Chapter 12

One-Way Analysis of Variance and Nonparametric Alternatives

In the last chapter, we discussed confidence intervals for the difference of two population means $\mu_1 - \mu_2$ and population medians $M_1 - M_2$. Perhaps more importantly, we also saw that the design of the experiment or study completely determined how the analysis should proceed.

- When the two samples are independent, this is called a (two) independent-sample design.
- When the two samples are obtained on the same individuals (so that the samples are dependent), this is called a matched pairs design.
- Confidence interval procedures for $\mu_1 - \mu_2$ depend on the design of the study.

More generally, the purpose of an experiment is to investigate differences between or among two or more treatments. In a statistical framework, we do this by comparing the population means of the responses to each treatment.

In order to detect treatment mean differences, we must try to control the effects of error so that any variation we observe can be attributed to the effects of the treatments rather than to structural differences among the individuals.

In Example 11.2, where we explored data on the amount of white paper disposed of at two plants, there may be a systematic source of variation arising from the ages of employees in the recycling project (e.g., younger employees may be more inclined to recycle paper instead of discarding it). Our two-independent sample design (one sample from Plant 1 and one sample from Plant 2) did not consider this potential confounding effect. In other words, even if age of the employee is a significant source of variability, our independent sample analysis does not acknowledge it.

Designs involving meaningful grouping of individuals, that is, **blocking**, can help reduce the effects of experimental error by identifying systematic components of variation among individuals. The matched pairs design for comparing two treatments is an example of such a design. In this situation, the "meaningful grouping of individuals" involves the individuals themselves. Responses to two different treatments on the same individual "blocks out" the variation that would arise had we observed one individual's response to the first treatment and a different individual's response to the second treatment.

Remark: Aside from matched pairs experiments, the analysis of data from experiments involving blocking will not be covered in this course. When there are more than two treatments (populations), we pursue the one-way classification model. This is basically an extension of the two independent sample design to two or more populations.

Consider an experiment to compare $t \geq 2$ treatments set up as follows:

- We obtain one random sample of individuals and then randomly assign individuals to treatments (i.e., different experimental conditions). Samples corresponding to the treatment groups are independent.
- In an observational study (where no treatment is physically applied to individuals), individuals are inherently different to begin with. We therefore simply take random samples from each treatment population.
- We do not attempt to group individuals according to some other factor (e.g., location, gender, weight, variety, etc.). This would be an example of blocking.

In a one-way classification, the only way individuals are "classified" is by the treatment group assignment. When individuals are thought to be "basically alike" (other than the possible effect due to treatment), experimental error consists only of the variation among the individuals themselves. There are no other systematic sources of variability.

Example 12.1. Mortar mixes are usually classified on the basis of compressive strength and their bonding properties and flexibility. In a building project, engineers wanted to compare specifically the population mean strengths of four types of mortars:

1. ordinary cement mortar (OCM)
2. polymer impregnated mortar (PIM)
3. resin mortar (RM)
4. polymer cement mortar (PCM).

Random samples of specimens of each mortar type were taken; each specimen was subjected to a compression test to measure strength in megapascal (MPa) which is millions of newtons per square meter. The strength measurements taken on different mortar specimens (36 in all) are reported in the table below.

OCM:	51.45	42.96	41.11	48.06	38.27	38.88	42.74	49.62		
PIM:	64.97	64.21	57.39	52.79	64.87	53.27	51.24	55.87	61.76	67.15
RM:	48.95	62.41	52.11	60.45	58.07	52.16	61.71	61.06	57.63	56.80
PCM:	35.28	38.59	48.64	50.99	51.52	52.85	46.75	48.31		

First note that this is an example of an observational study. This is not what statisticians would call an experiment, because the "individuals" (here, the mortar specimens) are not treated or influenced by different experimental conditions. We are simply observing individuals from different groups (populations) to begin with.

There is no form of blocking here either. For example, we do not attempt to further classify individual mortar specimens according to different manufacturers or subject individual mortar

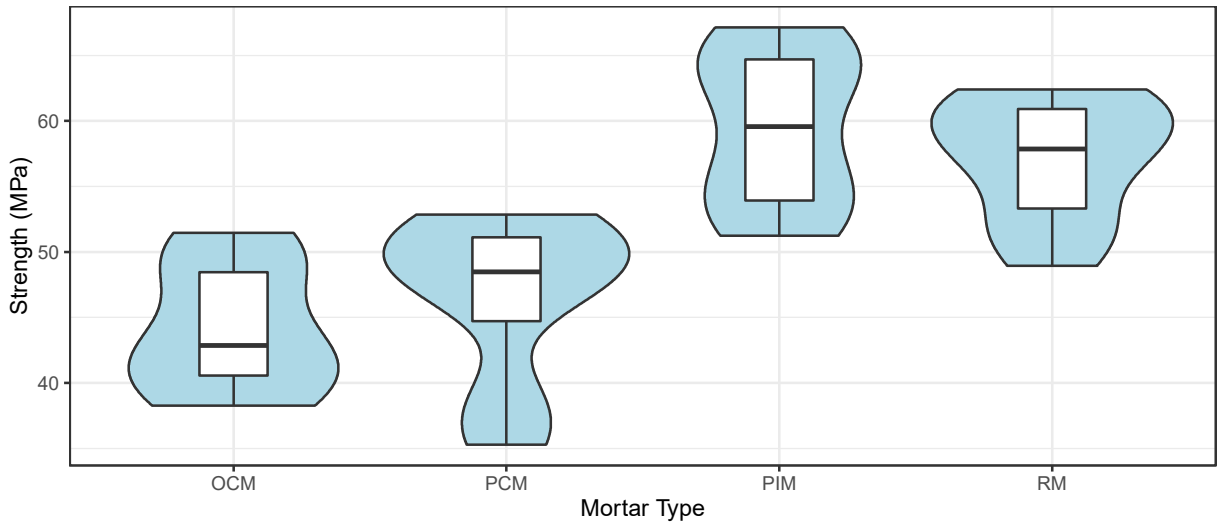

Figure 12.0.1: Violin plots of strength data (MPa) for four mortar types..

specimens to different environmental conditions (e.g., high/low temperature, etc.). If the study's purpose was investigate these potential sources of variability, then this would not be a one-way classification.

Side by side violin plots of these data are given in Figure 12.0.1 and created with the following `R` code.

```
> dat.mortar<-data.frame(strength=c(51.45,42.96,41.11,48.06,38.27,38.88,42.74,49.62,
+                          64.97,64.21,57.39,52.79,64.87,53.27,51.24,55.87,61.76,67.15,
+                          48.95,62.41,52.11,60.45,58.07,52.16,61.71,61.06,57.63,56.80,
+                          35.28,38.59,48.64,50.99,51.52,52.85,46.75,48.31),
+                         type=c(rep("OCM",8),rep("PIM",10),rep("RM",10),rep("PCM",8)))
> ggplot(data=dat.mortar,aes(x=type, y=strength))+
+   geom_violin(fill="lightblue")+
+   geom_boxplot(width=0.25)+
+   theme_bw()+
+   xlab("Mortar Type")+
+   ylab("Strength (MPa)")
```

An initial question that engineers may have is the following:

> "Are the population mean mortar strengths equal among the four types of mortars? Or, are the population means different?"

This initial question can be framed statistically as the following hypothesis test:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$
$$H_a : \text{the population means } \mu_i \text{ are not all equal.}$$

Remark: The alternative simply states that at least one of the means is different.

We now develop a statistical inference procedure that allows us to test this type of hypothesis in a one-way classification.

12.1 ANOVA – The Overall F test

Let t denote the number of treatments (populations) to be compared. Define

$$X_{ij} = \text{response on the } j\text{th individual in the } i\text{th treatment group}$$

for $i = 1, 2, ..., t$ and $j = 1, 2, ..., n_i$ where n_i is the number of observations for the ith treatment (population). In Example 12.1, these are $n_1 = 8$, $n_2 = 10$, $n_3 = 10$, and $n_4 = 8$.

When $n_1 = n_2 = ... = n_t = n$ we say the design is **balanced**; otherwise, the design is **unbalanced**. Let $N = n_1 + n_2 + ... + n_t$ denote the total number of measured individuals. If the design is balanced, then $N = nt$.

Define the statistics

$$\bar{X}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij} \qquad \textbf{[Group Mean]}$$

$$S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\cdot})^2 \qquad \textbf{[Group Variance]}$$

$$\bar{X}_{\cdot\cdot} = \frac{1}{N}\sum_{i=1}^{t}\sum_{j=1}^{n_i} X_{ij} \qquad \textbf{[Overall Mean]}$$

The statistics $\bar{X}_{i\cdot}$ and S_i^2 denote the sample mean and the sample variance, respectively, of the ith sample. The overall sample mean $\bar{X}_{\cdot\cdot}$ is the sample mean of all the data (aggregated across all t treatment groups).

Our goal is to develop a procedure to test

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_t$$
$$H_a : \text{the population means } \mu_i \text{ are not all equal.}$$

The null hypothesis H_0 says that there is "no treatment difference," that is, all t population means are the same. The alternative hypothesis H_a says that a difference among the t population means exists "somewhere." It does not specify how the means are different. When performing a hypothesis test, we basically decide which hypothesis is more supported by the data.

Suppose that we have t independent random samples:

$$\begin{aligned}
\text{Sample 1:} \quad & X_{11}, X_{12}, ..., X_{1n_1} \sim \text{Gaussian}(\mu_1, \sigma^2) \\
\text{Sample 2:} \quad & X_{21}, X_{22}, ..., X_{2n_2} \sim \text{Gaussian}(\mu_2, \sigma^2) \\
\vdots \quad & \quad \vdots \\
\text{Sample } t\text{:} \quad & X_{t1}, X_{t2}, ..., X_{tn_t} \sim \text{Gaussian}(\mu_t, \sigma^2)
\end{aligned}$$

Note the statistical assumptions we are making:

1. the t random samples are independent
2. the t population distributions are Gaussian (normal)

3. the t population distributions have the same variance σ^2

Note also that these are the same assumptions we made for the two independent-sample design in the last chapter; i.e., the special case when $t = 2$.

Question: If we are trying to learn about how the population means compare, why is the statistical inference procedure designed to do this called "the analysis of variance?"

Answer: We learn about the population means by estimating the common variance σ^2 in two different ways. These two estimators are formed by

- measuring variability of the observations within each sample
- measuring variability of the sample means across the samples

These two estimates tend to be similar when H_0 is true. The second estimate tends to be larger than the first estimate when H_a is true.

Within Estimator: Calculate the residual sum of squares:

$$\begin{aligned} SS_{res} &= (n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + \cdots + (n_t - 1)S_t^2 \\ &= \sum_{i=1}^{t} \underbrace{\sum_{i=1}^{n_i} (X_{ij} - \bar{X}_{i\cdot})^2}_{(n_i - 1)S_i^2} . \end{aligned}$$

The sample variance S_i^2 estimates the population parameter σ^2, which is assumed to be common across all t populations) from within the ith sample.

The weighted average of these estimates

$$\begin{aligned} MS_{res} &= \left(\frac{n_1 - 1}{N - t}\right) S_1^2 + \left(\frac{n_2 - 1}{N - t}\right) S_2^2 + \cdots + \left(\frac{n_t - 1}{N - t}\right) S_t^2 \\ &= \frac{SS_{res}}{N - t} \end{aligned}$$

is called the residual mean squares. It is an unbiased estimator of σ^2 regardless of whether H_0 or H_a is true.

Remark: The within estimator MS_{res} is a generalization of the pooled sample variance estimator S_p^2 we discussed last chapter with $t = 2$ populations.

Across Estimator: We assume a common sample size $n_1 = n_2 = \cdots = n_t = n$ to simplify notation (i.e., a balanced design).

We know that if a sample arises from a Gaussian population, then the sample mean is also Gaussian distributed. Therefore, the sample mean of the ith sample

$$\bar{X}_{i\cdot} \sim \text{Gaussian}\left(\mu_i, \frac{\sigma^2}{n}\right).$$

Therefore, when the null hypothesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_t$ is true, we have the following sampling

distributions for each sample mean:

$$\begin{aligned}
\bar{X}_{1\cdot} &\sim \text{Gaussian}\left(\mu, \frac{\sigma^2}{n}\right) \\
\bar{X}_{2\cdot} &\sim \text{Gaussian}\left(\mu, \frac{\sigma^2}{n}\right) \\
&\vdots \\
\bar{X}_{t\cdot} &\sim \text{Gaussian}\left(\mu, \frac{\sigma^2}{n}\right)
\end{aligned}$$

where μ is the common population mean under H_0. Now, think of

$$\bar{X}_{1\cdot}, \bar{X}_{2\cdot}, ..., \bar{X}_{t\cdot}$$

as a random sample from the Gaussian(μ, σ^2/n) population distribution. The sample variance of this "random sample" is

$$\frac{1}{t-1}\sum_{i=1}^{t}\left(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot}\right)^2$$

and is an unbiased estimator of σ^2/n. Therefore,

$$MS_{trt} = \frac{1}{t-1}\underbrace{\sum_{i=1}^{t} n\left(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot}\right)^2}_{SS_{trt}}$$

is an unbiased estimator of σ^2.

We call

$$\begin{aligned}
SS_{trt} &= \text{"treatment sums of squares"} \\
MS_{trt} &= \text{"treatment mean squares."}
\end{aligned}$$

The across estimator MS_{trt} is an unbiased estimator of σ^2 when H_0 is true.

Remark: Our derivation of the across estimator assumed a balanced design (this was done for simplicity). If we have different sample sizes n_i, we simply adjust MS_{trt} to

$$MS_{trt} = \frac{1}{t-1}\underbrace{\sum_{i=1}^{t} n_i\left(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot}\right)^2}_{SS_{trt}}$$

This is still an unbiased estimator for σ^2 when H_0 is true.

When H_0 is true: (i.e., the population means are equal), then

$$\begin{aligned}
E(MS_{trt}) &= \sigma^2 \\
E(MSres) &= \sigma^2
\end{aligned}$$

These two facts suggest that when H_0 is true,

$$F = \frac{MS_{trt}}{MS_{res}} \approx 1.$$

When H_a is true: (i.e., the population means are not equal), then

$$E(MS_{trt}) > \sigma^2$$
$$E(MSres) = \sigma^2$$

These two facts suggest that when H_a is true,

$$F = \frac{MS_{trt}}{MS_{res}} > 1.$$

Sampling Distribution: When H_0 is true, the statistic

$$F = \frac{MS_{trt}}{MS_{res}} \sim F(t-1, N-t).$$

Recall that the mean of an F distribution is around 1. Therefore,

- Values of F in the center of this distribution are consistent with H_0.
- Large values of F (i.e., out in the right tail) are consistent with H_a.

Remark: Unusually small values of F (i.e., close to zero) are not necessarily consistent with either hypothesis. This is more likely to occur when there is a violation of our statistical assumptions; e.g., correlated individuals within/across samples (most likely), unequal population variances, normality departures, etc.

Example 12.2. For the mortar data in Example 12.1, we had four types of mortar (treatments) making $t = 4$, and $n_1 = 8$, $n_2 = 10$, $n_3 = 10$, and $n_4 = 8$ making $N = 36$. Below, we go through all the steps of calculating the overall F test.

Within Estimator: We calculate the residual sum of squares:

$$\begin{aligned} SS_{res} &= (n_1-1)S_1^2 + (n_2-1)S_2^2 + (n_3-1)S_3^2 + (n_4-1)S_4^2 \\ &= (8-1)24.78437 + (8-1)40.27851 + (10-1)34.89406 + (10-1)21.48645 \\ &= 962.8648 \end{aligned}$$

These values are calculated in R as follows.

```
> N<-nrow(dat.mortar)
> t<-4
> (overall.mean<-mean(dat.mortar$strength))
[1] 52.52472
> ##Within Estimator -- weighted estimate of t variances
> (var.i<-tapply(X=dat.mortar$strength,INDEX=dat.mortar$type,FUN=var))
OCM      PCM      PIM       RM
24.78437 40.27851 34.89406 21.48645
> (n.i<-tapply(X=dat.mortar$strength,INDEX=dat.mortar$type,FUN=length))
OCM PCM PIM  RM
8   8  10  10
> (SS.res<-sum((n.i-1)*var.i))
[1] 962.8648
```

The sample variance S_i^2 estimates the population parameter σ^2, which is assumed to be common across all t populations) from within the ith sample.

The weighted average of these estimates is called the residual mean squares, which is an unbiased estimator of σ^2 regardless of whether H_0 or H_a is true.

$$\begin{aligned} MS_{res} &= \frac{SS_{res}}{N-t} \\ &= \frac{962.8648}{36-4} \\ &= 30.08952 \end{aligned}$$

This value is calculated in R as follows.

```
> (MS.res<-SS.res/(N-t))
[1] 30.08952
```

Across Estimator: We can think of

$$\bar{X}_{1\cdot}, \bar{X}_{2\cdot}, ..., \bar{X}_{t\cdot}$$

as a random sample from the Gaussian(μ, σ^2/n) population distribution. The sample variance of this "random sample" is

$$\frac{1}{t-1}\sum_{i=1}^{t}\left(\bar{X}_{i\cdot}-\bar{X}_{\cdot\cdot}\right)^2$$

and is an unbiased estimator of σ^2/n. Therefore, an unbiased estimator for σ^2 when H_0 is true is

$$\begin{aligned} MS_{trt} &= \frac{1}{t-1}\underbrace{\sum_{i=1}^{t} n_i\left(\bar{X}_{i\cdot}-\bar{X}_{\cdot\cdot}\right)^2}_{SS_{trt}} \\ &= \frac{(44.13625-52.52472)^2+(46.61625-52.52472)^2+(59.35200-52.52472)^2+(57.13500-52.52472)^2}{4-1} \\ &= \frac{1520.876}{3} \\ &= 506.9586 \end{aligned}$$

These values are calculated in R as follows.

```
> ##Across Estimator -- one overall variance
> (xbar.i<-tapply(X=dat.mortar$strength,INDEX=dat.mortar$type,FUN=mean))
OCM      PCM      PIM      RM
44.13625 46.61625 59.35200 57.13500
> (SS.trt<- sum(n.i*(xbar.i-overall.mean)^2))
[1] 1520.876
> (MS.trt<-1/(t-1) * SS.trt)
[1] 506.9586
```

Thus, the F statistic is

$$F = \frac{MS_{trt}}{MS_{res}} = \frac{506.9586}{30.08952} = 16.84834$$

These values are calculated in R as follows.

```
> ##Test statistic
> (f.stat<- MS.trt/MS.res)
[1] 16.84834
```

Then the p-value is the right-tail probability of observing F where,

$$F = \frac{MS_{trt}}{MS_{res}} \sim F(4-1, 36-4).$$

This value is calculated in R as follows.

```
> ##P-value
> pf(q=f.stat,df1=3,df2=32,lower.tail=FALSE)
[1] 9.576449e-07
```

Instead of completing all of the calculations "by hand" we can ask R to do all the steps of the overall F test and create the ANOVA table directly using the following R code.

```
> anova(lm(strength~type,data=ggdat))
Analysis of Variance Table

Response: strength
          Df  Sum Sq Mean Sq F value    Pr(>F)
type       3 1520.88  506.96  16.848 9.576e-07 ***
Residuals 32  962.86   30.09
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

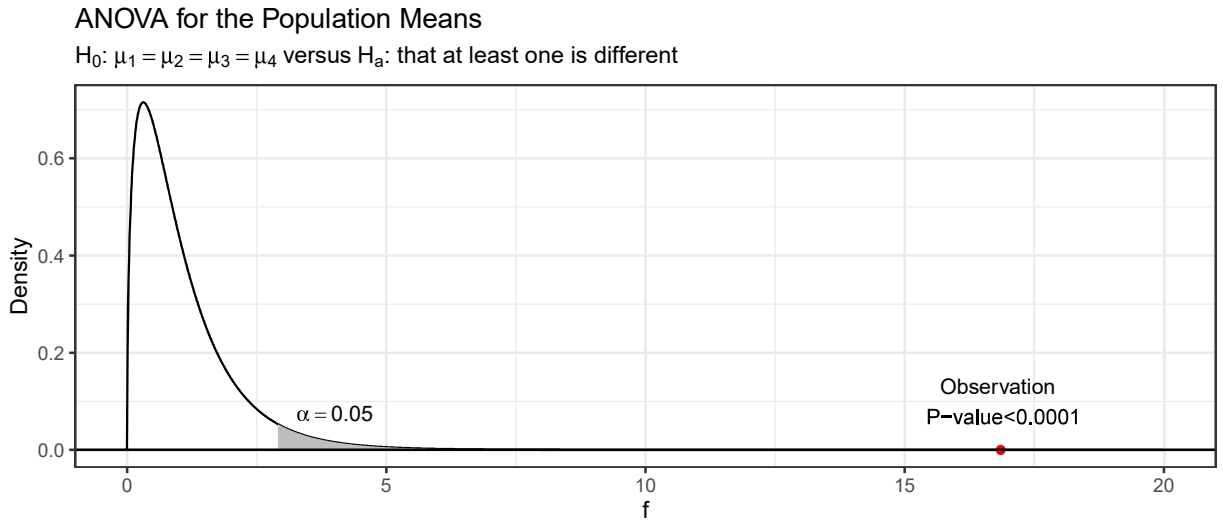

Figure 12.1.2: $F(3, 32)$ PDF. This is the sampling distribution of F when H_0 is true. A red point at $F = 16.848$ has been added.

There is very strong evidence that at least one of the four mortar strength population means is different. This value of F is not consistent with H_0. It is much more consistent with H_a.

It is common to display one-way classification results in an ANOVA table, which is printed as output of the `anova()` function in `R`. The form of the ANOVA table for the one-way classification is given below:

Source	df	SS	MS	F	p-value
Treatments	$t-1$	SS_{trt}	$MS_{trt} = \frac{SS_{trt}}{t-1}$	$F^* = \frac{MS_{TRT}}{MS_{res}}$	$P(F > F^*)$
Residuals	$N-t$	SS_{res}	$MS_{res} = \frac{SS_{res}}{N-t}$	–	–
Total	N-1	SS_{total}	–	–	–

This table shows us the following facts.

- In general, mathematics can show that

$$SS_{total} = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{\cdot\cdot})^2 = \sum_{i=1}^{t} n_i(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2 + \sum_{i=1}^{t}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\cdot})^2$$
$$= SS_{trt} + SS_{res}.$$

- SS_{total} measures how observations vary about the overall mean, without regard to treatment groups; that is, SStotal measures the total variation in all the data.
- SS_{total} can be partitioned into two components:
  - SS_{trt} measures how much of the total variation is due to the treatment groups.
  - SS_{res} measures what is "left over," which we attribute to inherent variation among the individuals.
- Degrees of freedom (df) add down.
- Mean squares (MS) are formed by dividing sums of squares by the corresponding degrees of freedom.
- The ratio of the mean squares (MS) gives the F statistic.

There are three main assumptions when performing an analysis of variance:

1. **Independent Random Samples**
   This assumption holding is largely up to the experimenter/investigator; i.e., drawing random samples from the different populations independently (in the case of an observational study) or using randomization to assign individuals to treatments (in an experiment).
2. **Normality** Each of the t population distributions is Gaussian (normal). This assumption can be assessed empirically using violin or qq plots for each sample separately. Of course, if the sample sizes are small (as in the mortar strength study), these plots may not be as useful. Thankfully, as with other statistical inference procedures involving means, a oneway ANOVA analysis is robust to normality departures.
3. **Equal population Variances** This is the most important assumption. A one-way ANOVA analysis is not robust to departures from this assumption, and it is very critical. Therefore, if you suspect the population variances may be markedly different, then you should not use a one-way ANOVA analysis.

There is a statistical inference procedure that is designed to test the equality of the population variances; i.e., to test

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_t^2$$
$$H_a : \text{the population variances } \sigma_i^2 \text{ are not all equal.}$$

The test is called Bartlett's test. However, I almost never use this test because it depends critically on the normality assumption. A nonparametric version of this test (i.e., one that does not assume normality) is available; it is called **Levene's test**.

12.1.1 Multiple comparisons/Follow-up analysis

In a one-way classification, the overall F test is used to test:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_t$$
$$H_a : \text{the population means } \mu_i \text{ are not all equal.}$$

If we do "reject H_0" in favor of H_a, we conclude that at least one population mean is different. However, we do not know which one(s) or how many. In this light, the decision to reject H_0 is not all that informative or useful.

Follow-up analysis: If H_0 is rejected, the obvious game becomes determining which population mean(s) is(are) different and how they are different. To do this, we will construct Tukey pairwise confidence intervals for all population treatment mean differences $\mu_i - \mu_j$, where $1 \leq i \leq j \leq t$. If there are t treatments, then there are

$$\binom{t}{2} = \frac{t(t-1)}{2}$$

pairwise confidence intervals to construct.

For example, in the mortar strength study (Example 12.1), there are $t = 4$ populations and therefore 6 pairwise intervals:

$$\mu_1 - \mu_2, \quad \mu_1 - \mu_3, \quad \mu_1 - \mu_4, \quad \mu_2 - \mu_3, \quad \mu_2 - \mu_4, \quad \mu_3 - \mu_4,$$

where

$$\mu_1 = \text{population mean strength for mortar type OCM}$$
$$\mu_2 = \text{population mean strength for mortar type PIM}$$
$$\mu_3 = \text{population mean strength for mortar type RM}$$
$$\mu_4 = \text{population mean strength for mortar type PCM.}$$

If we construct multiple confidence intervals (here, 6 of them), and if we construct each one using a $100(1-\alpha)$ percent confidence level, then the overall confidence level in the 6 intervals together will be less than $100(1-\alpha)$ percent. In statistics, this is known as the multiple comparisons problem.

There is a well-known inequality in probability called Bonferroni's Inequality, which states that if we have events $A_1, A_2, ..., A_J$, the probability that each event occurs

$$P\left(\bigcap_{j=1}^{J} A_j\right) \geq \sum_{j=1}^{J} P(A_j) - (J-1).$$

To see how this inequality can be used in our current discussion, define the event

$$A_j = j\text{th confidence interval includes its population mean difference,}$$

for $j = 1, 2, ..., J$. The event

$$\bigcap_{j=1}^{J} A_j = \text{each of the J intervals includes its population mean difference.}$$

In this light, consider the following table, which contains a lower bound on how small this probability can be (for different values of t and J). This table assumes that each pairwise interval has been constructed at the nominal $1 - \alpha = 0.95$ level.

# of treatments t	# of interval $J = \binom{t}{2}$	Lower Bound
3	3	3(0.95) - 2 = 0.85
4	6	6(0.95) - 5 = 0.70
5	10	10(0.95) - 9 = 0.50
6	15	15(0.95) - 14 = 0.25
$\vdots$	$\vdots$	$\vdots$
10	45	45(0.95) - 44 = -1.25

Therefore, with $t = 4$ treatments (populations), the probability that each of the six 95 percent intervals will contain its population mean difference can be as low as 0.7! For larger experiments with more treatments, this probability is even lower!! Clearly, we have to do something to address this.

Goal: Construct confidence intervals for all pairwise intervals $\mu_i - \mu_j$, $1 \leq i \leq j \leq t$, and have our family-wise confidence level still be at $100(1-\alpha)$ percent. By "family-wise," we mean that our level of confidence applies to the collection of all $\binom{t}{2}$ intervals (not to the intervals individually).

Solution: Increase the confidence level associated with each individual interval. Tukey's method is designed to do this. The intervals are of the form:

$$\left(\bar{X}_{i\cdot} - \bar{X}_{j\cdot} \pm q_{t,N-t,\alpha} \sqrt{MS_{res}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \right),$$

where $q_{t,N-t,\alpha}$ is the Tukey quantile that guarantees a family-wise confidence level of $100(1-\alpha)$ percent.

Example 12.3. Recall Example 12.1. We use `R` to construct the Tukey confidence intervals. The family-wise confidence level is 95 percent:

```
> TukeyHSD(aov(lm(strength~type,data=dat.mortar)),conf.level=0.95)
Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = lm(strength ~ type, data = dat.mortar))

$type
            diff       lwr       upr     p adj
```

```
PCM-OCM  2.48000 -4.950955  9.910955 0.8026758
PIM-OCM 15.21575  8.166127 22.265373 0.0000097
RM-OCM  12.99875  5.949127 20.048373 0.0001138
PIM-PCM 12.73575  5.686127 19.785373 0.0001522
RM-PCM  10.51875  3.469127 17.568373 0.0016850
RM-PIM  -2.21700 -8.863448  4.429448 0.8029266
```

In the R output, the columns labeled lwr and upr give, respectively, the lower and upper limits of the pairwise confidence intervals. The p adj column gives an adjusted p-value, which is adjusted so that we retain 0.05 significance ($\alpha = 0.05$) across the six tests, instead of individually.

- PCM-OCM: We are (at least) 95 percent confident that the difference in the population mean strengths for the PCM and OCM mortars is between -4.95 and 9.91 MPa. This confidence interval includes "0," so we cannot conclude these two population means are different. An equivalent finding is that the adjusted p-value for these two mortar types, given in the p adj column, is large (0.803).
- PIM-OCM: We are (at least) 95 percent confident that the difference in the population mean strengths for the PIM and OCM mortars is between 8.17 and 22.27 MPa. This confidence interval does not include "0" and contains only positive values. This suggests that the population mean strength of the PIM mortar is greater than the population mean strength of the OCM mortar An equivalent finding is that the adjusted p-value for these two mortar types, given in the p adj column, is very small (¡0.001).

Interpretations for the remaining four confidence intervals are written similarly. The main point is this:

- If a pairwise confidence interval (for two population means) includes "0," then these population means are not declared to be different.
- If a pairwise interval does not include "0," then the population means are declared to be different.

The conclusions we make for all possible pairwise comparisons are at the $100(1-\alpha)$ percent confidence level.

The following pairs of population means are declared to be different:PIM-OCM, RM-OCM, PIM-PCM, and RM-PCM. The following pairs of population means are declared to be not different: PCM-OCM, RM-PIM.

We can therefore conclude: The PIM and RM population mean strengths are larger than the OCM and PCM population mean strengths. The PIM and RM population mean strengths are not different. The OCM and PCM population mean strengths are not different.

When the number of tests is large, we can ask `R` to think about these comparisons for us and display a compact letter display that groups the mortar types that are not significantly different by assigning each group of similar mortar types a letter.

```
> library(multcomp)
> ph <- glht(aov(lm(strength~type,data=dat.mortar)), linfct = mcp(type = "Tukey"))
```

```
> cld(ph)
OCM PCM PIM  RM
"a" "a" "b" "b"
```

Furthermore, we have an overall (family-wise) confidence level of 95 percent that all of our conclusions are correct. Had we not used an adjusted analysis based on Tukey's method (e.g., just calculate all unadjusted pairwise intervals), our overall confidence level would have been much lower (as low as 70 percent)

We visualize the Tukey HSD intervals in Figure 12.1.3, with the following code in R. Matching the numerical output, we see that 0 is on the confidence intervals for the "PCM-OCM" and "RM-PIM" differences, where the p-values are large.

```
> tukey.intervals<-TukeyHSD(aov(lm(strength~type,data=dat.mortar)),conf.level=0.95)$type
> ggdat<-cbind(data.frame(difference=rownames(tukey.intervals)),
+              tukey.intervals)
> ggplot(data=ggdat,aes(x=difference,y=diff))+
+   geom_pointrange(aes(ymin = lwr, ymax = upr))+
+   geom_errorbar(aes(ymin = lwr, ymax = upr), width = 0.2)+
+   geom_hline(yintercept = 0, linetype="dashed")+
+   theme_bw()+
+   xlab("Pairwise Difference")+
+   ylab("Estimated Difference in Strength (MPa)")
```

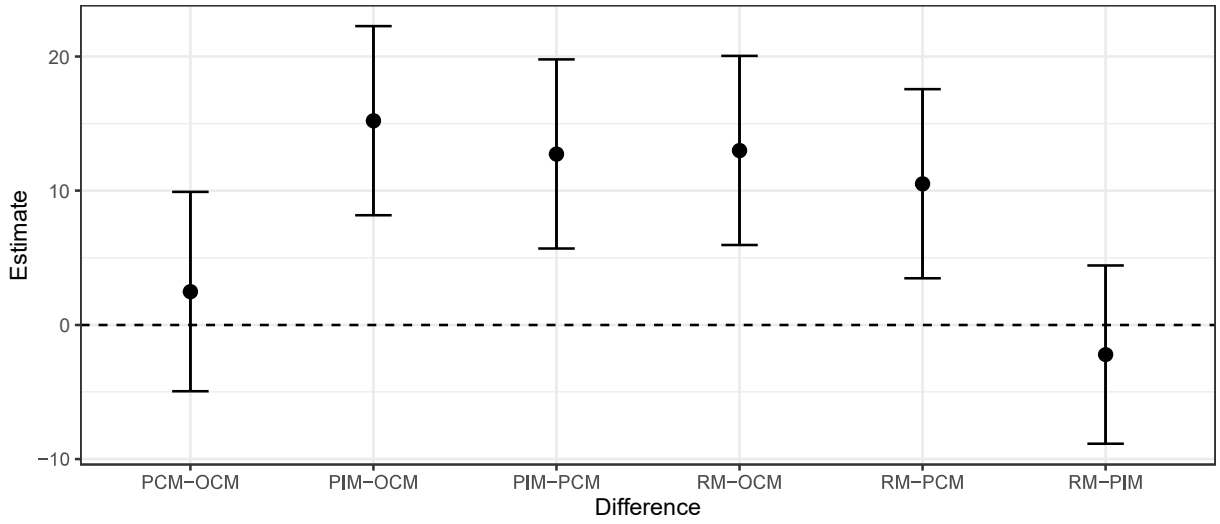

Figure 12.1.3: The Tukey HSD confidence intervals for the pairwise differences of strength by mortar type.

12.2 Mood's Median Test

When looking at the mortar data plotted in Figure 12.0.1 we see some skew in the strength observations among three of the four mortor types. This might motivate us to ask if there's a median alternative to the ANOVA procedure discussed so far. **Mood's Median Test** can be applied up

	$x_{j\cdot} < \hat{m}$	$x_{j\cdot} > \hat{m}$
Sample 1	O_{11}	O_{12}
Sample 2	O_{21}	O_{22}
...	...	...
Sample t	O_{t1}	O_{t2}

to t groups. We still don't yet have the information needed to discuss the theoretical details of this test, which can be carried out in `R` using the `mood.medtest()` function from the "RVAideMemoire" package (Hervé, 2019).

Mood's Median test only provides insight for the two sided test where the medians are the same; e.g., $M_1 = M_2 = M_3 = ... = M_t$; e.g.,

$$H_0 : M_1 = M_2 = M_3 = \cdots = M_t$$
$$H_a : \text{the population medians } M_i \text{ are not all equal.}$$

The intuition behind this test is a simple extension of the two-sample Mood's median test discussed in Chapter 11. The added complexity of considering t samples, instead of two, only requires a larger table as below, where $\hat{m}$ is the overall sample median. The idea is that if the t populations have the same median, then we would expect half of the observations from both samples to less than or greater than the median; i.e., we would expect $O_{11} \approx O_{12}$, $O_{21} \approx O_{22}$, ..., $O_{t1} \approx O_{t2}$. We can observe this by making a contingency table for our data as above, but the statistical mechanism for deciding when the data is "different enough," will be discussed in a later chapter.

Example 12.4. Revisiting the mortar data from Example 12.1. Suppose the engineers had a different question; e.g.

> "Are the population median mortar strengths equal among the four types of mortars? Or, are the population medians different?"

This initial question can be framed statistically as the following hypothesis test:

$$H_0 : M_1 = M_2 = M_3 = M_4$$
$$H_a : \text{the population medians } M_i \text{ are not all equal.}$$

Remark: The alternative simply states that at least one of the medians is different.

For the mortar data in Example 12.1, we can create the following table.

```
> (overall.median <- median(dat.mortar$strength)) #overall median
[1] 52.135
> table(dat.mortar$type,dat.mortar$strength<overall.median)
       FALSE TRUE
OCM        0    8
PCM        1    7
PIM        9    1
RM         8    2
> prop.table(table(dat.mortar$type,dat.mortar$strength<overall.median),margin=1)
       FALSE  TRUE
OCM 0.000 1.000
```

```
PCM 0.125 0.875
PIM 0.900 0.100
RM  0.800 0.200
```

From the tables we created, we see that there's different behavior. The overall median, is too large for the OCM and PCM mortar types where substantially more than 50% of observations were less than the overall median, and too small for the PIM and RM mortar types where substantially less than 50% of observations were less than the overall median.

This is reflected in a small p-value resulting from Mood's median test.

```
> library(RVAideMemoire)
> mood.medtest(strength~type,data=dat.mortar)

Mood's median test

data:  strength by type
p-value = 1.308e-05
```

Here, we find evidence that the population median strength of the four mortar types are different ($p < 0.0001$). Considering the violin plots in Figure 12.0.1 we see that this "difference" is that the median strengths of OCM and PCM is lower than the median strengths for PIM and RM, but we're not sure exactly which differences are "significant".

12.2.1 Multiple comparisons/Follow-up analysis

In a one-way classification, Mood's median test is used to test:

$$H_0 : M_1 = M_2 = \cdots = M_t$$
$$H_a : \text{the population medians } M_i \text{ are not all equal.}$$

If we do "reject H_0" in favor of H_a, we conclude that at least one population medians is different. However, we do not know which one(s) or how many. In this light, the decision to reject H_0 is not all that informative or useful.

Follow-up analysis: If H_0 is rejected, the obvious game becomes determining which population median(s) is(are) different and how they are different. To do this, we will conduct pairwise Mood's median tests across groups to check for pairwise differences and controlling for multiple comparisons. By doing this, we conduct tests $H_0 : M_i = M_j$ versus $H_a : M_i \neq M_j$ where $1 \leq i \leq j \leq t$. If there are t treatments, then there are

$$\binom{t}{2} = \frac{t(t-1)}{2}$$

pairwise comparisons to complete.

For example, in the mortar strength study (Example 12.1), there are $t = 4$ populations and therefore 6 pairwise comparisons:

$$M_1 - M_2, \quad M_1 - M_3, \quad M_1 - M_4, \quad M_2 - M_3, \quad M_2 - M_4, \quad M_3 - M_4,$$

where

$$M_1 = \text{population median strength for mortar type OCM}$$
$$M_2 = \text{population median strength for mortar type PIM}$$
$$M_3 = \text{population median strength for mortar type RM}$$
$$M_4 = \text{population median strength for mortar type PCM.}$$

If we construct multiple Mood's median tests (here, 6 of them), and if we conduct each one using a significance level of α, then the overall confidence level in the 6 tests together will be less than $100(1-\alpha)$ percent. This is another multiple comparisons problem. We can pass a p-value correction method as an argument into Mood's median test, these methods adjust p-values to correct for multiple corrections and control the probability or rate of Type I error.

Goal: Conduct hypothesis tests for all pairwise intervals $M_i - M_j$, $1 \leq i \leq j \leq t$, and have our family-wise confidence level still be at $100(1-\alpha)$ percent. By "family-wise," we mean that our level of confidence applies to the collection of all $\binom{t}{2}$ tests (not to the tests individually).

Solution: There are many approaches to this issue; for example there are six methods for adjusting p-values using the `p.adjust()` function in `R`. Two popular approaches are the approach of Bonferroni, based on Bonferroni (1936), and the approach of Benjamini and Hochberg (1995); these are the "bonferroni" and "BH" methods for `p.adjust()`.

The Bonferroni procedure for p-value calculation punishes all p-values in a model equally to control the probability that we make at least one error or Type I, whereas Benjamini-Hochberg punishes p-values accordingly to their ranking, as above, to control the rate of Type I errors. We can think of these corrections as controlling the following.

$$P(\#\text{ of Type I errors} \geq 1) \leq 0.05 \qquad \textbf{[Bonferroni]}$$
$$\frac{\#\text{ of Type I errors}}{\#\text{ of significant tests}} \leq 0.05 \qquad \textbf{[Benjamini-Hochberg]}$$

In this text, we will control the false discovery rate via the approach of Benjamini and Hochberg. The Bonferroni procedure is conservative in that it requires more extreme observations for statistical significance, which leads to more false negatives – it will disregard significant observations as insignificant. This may not be clearly visible in a handful of tests, but as the number of tests increase the Boneferroni procedure becomes more prohibitive. The Benjamini-Hochberg procedure is less conservative in that it aims to control the rate of Type I error, instead of the probability of at least one Type I error. In this sense, it should also seem more logical to consider the rate or proportion of errors instead of whether or not any errors were made; five false discoveries in ten hypothesis tests is much worse than the same number out of one hundred tests.

Remark: The Bonferroni approach is an excellent choice if an error of Type I is truly catastrophic.

These adjusted p-values are calculated in Table 12.2.1, where m is the number of tests, i is the rank of the p-value, p is the unadjusted p-value, and `cummin()` returns the cumulative minimum (from the bottom of Table 12.2.1 to the top).

$$p_{\text{adjusted}} = \text{cummin}\left((m/i)p\right) \qquad \textbf{[Benjamini-Hochberg]}$$
$$p_{\text{adjusted}} = mp \qquad \textbf{[Bonferroni]}$$

Remark: For adjusted p-values greater than on we simply report 1.

We report the adjusted p-values for the mortar data below. While we calculate both the Benjamini-Hochberg and Bonferroni adjusted p-values below, in practice we would choose one to proceed with.

Difference	p	m	i	$(m/i) \times p$	BH	mp	Bonferroni
OCM - PIM	0.0002262	6	1	0.0013572	0.0006786	0.0013572	0.001357
OCM - RM	0.0002262	6	2	0.0006786	0.0006786	0.0013572	0.001357
PCM - PIM	0.005677	6	3	0.0113540	0.0085155	0.0340620	0.034060
PCM - RM	0.005677	6	4	0.0085155	0.0085155	0.0340620	0.034060
OCM - PCM	0.3329	6	5	0.3994800	0.3994800	1.9974	1
PIM - RM	1	6	6	1	1	6	1

To control the false discovery rate at 0.05, we simply treat the BH adjusted p-values as usual and compare to $\alpha = 0.05$. To control the probability of making at least one error of Type I, we simply treat the Bonferroni adjusted p-values as usual and compare to $\alpha = 0.05$. Note that it is possible for the unadjusted p-value to be significant while the adjusted p-value is not significant.

Example 12.5. Recall Example 12.1. We use `R` to construct the pairwise Mood's median tests. The false discovery rate is controlled to be 0.05 via the Benjamini-Hochberg p-value adjustment.

```
> library(rcompanion)
> PT<-pairwiseMedianTest(strength~type,
+                        data   = dat.mortar,
+                        method = "BH")
> PT
     Comparison   p.value  p.adjust
1 OCM - PCM = 0    0.3329 0.3995000
2 OCM - PIM = 0 0.0002262 0.0006786
3  OCM - RM = 0 0.0002262 0.0006786
4 PCM - PIM = 0  0.005677 0.0085150
5  PCM - RM = 0  0.005677 0.0085150
6  PIM - RM = 0         1 1.0000000
```

In the `R` output, the columns labeled p.value and p.adjust give, respectively, the unadjusted and adjusted p-values.

- OCM-PCM: The adjusted p-value for these two mortar types, given in the p.adjust column, is large (0.3995), thus we do not have significant evidence that their median strengths are different.
- OCM-PIM: The adjusted p-value for these two mortar types, given in the p.adjust column, is very small (0.0007), thus we have significant evidence that their median strengths are different.

Interpretations for the remaining four p values are written similarly. The main point is this:

- If an adjusted p-value is greater than 0.05, then these population medians are not declared to be different.

- If an adjusted p-value is less than 0.05, then these population medians are declared to be different.

The following pairs of population medians are declared to be different:PIM-OCM, RM-OCM, PIM-PCM, and RM-PCM. The following pairs of population medians are declared to be not different: PCM-OCM, RM-PIM.

When the number of tests is large, we can ask `R` to think about these comparisons for us and display a compact letter display that groups the mortar types that are not significantly different by assigning each group of similar mortar types a letter.

```
> cldList(p.adjust ~ Comparison,
+         data = PT,
+         threshold = 0.05)
  Group Letter MonoLetter
1   OCM      a          a
2   PCM      a          a
3   PIM      b           b
4    RM      b           b
```

We can therefore conclude: The PIM and RM population median strengths are larger than the OCM and PCM population median strengths. The PIM and RM population median strengths are not different. The OCM and PCM population median strengths are not different.

Furthermore, we have controlled the rate of Type I error to be 0.05 or less in our conclusions. Had we not used an adjusted analysis based on the Benjamini-Hochberg procedure (e.g., just completed all the pairwise Mood's median tests), the overall rate of Type I error would have been higher.

Goal: We also may want to construct confidence intervals for all pairwise intervals $M_i - M_j$, $1 \leq i \leq j \leq t$, and have our confidence level still be at $100(1-\alpha)$, or control for the Type I error rate.

Solution: Increase the confidence level associated with each individual interval. While we were able to use Tukey's method for ANOVA, we will use Bonferroni and Benjamini-Hochberg correction for confidence intervals which provide us ways for selecting the appropriate confidence levels for our usual interval procedure.

The Bonferroni adjustment tells us to change our $100(1-\alpha)$ percent confidence intervals to $100(1-\alpha/m)$ percent confidence intervals, where m is the number of test. We see that $100(1-\alpha/m) \geq 100(1-\alpha)$ means that to correct for multiple comparisons we make the intervals wider – more confidence is required for individual intervals to retain 95% confidence overall.

The Benjamini-Hochberg approach tells us to calculate $100(1-\alpha)$ percent confidence intervals for the comparisons that are significant to $100(1-R(\alpha/m))$ percent confidence intervals, where m is the number of tests, by following these steps.

1. Sort the p values used for testing as we did in Table 12.2.1.
2. Let R be the row of the last significant result.
3. Select the R significant comparisons.

4. Construct $100(1 - R(\alpha/m))$ percent confidence intervals for the R significant comparisons selected in step (3).

We see that $100(1-R(\alpha/m)) \geq 100(1-\alpha)$ means that to correct for multiple comparisons we make the intervals wider – more confidence is required for individual intervals to control the proportion of intervals that result in false discovery.

Remark: Note that $100(1-\alpha/m) \geq 100(1-R(\alpha/m)) \geq 100(1-\alpha)$ so, like the p-value adjustments, the Bonferroni adjusted confidence intervals are more conservative than the Benjamini-Hochberg confidence intervals. The Bonferroni procedure is conservative (wider) in that it requires more extreme observations for the interval not to include 0, which leads to more false negatives. The Benjamini-Hochberg procedure is less conservative (narrower).

Example 12.6. Recall Example 12.1. We use R to construct the Bonferroni adjusted bootstrapping confidence intervals. The family-wise confidence level is 95 percent:

```
> alpha<-0.05
> m<-6
> (bonf.conf<-1-alpha/m)
[1] 0.9916667
> library("simpleboot")
> library("boot")
> boot.median<-function(data,indices){
+   d<-data[indices] # allows boot to select sample
+   return(median(d))
+ }
> ## Row 1
> boot1<-two.boot(sample1=dat.mortar$strength[which(dat.mortar$type=="OCM")],
+                 sample2=dat.mortar$strength[which(dat.mortar$type=="PCM")],
+                 FUN=boot.median, R=1000)
> ci1<-boot.ci(boot.out=boot1,conf=bonf.conf,type="perc")
> estimate1<-median(dat.mortar$strength[which(dat.mortar$type=="OCM")])-
+   median(dat.mortar$strength[which(dat.mortar$type=="PCM")])
> ## Row 2
> boot2<-two.boot(sample1=dat.mortar$strength[which(dat.mortar$type=="OCM")],
+                 sample2=dat.mortar$strength[which(dat.mortar$type=="PIM")],
+                 FUN=boot.median, R=1000)
> ci2<-boot.ci(boot.out=boot2,conf=bonf.conf,type="perc")
> estimate2<-median(dat.mortar$strength[which(dat.mortar$type=="OCM")])-
+   median(dat.mortar$strength[which(dat.mortar$type=="PIM")])
> ## Row 3
> boot3<-two.boot(sample1=dat.mortar$strength[which(dat.mortar$type=="OCM")],
+                 sample2=dat.mortar$strength[which(dat.mortar$type=="RM")],
+                 FUN=boot.median, R=1000)
> ci3<-boot.ci(boot.out=boot3,conf=bonf.conf,type="perc")
> estimate3<-median(dat.mortar$strength[which(dat.mortar$type=="OCM")])-
+   median(dat.mortar$strength[which(dat.mortar$type=="RM")])
> ## Row 4
> boot4<-two.boot(sample1=dat.mortar$strength[which(dat.mortar$type=="PCM")],
+                 sample2=dat.mortar$strength[which(dat.mortar$type=="PIM")],
+                 FUN=boot.median, R=1000)
```

```
> ci4<-boot.ci(boot.out=boot4,conf=bonf.conf,type="perc")
> estimate4<-median(dat.mortar$strength[which(dat.mortar$type=="PCM")])-
+   median(dat.mortar$strength[which(dat.mortar$type=="PIM")])
> ## Row 5
> boot5<-two.boot(sample1=dat.mortar$strength[which(dat.mortar$type=="PCM")],
+                 sample2=dat.mortar$strength[which(dat.mortar$type=="RM")],
+                 FUN=boot.median, R=1000)
> ci5<-boot.ci(boot.out=boot5,conf=bonf.conf,type="perc")
> estimate5<-median(dat.mortar$strength[which(dat.mortar$type=="PCM")])-
+   median(dat.mortar$strength[which(dat.mortar$type=="RM")])
> ## Row 6
> boot6<-two.boot(sample1=dat.mortar$strength[which(dat.mortar$type=="PIM")],
+                 sample2=dat.mortar$strength[which(dat.mortar$type=="RM")],
+                 FUN=boot.median, R=1000)
> ci6<-boot.ci(boot.out=boot6,conf=bonf.conf,type="perc")
> estimate6<-median(dat.mortar$strength[which(dat.mortar$type=="PIM")])-
+   median(dat.mortar$strength[which(dat.mortar$type=="RM")])
> B.conf.intervals<- data.frame(Comparison=c("OCM-PCM","OCM-PIM","OCM-RM",
+                                            "PCM-PIM","PCM-RM","PIM-RM"),
+                               Estimate=c(estimate1,estimate2,estimate3,
+                                          estimate4,estimate5,estimate6),
+                              lower=c(ci1$percent[4],ci2$percent[4],ci3$percent[4],
+                                      ci4$percent[4],ci5$percent[4],ci6$percent[4]),
+                               upper=c(ci1$percent[5],ci2$percent[5],ci3$percent[5],
+                                      ci4$percent[5],ci5$percent[5],ci6$percent[5]))
```

The Benjamini-Hochberg intervals are calculated the same way, we just replaced bonf.conf with bh.conf and we don't calculate intervals for the first and last comparisons, which were not significant.

```
> alpha<-0.05
> m<-6
> length(which(PT$p.value<0.05))
[1] 4
> (bh.conf<-1-4*(alpha/m))  #4 significant results
[1] 0.9666667
```

In the R output displayed below, the columns labeled lower and upper give, respectively, the lower and upper limits of the pairwise Bonferroni corrected and Benjamini-Hochberg corrected confidence intervals.

```
> B.conf.intervals
  Comparison Estimate      lower     upper
1    OCM-PCM   -5.625 -11.260000  7.590000
2    OCM-PIM  -16.725 -25.655360 -4.485685
3     OCM-RM  -15.000 -21.536589 -3.431370
4    PCM-PIM  -11.100 -25.453873 -1.754640
5     PCM-RM   -9.375 -22.122308 -1.145000
6     PIM-RM    1.725  -7.733387 12.070360
> BH.conf.intervals
```

```
  Comparison Estimate     lower     upper
1    OCM-PCM   -5.625        NA        NA
2    OCM-PIM  -16.725 -23.43000 -6.980000
3     OCM-RM  -15.000 -20.40333 -6.199671
4    PCM-PIM  -11.100 -21.94466 -4.315000
5     PCM-RM   -9.375 -18.69344 -2.820009
6     PIM-RM    1.725        NA        NA
```

- OCM-PCM: We are (at least) 95 percent confident that the difference in the population median strengths for the OCM and PCM mortars is between -11.26 and 7.95 MPa. This confidence interval includes "0," so we cannot conclude these two population medians are different. An equivalent finding is that the p-value for the pairwise Mood's median test was large (0.3995).
- OCM-PIM: We are (at least) 95 percent confident that the difference in the population median strengths for the OCM and PIM mortars is between -25.6554 and -4.4857 MPa. This confidence interval does not include "0" and contains only negative values. This suggests that the population median strength of the PIM mortar is greater than the population median strength of the OCM mortar. An equivalent finding is that the p-value for the pairwise Mood's median test was small (0.0007).

Interpretations for the remaining four Bonferroni-adjusted confidence intervals are written similarly. The main point is this:

- If a pairwise confidence interval (for two population medians) includes "0," then these population medians are not declared to be different.
- If a pairwise interval does not include "0," then the population medians are declared to be different.

Remark: The Benjamini-Hochberg intervals are only calculated for the significant differences found in the pairwise Mood's median tests, which aims to correct for the overly conservative Bonferroni procedure which yields longer intervals. The interpretations are the same, except we remove "(at least)."

We visualize the Bonferroni and Benjamini-Hochberg adjusted intervals in Figure 12.2.4, with the following code in `R`. Matching the numerical output, we see that 0 is on the Bonferroni-adjusted confidence intervals for the "PCM-OCM" and "RM-PIM" differences, where the p-values were large.

```
> conf.intervals<-rbind(B.conf.intervals,BH.conf.intervals)
> conf.intervals<-cbind(conf.intervals,Type=c(rep("Bonferonni",6),rep("BH",6)))
> ggdat<-data.frame(conf.intervals)
> ggplot(data=ggdat,aes(x=Comparison,y=Estimate,color=Type))+
+   geom_pointrange(aes(ymin=lower,ymax=upper),position=position_dodge(0.25))+
+   geom_errorbar(aes(ymin=lower,ymax=upper),width=0.2,position=position_dodge(0.25))+
+   geom_hline(yintercept=0,linetype="dashed")+
+   theme_bw()+
+   xlab("Pairwise Difference")+
+   ylab("Estimated Difference in Strength (MPa)")
```

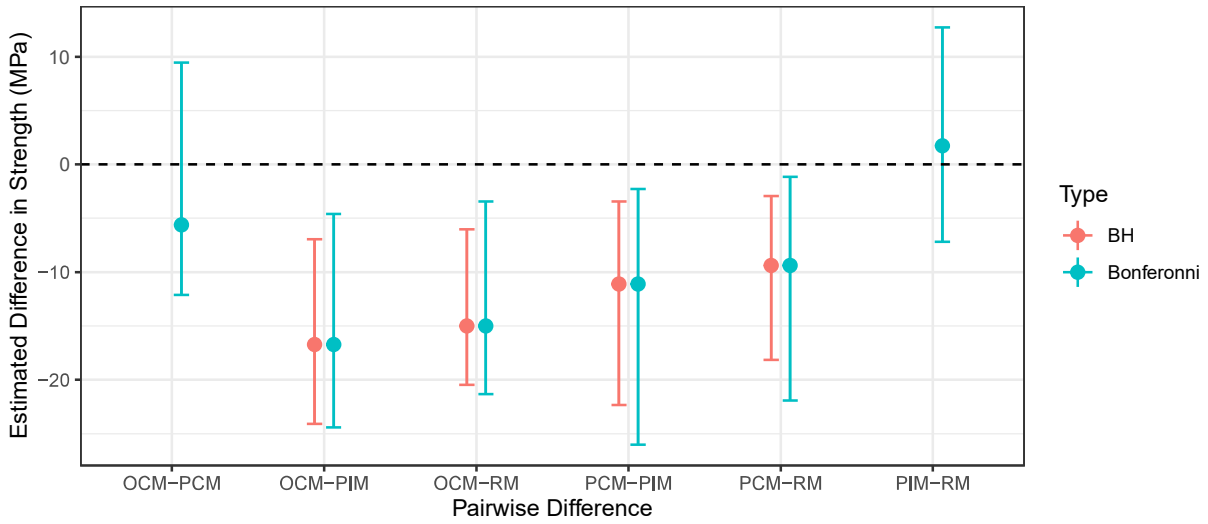

Figure 12.2.4: The Bonferroni and Benjamini-Hochberg adjusted intervals for the pairwise differences of strength by mortar type.

12.3 Kruskal Wallis

When looking at the mortar data plotted in Figure 12.0.1 we saw some skew in the strength observations among three of the four mortor types. In the last section, this motivated us to ask if there's a median difference among the different mortar types. An alternative to Mood's median test is the Kruskal Wallis test, which tests that the mean ranks of the groups are the same.

$$H_0 : \mu_1^R = \mu_2^R = \mu_3^R = \cdots = \mu_t^R$$
$$H_a : \text{the population mean ranks } \mu_i^R \text{ are not all equal,}$$

where μ_i^R is the population mean rank for the ith treatment (population).

The test is completed by replacing the data with their ranks, 1 for the smallest observation to n for the largest operation. We can ask for the ranks in R using the `rank()` function.

```
> rank(dat.mortar$strength)
[1] 16  7  5  9  2  4  6 13 35 33 25 20 34 22 15 23 31 36 12
[20] 32 18 28 27 19 30 29 26 24  1  3 11 14 17 21  8 10
```

The ranks are listed below the observations in the table of the mortar data.

OCM:	51.45	42.96	41.11	48.06	38.27	38.88	42.74	49.62		
Ranks	16	7	5	9	2	4	6	13		
PIM:	64.97	64.21	57.39	52.79	64.87	53.27	51.24	55.87	61.76	67.15
Ranks	35	33	25	20	34	22	15	23	31	36
RM:	48.95	62.41	52.11	60.45	58.07	52.16	61.71	61.06	57.63	56.80
Ranks	12	32	18	28	27	19	30	29	26	24
PCM:	35.28	38.59	48.64	50.99	51.52	52.85	46.75	48.31		
Ranks	1	3	11	14	17	21	8	10		

An ANOVA-like procedure, with corrections for ties is performed on the ranks. That is, the Kruskal Wallis procedure compares the average mean ranks per group to the overall mean rank. For example, we provide the group and overall mean ranks for the mortar data below.

	OCM	PIM	RM	PCM	All
Observations	8	8	10	10	36
Mean Rank	7.750	10.625	27.4	24.5	18.5

We calculate

$$\begin{aligned} SS_{trt(R)} &= \sum_{i=1}^{t} n_i(\mu_i^R - \mu_\cdot^R)^2 \\ &= 8(7.75-18.5)^2 + 8(10.625-18.5)^2 + 10(27.4-18.5)^2 + 10(24.5-18.5)^2 \\ &= 2572.725 \end{aligned}$$

The Kruskal-Wallis H statistic is

$$H = \frac{SS_{trt(R)}}{\frac{n(n+1)}{12}} \sim \chi^2_{t-1},$$

where the sampling distribution is taken to be χ^2 with degrees of freedom equal to $t-1$. When the null is true, we expect small values of H and when the alternative is true, we expect large values of H.

For the mortar data, where $t = 4$,

$$H = \frac{SS_{trt(R)}}{\frac{n(n+1)}{12}} \sim \chi^2_3.$$

The observation yields a test statistic of

$$h = \frac{SS_{trt(R)}}{\frac{n(n+1)}{12}} = \frac{2572.725}{\frac{36(36+1)}{12}} = 23.1777$$

which is calculated in R as follows.

```
> dat.mortar.kw<-cbind(dat.mortar,strength.rank=rank(dat.mortar$strength))
> (counts<-tapply(dat.mortar.kw$strength.rank,FUN=length,INDEX=dat.mortar.kw$type))
OCM PCM PIM  RM
8   8  10  10
> (means<-tapply(dat.mortar.kw$strength.rank,FUN=mean,INDEX=dat.mortar.kw$type))
OCM    PCM    PIM     RM
7.750 10.625 27.400 24.500
> mean(dat.mortar.kw$strength.rank)
[1] 18.5
> (H<-sum(counts*(means-mean(dat.mortar.kw$strength.rank))^2)/(n*(n+1)/12))
[1] 23.1777
```

and a p-value of

```
> 1-pchisq(q=H,df=3)
[1] 3.708119e-05
```

We can ask R to conduct the Kruskal Wallis test directly as follows.

```
> kruskal.test(strength~type,data=dat.mortar)

Kruskal-Wallis rank sum test

data:  strength by type
Kruskal-Wallis chi-squared = 23.178, df = 3, p-value = 3.708e-05
```

Here, we find evidence that the population mean rank strength of the four mortar types are different ($p < 0.0001$). Considering the table of mean ranks above we see the sample mean ranks are all different; e.g.,

$$\bar{x}^R_{OCM} < \bar{x}^R_{PIM} < \bar{x}^R_{PCM} < \bar{x}^R_{RM}.$$

We're not sure, however, which differences are "significant".

12.3.1 Multiple comparisons/Follow-up analysis

In a one-way classification, the Kruskal Wallis test is used to test:

$$\begin{aligned} H_0 &: \mu^R_1 = \mu^R_2 = \mu^R_3 = \cdots = \mu^R_t \\ H_a &: \text{the population mean ranks } \mu^R_i \text{ are not all equal,} \end{aligned}$$

where mu^R_i is the mean of the ranks of the data from group i.

If we do "reject H_0" in favor of H_a, we conclude that at least one population mean ranks is different. However, we do not know which one(s) or how many. In this light, the decision to reject H_0 is not all that informative or useful.

Follow-up analysis: If H_0 is rejected, the obvious game becomes determining which population mean rank(s) is(are) different and how they are different. We will use Dunn's Test to answer this question.

$$\begin{aligned} H_0 : P(X_i > X_j) = 0.50 &= \text{it is equally likely a random observation from treatement i is} \\ &\qquad \text{larger than a random observation from treatment j} \\ H_a : P(X_i > X_j) \neq 0.50 &= \text{it is not equally likely a random observation from treatement i is} \\ &\qquad \text{larger than a random observation from treatment j} \end{aligned}$$

The test statistic for this test is

$$z_{ij} = \frac{\bar{X}^R_i - \bar{X}^R_j}{\sqrt{\left[\frac{N(N+1)}{12} - \frac{\sum_{s=1}^{r} \tau_s^3 - \tau_s}{12(N-1)}\right]\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \sim \mathcal{AG}(\mu_z = 0, \sigma_z = 1),$$

where

$$\begin{aligned} r &= \text{the number of tied ranks across all } t \text{ groups} \\ \tau_s &= \text{the number of observations across all } t \text{ groups with the rank of } s. \end{aligned}$$

To check for pairwise differences and controlling for multiple comparisons we will conduct Dunn's Test across groups. By doing this, we conduct tests $H_0 : \mu_i^R = \mu_j^R$ versus $H_a : \mu_i^R \neq \mu_j^R$ where $1 \leq i \leq j \leq t$. If there are t treatments, then there are

$$\binom{t}{2} = \frac{t(t-1)}{2}$$

pairwise comparisons to complete.

For example, in the mortar strength study (Example 12.1), there are $t = 4$ populations and therefore 6 pairwise comparisons:

$$\mu_1^R - \mu_2^R, \quad \mu_1^R - \mu_3^R, \quad \mu_1^R - \mu_4^R, \quad \mu_2^R - \mu_3^R, \quad \mu_2^R - \mu_4^R, \quad \mu_3^R - \mu_4^R,$$

where

$$\begin{aligned}
\mu_1^R &= \text{population mean rank strength for mortar type OCM} \\
\mu_2^R &= \text{population mean rank strength for mortar type PIM} \\
\mu_3^R &= \text{population mean rank strength for mortar type RM} \\
\mu_4^R &= \text{population mean rank strength for mortar type PCM.}
\end{aligned}$$

If we construct multiple Dunn's tests (here, 6 of them), and if we conduct each one using a significance level of α, then the overall confidence level in the 6 tests together will be less than $100(1-\alpha)$ percent. This is another multiple comparisons problem. We can pass a p-value correction method as an argument into Dunn's median test.

Goal: Conduct hypothesis tests for all pairwise tests $H_0 : \mu_i^R - \mu_j^R = 0$, $1 \leq i \leq j \leq t$, and have our family-wise confidence level still be at $100(1-\alpha)$ percent. By "family-wise," we mean that our level of confidence applies to the collection of all $\binom{t}{2}$ tests (not to the tests individually).

Solution: We will conduct Dunn's test and adjust p-values using the approaches of Benjamini and Hochberg, and Bonferroni.

Example 12.7. Recall Example 12.1. We use the "FSA" package for R to complete Dunn's test for the pairwise differences.

```
> library(FSA)
> (DT<-dunnTest(strength~type,data=dat.mortar,method="bh"))
Dunn (1964) Kruskal-Wallis multiple comparison
p-values adjusted with the Benjamini-Hochberg method.
  Comparison          Z      P.unadj        P.adj
1  OCM - PCM -0.5457658 5.852269e-01 0.5852269022
2  OCM - PIM -3.9319665 8.425382e-05 0.0005055229
3  PCM - PIM -3.3566788 7.888471e-04 0.0023665413
4   OCM - RM -3.3516763 8.032389e-04 0.0016064778
5   PCM - RM -2.7763885 5.496648e-03 0.0082449717
6   PIM - RM  0.6154907 5.382307e-01 0.6458768336
```

In the R output, the columns labeled P.unadj and P.adj give, respectively, the unadjusted and adjusted p-values.

- OCM-PCM: The adjusted p-value for these two mortar types, given in the p.adjust column, is large (0.5852), thus we do not have significant evidence that their mean rank strengths are different.

- OCM-PIM: The adjusted p-value for these two mortar types, given in the p.adjust column, is very small (0.0005), thus we have significant evidence that their mean rank strengths are different.

Interpretations for the remaining four p values are written similarly. The main point is this:

- If an adjusted p-value is greater than 0.05, then these population mean ranks are not declared to be different.
- If an adjusted p-value is less than 0.05, then these population mean ranks are declared to be different.

The following pairs of population mean ranks are declared to be different:PIM-OCM, RM-OCM, PIM-PCM, and RM-PCM. The following pairs of population mean ranks are declared to be not different: PCM-OCM, RM-PIM.

When the number of tests is large, we can ask `R` to think about these comparisons for us and display a compact letter display that groups the mortar types that are not significantly different by assigning each group of similar mortar types a letter.

We can therefore conclude: The PIM and RM population mean rank strengths are larger than the OCM and PCM population mean rank strengths. The PIM and RM population mean rank strengths are not different. The OCM and PCM population mean rank strengths are not different.

```
> cldList(P.adj ~ Comparison,
+         data = DT$res,
+         threshold = 0.05)
  Group Letter MonoLetter
1   OCM      a          a
2   PCM      a          a
3   PIM      b           b
4    RM      b           b
```

Furthermore, we have controlled the rate of Type I error to be 0.05 in our conclusions. Had we not used an adjusted analysis based on the Benjamini-Hochberg procedure (e.g., if we just completed all the pairwise tests), the overall rate of Type I error would have been higher.