

Chapter 5 - Comparisons among several population means

- *Analysis of Variance* is abbreviated as ANOVA.
- A one-way ANOVA considers $I \geq 2$ populations where the populations. The mean of the i^{th} population is μ_i . The variance of the i^{th} population is σ_i^2 .
- ANOVA compares the I population means, $\mu_1, \mu_2, \dots, \mu_I$.
- ANOVA with $I = 2$ is the same as a two-sided two-sample pooled t -test.
- “One-way” means that the levels of a single factor define the I populations being compared. In other words, the categories of a categorical variable define the populations. A k -way ANOVA still considers I populations but now the populations are defined by the combinations of the levels of k factors.

Setting:

1. A RS of size n_i has been chosen from population i for $i = 1, 2, \dots, I$.
2. Each RS is independent of the others.
3. *Homogeneity of Variance*: $\sigma^2 = \sigma_1^2 = \sigma_2^2 = \dots = \sigma_I^2$. This is usually assumed because:
 - It is technically challenging to compare populations with unequal variability
 - For many study designs, a factor’s effect (or association) on the mean response is of interest. Many times there is no effect on the SD.
 - For problems where the variance is not equal for different groups, we can often transform to a scale where the variance is approximately the same. An ANOVA can then be performed on the transformed data. If the transformed data are symmetric, then an ANOVA that tests the means of the transformed data is interpreted as tests of the medians of the original data.
4. Each population is normally distributed, so the distribution of the i^{th} population is $N(\mu_i, \sigma)$. If $n_i \geq 30$ for all I samples, then the assumption of normality can be dropped.

Scope of inference depends on study design (as usual):

- Randomized Experiment
 - The I treatment groups from an experiment are considered to be independent samples from I populations.
 - If there is a statistically significant difference among the I sample means, it is appropriate to claim that the factor (treatment) caused the difference in means.
 - If the groups of individuals in the experiment were chosen from a RS, then conclusions about I means can be extended to the populations from which the individuals were drawn. However, if the individuals were not from a RS, then conclusions to larger populations are dubious.
- Observational Study
 - The I samples to be compared in an observational study are considered independent if individuals in each sample were randomly chosen from each respective population.
 - Do not claim that the factor (explanatory variable) caused the difference in means.

The one-way (fixed effect) ANOVA Model

$$X_{ij} = \mu_i + \epsilon_{ij} \quad \text{where } \epsilon_{ij} \sim N(0, \sigma)$$

- X_{ij} is the response of the j^{th} individual in the RS from the i^{th} population.
- μ_i is mean of i^{th} population, σ is the standard deviation of all I populations.
 - The model where the i^{th} group has a different mean μ_i is called the *full model*.
 - The model where all groups have the same mean $\mu = \mu_1 = \mu_2 = \dots = \mu_I$ is called the *reduced model*. This is the model assumed under the null hypothesis.

Your textbook (p. 123) provides a table showing the means for each group when using these two models:

Group	1	2	...	I
Full Model	μ_1	μ_2	...	μ_I
Reduced Model	μ	μ	...	μ

- ϵ_{ij} is the error term, $x_{ij} - \mu_i$, also called the *deviation* of x_{ij} from μ_i
- $\epsilon_{ij} \sim N(0, \sigma)$ is equivalent to the assumption that the i^{th} population is normal, $x_{ij} \sim N(\mu_i, \sigma)$.

The “Estimated” one-way ANOVA Model

$$X_{ij} = \bar{X}_i + e_{ij}$$

- \bar{X}_i is the sample mean of the RS of size n_i from the i^{th} population, an unbiased point estimator of μ_i :

Group	1	2	...	I
Full Model	\bar{X}_1	\bar{X}_2	...	\bar{X}_I
Reduced Model	\bar{X}	\bar{X}	...	\bar{X}

\bar{X} is the grand mean over the data set.

- s_i is the sample standard deviation of the RS from the i^{th} population. s_i^2 has degrees of freedom $df = n_i - 1$ (this is the same df that we associated with the t -test in Chapter 3). s_i^2 is an unbiased point estimator of σ^2 if the assumption of homogeneity of variance is true.
- $s_p^2 = \frac{\sum_{i=1}^I (n_i - 1) s_i^2}{\sum_{i=1}^I (n_i - 1)}$ is the unbiased pooled estimate of σ^2 . This formula is an extension of the formula for the variance used by the pooled two-sample t -test from Chapters 2-3.
- $e_{ij} = X_{ij} - \bar{X}_i$ is the residual, the deviation of X_{ij} from \bar{X}_i

5.3 HYPOTHESIS TEST TO COMPARE I POPULATION MEANS

The Overall Test:

The Idea: We will compare the variability **between** the group means (MSG) to the variability **within** each group (MSE).

- If the variability among the \bar{x}_i 's is **large** relative to the variability within each group ($MSG \gg MSE$), then we will claim that there is a difference among the μ_i 's.

- If the variability among the \bar{x}_i 's is **not large** relative to the variability within each group, then we will fail to claim that there is a difference among the μ_i 's.
- The statistic $MSE = s_p^2$ is an unbiased estimator of the constant variance σ^2 .

1. Hypotheses:

$$H_0: \mu_1 = \mu_2 = \dots = \mu_I$$

$$H_a: \mu_i \neq \mu_j \text{ for some } i \text{ and } j \quad (\text{at least one of the } \mu_i \text{ are different than the others})$$

2. Check Assumptions:

- The i^{th} group contains individuals that are independent of each other such as when a RS of individuals is collected from the i^{th} population.
- The groups are independent of each other.
- The residuals are normally distributed. Check via residual plots as in Chapter 3 notes.
- Constant variance. Check via residual plots and/or $\frac{\text{largest } s_i}{\text{smallest } s_i} < 2$. See Display 5.15.

Use the R function `diagANOVA.r`, available at the STAT411 website, to generate residual plots.

Perform Steps 3 and 4 assuming that H_0 is true!

3. Test Statistic:

$$F = \frac{MSG}{MSR}, \text{ where } MSG \text{ and } MSE \text{ are from the following ANOVA table:}$$

One-way ANOVA Table

Source	DF	Sum of Squares (SS)	Mean Squares (MS)	F	p-value
Group	$DFG = I - 1$	$SSG = \sum_{i=1}^I n_i (\bar{x}_i - \bar{x})^2$	$MSG = \frac{SSG}{DFG}$	$F^* = \frac{MSG}{MSE}$	$P(F > F^*)$
Residual	$DFE = n - I$	$SSE = \sum_{i=1}^I (n_i - 1) s_i^2$	$MSE = \frac{SSE}{DFE}$		
Total	$DFT = n - 1$	$SST = \sum_{\text{all } x} (x_{ij} - \bar{x})^2$			

Regarding the ANOVA table:

- Two functions can be used to perform an ANOVA: `lm()` as in the EXAMPLE below or `aov()`.
- $n = \sum n_i$ is the *total sample size*
- $\bar{x} = \frac{\sum n_i \bar{x}_i}{n}$ is the *grand mean* or *grand average*.
- The “Residual” row refers to the variability within groups, or the “error.” The “E” in *DFE*, *SSE* and *MSE* stands for “error”. Your book describes this row as describing the *full separate means model*. Hence the abbreviations *DFE*, *SSE* and *MSE* refer to the degrees of freedom, sum of squares and mean squares for the full model.
- The “Total” row refers to the “total variability” of the response when there is only a single mean (i.e., when H_0 is true). The “T” in *DFT*, *SST* and *MST* stands for “total variability”. Your book refers to this row as describing the *reduced equal means model* (under H_0). Hence *DFT* and *SST* refer to the degrees of freedom and sum of squares for the reduced model.
- The “Group” row refers to the variability among the group means. This row summarizes the comparison of the *full model* to the *reduced model* (when H_0 is true). The “G” in *DFG*, *SSG* and *MSG* stands for “among or between groups”. Your book refers to *DFG*, *SSG* and *MSG* as the extra degrees of freedom, the extra sum of squares and the extra mean squares related to the full vs. reduced model comparison.

- The degrees of freedom add up: $DFT = DFG + DFE$.
- The sums of squares add up, $SST = SSG + SSE$.
- The among group (or extra) sum of squares SSG is interpreted as the variance explained by the ANOVA (full separate means) model.
- $R^2 = \frac{SST - SSE}{SST} = \frac{SSG}{SST}$ is the percentage of variance in the data that is explained by the ANOVA (full separate means) model.
- When $H_0 : \mu_1 = \mu_2 = \dots = \mu_I$ is true, then on the average $MSG \approx MSE$ and therefore $F = \frac{MSG}{MSE} \approx 1$.
- When H_0 is false, then on the average $MSG \gg MSE$ and therefore $F = \frac{MSG}{MSE} \gg 1$.
- Large $F = \frac{MSG}{MSE}$ values are strong evidence against H_0 in favor of H_a .

4. p-value: The p -value = $P(F > F^*)$.

- The test statistic has an F distribution, $F \sim F(DFG, DFE)$, when H_0 is true.
- An F distribution $F(DFG, DFE)$ has two parameters, the *numerator degrees of freedom* DFG and the *denominator degrees of freedom* DFE .
- An F distribution is a beautiful right-skewed distribution - see Display 5.9. Probabilities can be found using R's `pf(F, df1=DFG, df2=DFE, lower.tail=FALSE)` function.

5. and 6. Make a Decision and give a Conclusion.

EXAMPLE:

An archeologist is interested in studying head breadths from human skulls from different epochs. Significant changes in skull shape over time would suggest that interbreeding occurred with immigrant human populations. A sample of 27 head breadths were obtained by measuring skulls of Egyptian males from three different epochs: 4000BC, 1850BC, and 150AD. The data are from *Ancient Races of the Thebaid*, by Thomson and Randall-Maciver).

Display the data

```
D = read.table("http://www.math.montana.edu/parker/courses/STAT411/Chapter5.5.headbreadth.txt",
              header=T)
```

```
D[1:10,]
##   Case Epoch HeadBreadth
##1    1 4000BC          131
##2    2 4000BC          138
##3    3 4000BC          125
##4    4 4000BC          129
##5    5 4000BC          132
##6    6 4000BC          135
##7    7 4000BC          132
##8    8 4000BC          134
##9    9 4000BC          138
##10  10 1850BC          129
```

```
summary(D)
##      Case      Epoch      HeadBreadth
## Min.   : 1.0    150AD :9      Min.     :125.0
```

```

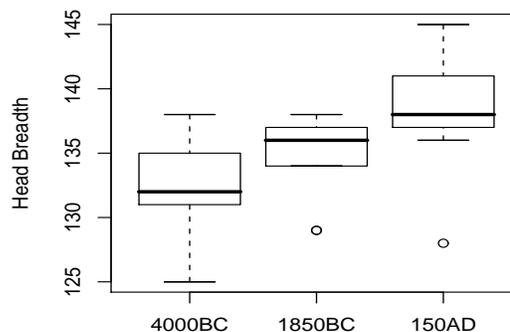
## 1st Qu.: 7.5   1850BC:9   1st Qu.:132.0
## Median :14.0  4000BC:9   Median :136.0
## Mean   :14.0                Mean   :135.1
## 3rd Qu.:20.5                3rd Qu.:138.0
## Max.   :27.0                Max.   :145.0

# R lists the levels of Epoch alpha-numerically
levels(D$Epoch)
##[1] "150AD" "1850BC" "4000BC"

# Put the categories in chronological order
D$Epoch = factor(D$Epoch, levels = c("4000BC","1850BC","150AD"))
levels(D$Epoch)
##[1] "4000BC" "1850BC" "150AD"

# Generate a beautiful side-by-side bixplot
boxplot(HeadBreadth ~ Epoch,data=D,ylab="Head Breadth")

```



Fit the One-way ANOVA Model

```

hb.lm=lm(HeadBreadth~Epoch,data=D)
anova(hb.lm)
##Analysis of Variance Table
##
##Response: HeadBreadth
##          Df Sum Sq Mean Sq F value Pr(>F)
##Epoch    2  138.74   69.37   4.0497 0.03052 *
##Residuals 24  411.11   17.13

```

Or, you get the same output by using the `aov()` function. In Chapter 6 we'll see an advantage of using `aov()`:

```

hb.aov = aov(HeadBreadth~Epoch,data=D)
summary(hb.aov)

```

Check some parts of the ANOVA table by hand (using the formulas provided earlier in these notes):

```

# Group means
xbari = tapply(D$HeadBreadth,D$Epoch,mean)
xbari
## 4000BC  1850BC  150AD
##132.6667 134.4444 138.1111

# Group variances
vari=tapply(D$HeadBreadth,D$Epoch,var)
vari
## 4000BC  1850BC  150AD
##17.50000 11.27778 22.61111

# Group sample sizes
ni = tapply(D$HeadBreadth,D$Epoch,length)
ni
##4000BC 1850BC  150AD
##      9      9      9

# Grand mean
gmean = mean(D$HeadBreadth)
##[1] 135.0741

```

```

# SSG = extra sum of squares explained by Epoch means
sum(ni*(xbari-gmean)^2)
##[1] 138.7407

# SSE = residual sum of squares for the full separate means model
sum((ni-1)*vari)
##[1] 411.1111

# p-value
1-pf(4.0497,2,24)
##[1] 0.03051911

```

Check Your Assumptions

1. *Independent data in each group*
2. *Independent groups*
3. *Normal Distribution*

```

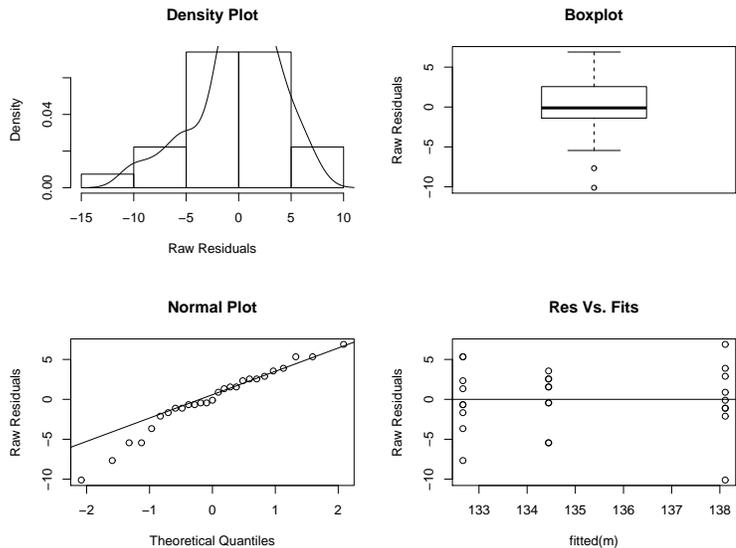
# Load in a function that allows you to graphically assess the normal
# and constant variance assumptions of the ANOVA model
source("http://www.math.montana.edu/parker/courses/STAT411/diagANOVA.r")

```

```

diagANOVA(hb.lm,do.tests=T)
##[1] "In this sample of size n=27, correlation of the residuals in the qq-plot is r=0.979712"
##[1] "In the following table, if r < critical.r, then the qq-plot suggests the residuals
## are not normal:"
## n critical.r
##1 5 0.832
##2 10 0.880
##3 15 0.911
##4 20 0.929
##5 25 0.941
##6 30 0.949
##7 40 0.960
##8 50 0.966
##9 60 0.971
##10 75 0.976

```



4. *Check Constant Variance*

The residuals have very similar spread in the residual vs. fits plot. Also:

```

> tapply(D$HeadBreadth,D$Epoch,sd)
 4000BC 1850BC 150AD
4.183300 3.358240 4.755114

```

$\frac{\text{largest } s_i}{\text{smallest } s_i} \approx \frac{4.76}{3.35} \approx 1.42 < 2$, so the constant variance assumption appears to hold.

1. Perform the **overall or omnibus Hypothesis Test**

(a) Hypotheses:

(b) Test statistic value:

(c) Distribution of the test statistic given that H_0 is true:

(d) p-value:

(e) Decision at $\alpha = .05$:

(f) Conclusion:

2. Give an unbiased estimate of the constant variance σ^2 .

5.2 PERFORMING A SINGLE COMPARISON OF ANY TWO OF THE MEANS AFTER ANOVA

- Even though you may only be interested in comparing two means, it can be advantageous to perform an ANOVA on several groups (if the ANOVA assumptions are satisfied) and then compare the two group means based on the ANOVA output.
- The idea is that the ANOVA provides a better estimate s_p^2 of the pooled variance σ^2 when using many groups (with more degrees of freedom $DFE = n - I$) versus when directly comparing any two groups by themselves.
- If you do NOT have know which two means to compare after an ANOVA, and instead want to compare all pairs of means, then do not perform the following test many times! If you do, then the actual confidence level will be less than your target of, say, 95%. Instead, in Chapter 6, we will implement a multiple comparison correction.

A pooled follow-up t-test and t-CI for $\mu_i - \mu_j$

- **Facts about the sampling distribution of $\bar{X}_i - \bar{X}_j$ from Chapter 2:**

– $\bar{X}_i - \bar{X}_j$ is our best (unbiased) estimator of $\mu_i - \mu_j$.

$$- SD(\bar{X}_i - \bar{X}_j) = \sqrt{\frac{\sigma_i^2}{n_i} + \frac{\sigma_j^2}{n_j}}$$

– If the sampling distributions of \bar{X}_i and \bar{X}_j are approximately normal, or if the sample sizes n_i and n_j are large, then the sampling distribution of $\bar{X}_i - \bar{X}_j$ is approximately normal, i.e.,

$$\bar{X}_i - \bar{X}_j \sim N\left(\mu_i - \mu_j, \sqrt{\frac{\sigma_i^2}{n_i} + \frac{\sigma_j^2}{n_j}}\right).$$

- Because of the constant variance assumption, $\sigma = \sigma_i = \sigma_j$, then $SD(\bar{X}_i - \bar{X}_j) = \sqrt{\sigma^2 \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$.
- Because we estimate σ^2 with the ANOVA pooled estimate MSE with DFE degrees of freedom, then we will use

$$SE(\bar{X}_i - \bar{X}_j) = \sqrt{MSE \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

in conjunction with t -tools with DFE degrees of freedom to perform tests and construct CIs regarding $\mu_i - \mu_j$.

A **Confidence Interval** for for $\mu_i - \mu_j$ is:

$$\bar{x}_i - \bar{x}_j \pm t_{1-\alpha/2, df=DFE} \sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

Hypothesis Testing for $\mu_1 - \mu_2$:

1. Hypotheses:

$H_0 : \mu_1 - \mu_2 = \Delta_0$ where Δ_0 is a specific value (usually 0).

$H_a : \mu_1 - \mu_2 > \Delta_0$

Choose one: $H_a : \mu_1 - \mu_2 < \Delta_0$

$H_a : \mu_1 - \mu_2 \neq \Delta_0$

NOTE: Perform steps 2-4 assuming that H_0 is true!

2. Assumptions: These are the same as for the original ANOVA that was already performed. If the assumptions of the ANOVA were satisfied, then they are satisfied for this follow-up test.

3. Test Statistic:

- $T = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{MSE\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

- If H_0 is true, $T \sim t(df)$ where $df = DFE$. For one-way ANOVA with I groups, $DFE = n - I$. For more general ANOVA, DFE is something else.

4. , 5. and 6. Find the p-value, make a Decision, and give a Conclusion as for any of the two-sample tests in Chapter 2.

EXAMPLE continued: Were brains smaller thousands of years ago? Use the skull data to test whether the 150AD head breadths are LARGER than the 4000BC head breadths on average.

1. Hypotheses:

2. Test statistic value:

3. Distribution of the test statistic given that H_0 is true:

4. p-value:

5. Decision at $\alpha = .05$:

6. Conclusion:

Relevant R code:

```

# ANOVA from earlier
hb.lm=lm(HeadBreadth~Epoch,data=D)

# Extract MSE and DFE
SSE=sum(hb.lm$residuals^2)
DFE=hb.lm$df.residual
MSE=SSE/DFE
cbind(DFE,SSE,MSE) # Compare with 1st row of ANOVA table
##      DFE      SSE      MSE
##[1,]  24 411.1111 17.12963

# Group means
xbari = tapply(D$HeadBreadth,D$Epoch,mean)
xbari
## 4000BC 1850BC 150AD
##132.6667 134.4444 138.1111

# Group sample sizes
ni = tapply(D$HeadBreadth,D$Epoch,length)
ni
##4000BC 1850BC 150AD
##      9      9      9

# Difference in means (150AD - 4000BC)
as.numeric(xbari[3]-xbari[1])
##[1] 5.444444

# SE(Difference in means) (4000BC - 150AD)
as.numeric(sqrt(MSE*(1/ni[1]+1/ni[3])))
##[1] 1.951047

# Perform the t-test by hand:
# test-stat for Ha: mu150AD - mu4000BC > 0
test.stat = as.numeric((xbari[3]-xbari[1])/sqrt(MSE*(1/ni[1] + 1/ni[3])))
test.stat
##[1] 2.790525

# Upper one-sided p-value for Ha: mu150AD - mu4000BC > 0
1 - pt(test.stat,DFE)
## [1] 0.005074997

# two-sided p-value for Ha: mu150AD - mu4000BC not equal to 0
2*(1-pt(abs(test.stat),DFE))
##[1] 0.01014999

# 95% CI for mu150AD - mu4000BC
xbari[3]-xbari[1] + c(-1,1)*qt(.975,DFE)*sqrt(MSE*(1/ni[1]+1/ni[3]))
##[1] 1.417681 9.471208

###
# Let R do the work for the 2-sided test of Ha: mu150AD - mu4000BC not equal to 0
summary(hb.lm)
##Call:
##lm(formula = HeadBreadth ~ Epoch, data = D)
##
##Coefficients:

```

```

##           Estimate Std. Error t value Pr(>|t|)
##(Intercept) 132.667      1.380  96.163 <2e-16 *** # Ha: mu4000BC not equal 0
##Epoch1850BC  1.778      1.951   0.911  0.3713 # Ha: mu1850BC - mu4000BC not equal 0
##Epoch150AD   5.444      1.951   2.791  0.0101 * # Ha: mu150AD - mu4000BC not equal 0 (as above)
##
##Residual standard error: 4.139 on 24 degrees of freedom
##Multiple R-squared:  0.2523, Adjusted R-squared:  0.19
##F-statistic:  4.05 on 2 and 24 DF, p-value: 0.03052

###
# Let R do the work for the 2-sided CI for mu150AD - mu4000BC
confint(hb.lm, level=0.95)
##           2.5 %      97.5 %
##(Intercept) 129.819315 135.514018 # CI for mu4000BC
##Epoch1850BC -2.248985  5.804541 # CI for mu1850BC - mu4000BC
##Epoch150AD  1.417681  9.471208 # CI for mu150AD - mu4000BC (as above)

###
# To change the 'reference level' from 4000BC to something else, e.g. to test or
# construct a CI for mu1850 - mu150AD
D$Epoch=relevel(D$Epoch,ref="1850BC")
hb.lm2 = lm(HeadBreadth ~ Epoch,data=D)

```

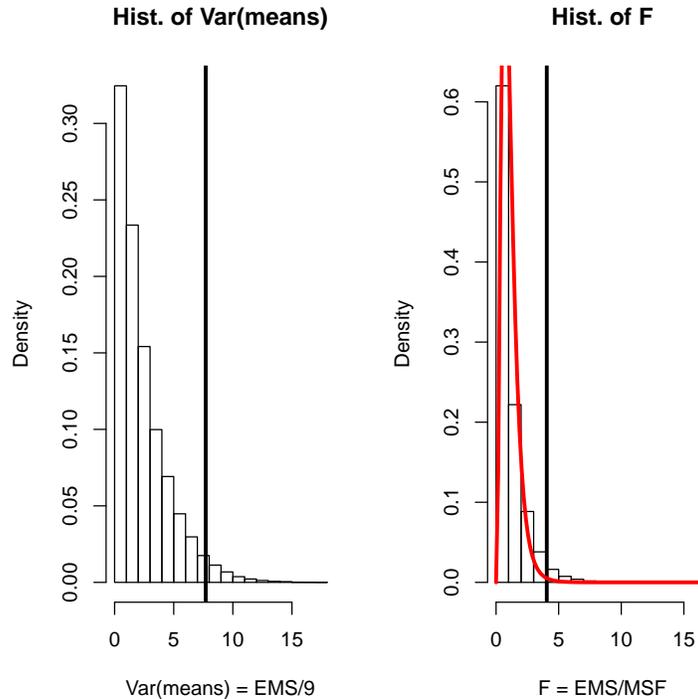
5.5 Robustness and resistance of ANOVA

1. Normality becomes critical when there are small and/or different sample sizes coupled with extremely long-tailed or skewed distributions.
2. Assumptions of independence within and across groups are critical.
3. Assumption of equal population standard deviations is crucial.
 - We pool the estimates of variability to make the most powerful comparisons possible. However, if one has a very different spread, then serious problems may result.
 - Unlike the 2-sample t -tools the effect of unequal SD's (when there's an SD more a factor of 2 or more different than the others) can be serious even if sample sizes are equal (See Display 5.13)
4. ANOVA is *not* resistant to severely outlying observations.

Randomization and permutation distributions when comparing several means

The skull data are from an observational study, not a randomized experiment, so we can simulate the data to estimate the permutation distribution of different test statistics under the assumption that there are no mean differences between the Epochs. The permutation distribution is constructed just as we did for the 2 sample case: the Epoch associated with each of the different head breadths is randomly assigned to give a new balanced data set with 9 head breadths in each Epoch.

- The left pane shows a histogram that approximates the permutation distribution of the variance in the Epoch means for the 10^5 simulated data sets. The variance in the actual 3 Epoch means is 7.71. as indicated by the thick black vertical line. The area to the right of this line is the p -value = 0.0307 for the test of $H_0 : \mu_{4000BC} = \mu_{1850BC} = \mu_{150AD}$ vs. H_a : at least one of the means is different. This agrees remarkably well with the p -value = 0.0305 from the ANOVA.



- The right pane shows a histogram that approximates the permutation distribution of the F statistic calculated for the 10^5 simulated data sets. Recall that the F statistic is basically the variance in the Epoch means (as in the left pane) divided by the pooled within-Epoch variance, $F = MSG/MSE$. The F statistic from the ANOVA for the actual data set is 4.05 as indicated by the thick black vertical line. The area to the right of this line is the p -value = 0.0303 for the test of H_0 vs. H_a . This agrees remarkably well with the p -value = 0.0305 from the ANOVA and also agrees with the p -value=0.0307 from the approximate permutation distribution for the variance in Epoch means. The red curve overlays the $F(2, 24)$ density indicating good agreement with the permutation distribution.

```
# Two permutation distributions for the skull data
names(D)
## [1] "Case"      "Epoch"      "HeadBreadth"
dim(D)
##[1] 27  3
n_grp = 27/3 # in this problem, there are 9 skulls in each of the 3 Epochs

var.test.stat=var(tapply(D$HeadBreadth,D$Epoch,mean))
var.test.stat
##[1] 7.707819

F.stat = 4.05 # from the ANOVA that we did earlier

num_sim = 1e5 # Draw num_sim randomizations
var.mean<-numeric(num_sim) # storage vector
MSF<-numeric(num_sim) # storage vector

# generate num_sim random assignments and calculate the variance in means
for(i in 1:num_sim)
{
  grp<-sample(D$Epoch,27,replace=F)
  var.mean[i]<- var(tapply(D$HeadBreadth,grp,mean))
  MSF[i] <- mean(tapply(D$HeadBreadth,grp,var)) # When equal sample sizes, mean(vars) = pooled var
```

```

}

EMS = var.mean*n_grp # only true when the sample sizes are equal
F=EMS/MSF

# Graph the approximate permutation distribution of the variance in means
par(mfrow=c(1,2))
hist(var.mean,prob=T,main="Hist. of Var(means)",
      xlab="Var(means) = EMS/9") # a density histogram
abline(v=var.test.stat,lwd=3) # puts a vertical line at the observed variance

# Graph the approximate permutation distribution of the F-stat
hist(F,prob=T, main="Hist. of F",xlab = "F = EMS/MSF") # a density histogram
abline(v=F.stat,lwd=3) # puts a vertical line at the observed variance
curve(df(x,2,24), add=TRUE, col=2, lwd=3)

# Get the p-value wrt var.test.stat
sum(var.mean>=var.test.stat)/num_sim
## [1] 0.03066

# Get the p-value wrt F.test.stat
sum(F>=F.stat)/num_sim
## [1] 0.03031

```

5.6.2 Kruskal-Wallis Nonparametric ANOVA

This approach is “non-parametric” because there is no (“parametric”) assumption of normality. This is one approach (besides simulating the randomization of permutation distribution directly) that deals with serious outliers. Kruskal-Wallis ANOVA still requires three assumptions:

1. RSs from each of the I groups;
2. independent groups;
3. the distributions for the groups have similar shapes but different means.

The Kruskal-Wallis test is performed in R by `kruskal.test()`. The steps to the test are:

1. List the data of total size n from lowest to highest. You can use R’s `sort()` function.
2. Assign ranks 1 to n to the list, with the smallest datum getting rank 1 and the largest datum getting rank n .
3. Perform an ANOVA on these ranks. The ANOVA calculations are simplified compared to what’s shown earlier in Chapter 5 because the variance of ranks is known (and does not need to be estimated by MSE). As a result, instead of an F test statistic when comparing H_0 vs. H_a , a χ^2 test statistic is used.

EXAMPLE continued:

```

kruskal.test(HeadBreadth ~ Epoch,data=D)
##
##           Kruskal-Wallis rank sum test
##
##data:  HeadBreadth by Epoch
##Kruskal-Wallis chi-squared = 6.698, df = 2, p-value = 0.03512

```

5.6.3 Random Effects ANOVA

There are two kinds of effects that statisticians try to model:

- **FIXED EFFECTS:** there is direct interest in the groups chosen.

- **RANDOM EFFECTS:** the groups are a random sample from a population of groups, and there is *not* direct interest in the particular groups chosen. Random effects may be used to model the **CLUSTER EFFECTS** described in Chapter 3. Unlike cluster effects however, random effects are purposely selected to be representative of some broader population of groups and it is of interest to make inference to the broader population.

The one-way random effects ANOVA Model

$$X_{ij} = \mu_i + \epsilon_{ij} \quad \text{where } \epsilon_{ij} \sim N(0, \sigma)$$

- X_{ij} is the response of the j^{th} individual in the RS from the i^{th} population.
- There are I populations and μ_i is the mean of the i^{th} population. Unlike the fixed effects ANOVA, μ_i for the different groups is no longer fixed but rather $\mu_i \sim N(\mu, \sigma_\mu)$ with μ_i independent of $\mu_{i'}$. Put another way, we assume that we have a RS of groups. σ_μ^2 is the group-to-group variance.
- $\epsilon_{ij} \sim N(0, \sigma)$ where σ^2 is the within group variance
- The random effects model implies that $x_{ij} \sim N(\mu, \sqrt{\sigma^2 + \sigma_\mu^2})$. The mean of x_{ij} is μ and $\sigma^2 + \sigma_\mu^2$ is the *total variance*.
- The random effects model allows the researcher to model the correlation of data from the same group. Two data x_{ij} and $x_{ij'}$ in the same group (i) are correlated by Within-group correlation $= \rho = \frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma^2}$.
- Because the within-group correlation can also be written as $\frac{\sigma_\mu^2}{\text{total variance}}$, the within-group can be interpreted as the proportion of the total variance of the measurements due to between-group sources.
- The use of random effects in a model may allow one to model cluster effects in a non-random sample.
- Unlike fixed effects ANOVA where we wish to test $H_0 : \mu_1 = \dots = \mu_I$, in the random effects ANOVA it is of interest to perform a hypothesis test or construct a CI for μ . In a balanced design with k individuals in each group, then the grand mean \bar{x} is the “best” estimator for μ and $SD(\bar{x}) = \sqrt{\frac{\sigma^2}{k \times I} + \frac{\sigma_\mu^2}{I}}$.
- When estimating σ^2 and σ_μ^2 from data to give $SE(\bar{x})$, then:
 - when testing $H_0 : \mu = \mu_0$ vs some H_a , you can use a t -test with $df=I-1$ because the test statistic $t = \frac{\bar{x} - \mu_0}{SE(\bar{x})} \sim t(I-1)$.
 - when constructing a CI for μ , a $100 \times (1 - \alpha)\%$ CI is $\bar{x} \pm t_{1-\alpha/2} SE(\bar{x})$.

Some examples:

- In a multi-hospital trial studying infections, hospital is the random effect. μ is the overall mean infection and μ_i is the infection rate at hospital i . ρ is the proportion of the total variability in infection rates due to hospital-to-hospital sources such as different patient populations, staff, equipment, and location.
- When checking the tolerance of an industrial product made by different machines, machine is the random effect. μ is the overall mean tolerance and μ_i is the tolerance of machine i . ρ is the proportion of the total variability in the products’ tolerance due to machine-to-machine sources such as the age of the machine, different maintenance schedules, and different operators.
- In a study of grizzly bears in Yellowstone National Park, I different locations are randomly selected in the Park and monitored for griz activity for a month, location is the random effect. μ is the overall mean number of grizzly bears in a month and μ_i is the mean monthly number of grizzly bears at location i . ρ is the proportion of the total variability in the monthly grizzly bear numbers due to location-to-location sources such as close proximity to water sources and prey and other bears, use by humans, cover, and elevation.

EXAMPLE continued:

```
library(lme4)      # install the lme4 pckage from CRAN
hb.lmer = lmer(HeadBreadth ~ (1|Epoch), data=D)
summary(hb.lmer)
##Linear mixed model fit by REML ['lmerMod']
##Formula: HeadBreadth ~ (1 | Epoch)
```

```

##
##Random effects:
## Groups   Name          Variance Std.Dev.
## Epoch   (Intercept)  5.805   2.409
## Residual                17.130   4.139
##Number of obs: 27, groups: Epoch, 3
##
##Fixed effects:
##              Estimate Std. Error t value
##(Intercept)  135.074      1.603   84.27

# p-value for Ha: overall mean not equal to 0
2*(1-pt(abs(84.27),2))
##[1] [1] 0.0001407869

# 95% CI for overall mean
135.074 + c(-1,1)*qt(.975,2)*1.603
##[1] 128.1768 141.9712

```

1. Assuming that the Epochs were a random sample of all Epochs, find the variability of the mean head breadth across Epochs.
2. Confirm that the estimate of σ^2 is the same as MSE from the 1-way fixed effect ANOVA.
3. Give the total variance of head breadths across Epochs.
4. What percentage of the total variability was attributable to Epoch-to-Epoch variability?
5. What is the correlation of headbreadths from the same Epoch?
6. Give \bar{x} an estimate of the overall mean μ of head breadth across all Epochs.
7. Write down the formula that shows how $SE(\bar{x})$ was calculated.
8. Give a 95% CI for the overall mean μ of head breadth across all Epochs.