y})^2$) that took all the observations, subtracted the mean, squared the differences, and then added up the results over all the observations to generate a measure of total variability. With multiple groups, we will focus on decomposing that total variability (***Total Sums of Squares***) into variability among the means (we'll call this ***Explanatory Variable* A's *Sums of Squares***) and variability in the residuals or errors (***Error Sums of Squares***). We define each of these quantities in the One-Way ANOVA situation as follows:

- $\mathbf{SS_{Total}}$ = Total Sums of Squares = $\Sigma_{j=1}^{J}\Sigma_{i=1}^{n_j}(y_{ij} - \bar{\bar{y}})^2$
  - This is the total variation in the responses around the overall or ***grand mean*** ($\bar{\bar{y}}$, the estimated mean for all the observations and available from the mean-only model).
  - By summing over all n_j observations in each group, $\Sigma_{i=1}^{n_j}(\ )$, and then adding those results up across the groups, $\Sigma_{j=1}^{J}(\ )$, we accumulate the variation across all N observations.

  - Note: this is the residual variation if the null model is used, so there is no further decomposition possible for that model.
  - This is also equivalent to the numerator of the sample variance, $\Sigma_1^N(y_i - \bar{y})^2$ which is what you get when you ignore the information on the potential differences in the groups.
- $\mathbf{SS_A}$ = Explanatory Variable A's Sums of Squares $= \Sigma_{j=1}^J \Sigma_{i=1}^{n_j}(\bar{y}_i - \bar{\bar{y}})^2 = \Sigma_{j=1}^J n_j(\bar{y}_i - \bar{\bar{y}})^2$
  - This is the variation in the group means around the grand mean based on the explanatory variable A.
  - Also called sums of squares for the treatment, regression, or model.
- $\mathbf{SS}_E$ = Error (Residual) Sums of Squares $= \Sigma_{j=1}^J \Sigma_{i=1}^{n_j}(y_{ij} - \bar{y})^2 = \Sigma_{j=1}^J \Sigma_{i=1}^{n_j}(e_{ij})^2$
  - This is the variation in the responses around the group means.
  - Also called the sums of squares for the residuals, with the second version of the formula showing that it is just the squared residuals added up across all the observations.

The possibly surprising result given the mass of notation just presented is that the total sums of squares is **ALWAYS** equal to the sum of explanatory variable A's sum of squares and the error sums of squares,

$$\mathbf{SS_{Total}}\mathbf{=}\mathbf{SS_A}\mathbf{+}\mathbf{SS_E}.$$

This equality means that if the $\mathbf{SS_A}$ goes up, then the $\mathbf{SS_E}$ must go down if $\mathbf{SS_{Total}}$ remains the same. This result is called the ***sums of squares decomposition formula***. We use these results to build our test statistic and organize this information in what is called an ***ANOVA table***. The ANOVA table is generated using the `anova` function applied to the reference-coded model, `lm2` :

```
lm2<-lm(Years ~ Attr, data=MockJury)
anova(lm2)
```

```
## Analysis of Variance Table
##
## Response: Years
##            Df  Sum Sq Mean Sq F value Pr(>F)
## Attr        2   70.94  35.469    2.77  0.067
## Residuals 111 1421.32  12.805
```

Note that the ANOVA table has a row labelled `Attr`, which contains information for the grouping variable (we'll generally refer to this as explanatory variable A but here it is the picture group that was randomly assigned), and a row labelled `Residuals`, which is synonymous with "Error". The Sums of Squares (SS) are available in the `Sum Sq` column. It doesn't show a row for "Total" but the $\mathbf{SS_{Total}}\mathbf{=}\mathbf{SS_A}\mathbf{+}\mathbf{SS_E} = 1492.26$.

```
70.94 + 1421.32
```

```
## [1] 1492.26
```

It may be easiest to understand the *sums of squares decomposition* by connecting it to our permutation ideas. In a permutation situation, the total variation (SS_{Total}) cannot change – it is the same responses varying around the grand mean. However, the amount of variation attributed to variation among the means and in the residuals can change if we change which observations go with which group. In Figure 3.4 (panel a), the means, sums of squares, and 95% confidence intervals for each mean are displayed for the three treatment levels from the original prisoner rating data. Three permuted versions of the data set are summarized in panels (b), (c), and (d). The SS_A is 70.9 in the real data set and between 6.6 and 11 in the permuted data sets. If you had to pick among the plots for the one with the most evidence of a difference in the means, you hopefully would pick panel (a). This visual "unusualness" suggests that this observed result is unusual relative to the possibilities under permutations, which are, again, the possibilities tied to having the null hypothesis being true. But note that the differences here are not that great between these three permuted data sets and the real one. It is likely that at least some of you might have selected panel (d) as also looking like it shows some evidence of differences (maybe not the most?) as it also looks like it shows some evidence differences.

One way to think about $\mathbf{SS_A}$ is that it is a function that converts the variation in the group means into a single value. This makes it a reasonable test statistic in a permutation testing context. By comparing the observed $\text{SS}_A = 70.9$ to the permutation results of 6.5, 9.7, and 40.5 we see that the observed result is much more extreme than the three alternate versions. In contrast to our previous test statistics where positive and negative differences were possible, SS_A is always positive with a value of 0 corresponding to no variation in the means. The larger the SS_A, the more variation there is in the means. The permutation p-value for the alternative hypothesis of **some** (not of greater or less than!) difference in the true means of the groups will involve counting the number of permuted SS_A^* results that are larger than what we observed.

```
## [1] 70.93836
```

To do a permutation test, we need to be able to calculate and extract the SS_A value. In the ANOVA table, it

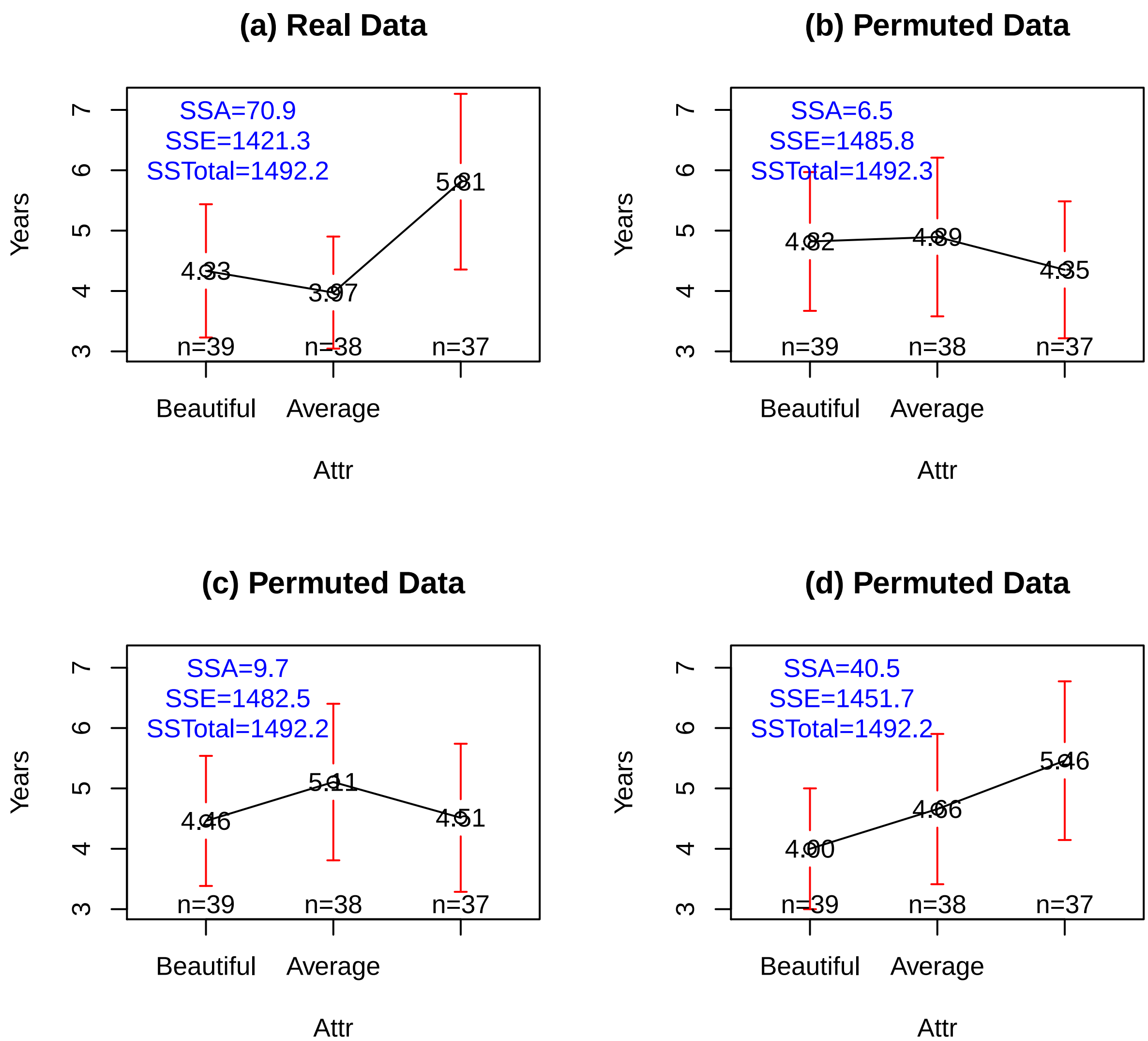


Figure 3.4: Plot of means and 95% confidence intervals for the three groups for the real data (a) and three different permutations of the treatment labels to the same responses in (b), (c), and (d). Note that SSTotal is always the same but the different amounts of variation associated with the means (SSA) or the errors (SSE) changes in permutation.

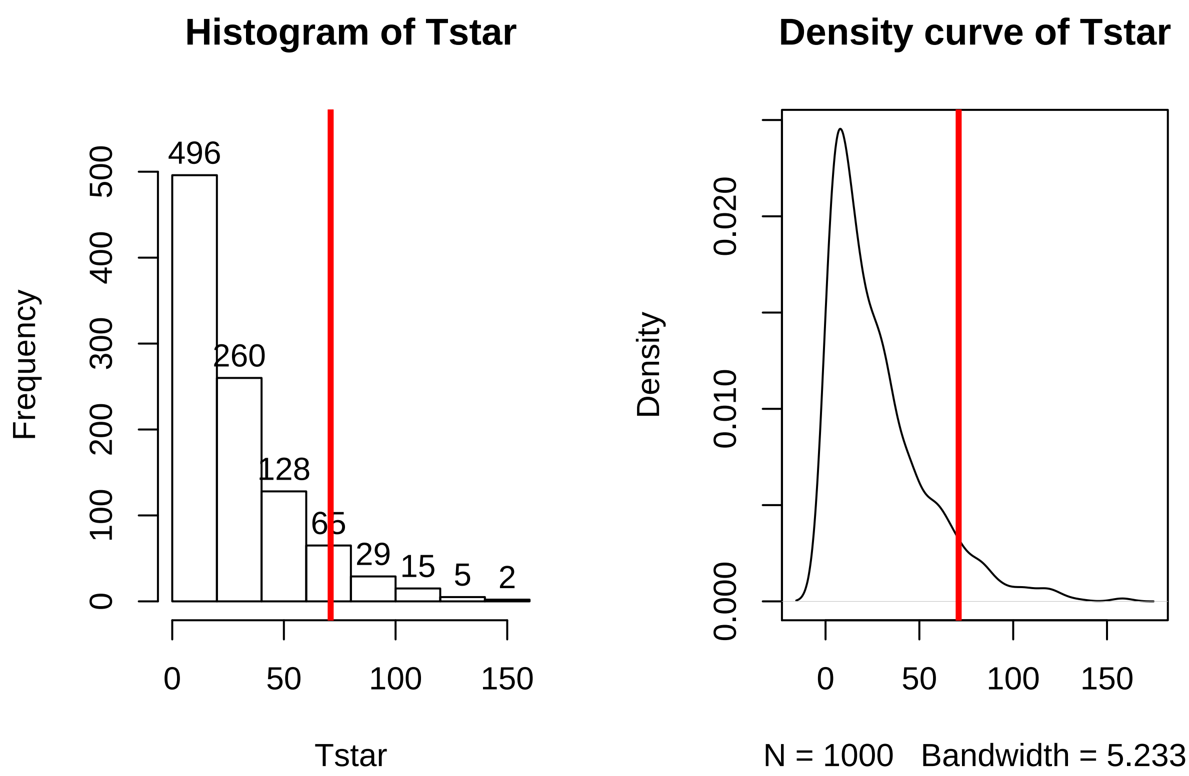


Figure 3.5: Histogram and density curve of permutation distribution of SS_A with the observed value of SS_A displayed as a bold, vertical line. The proportion of results that are larger than the observed value of SS_A provides an estimate of the p-value.

is in the first row and is the second number and we can use the bracket, `[, ]`, referencing to extract that number from the ANOVA table that `anova` produces with `anova(lm(Years~Attr, data=MockJury))[1, 2]`. We'll store the observed value of SS_A in `Tobs`, reusing some ideas from Chapter @ref{chapter2}.

```
Tobs <- anova(lm(Years~Attr,data=MockJury))[1,2]; Tobs
```

```
## [1] 70.93836
```

The following code performs the permutations `B=1,000` times using the `shuffle` function, builds up a vector of results in `Tobs`, and then makes a plot of the resulting permutation distribution:

```
par(mfrow=c(1,2))
B<- 1000
Tstar<-matrix(NA,nrow=B)
for (b in (1:B)){
  Tstar[b]<-anova(lm(Years~shuffle(Attr),data=MockJury))[1,2]
  }
hist(Tstar,labels=T,ylim=c(0,550))
abline(v=Tobs,col="red",lwd=3)
plot(density(Tstar),main="Density curve of Tstar")
abline(v=Tobs,col="red",lwd=3)
```

The right-skewed distribution (Figure 3.5) contains the distribution of SS_A's under permutations (where all the groups are assumed to be equivalent under the null hypothesis). While the observed result is larger than many of the SS_A's, there are also many permuted results that are much larger than observed. The proportion of permuted results that exceed the observed value is found using `pdata` as before, except only for the area to the right of the observed result. We know that `Tobs` will always be positive so no absolute values are required here.

```
pdata(Tstar,Tobs,lower.tail=F)
```

```
## [1] 0.072
```

This provides a permutation-based p-value of 0.072 and suggests marginal evidence against the null hypothesis of no difference in the true means. We would interpret this p-value as saying that there is a 7.2% chance of

Table 3.3: General One-Way ANOVA table.

Source	DF	Sums of Squares	Mean Squares	F-ratio
Variable A	$J-1$	SS_A	$\text{MS}_A=\text{SS}_A/(J-1)$	$F=\text{MS}_A/\text{
Residuals	$N-J$	SS_E	$\text{MS}_E = \text{SS}_E/(N-J)$	
Total	$N-1$	SS_{Total}		

getting a SS_A as large or larger than we observed, given that the null hypothesis is true.

It ends up that some nice parametric statistical results are available (if our assumptions are met) for the ratio of estimated variances, which are called ***Mean Squares***. To turn sums of squares into mean square (variance) estimates, we divide the sums of squares by the amount of free information available. For example, remember the typical variance estimator introductory statistics, $\Sigma^N_1(y_i - \bar{y})^2/(N-1)$? Your instructor spent some time trying various approaches to explaining why we have a denominator of $N-1$. The most useful for our purposes moving forward is that we "lose" one piece of information to estimate the mean and there are N deviations around the single mean so we divide by $N-1$. The main point is that the sums of squares were divided by something and we got an estimator for the variance, here of the observations.

Now consider $\text{SS}_E = \Sigma^J_{j=1}\Sigma^{n_j}_{i=1}(y_i - \bar{y})^2$ which still has N deviations but it varies around the J means, so the

$$\textbf{Mean Square Error} = \text{MS}_E = \text{SS}_E/(N-J).$$

Basically, we lose J pieces of information in this calculation because we have to estimate J means.

The similar calculation of the ***Mean Square for variable* A** (MS_A) is harder to see in the formula ($\text{SS}_A = \Sigma^J_{j=1}n_j(\bar{y}_i - \bar{\bar{y}})^2$), but the same reasoning can be used to understand the denominator for forming MS_A: there are J means that vary around the grand mean so

$$\text{MS}_A = \text{SS}_A/(J-1).$$

In summary, the two mean squares are simply:

- $\text{MS}_A = \text{SS}_A/(J-1)$, which estimates the variance of the group means around the grand mean.
- $\text{MS}_{\text{Error}} = \text{SS}_{\text{Error}}/(N-J)$, which estimates the variation of the errors around the group means.

These results are put together using a ratio to define the ***ANOVA F-statistic*** (also called the ***F-ratio***) as

$$F = \text{MS}_A/\text{MS}_{\text{Error}}.$$

If the variability in the means is "similar" to the variability in the residuals, the statistic would have a value around 1. If that variability is similar then there be no evidence of a difference in the means. If the MS_A is much larger than the MS_E, the F-statistic will provide evidence against the null hypothesis. The "size" of the F-statistic is formalized by finding the p-value. The F-statistic, if assumptions discussed below are met and we assume the null hypothesis is true, follows what is called an F-distribution. The ***F-distribution*** is a right-skewed distribution whose shape is defined by what are called the ***numerator degrees of freedom*** $(J-1)$ and the ***denominator degrees of freedom*** $(N-J)$. These names correspond to the values that we used to calculate the mean squares and where in the F-ratio each mean square was used; F-distributions are denoted by their degrees of freedom using the convention of F (*numerator df*, *denominator df*). Some examples of different F-distributions are displayed for you in Figure 3.6.

The characteristics of the F-distribution can be summarized as:

- Right skewed,
- Nonzero probabilities for values greater than 0,
- Its shape changes depending on the **numerator** and **denominator DF**, and
- **Always use the right-tailed area for p-values.**

Now we are ready to discuss an ANOVA table since we know about each of its components. Note the general format of the ANOVA table is[3]:

[3] Make sure you can work from left to right and up and down to fill in the ANOVA table given just the necessary information to determine the other components – there is always a question like this on the exam...

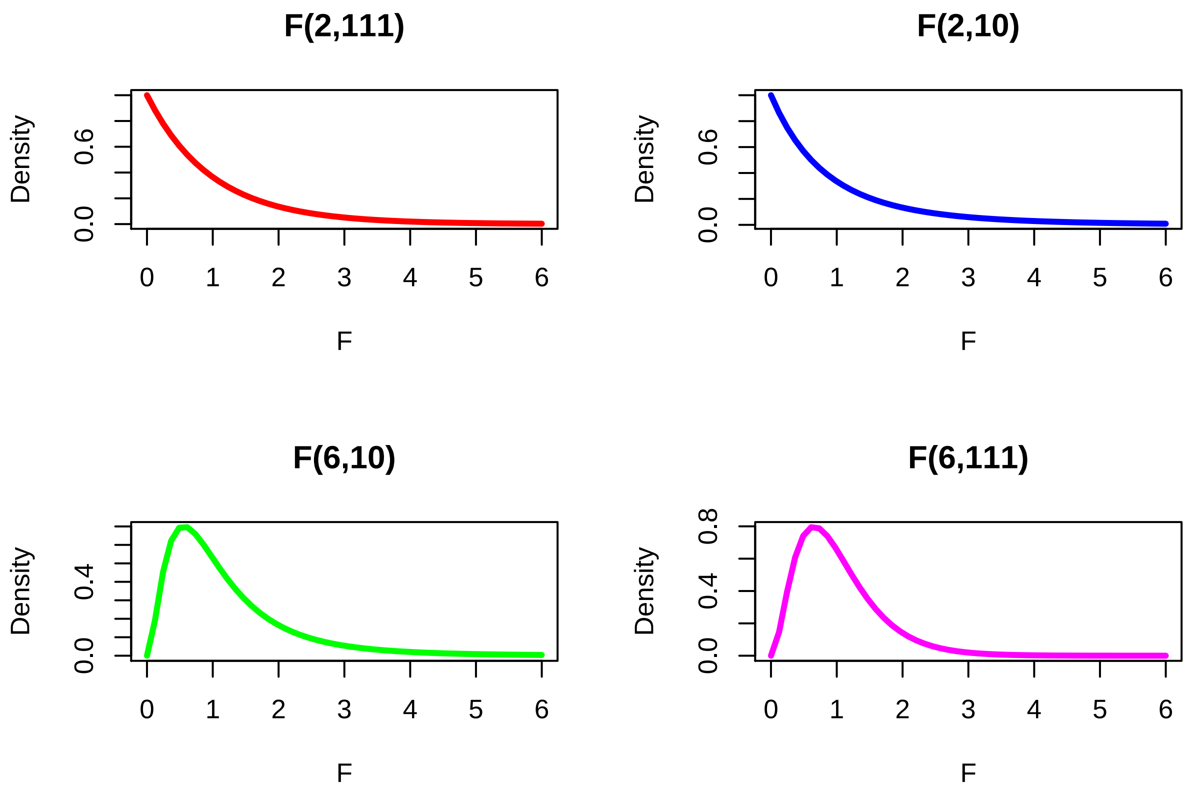


Figure 3.6: Density curves of four different F-distributions. Upper left is an $F(2, 111)$, upper right is $F(2, 10)$, lower left is $F(6, 10)$, and lower right is $F(6, 111)$. P-values are found using the areas to the right of the observed F-statistic value.

The table is oriented to help you reconstruct the F-ratio from each of its components. The output from R is similar although it does not provide the last row and sometimes switches the order of columns. The R version of the table for the type of picture effect (`Attr`) with $J = 3$ levels and $N = 114$ observations, repeated from above, is:

```
anova(lm2)
```

```
## Analysis of Variance Table
##
## Response: Years
##            Df  Sum Sq Mean Sq F value Pr(>F)
## Attr        2   70.94  35.469    2.77  0.067
## Residuals 111 1421.32  12.805
```

The p-value from the F-distribution is 0.067. We can verify this result using the observed F-statistic of 2.77 (which came from taking the ratio of the two mean squares, F=35.47/12.8) which follows an $F(2, 111)$ distribution if the null hypothesis is true and some other assumptions are met. Using the `pf` function provides us with areas in the specified F-distribution with the `df1` provided to the function as the numerator *df* and `df2` as the denominator *df* and `lower.tail=F` reflecting our desire for a right tailed area.

```
pf(2.77,df1=2,df2=111,lower.tail=F)
```

```
## [1] 0.06699803
```

The result from the F-distribution using this parametric procedure is similar to the p-value obtained using permutations with the test statistic of the SS_A, which was 70.9. The F-statistic obviously is another potential test statistic to use as a test statistic in a permutation approach, now that we know about it. We should check that we get similar results from it with permutations as we did from using SS_A as a permutation test test statistic. The following code generates the permutation distribution for the F-statistic (Figure 3.7) and assesses how unusual the observed F-statistic of 2.77 was in this permutation distribution. The only change in the code involves moving from extracting SS_A to extracting the F-ratio which is in the 4th column of the `anova` output:

```
Tobs <- anova(lm(Years ~ Attr, data=MockJury))[1,4]; Tobs
```

```
## [1] 2.770024
```

```
par(mfrow=c(1,2))
B<- 1000
Tstar<-matrix(NA,nrow=B)
for (b in (1:B)){
  Tstar[b]<-anova(lm(Years~shuffle(Attr), data=MockJury))[1,4]
}

pdata(Tstar, Tobs, lower.tail=F)
```

```
## [1] 0.064
```

```
hist(Tstar, labels=T)
abline(v=Tobs, col="red", lwd=3)
plot(density(Tstar), main="Density curve of Tstar")
abline(v=Tobs, col="red", lwd=3)
```

The permutation-based p-value is 0.064 which, again, matches the other results closely. The first conclusion is that using a test statistic of either the F-statistic or the SS_A provide similar permutation results. However, we tend to favor using the F-statistic because it is more commonly used in reporting ANOVA results, not because it is any better in a permutation context.

It is also interesting to compare the permutation distribution for the F-statistic and the parametric $F(2, 111)$ distribution (Figure 3.8). They do not match perfectly but are quite similar. Some the differences around 0 are due to the behavior of the method used to create the density curve and are not really a problem for the methods. The similarity in the two curves explains why both methods give similar results. In some situations, the correspondence will not be quite so close.

So how can we rectify this result (p-value $\approx$ 0.06) and the Chapter **??** result that detected a difference between *Average* and *Unattractive* with a p-value $\approx$ 0.03? I selected the two groups to compare in Chapter **??** because they were furthest apart. "Cherry-picking" the comparison that is likely to be most different creates a false sense of the real situation and inflates the Type I error rate because of the selection. If the entire suite of pairwise comparisons are considered, this result may lose some of its luster. In other words, if we consider the suite of three pair-wise differences (and the tests) implicit in comparing all of them, we may need stronger evidence in the most different pair than a p-value of 0.033 to suggest overall differences. In this situation, the *Beautiful* and *Average* groups are not that different from each other so their difference does not contribute much to the will revisit this topic and consider a method that is statistically valid for performing all possible pair-wise comparisons that is also consistent with our overall test results.

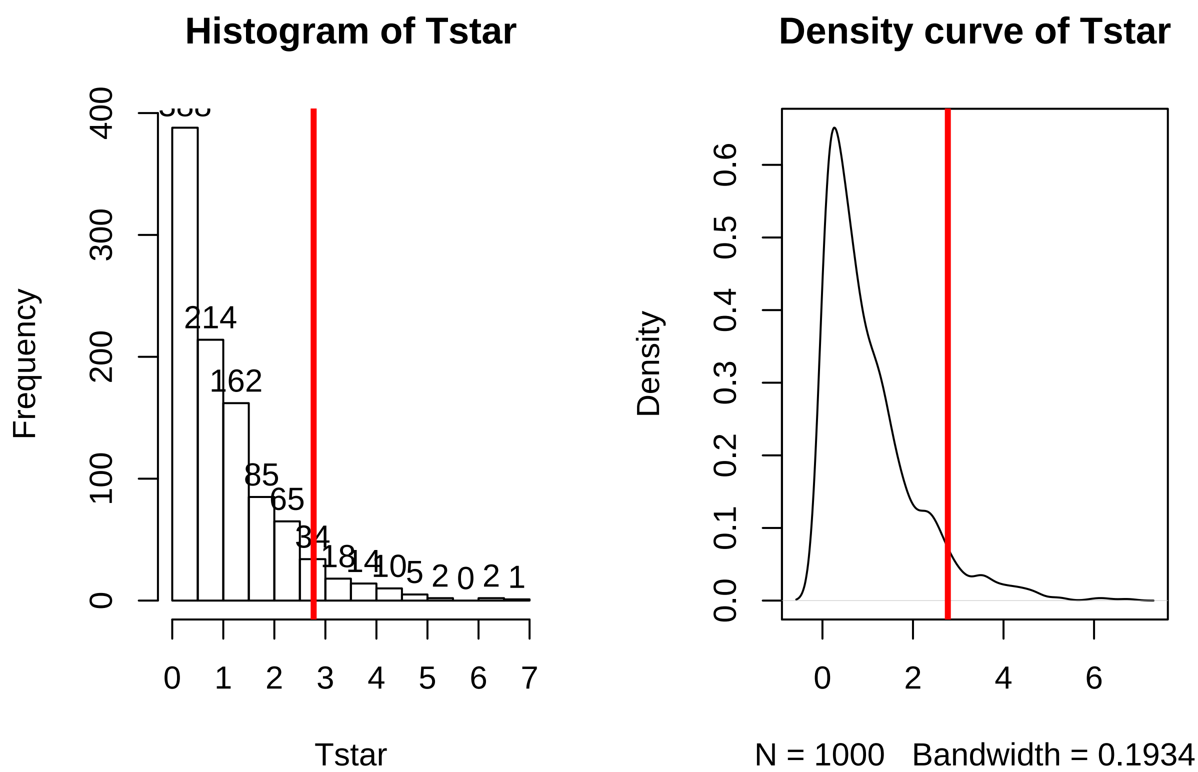


Figure 3.7: Histogram and density curve of the permutation distribution of the F-statistic with bold, vertical line for observed value of the test statistic of 2.77.

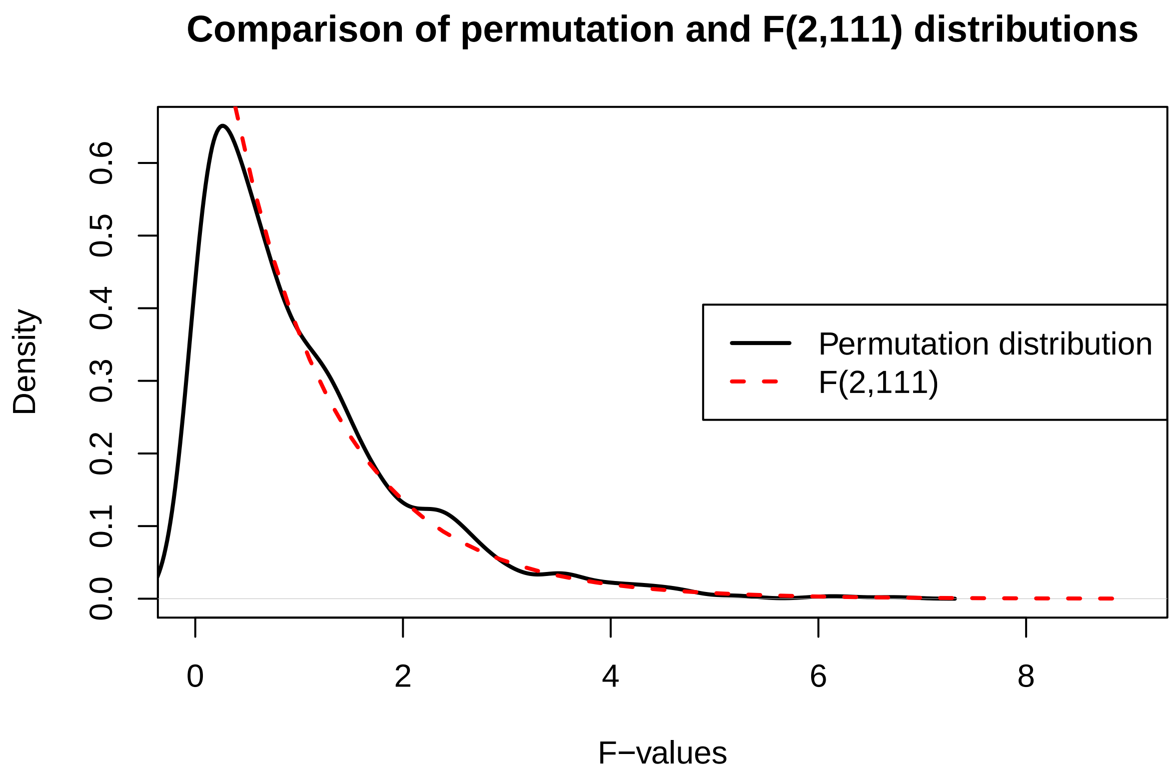


Figure 3.8: Comparison of $F(2, 111)$ (dashed line) and permutation distribution (solid line).

3.4 ANOVA model diagnostics including QQ-plots

The requirements for a One-Way ANOVA F-test are similar to those discussed in Chapter **??**, except that there are now J groups instead of only 2. Specifically, the linear model assumes:

1. **Independent observations**,
2. **Equal variances**, and
3. **Normal distributions**.

For assessing equal variances across the groups, it is best to use plots to assess this. We can use boxplots and beanplots to compare the spreads of the groups, which were provided in Figure 3.1. The range and IQRs should be relatively similar across the groups if you do not find evidence of a problem with this assumption. You should start with noting how clear or big the violation of the assumption might be but remember that there will always be some differences in the variation among groups even if the true variability is exactly equal in the populations. In addition to our direct plotting, there are some diagnostic plots available from the `lm` function that can help us more clearly assess potential violations of the previous assumptions.

We can obtain a suite of four diagnostic plots by using the `plot` function on any linear model object that we have fit. To get all the plots together in four panels we need to add the `par(mfrow=c(2, 2))` command to tell R to make a graph with 4 panels[4].

```
par(mfrow=c(2,2))
plot(lm2,pch=16)
```

There are two plots in Figure 3.9 with useful information for the equal variance assumption. The "Residuals vs Fitted" panel in the top left displays the residuals ($e_{ij} = y_{ij} - \hat{y}_ij$) on the y-axis and the fitted values ($\hat{y}_{ij}$) on the x-axis. This allows you to see if the variability of the observations differs across the groups as a function of the mean of the groups because all the observations in the same group get the same fitted value, the mean of the group. In this plot, the points seem to have fairly similar spreads at the fitted values for the three groups with fitted values of 4, 4.3, and 6. The "Scale-Location" plot in the lower left panel has the same x-axis but the y-axis contains the square-root of the absolute value of the standardized residuals. The absolute value transforms all the residuals into a magnitude scale (removing direction) and the square-root helps you see differences in variability more accurately. The standardization scales them to have a variance of 1 so help you in other displays to get a sense of how many standard deviations you are away from the mean in the residual distribution. The visual assessment is similar in the two plots – you want to consider whether it appears that the groups have somewhat similar or noticeably different amounts of variability. If you see a clear funnel shape in the Residuals vs Fitted or an increase or decrease in the upper edge of points in the Scale-Location plot that may indicate a violation of the constant variance assumption. Remember that some variation across the groups is expected and is OK, but large differences in spreads are problematic for all the procedures that involve linear models. When discussing these results, you want to discuss how clearly the differences in variation are and whether that *shows a clear violation of the assumption* of equal variance for all observations. Like in hypothesis testing, you can't prove that you've met assumptions based on a plot "looking OK", but you can say that there is no clear evidence that the assumption is violated!

The linear model assumes that all the random errors (ε_{ij}) follow a normal distribution. To gain insight into the validity of this assumption, we can explore the original observations as displayed in the beanplots, mentally subtracting off the differences in the means and focusing on the shapes of the distributions of observations in each group. These plots are especially good for assessing whether there is there a skew or outliers present in each group. If so, by definition, the normality assumption is violated. But our assumption is about the distribution of all the errors after the remove the differences in the means and so we want an overall assessment technique to understand how reasonable our assumption is overall for our model. The residuals from the entire model provide us with estimates of the random errors and if the normality assumption is met, then the residuals all-together should approximately follow a normal distribution. The ***Normal Q-Q Plot*** in upper right panel of Figure 3.9 is a direct visual assessment of how well our residuals match what we would expect from a normal distribution. Outliers, skew, heavy and light-tailed aspects of distributions (all violations of normality) show up in this plot once you learn to read it – which is our next task. To make it easier to read QQ-plots, it is nice to start with just considering histograms and/or density plots of the residuals and to see how that maps into this new display. We can obtain the residuals from the linear model

[4] We have been using this function quite a bit to make multi-panel graphs but did not show you that line of code. But you need to use this command for linear model diagnostics or you won't get the plots we want from the model. And you really just need `plot(lm2)` but the `pch=16` option makes it easier to see some of the points in the plots.

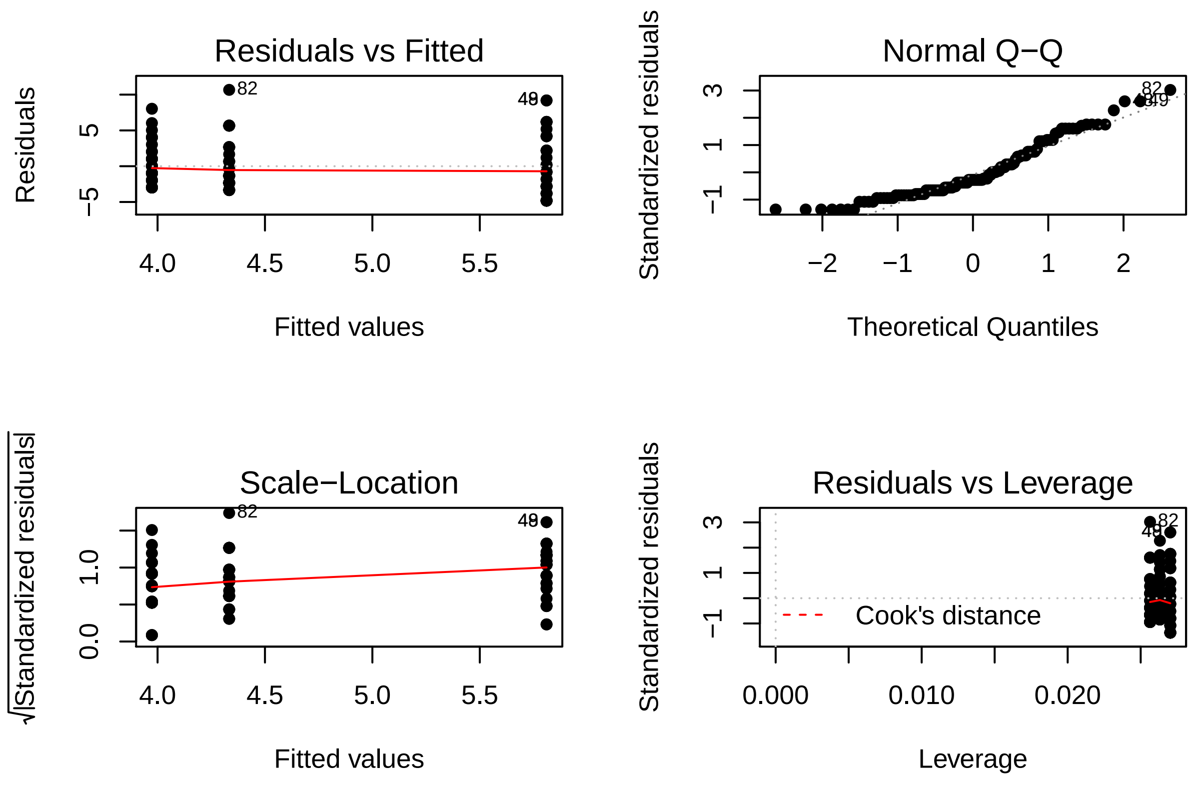


Figure 3.9: Default diagnostic plots for the linear model.

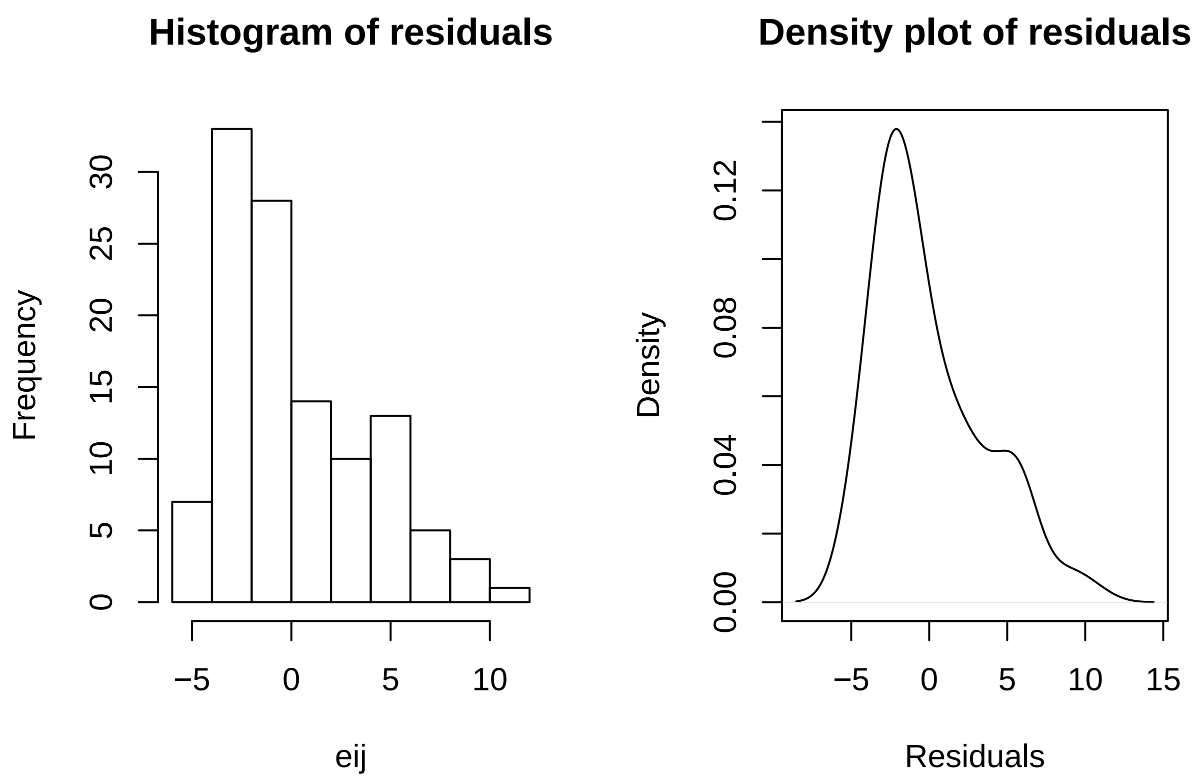


Figure 3.10: Histogram and density curve of the linear model raw residuals.

using the `residuals` function on any linear model object.

```
par(mfrow=c(1,2))
eij<-residuals(lm2)
hist(eij, main="Histogram of residuals")
plot(density(eij), main="Density plot of residuals", ylab="Density",
     xlab="Residuals")
```

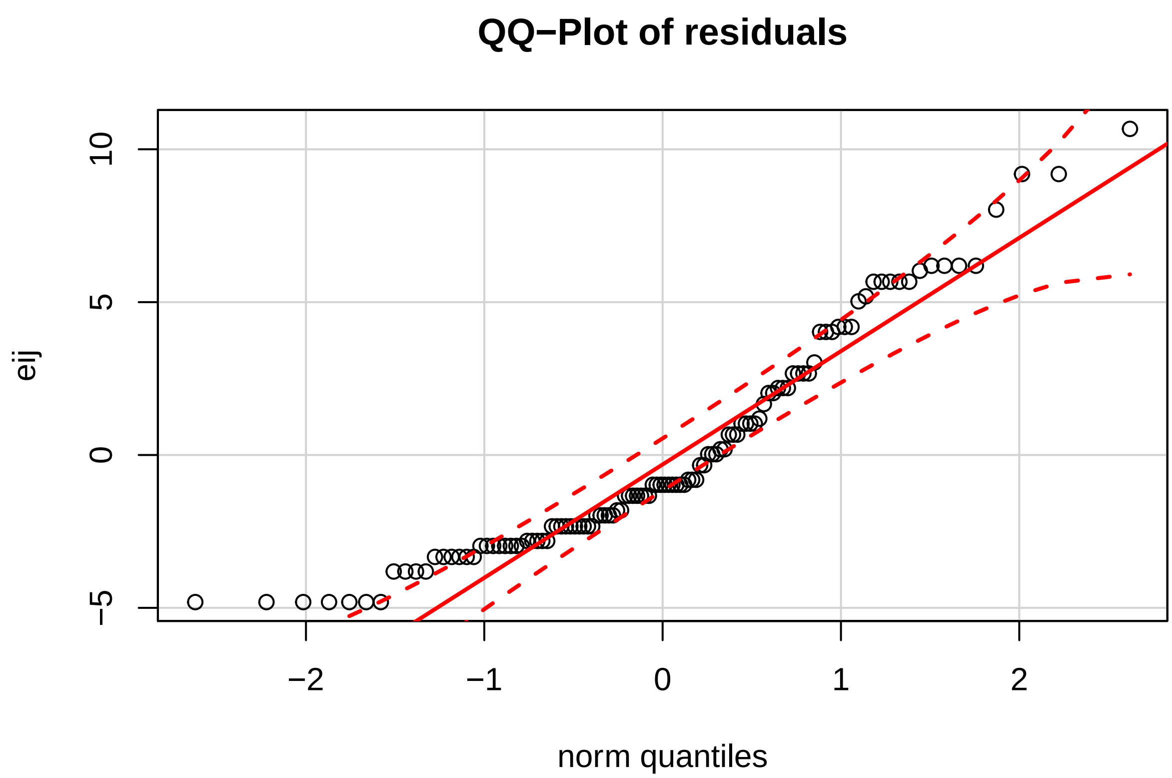
Figure 3.10 shows that there is a right skew present in the residuals for the prisoner rating data model that accounted for different means in the three groups, which is consistent with the initial assessment of some right skew in the plots of observations in each group.

A Quantile-Quantile plot (***QQ-plot***) shows the "match" of an observed distribution with a theoretical distribution, almost always the normal distribution. They are also known as Quantile Comparison, Normal Probability, or Normal Q-Q plots, with the last two names being specific to comparing results to a normal distribution. In this version[5], the QQ-plots display the value of observed percentiles in the residual distribution on the y-axis versus the percentiles of a theoretical normal distribution on the x-axis. If the observed **distribution of the residuals matches the shape of the normal distribution, then the plotted points should follow a 1-1 relationship.** If the points follow the displayed straight line then that suggests that the residuals have a similar shape to a normal distribution. Some variation is expected around the line and some patterns of deviation are worse than others for our models, so you need to go beyond saying "it does not match a normal distribution". It is best to be specific about the type of deviation you are detecting. And to do that, we need to practice interpreting some QQ-plots.

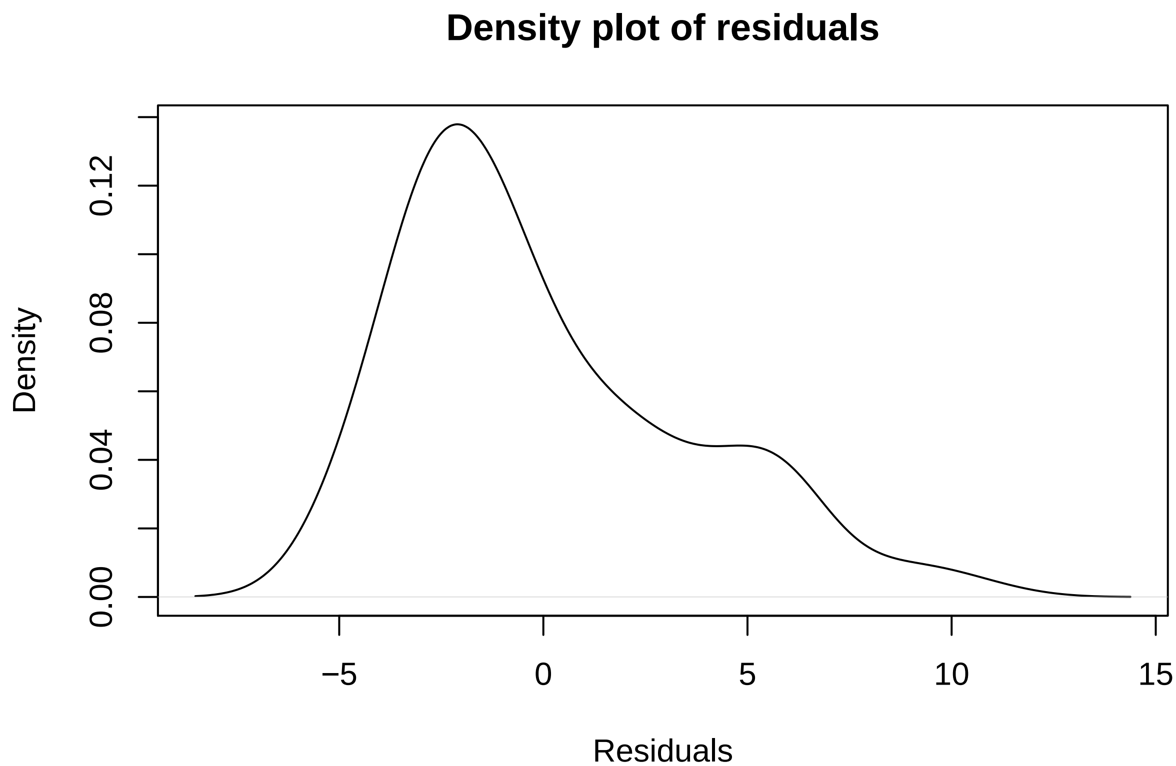
The QQ-plot of the linear model residuals from Figure 3.9 is extracted and enhanced it a little to make Figure **??** so we can just focus on it. We know from looking at the histogram that this is a slightly right skewed

[5]Along with multiple names, there is variation of what is plotted on the x and y axes and the scaling of the values plotted, increasing the challenge of interpreting QQ-plots. We are consistent about the x and y axis choices but different functions that make these plots in R do switch the axes.

distribution. The QQ-plot places the observed ***standardized***[6] ***residuals*** on the y-axis and the theoretical normal values on the x-axis. The most noticeable deviation from the 1-1 line is in the lower left corner of the plot. These are for the negative residuals (left tail) and there are many residuals at around the same value that are a little smaller than -1. If the distribution had followed the normal distribution here, the points would be on the 1-1 line and there would be some standardized residuals much smaller than -1.5. So we are not getting as much spread in the smaller residuals as we would expect in a normal distribution. If you go back to the histogram you can see that the smallest residuals are all stacked up and do not spread out like the left tail of a normal distribution should. In the right tail (positive) residuals, there is also a systematic lifting from the 1-1 line to larger values in the residuals than the normal would generate. For example, the point labeled as "82" (the 82nd observation in the data set) has a value of 3 in residuals but should actually be smaller (maybe 2.5) if the distribution was normal. Put together, this pattern in the QQ-plot suggests that the left tail is too compacted (too short) and the right tail is too spread out – this is the right skew we identified from the histogram and density curve!



[6]Here this means re-scaled so that they should have similar scaling to a standard normal with mean 0 and standard deviation 1. This does not change the shape of the distribution but can make outlier identification simpler – having a standardized residual more extreme than 5 or -5 would suggest a deviation from normality since we rarely see values that many standard deviations from the mean in a normal distribution. But mainly focus on the shape of the pattern in the QQ-plot.



Generally, when both tails deviate on the same side of the line (forming a sort of quadratic curve, especially in more extreme cases), that is evidence of a skew. To see some different potential shapes in QQ-plots, six different data sets are displayed in Figures 3.11 and 3.12. In each row, a QQ-plot and associated density curve are displayed. If the points are both above the 1-1 line in the lower and upper tails as in Figure 3.11(a), then the pattern is a right skew, here even more extreme than in the real data set. If the points are below the 1-1 line in both tails as in Figure 3.11(c), then the pattern is identified as a left skew. Skewed residual distributions (either direction) are problematic for models that assume normally distributed responses but not necessarily for our permutation approaches if all the groups have similar skewed shapes. The other problematic pattern is to have more spread than a normal curve as in Figure 3.11(e) and (f). This shows up with the points being below the line in the left tail (more extreme negative than expected by the normal) and the points being above the line for the right tail (more extreme positive than the normal predicts). We call these distributions ***heavy-tailed*** which can manifest as distributions with outliers in both tails or just a bit more spread out than a normal distribution. Heavy-tailed residual distributions can be problematic for our models as the variation is greater than what the normal distribution can account for and our methods might under-estimate the variability in the results. The opposite pattern with the left tail above the line and the right tail below the line suggests less spread (***lighter-tailed***) than a normal as in Figure 3.11(g) and (h). This pattern is relatively harmless and you can proceed with methods that assume normality safely as they will just be a little conservative.

Finally, to help you calibrate expectations for data that are actually normally distributed, two data sets simulated from normal distributions are displayed in Figure 3.12. Note how neither follows the line exactly but that the overall pattern matches fairly well. **You have to allow for some variation from the line in real data sets** and focus on when there are really noticeable issues in the distribution of the residuals such as those displayed above. Again, you will never be able to prove that you have normally distributed residuals even if the residuals are all exactly on the line, but if you see QQ-plots as in Figure 3.11 you can encounter situations that provide evidence of clear violations of the normality assumption.

The last issues with assessing the assumptions in an ANOVA relates to situations where the methods are

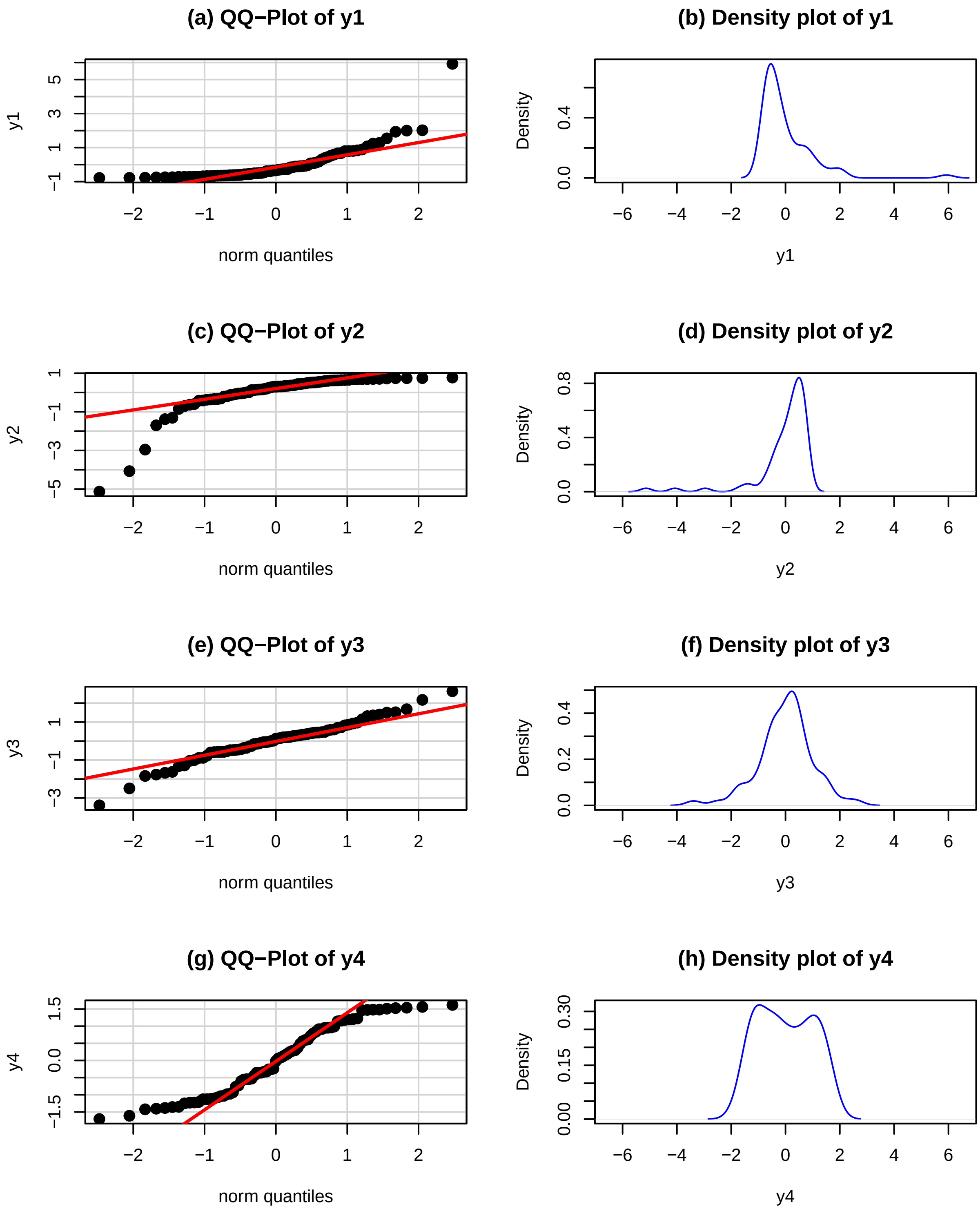


Figure 3.11: QQ-plots and density curves of four simulated distributions with different shapes.

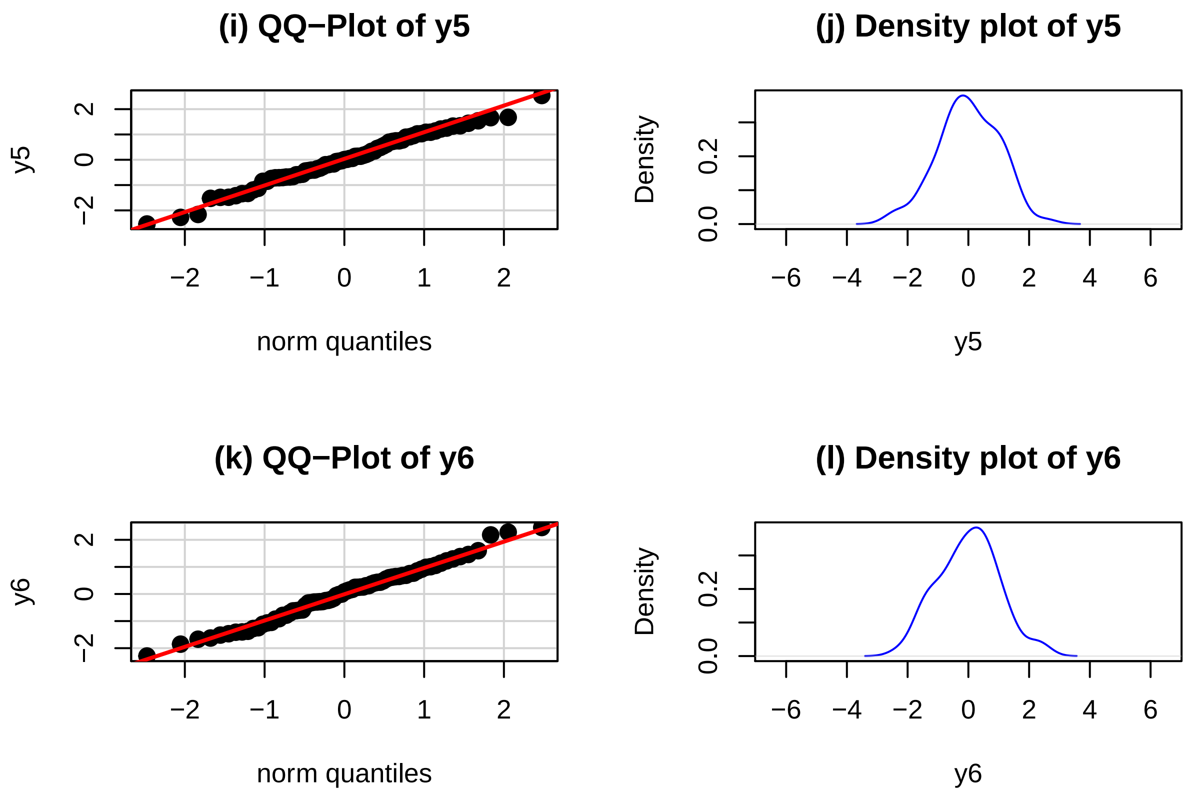


Figure 3.12: Two more simulated data sets, generated from normal distributions.

more or less ***resistant***[7] to violations of assumptions. For reasons beyond the scope of this book, the parametric ANOVA F-test is more resistant to violations of the assumptions of the normality and equal variance assumptions if the design is balanced. A ***balanced design*** occurs when each group is measured the same number of times. The resistance decreases as the data set becomes less balanced, as the sample sizes in the groups are more different, so having close to balance is preferred to a more imbalanced situation if there is a choice available. There is some intuition available here – it makes some sense that you would have better results in comparing groups if the information available is similar in all the groups and none are relatively under-represented. We can check the number of observations in each group to see if they are equal or similar using the `tally` function from the `mosaic` package. This function is useful for being able to get counts of observations, especially for cross-classifying observations on two variables that is used in Chapter 5. For just a single variable, we use `tally(~x, data=...)`:

```
require(mosaic)
tally(~Attr, data=MockJury)
```

```
## Attr
##    Beautiful      Average Unattractive
##           39           38           37
```

So the sample sizes do vary among the groups and the design is technically not balanced, but it is also very close to being balanced with only two more observations in the largest group compared to the smallest group size. This tells us that the F-test should have some resistance to violations of assumptions. This nearly balanced design, and the moderate sample size (over 37 per group is considered a good but not large sample), make the parametric and nonparametric approaches provide similar results in this data set even in the presence of the skewed residual error distribution.

3.5 Guinea pig tooth growth One-Way ANOVA example

A second example of the One-way ANOVA methods involves a study of length of odontoblasts (cells that are responsible for tooth growth) in 60 Guinea Pigs (measured in microns) from Crampton (1947). $N = 60$ Guinea Pigs were obtained from a local breeder and each received one of three dosages (0.5, 1, or 2 mg/day) of Vitamin C via one of two delivery methods, Orange Juice (OJ) or ascorbic acid (the stuff in vitamin C capsules, called VC below) as the source of Vitamin C in their diets. Each guinea pig was randomly assigned to receive one of the six different treatment combinations possible (OJ at 0.5 mg, OJ at 1 mg, OJ at 2 mg, VC at 0.5 mg, VC at 1 mg, and VC at 2 mg). The animals were treated similarly otherwise and we can assume lived in separate cages and only one observation was taken for each guinea pig, so we can assume the observations are independent. We need to create a variable that combines the levels of delivery type (OJ, VC) and the dosages (0.5, 1, and 2) to use our One-Way ANOVA on the six levels. The `interaction` function can be used create a new variable that is based on combinations of the levels of other variables. Here a new variable is created in the `ToothGrowth` data.frame that we called `Treat` that provides a six-level grouping variable for our One-Way ANOVA to compare the combinations of treatments. To get a sense of the pattern of observations in the data set, the counts in `supp` (supplement type) and `dose` are provided.

```
data(ToothGrowth) #Available in Base R
require(mosaic)
tally(~supp,data=ToothGrowth) #Supplement Type (VC or OJ)
```

```
## supp
## OJ VC
## 30 30
```

```
tally(~dose,data=ToothGrowth) #Dosage level
```

```
## dose
## 0.5   1   2
##  20  20  20
```

```
#Creates a new variable Treat with 6 levels
ToothGrowth$Treat=with(ToothGrowth,interaction(supp,dose))

#New variable that combines supplement type and dosage
tally(~Treat,data=ToothGrowth)
```

```
## Treat
## OJ.0.5 VC.0.5   OJ.1   VC.1   OJ.2   VC.2
##     10     10     10     10     10     10
```

The `tally` function helps us to check for balance; this is a balanced design because the same number of guinea pigs ($n_j = 10$ for $j = 1, 2, \ldots, 6$) were measured in each treatment combination.

With the variable `Treat` prepared, the first task is to visualize the results using boxplots and beanplots[8]

[7] A resistant procedure is one that is not severely impacted by a particular violation of an assumption. For example, the median is resistant to the impact of an outlier.

[8] Note that to see all the group labels in the plot when making the figure, you have to widen the plot window before copying

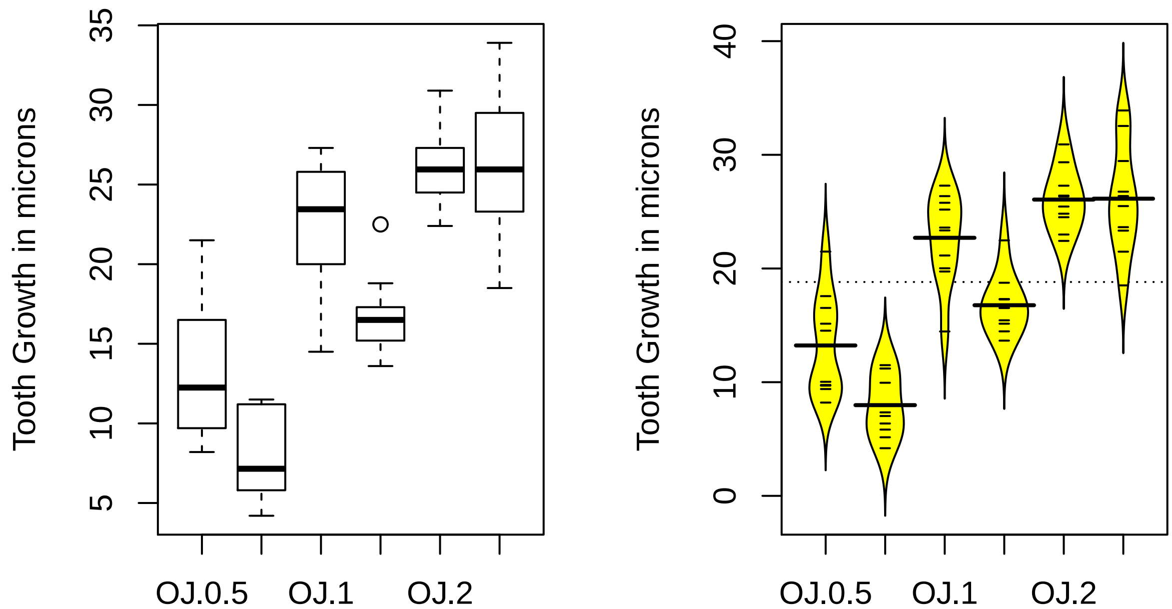


Figure 3.13: Boxplot and beanplot of tooth growth responses for the six treatment level combinations.

(Figure 3.13) and generate some summary statistics for each group using `favstats`.

```
par(mfrow=c(1,2))
boxplot(len~Treat,data=ToothGrowth,ylab="Tooth Growth in microns")
beanplot(len~Treat,data=ToothGrowth,log="",col="yellow",
         method="jitter",ylab="Tooth Growth in microns")

favstats(len~Treat,data=ToothGrowth)
```

```
##    Treat  min     Q1 median     Q3  max  mean       sd  n missing
## 1 OJ.0.5  8.2  9.700  12.25 16.175 21.5 13.23 4.459709 10       0
## 2 VC.0.5  4.2  5.950   7.15 10.900 11.5  7.98 2.746634 10       0
## 3   OJ.1 14.5 20.300  23.45 25.650 27.3 22.70 3.910953 10       0
## 4   VC.1 13.6 15.275  16.50 17.300 22.5 16.77 2.515309 10       0
## 5   OJ.2 22.4 24.575  25.95 27.075 30.9 26.06 2.655058 10       0
## 6   VC.2 18.5 23.375  25.95 28.800 33.9 26.14 4.797731 10       0
```

Figure 3.13 suggests that the mean tooth growth increases with the dosage level and that OJ might lead to higher growth rates than VC except at a dosage of 2 mg/day. The variability around the means looks to be small relative to the differences among the means, so we should expect a small p-value from our F-test. The design is balanced as noted above ($n_j = 10$ for all six groups) so the methods are some what resistant to impacts from non-normality and non-constant variance. There is some suggestion of non-constant variance in the plots but this will be explored further below when we can remove the difference in the means and combine all the residuals together. There might be some skew in the responses in some of the groups but there are only 10 observations per group so skew in the boxplots could be generated by impacts of very few of the observations.

Now we can apply our 6+ steps for performing a hypothesis test with these observations. The initial step is deciding on the claim to be assessed and the test statistic to use. This is a six group situation with a quantitative response, identifying it as a One-Way ANOVA where we want to test a null hypothesis that all the groups have the same population mean, at least to start. We will use a 5% significance level.

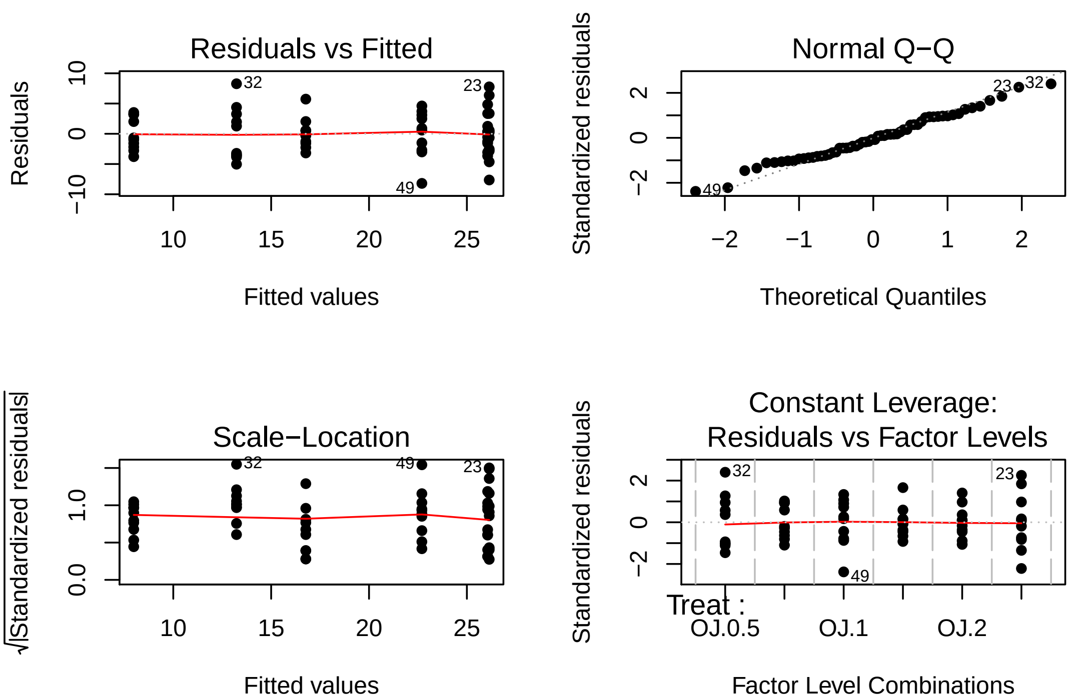
---

the figure out of R. You can resize the plot window using the small vertical and horizontal "=" signs in the grey bars that separate the different panels in RStudio.

1. **Hypotheses**: $\boldsymbol{H_0 : \mu_{OJ0.5} = \mu_{VC0.5} = \mu_{OJ1} = \mu_{VC1} = \mu_{OJ2} = \mu_{VC2}}$ **vs** $\boldsymbol{H_A :}$ **Not all** $\boldsymbol{\mu_j}$ **equal**
   - The null hypothesis could also be written in reference-coding as below since OJ.0.5 is chosen as the baseline group (discussed below).
     - $\boldsymbol{H_0 : \tau_{VC0.5} = \tau_{OJ1} = \tau_{VC1} = \tau_{OJ2} = \tau_{VC2} = 0}$
   - The alternative hypothesis can be left a bit less specific:
     - $\boldsymbol{H_A :}$ **Not all** $\boldsymbol{\tau_j}$ **equal 0**
2. **Validity conditions**:
   - Independence:
     - This is where the separate cages note above is important. Suppose that there were cages that contained multiple animals and they competed for food or could share illness or levels of activity. The animals in one cage might be systematically different from the others and this "clustering" of observations would present a potential violation of the independence assumption. If the experiment had the animals in separate cages, there is no clear dependency in the design of the study and we can assume that there is no problem with this assumption.
   - Constant variance:
     - As noted above, there is some indication of a difference in the variability among the groups in the boxplots and beanplots but the sample size was small in each group. We need to fit the linear model to get the other diagnostic plots to make an overall assessment.

   (ref:fig3-15) Diagnostic plots for the toothgrowth model.

```
m2<-lm(len~Treat,data=ToothGrowth)
par(mfrow=c(2,2))
plot(m2,pch=16)
```



     - The Residuals vs Fitted panel in Figure **??** shows some difference in the spreads but the spread is not that different between the groups.
     - The Scale-Location plot also shows just a little less variability in the group with the smallest fitted value but the spread of the groups looks fairly similar in this alternative scaling.
     - Put together, the evidence for non-constant variance is not that strong and we can assume that there is at least not a major problem with this assumption.
   - Normality of residuals:
     - The Normal Q-Q plot shows a small deviation in the lower tail but nothing that we wouldn't expect from a normal distribution. So there is no evidence of a problem with the normality

assumption in the upper right panel of Figure **??**.

3. **Calculate the test statistic**:
   - The ANOVA table for our model follows, providing an F-statistic of 41.557:

```
anova(m2)
```

```
## Analysis of Variance Table
##
## Response: len
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Treat      5 2740.10  548.02  41.557 < 2.2e-16
## Residuals 54  712.11   13.19
```

4. **Find the p-value**:
   - There are two options here, especially since it seems that our assumptions about variance and normality are not violated (note that we do not say "met" – we just have no clear evidence against them). The parametric and nonparametric approaches should provide similar results here.
   - The parametric approach is easiest – the p-value comes from the previous ANOVA table as `<2e-16`. First, note that this is in scientific notation that is a compact way of saying that the p-value here is $2.2 * 10^{-16}$ or 0.00000000000000022. When you see `2.2e-16` in R output, it also means that the calculation is at the numerical precision of the computer. What R is really trying to report is that this is a very small number. **When you encounter p-values that are smaller than 0.0001, you should just report that the p-value<0.0001** Do not report that it is 0 as this gives the false impression that there is no chance of the result occurring when it is just a really small probability. This distribution (the distribution of the test statistic if the null hypothesis is true).
   - The nonparametric approach is not too hard so we can compare the two approaches here as well:

(ref:fig3-16) Histogram and density curve of permutation distribution for F-statistic for tooth growth data. Observed test statistic in bold, vertical line at 41.56.

```
Tobs <- anova(lm(len~Treat,data=ToothGrowth))[1,4]; Tobs
```

```
## [1] 41.55718
```

```
par(mfrow=c(1,2))
B<- 1000
Tstar<-matrix(NA,nrow=B)
for (b in (1:B)){
  Tstar[b]<-anova(lm(len~shuffle(Treat),data=ToothGrowth))[1,4]
}
pdata(Tstar,Tobs,lower.tail=F)
```

```
## [1] 0
```

```
hist(Tstar,xlim=c(0,Tobs+3))
abline(v=Tobs,col="red",lwd=3)
plot(density(Tstar),xlim=c(0,Tobs+3),main="Density curve of Tstar")
abline(v=Tobs,col="red",lwd=3)
```

   - **The permutation p-value was reported as 0. This should be reported as p-value<0.001** since we did 1000 permutations and found that none of the permuted F-statistics, F^*, were larger than the observed F-statistic of 41.56. The permuted results do not exceed 6 as seen in Figure 3.14, so the observed result is *really unusual* relative to the null hypothesis. As suggested previously, the parametric and nonparametric approaches should be similar here and they were.

5. **Make a decision**:
   - Reject H_0 since the p-value is less than 5%.

6. **Write a conclusion**:
   - There is evidence at the 5% significance level that the different treatments (combinations of OJ/VC and dosage levels) **cause some** difference in the **true** mean tooth growth for **these** guinea pigs.
     - We can make the causal statement of the treatment causing differences because the treatments were randomly assigned but these inferences only apply to these guinea pigs since they were not randomly selected from a larger population.
     - Remember that we are making inferences to the population or true means and not the sample means and want to make that clear in any conclusion. When there is not a random sample from a population it is more natural to discuss the true means since we can't extend to the population values.
     - The alternative is that there is some difference in the true means – be sure to make the wording clear that you aren't saying that all the means differ. In fact, if you look back at Figure 3.13, the means for the 2 mg dosages look almost the same so we will have a tough time arguing that all groups differ. The F-test is about finding evidence of some means. The next section will provide some additional tools to get more specific about the source of those detected differences.

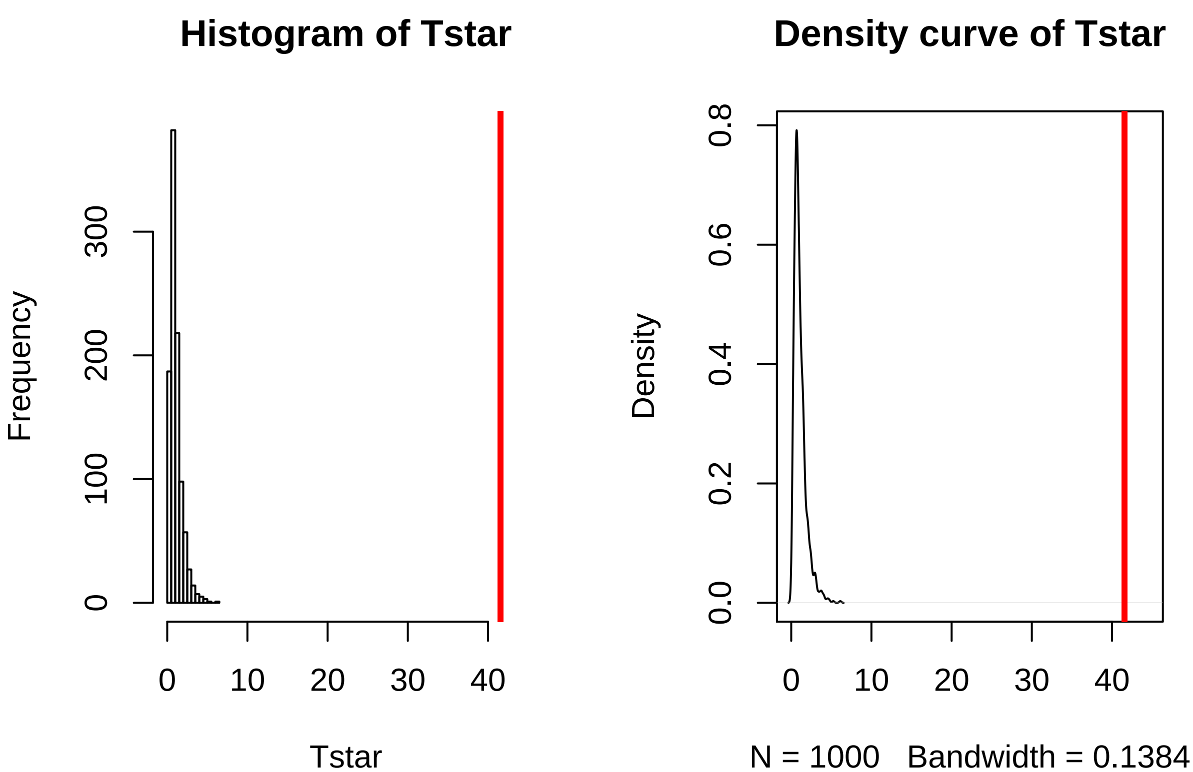


Figure 3.14: (ref:fig3-16)

Before we leave this example, we should revisit our model estimates and interpretations. The default model parameterization is into the reference-coding. Running the model `summary` function on `m2` provides the estimated coefficients:

```
summary(m2)
```

```
##
## Call:
## lm(formula = len ~ Treat, data = ToothGrowth)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8.20  -2.72  -0.27   2.65   8.27
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   13.230      1.148  11.521 3.60e-16
## TreatVC.0.5   -5.250      1.624  -3.233  0.00209
## TreatOJ.1      9.470      1.624   5.831 3.18e-07
## TreatVC.1      3.540      1.624   2.180  0.03365
## TreatOJ.2     12.830      1.624   7.900 1.43e-10
## TreatVC.2     12.910      1.624   7.949 1.19e-10
##
## Residual standard error: 3.631 on 54 degrees of freedom
## Multiple R-squared:  0.7937, Adjusted R-squared:  0.7746
## F-statistic: 41.56 on 5 and 54 DF,  p-value: < 2.2e-16
```

For some practice with the reference coding used in these models, let's find the estimates for observations for a couple of the groups. To work with the parameters, you need to start with diagnosing the baseline category that was used by considering which level is not displayed in the output. The `levels` function can list the groups in a categorical variable and their coding in the data set. The first level is usually the baseline category but you should check this in the model summary as well.

```
levels(ToothGrowth$Treat)
```

```
## [1] "OJ.0.5" "VC.0.5" "OJ.1"   "VC.1"   "OJ.2"   "VC.2"
```

There is a `VC.0.5` in the second row of the model summary, but there is no row for `OJ.0.5` and so this must be the baseline category. That means that the fitted value or model estimate for the OJ at 0.5 mg/day group is the same as the `(Intercept)` row or $\hat{\alpha}$, estimating a mean tooth growth of 13.23 microns when the pigs get OJ at a 0.5 mg/day dosage level. You should always start with working on the baseline level in a reference-coded model. To get estimates for any other group, then you can use the `(Intercept)` estimate and add the deviation for the group of interest. For `VC.0.5`, the estimated mean tooth growth is $\hat{\alpha}+\hat{\tau}_2 = \hat{\alpha}+\hat{\tau}_{VC0.5} = 13.23+(-5.25) = 7.98$ microns. It is also potentially interesting to directly interpret the estimated difference (or deviation) between `OJ0.5` (the baseline) and `VC0.5` (group 2) that is $\hat{\tau}_{VC0.5} = -5.25$: we estimate that the mean tooth growth in `VC0.5` is 5.25 microns shorter than it is in `OJ0.5`. This and many other direct comparisons of groups are likely of interest to researchers involved in studying the impacts of these supplements on tooth growth and the next section will show us how to do that (correctly!).

The reference-coding is still going to feel a little uncomfortable so the comparison to the cell-means model and exploring the effect plot can help to reinforce that both models patch together the same estimated means for each group. For example, we can find our estimate of 7.98 microns for the VC0.5 group in the output and Figure 3.15. Also note that Figure 3.15 is the same whether you plot the results from `m2` or `m3`.

```
m3<-lm(len~Treat-1,data=ToothGrowth)
summary(m3)
```

```
##
## Call:
## lm(formula = len ~ Treat - 1, data = ToothGrowth)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8.20  -2.72  -0.27   2.65   8.27
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## TreatOJ.0.5   13.230      1.148  11.521 3.60e-16
## TreatVC.0.5    7.980      1.148   6.949 4.98e-09
## TreatOJ.1     22.700      1.148  19.767  < 2e-16
## TreatVC.1     16.770      1.148  14.604  < 2e-16
## TreatOJ.2     26.060      1.148  22.693  < 2e-16
## TreatVC.2     26.140      1.148  22.763  < 2e-16
##
## Residual standard error: 3.631 on 54 degrees of freedom
## Multiple R-squared:  0.9712, Adjusted R-squared:  0.968
## F-statistic:   303 on 6 and 54 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(1,2))
plot(allEffects(m2))
```

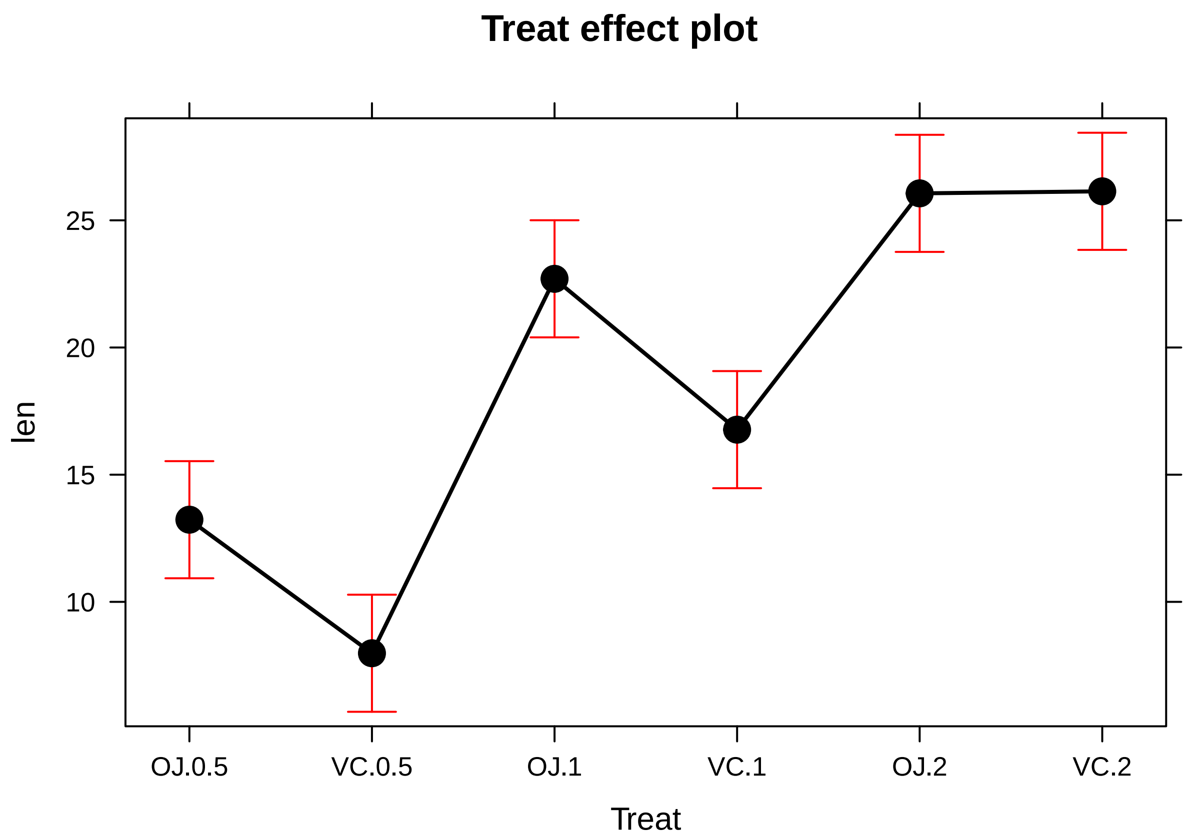


Figure 3.15: Effect plot of the One-Way ANOVA model for the toothgrowth data.

3.6 Multiple (pair-wise) comparisons using Tukey's HSD and the compact letter display

With evidence that the true means are likely not all equal, many researchers want to know which groups show evidence of differing from one another. This provides information on the source of the overall difference that was detected and detailed information on which groups differed from one another. Because this is a shot-gun/unfocused sort of approach, some people think it is an over-used procedure. Others feel that it is an important method of addressing detailed questions about group comparisons in a valid way. For example, we might want to know if OJ is dosage level and these methods will allow us to get an answer to this sort of question. It also will test for differences between the OJ,0.5 and VC,2 groups and every other pair of levels that you can construct. This method actually takes us back to the methods in Chapter **??** where we compared the means of two groups except that we need to deal with potentially many pair-wise comparisons, making an adjustment to account for that inflation in Type I errors that occurs due to many tests being performed at the same time. There are many different statistical methods to make all the pair-wise comparisons, but we will employ the most commonly used one, called ***Tukey's Honest Significant Difference*** (Tukey's HSD) method[9]. The name suggests that not using it could lead to a dishonest answer and that it will give you an honest result. It is more that if you don't do some sort of correction for all the tests you are performing, you might find some ***spurious***[10] results. There are other methods that could be used to do a similar correction and also provide "honest" inferences; we are just going to learn one of them.

Generally, the general challenge in this situation is that if you perform many tests at the same time, you inflate the Type I error rate. We can define the ***family-wise error rate*** as the probability that at least one error is made on a set of tests or, more compactly, Pr(At least 1 error is made) where Pr() is the probability of an event occurring. The family-wise error is meant to capture the overall situation in terms of measuring the likelihood of making a mistake if we consider many tests, each with some chance of making their own mistake, and focus on how often we make at least one error when we do many tests. A quick probability calculation shows the magnitude of the problem. If we start with a 5% significance level test, then Pr(Type I error on one test) =0.05 and the Pr(no errors made on one test) =0.95, by definition. This is our standard hypothesis testing situation. Now, suppose we have m independent tests, then

Figure **??**fig:Figure3-18) shows how the probability of having at least one false detection grows rapidly with the number of tests. The plot stops at 100 tests since it is effectively a 100% chance of at least on false detection. It might seem like doing 100 tests is a lot, but in Genetics research it is possible to consider situations where millions of tests are considered so these are real issues to be concerned about in many situations. Researchers want to make sure that when they report a "significant" result that it is really likely to be a real result and will show up as a difference in the next data set they collect.[11]

In pair-wise comparisons between all the pairs of means in a One-Way ANOVA, the number of tests is based on the number of pairs. We can calculate the number of tests using J choose 2, $\binom{J}{2}$, to get the number of unique pairs of size 2 that we can make out of J individual treatment levels. We don't need to explore the combinatorics formula for this, as the `choose` function can give us the answers:

```
choose(3,2)
## [1] 3
choose(4,2)
## [1] 6
choose(5,2)
## [1] 10
choose(6,2)
## [1] 15
```

So if you have three groups (prisoner rating study), there are 3 unique pairs to compare. For six groups, like in the guinea pig study, we have to consider 15 tests to compare all the unique pairs of groups. 15 tests seems like enough that we should be worried about inflated family-wise error rates. Fortunately, the Tukey's HSD method controls the family-wise error rate at your specified level (say 0.05) across any number of pair-wise

[9] When this procedure is used with unequal group sizes it is also sometimes called Tukey-Kramer's method.
[10] We often use "spurious" to describe falsely rejected null hypotheses, but there are also called false detections.
[11] Many researchers are now collecting multiple data sets to use in a single study and using one data set to identify interesting results and then using a validation or test data set that they withheld from initial analysis to verify that the first results are also present in that second data set.

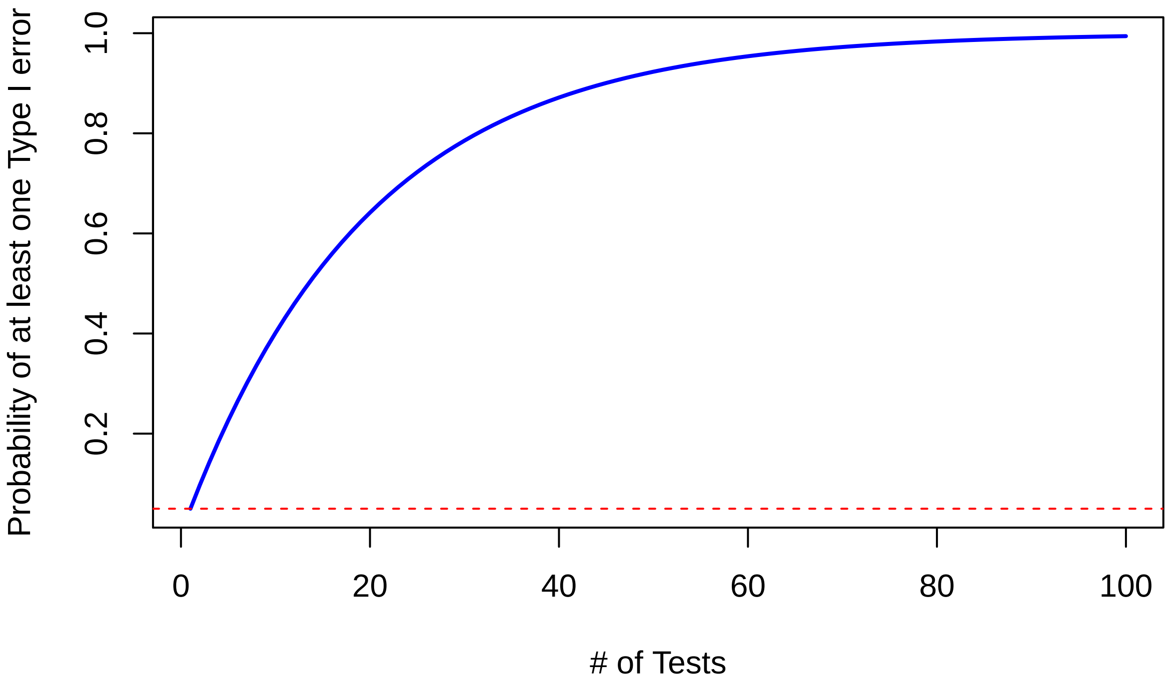


Figure 3.16: Plot of family-wise error rate as the number of tests performed increases. Dashed line indicates 0.05.

comparisons. This means that the overall rate of at least one Type I error is controlled at the specified significance level, often 5%. To do this, each test must use a slightly more conservative cut-off than if just one test is performed and the procedure helps us figure out how much more conservative we need to be.
Tukey's HSD starts with focusing on the difference between the groups with the largest and smallest means $(\bar{y}_{max} - \bar{y}_{min})$. If $(\bar{y}_{max} - \bar{y}_{min}) \leq$ Margin of Error for the difference in the means, then all other pairwise differences, say $|\bar{y}_j - \bar{y}_{j'}|$, for two groups j and j', will be less than or equal to that margin of error. This also means that any confidence intervals for any difference in the means will contain 0. Tukey's HSD selects a critical value so that $(\bar{y}_{max} - \bar{y}_{min})$ will be less than the margin of error in 95% of data sets drawn from populations with a common mean. This implies that in 95% of data sets in which all the population means are the same, all confidence intervals for differences in pairs of means will contain 0. Tukey's HSD provides confidence intervals for the difference in true means between groups j and j', $\mu_j - \mu_{j'}$, for all pairs where $j \neq j'$, using

$$(\bar{y}_j - \bar{y}_{j'}) \mp \frac{q^*}{\sqrt{2}} \sqrt{\text{MS}_E \left(\frac{1}{n_j} + \frac{1}{n_{j'}}\right)}$$

where $\frac{q^*}{\sqrt{2}} \sqrt{\text{MS}_E \left(\frac{1}{n_j} + \frac{1}{n_{j'}}\right)}$ is the margin of error for the intervals. The distribution used to find the multiplier, q^*, for the confidence intervals is available in the `qtukey` function and generally provides a slightly larger multiplier than the regular t^* from our two-sample t-based confidence interval discussed in Chapter **??**. We will use the `confint`, `cld`, and `plot` functions applied to output from the `glht` function (all from the `multcomp` package; Hothorn and Westfall (2008), (Hothorn et al., 2016)) to easily get the required comparisons from our ANOVA model. Unfortunately, its code format is a little complicated – but there are just two places to modify the code, by including the model name and after `mcp` (stands for multiple comparisons) in the `linfct` option, you need to include the explanatory variable name as `VARIABLENAME="Tukey"`. The last part is to get the TukeyHSD multiple comparisons run on our explanatory variable. Once we obtain the intervals, we can use them to test $H_0 : \mu_j = \mu_{j'}$ vs $HA : \mu_j \neq \mu j'$ by assessing whether 0 is in the confidence interval for each pair. If 0 is in the interval, then there is no evidence of a difference for that pair. If 0 is not in the interval, then we reject H_0 and have evidence *at the specified family-wise significance level* of a difference for that pair. You will see a switch to using the word "detection" to describe rejected null hypotheses of no difference as it can help to write up these results. The following code provides the numerical and graphical[12] results of applying Tukey's HSD to the linear model for the Guinea Pig data:

```
par(mfrow=c(1,1))
require(multcomp)
Tm2 <- glht(m2, linfct = mcp(Treat = "Tukey"))
confint(Tm2)
```

```
##
##   Simultaneous Confidence Intervals
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = len ~ Treat, data = ToothGrowth)
##
## Quantile = 2.9545
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##                      Estimate lwr      upr
## VC.0.5 - OJ.0.5 == 0  -5.2500 -10.0482  -0.4518
## OJ.1 - OJ.0.5 == 0     9.4700   4.6718  14.2682
## VC.1 - OJ.0.5 == 0     3.5400  -1.2582   8.3382
## OJ.2 - OJ.0.5 == 0    12.8300   8.0318  17.6282
## VC.2 - OJ.0.5 == 0    12.9100   8.1118  17.7082
## OJ.1 - VC.0.5 == 0    14.7200   9.9218  19.5182
## VC.1 - VC.0.5 == 0     8.7900   3.9918  13.5882
## OJ.2 - VC.0.5 == 0    18.0800  13.2818  22.8782
## VC.2 - VC.0.5 == 0    18.1600  13.3618  22.9582
## VC.1 - OJ.1 == 0      -5.9300 -10.7282  -1.1318
## OJ.2 - OJ.1 == 0       3.3600  -1.4382   8.1582
## VC.2 - OJ.1 == 0       3.4400  -1.3582   8.2382
## OJ.2 - VC.1 == 0       9.2900   4.4918  14.0882
## VC.2 - VC.1 == 0       9.3700   4.5718  14.1682
## VC.2 - OJ.2 == 0       0.0800  -4.7182   4.8782
```

```
old.par <- par(mai=c(1.5,2,1,1)) #Makes room on the plot for the group names
plot(Tm2)
```

Figure 3.17 contains confidence intervals for the difference in the means for all 15 pairs of groups. For example,

[12]The plot of results usually contains all the labels of groups but if the labels are long or there many groups, sometimes the row labels are hard to see even with re-sizing the plot to make it taller in RStudio. The numerical output is useful as a guide to help you read the plot.

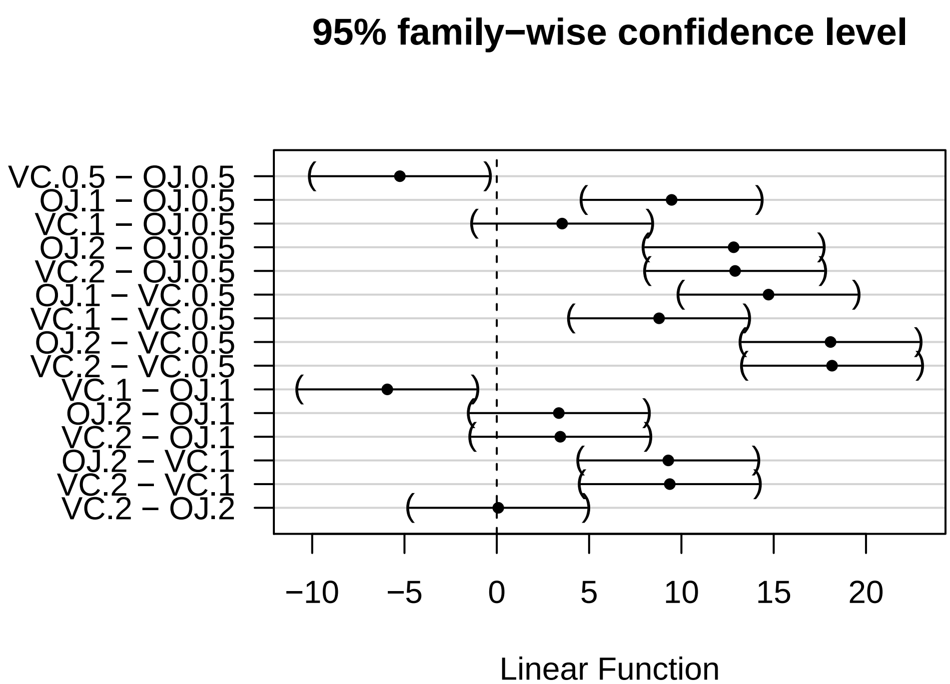


Figure 3.17: Graphical display of pair-wise comparisons from Tukey's HSD for the Guinea Pig data. Any confidence intervals that do not contain 0 provide evidence of a difference in the groups.

the first row in the plot contains the confidence interval OJ.0.5). In the numerical output, you can find that this 95% family-wise confidence interval goes from -10.05 to -0.45 microns (`lwr` and `upr` in the numerical output provide the CI endpoints). This interval does not contain 0 since its upper end point is -0.45 microns and so we can now say that there is evidence that OJ and VC have different true mean growth rates at the 0.5 mg dosage level. We can go further and say that we are 95% confident that the difference in the true mean tooth growth between VC.0.5 and OJ.0.5 (VC.0.5-OJ.0.5) is between -10.05 and -0.45 microns, after adjusting for comparing all the pairs of groups. But there are fourteen more similar intervals...

If you put all these pair-wise tests together, you can generate an overall interpretation of Tukey's HSD results that discusses sets of groups that are not detectably different from one another and those groups that were distinguished from other sets of groups. To do this, start with listing out the groups that do are not detectably different (CIs contain 0), which, here, only occurs for four of the pairs. The CIs that contain 0 are for the pairs VC.1 and OJ.0.5, OJ.2 and OJ.1, VC.2 and OJ.1, and, finally, VC.2 and OJ.2. So VC.2, OJ.1, and OJ.2 are all not detectably different from each other and VC.1 and OJ.0.5 are also not detectably different. If you look carefully, VC.0.5 is detected as different from every other group. So there are basically three sets of groups that can be grouped together as "similar": VC.2, OJ.1, and OJ.2; VC.1 and OJ.0.5; and VC.0.5. Sometimes groups overlap with some levels not being detectably different from other levels that belong to different groups and the story is not as clear as it is in this case. An example of this sort of overlap is seen in the next section.

There is a method that many researchers use to more efficiently generate and report these sorts of results that is called a ***compact letter display*** (CLD, Piepho (2004)). The `cld` function can be applied to the results from `glht` to generate the CLD that we can use to provide a "simple" summary of the sets of groups. In this discussion, we define a **set as a union of different groups that can contain one or more members** and the member of these groups are the different treatment levels.

```
cld(Tm2)
```

```
## OJ.0.5 VC.0.5   OJ.1   VC.1   OJ.2   VC.2
##    "b"    "a"    "c"    "b"    "c"    "c"
```

Groups with the same letter are not detectably different (are in the same set) and groups that are detectably different get different letters (are in different sets). Groups can have more than one letter to reflect "overlap" between the sets of groups and sometimes a set of groups contains only a single treatment level (VC.0.5 is a set of size 1). Note that if the groups have the same letter, this does not mean they are the same, just that there is **no evidence of a difference for that pair**. If we consider the previous output for the CLD, the "a" set contains VC.0.5, the "b" set contains OJ.1, OJ.2, and VC.2, and the "c" set contains OJ.0.5 and VC.1. These are exactly the groups of treatment levels that we obtained by going through all fifteen pairwise results. One benefit of this work is that the CLD letters can be added to a beanplot to help fully report the results and understand the sorts of differences Tukey's HSD detected.

The lines with text in them are involved in placing text on the figure but are something you could do in image editing software just as easily. Figure 3.18 enhances the discussion by showing that the "**a**" group with VC.0.5 had the lowest average tooth growth, the "**c**" group had intermediate tooth growth for treatments OJ.0.5 and VC.1, and the highest growth rates came from OJ.1, OJ.2, and VC.2. Even though VC.2 had the highest average growth rate, we are not able to prove that its true mean is any higher than the other groups labeled with "**b**". Hopefully the ease of getting to the story of the Tukey's HSD results from a plot like this explains why it is common to report results using these methods instead of reporting 15 confidence intervals.

There are just a couple of other details to mention on this set of methods. First, note that we interpret the set of confidence intervals simultaneously: We are 95% confident that **ALL** the intervals contain the respective differences in the true means (this is a ***family-wise interpretation***). These intervals are adjusted from our regular 2 sample t intervals from Chapter **??** to allow this stronger interpretation. Specifically, they are wider. Second, if sample sizes are unequal in the groups, Tukey's HSD is conservative and provides a family-wise error rate that is lower than the nominal (or specified) level. In other words, it fails less often than expected and the intervals provided are a little wider than needed, containing all the pairwise differences at higher than the nominal confidence level of (typically) 95%. Third, this is a parametric approach and violations of normality and constant variance will push the method in the other direction, potentially making the technique dangerously liberal. Nonparametric approaches to this problem are also possible, but will not be considered here.

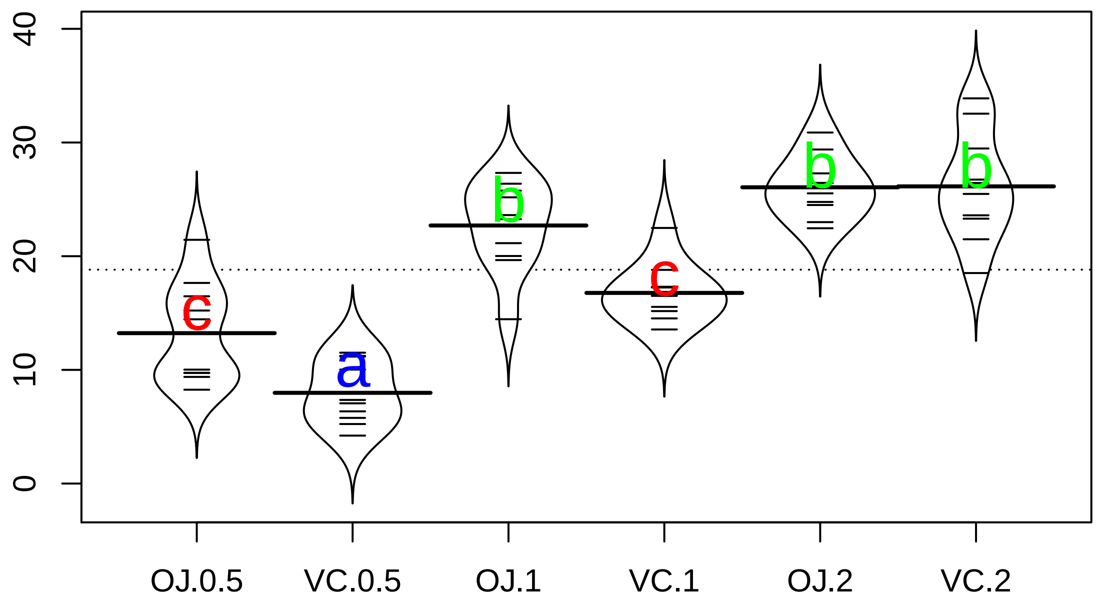


Figure 3.18: Beanplot of tooth growth by group with Tukey?s HSD compact letter display.

3.7 Pair-wise comparisons for Prisoner Rating data

In our previous work with the prisoner rating data, the overall ANOVA test provided only marginal evidence of some difference in the true means across the three groups with a p-value=0.067. Tukey's HSD does not require you to find a small from your overall F-test to employ the methods but if you apply it to situations with p-values larger than your *a priori* significance level, you are unlikely to find any pairs that are detected as being different. Some statisticians suggest that you shouldn't employ follow-up tests such as Tukey's HSD when there is not sufficient evidence to reject the overall null hypothesis and would be able to reasonably criticize the following results. But for the sake of completeness, we can find the pair-wise comparison results at our typical 95% family-wise confidence level in this situation, with the three confidence intervals displayed in Figure 3.19.

```
lm2<-lm(Years~Attr, data=MockJury)
require(multcomp)
Tm2 <- glht(lm2, linfct = mcp(Attr = "Tukey"))
confint(Tm2)
```

```
##
##   Simultaneous Confidence Intervals
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = Years ~ Attr, data = MockJury)
##
## Quantile = 2.3751
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##                               Estimate lwr     upr
## Average - Beautiful == 0      -0.3596  -2.2969  1.5776
## Unattractive - Beautiful == 0  1.4775  -0.4730  3.4280
## Unattractive - Average == 0    1.8371  -0.1258  3.8001
```

```
cld(Tm2)
```

```
##    Beautiful      Average Unattractive
##          "a"          "a"          "a"
```

```
old.par <- par(mai=c(1.5,2.5,1,1)) #Makes room on the plot for the group names
plot(Tm2)
```

At the family-wise 5% significance level, there are no pairs that are detectably different – they all get the same letter of "a". Now we will produce results for the reader that thought a 10% significance was suitable for this application before seeing any of the results. We just need to change the confidence level or significance level that the CIs or tests are produced with inside the functions. For the `confint` function, the `level` option is the confidence level and for the `cld`, it is the family-wise significance level. Note that 90% confidence corresponds to a 10% significance level.

```
confint(Tm2,level=0.9)
```

```
##
##   Simultaneous Confidence Intervals
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = Years ~ Attr, data = MockJury)
##
## Quantile = 2.0739
## 90% family-wise confidence level
##
##
## Linear Hypotheses:
##                               Estimate lwr     upr
## Average - Beautiful == 0      -0.3596  -2.0513  1.3320
## Unattractive - Beautiful == 0  1.4775  -0.2257  3.1806
## Unattractive - Average == 0    1.8371   0.1231  3.5512
```

```
cld(Tm2,level=0.1)
```

```
##    Beautiful      Average Unattractive
##         "ab"          "a"          "b"
```

```
old.par <- par(mai=c(1.5,2.5,1,1)) #Makes room on the plot for the group names
plot(confint(Tm2,level=.9))
```

With family-wise 10% significance and 90% confidence levels, the *Unattractive* and *Average* picture groups are detected as being different but the *Average* group is not detected as different from *Beautiful* and *Beautiful* is not detected to be different from *Unattractive*. This leaves the "overlap" of groups across the sets of groups that was noted earlier. The *Beautiful* level is not detected as being dissimilar from levels in two different sets and so gets two different letters.

The beanplot (Figure 3.21) helps to clarify some of the reasons for this set of results. The detection of a difference between *Average* and *Unattractive* just barely occurs and the mean for *Beautiful* is between the other two so it ends up not being detectably different from either one. This sort of overlap is actually a fairly common occurrence in these sorts of situations so be prepared a mixed set of letters for some levels.

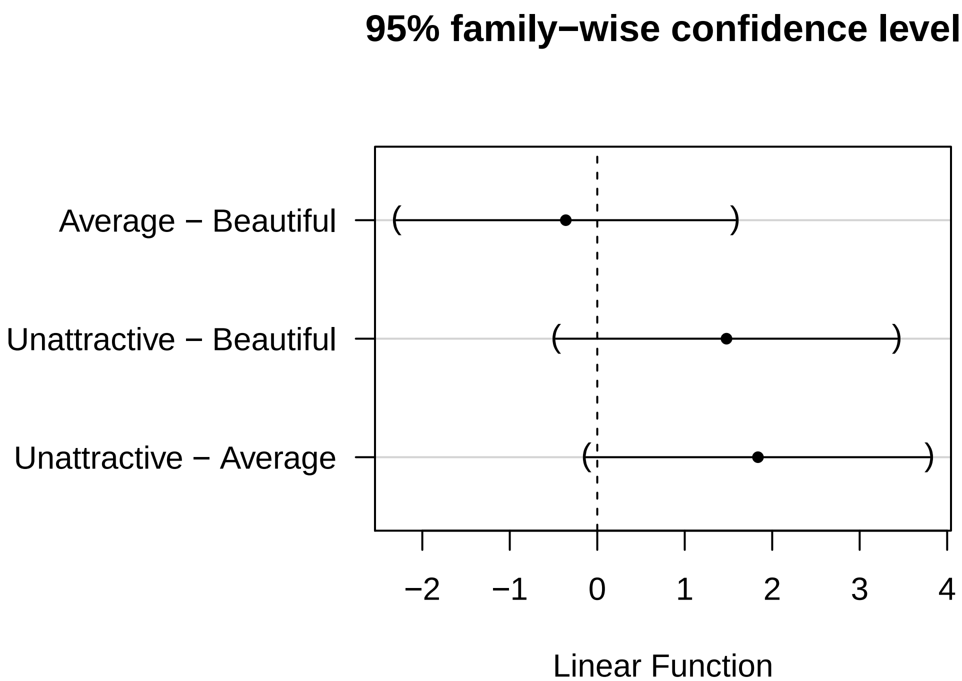


Figure 3.19: Tukey's HSD confidence interval results at the 95% family-wise confidence level.

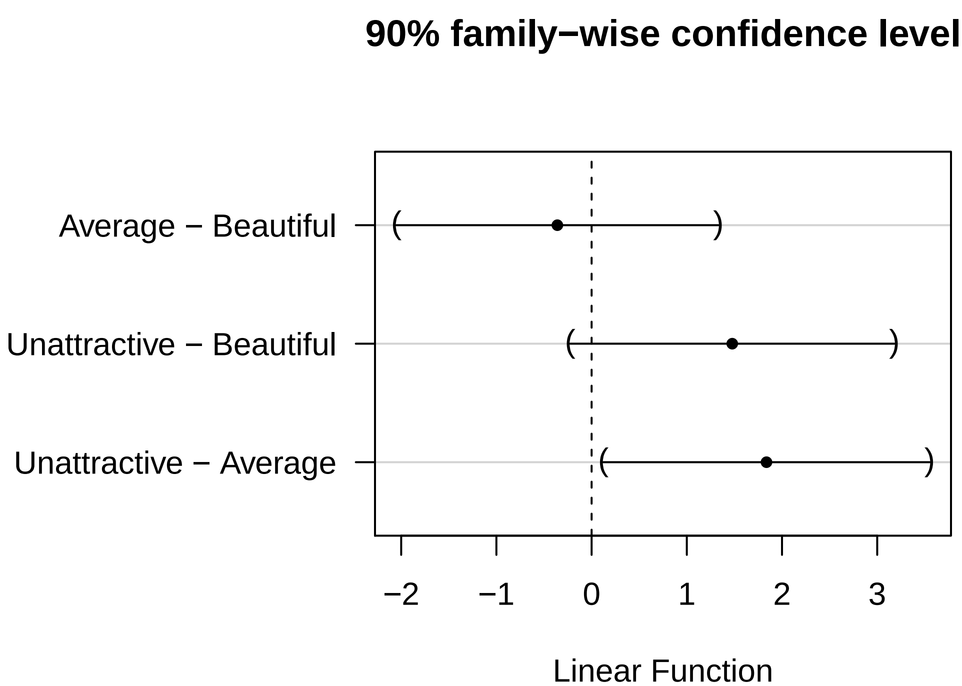


Figure 3.20: Tukey's HSD 90% family-wise confidence intervals.

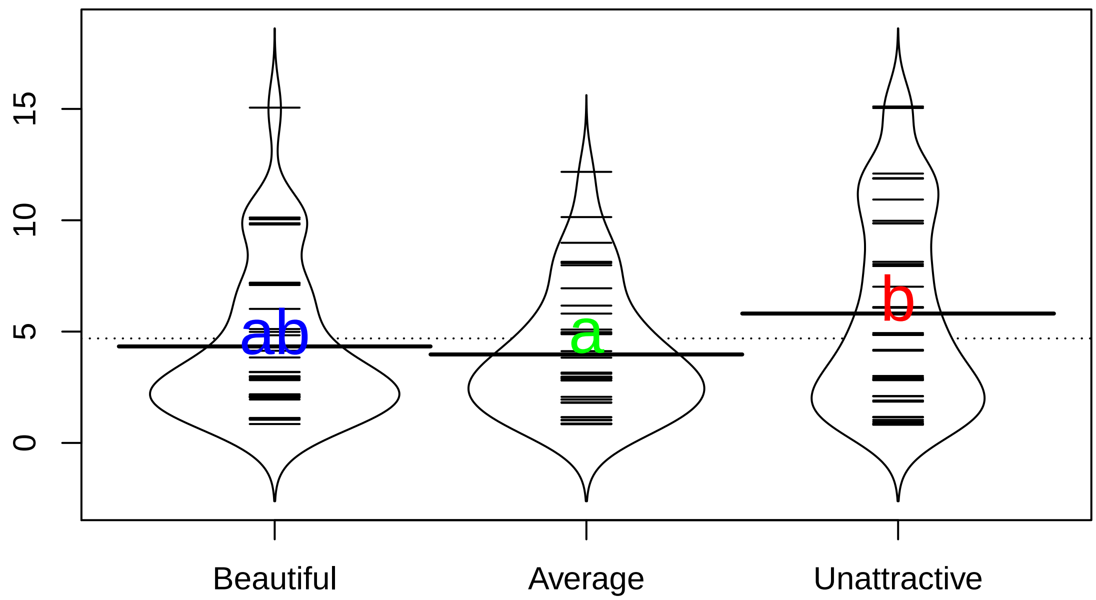


Figure 3.21: Beanplot of sentences with compact letter display results from 10% family-wise significance level Tukey's HSD. *Average* and *Unattractive* picture groups are detected as being different and are displayed as belonging to different groups. *Beautiful* picture responses are not detected as different from the other two groups.

```
beanplot(Years~Attr,data=MockJury,log="",col="white",method="jitter")
text(c(1),c(5),"ab",col="blue",cex=2)
text(c(2),c(4.8),"a",col="green",cex=2)
text(c(3),c(6.5),"b",col="red",cex=2)
```

3.8 Chapter Summary

In this chapter, we explored methods for comparing a quantitative response across J groups ($J \geq 2$), with what is called the One-Way ANOVA procedure. The initial test is based on assessing evidence against a null hypothesis of no groups. There are two different methods for estimating these One-Way ANOVA models: the cell-means model and the reference-coded versions of the model. There are times when either model will be preferred, but for the rest of the text, the reference coding is used (sorry!). The ANOVA F-statistic, often presented with underlying information in the ANOVA table, provides a method of assessing evidence against the null hypothesis either using permutations or via the F-distribution. Pair-wise comparisons using Tukey's HSD provide a method for comparing all the groups and are a nice complement to the overall ANOVA results. A compact letter display was shown that enhanced the interpretation of Tukey's HSD result.

In the guinea pig example, we are left with some lingering questions based on these results. It appears that the effect of *dosage* changes as a function of the *delivery method* (OJ, VC) because the size of the differences between OJ and VC change for different dosages. These methods can't directly assess the question of whether the effect of delivery method is the same or not across the different dosages. In chapter 4, the two variables, *Dosage* and *Delivery method* are modeled as two separate variables so we can consider their effects both separately and together. This allows more refined hypotheses, such as *Is the effect of delivery method the same for all dosages?*, to be tested. This will introduce new models and methods for analyzing data where there are two factors as explanatory variables in a model for a quantitative response variable in what is called the Two-Way ANOVA.

3.9 Summary of important R code

The main components of R code used in this chapter follow with components to modify in red, remembering that any R packages mentioned need to be installed and loaded for this code to have a chance of working:

- MODELNAME <- lm(Y~X, data=DATASETNAME)
  - Probably the most frequently used command in R.
  - Here it is used to fit the reference-coded One-Way ANOVA model with Y as the response variable and X as the grouping variable, storing the estimated model object in MODELNAME.
- MODELNAME <- lm(Y~X-1, data=DATASETNAME)
  - Fits the cell means version of the One-Way ANOVA model.
- summary(MODELNAME)
  - Generates model summary information including the estimated model coefficients, SEs, t-tests, and p-values.
- anova(MODELNAME)
  - Generatesthe ANOVA table but **must only be run on the reference-coded version of the model**.
  - Results are incorrect if run on the cell-means model since the reduced model under the null is that the mean of all the observations is 0!
- pf(FSTATISTIC,df1=NUMDF,df2=DENOMDF, lower.tail=F)
  - Finds the p-value for an observed F-statistic with NUMDF and DENOMDF degrees of freedom.
- par(mfrow=c(2,2)); plot(MODELNAME)
  - Generates four diagnostic plots including the Residuals vs Fitted and Normal Q-Q plot.
- plot(allEffects(MODELNAME))
  - Requires the `effects` package be loaded.
  - Plots the estimated model component.
- Tm2 <- glht(MODELNAME, linfct=mcp(X="Tukey")); confint(Tm2); plot(Tm2); cld(Tm2)
  - Requires the `multcomp` package to be installed and loaded.
  - Can only be run on the reference-coded version of the model.
  - Generates the text output and plot for Tukey's HSD as well as the compact letter display.

3.10 Practice problems

For these practice problems, you will work with the cholesterol data set from the `multcomp` package that was used to generate the Tukey's HSD results. To load the data set and learn more about the study, use the following code:

```
require(multcomp)
data(cholesterol)
help(cholesterol)
```

3.1. Graphically explore the differences in the changes in Cholesterol levels for the five levels using boxplots and beanplots.
3.2. Is the design balanced?
3.3. Complete all 6+ steps of the hypothesis test using the parametric F-test, reporting the ANOVA table and the distribution of the test statistic under the null.
3.4. Discuss the scope of inference using the information that the treatment levels were randomly assigned to volunteers in the study.
3.5. Generate the permutation distribution and find the p-value. Compare the parametric p-value to the permutation test results.
3.6. Perform Tukey's HSD on the data set. Discuss the results – which pairs were detected as different and which were not? Bigger reductions in cholesterol are good, so are there any levels you would recommend or

that might provide similar reductions?

3.7. Find and interpret the CLD and compare that to your interpretation of results from 3.6.

Chapter 4

Two-Way ANOVA

Chapter 5

Chi-square tests

Chapter 6

Correlation and Simple Linear Regression

Chapter 7

Simple linear regression inference

Chapter 8

Multiple linear regression

Chapter 9

Case studies

Chapter 10

Placeholder

Bibliography

Crampton, E. (1947). The growth of the odontoblast of the incisor teeth as a criterion of vitamin c intake of the guinea pig. *The Journal of Nutrition*, 33(5):491–504.

Fox, J. (2003). Effect displays in r for generalised linear models. *Journal of Statistical Software*, 8(15):1–27.

Fox, J., Weisberg, S., Friendly, M., and Hong, J. (2016). *effects: Effect Displays for Linear, Generalized Linear, and Other Models*. R package version 3.1-2.

Hothorn, T., B. F. and Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, 50(3):346–363.

Hothorn, T., Bretz, F., and Westfall, P. (2016). *multcomp: Simultaneous Inference in General Parametric Models*. R package version 1.4-6.

Piepho, H. (2004). An algorithm for a letter-based representation of all-pairwise comparisons. *Journal of Computational and Graphical Statistics*, 13(2):456–466.