Stat 427/527: Advanced Data Analysis I

Chapter 5: One-Way Analysis of Variance

Instructor: Yan Lu

THE UNIVERSITY OF NEW MEXICO.

Learning objectives

After completing this topic, you should be able to:

- **select** graphical displays that meaningfully compare independent populations.
- **assess** the assumptions of the analysis of variance (ANOVA) visually and by formal tests.
- **decide** whether the means between populations are different, and how.

One-way ANOVA

- The one-way analysis of variance is a generalization of the two sample t-test to $k > 2$ groups.
  —— Assume that the populations of interest have the following (unknown) population means and standard deviations:

	population 1	population 2	$\cdots$	population k
mean	μ_1	μ_2	$\cdots$	μ_k
std dev	σ_1	σ_2	$\cdots$	σ_k

- A usual interest in ANOVA is whether $\mu_1 = \mu_2 = \cdots = \mu_k$. If not, then we wish to know which means differ, and by how much.

Data Structure

- Let Y_{ij} denote the j^{th} observation in the i^{th} sample/group, $i = 1, 2, \cdots, k$ and $j = 1, 2, \cdots n_i$
- Select samples from each of the k populations,

	sample 1	sample 2	$\cdots$	sample k
size	n_1	n_2	$\cdots$	n_k
mean	$\bar{Y}_{1\cdot}$	$\bar{Y}_{2\cdot}$	$\cdots$	$\bar{Y}_{k\cdot}$
SE	S_1	S_2	$\cdots$	S_k

- total sample size $n_T = n_1 + n_2 + \cdots + n_k$, $\bar{Y}_{i\cdot} = \sum_{j=1}^{n_i} Y_{ij}/n_i$
- let $\bar{Y}_{\cdot\cdot}$ be the average response over all samples, that is

$$\bar{Y}_{\cdot\cdot} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}}{n} = \frac{\sum_{i=1}^{k} n_i \bar{Y}_{i\cdot}}{n}.$$

Note that $\bar{Y}_{\cdot\cdot}$ is *not* the average of the sample means, unless the sample sizes n_i are equal.

An F-statistic is used to test

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

against

$$H_A : \text{ not } H_0, \text{ that is, at least two means are different.}$$

The assumptions needed for the standard ANOVA F-test are analogous to the independent pooled two-sample t-test assumptions:

(1) Independent random samples from each population.

(2) The population frequency curves are normal.

(3) The populations have equal standard deviations, $\sigma_1 = \sigma_2 = \cdots = \sigma_k$.

Sum of Squares (SS)

- **Within SS**, often called the **Residual SS** or the **Error SS**, is the portion of the total spread due to variability *within* samples:

$$\text{SS(Within)} = (n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + \cdots + (n_k - 1)S_k^2 = \sum_{ij}(Y_{ij} - \bar{Y}_{i\cdot})^2.$$

- **Between SS**, often called the Model SS, measures the spread between the sample means

$$\text{SS(Between)} = n_1(\bar{Y}_{1\cdot} - \bar{Y}_{\cdot\cdot})^2 + n_2(\bar{Y}_{2\cdot} - \bar{Y}_{\cdot\cdot})^2 + \cdots + n_k(\bar{Y}_{k\cdot} - \bar{Y}_{\cdot\cdot})^2 = \sum_i n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2,$$

weighted by the sample sizes. These two SS add to give

- **SS(total)**

$$\text{SS(Total)} = \text{SS(Between)} + \text{SS(Within)} = \sum_{ij}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2.$$

Degrees of Freedom (df)

- The df(Between) is the number of groups minus one, $k - 1$.
- The df(Within) is the total number of observations minus the number of groups:
  $(n_1 - 1) + (n_2 - 1) + \cdots + (n_k - 1) = n - k$.
- These two df add to give df(Total)
  $= (k - 1) + (n - k) = n - 1$.

ANOVA Table

Source	df	SS	MS	F
Between Groups (Model)	$df_M = k - 1$	$SSM = \sum_i n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$	$MSM = \frac{SSM}{df_M}$	$F = \frac{MSM}{MSE}$
Within Groups (Error)	$df_E = n - k$	$SSE = \sum_i (n_i - 1)S_i^2$	$MSE = \frac{SSE}{df_E}$	
Total	$df_T = n - 1$	$SST = \sum_{ij}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2$	$MST = \frac{SST}{df_T}$	

The Mean Square for each source of variation is the corresponding SS divided by its *df*.

- The MS(Within)

$$\text{MS(Within)} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + \cdots + (n_k - 1)S_k^2}{n - k} = S^2_{\text{pooled}}$$

  is a weighted average of the sample variances.

- The MS(Within) is known as the pooled estimator of variance, and estimates the assumed common population variance.
- If all the sample sizes are equal, the MS(Within) is the average sample variance. The MS(Within) is identical to the **pooled variance estimator** in a two-sample problem when $k = 2$.

The MS(Between)

$$\mathrm{MS(Between)} = \frac{\sum_{i=1}^{k} n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2}{k-1}$$

is a measure of variability among the sample means.

- This MS is a multiple of the sample variance of $\bar{Y}_{1\cdot}, \bar{Y}_{2\cdot}, \ldots, \bar{Y}_{k\cdot}$.

The MS(Total)

$$\mathrm{MS(Total)} = \frac{\sum_{ij}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2}{n-1}$$

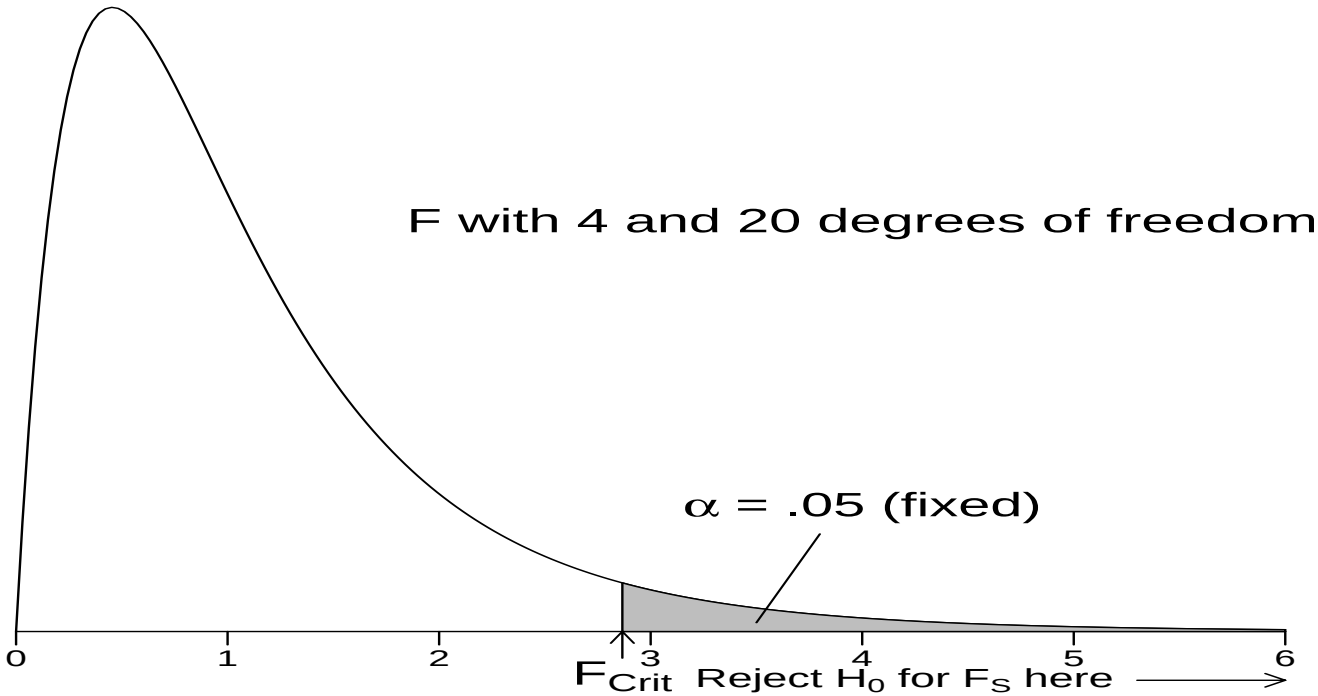
is the variance in the combined data set.

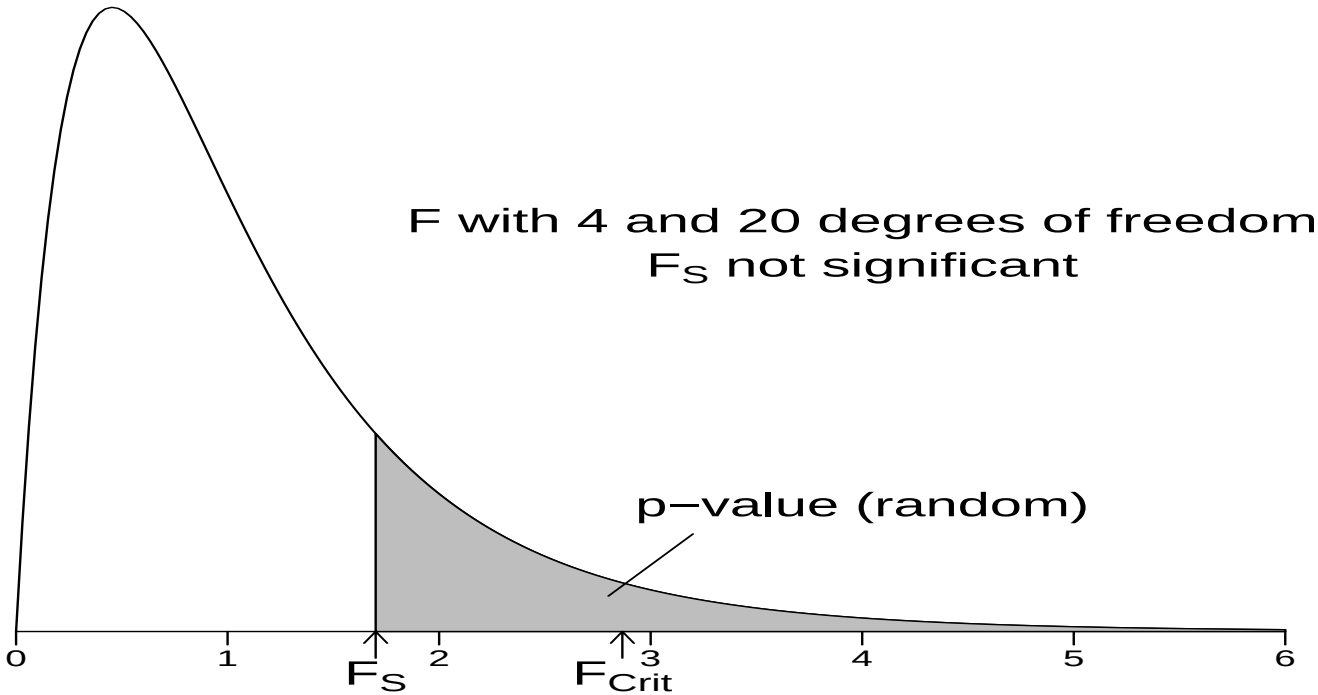
Test of equivalence of the means

The decision on whether to reject $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ is based on the ratio of the MS(Between) and the MS(Within):

$$F_s = \frac{\text{MS(Between)}}{\text{MS(Within)}}.$$

- Large values of F_s indicate large variability among the sample means $\bar{Y}_{1\cdot}, \bar{Y}_{2\cdot}, \ldots, \bar{Y}_{k\cdot}$ relative to the spread of the data within samples. That is, large values of F_s suggest that H_0 is false.
- Formally, for a size α test, reject H_0 if $F_s \geq F_{crit}$, —-where F_{crit} is the upper-α percentile from an $F(k-1, n-k)$ distribution with numerator degrees of freedom $k-1$ and denominator degrees of freedom $n-k$ (i.e., the *df* for the numerators and denominators in the F-ratio).
- The p-value for the test is the area under the F-probability curve to the right of F_s.


F with 4 and 20 degrees of freedom
α = .05 (fixed)
0
1
2
3
4
5
6
F Crit
Reject H0 for FS here


F with 4 and 20 degrees of freedom
F_S not significant
p−value (random)
0
1
F_S
2
F_{Crit}
3
4
5
6

- For $k = 2$ the ANOVA F-test is equivalent to the pooled two-sample t-test.
- We calculate a model object using lm() or aov() and extract the analysis of variance table with anova().

Example: Comparison of Fats

During cooking, doughnuts absorb fat in various amounts. A scientist wished to learn whether the amount absorbed depends on the type of fat.

- For each of 4 fats, 6 batches of 24 doughnuts were prepared.
- The data are grams of fat absorbed per batch.

Row	fat1	fat2	fat3	fat4
1	164	178	175	155
2	172	191	186	166
3	168	197	178	149
4	177	182	171	164
5	190	185	163	170
6	176	177	176	168

Read in the wide table

```
#### Example: Comparison of Fats
fat <- read.table(text="
Row  fat1  fat2  fat3  fat4
  1   164   178   175   155
  2   172   191   186   166
  3   168   197   178   149
  4   177   182   171   164
  5   190   185   163   170
  6   176   177   176   168
", header=TRUE)
fat

##   Row fat1 fat2 fat3 fat4
## 1   1  164  178  175  155
## 2   2  172  191  186  166
## 3   3  168  197  178  149
## 4   4  177  182  171  164
## 5   5  190  185  163  170
## 6   6  176  177  176  168
```

Convert the wide table into long format

Use melt() from the reshape2 package.

```
#### From wide to long format
library(reshape2)
fat.long <- melt(fat,
                 # id.vars: ID variables
                 #   all variables to keep but not split apart on
                 id.vars=c("Row"),
                 # measure.vars: The source columns
                 #(if unspecified then all other variables are measure.var
                 measure.vars = c("fat1", "fat2", "fat3", "fat4"),
                 # variable.name: Name of the destination column
                 #each original column that the measurement came from
                 variable.name = "type",
                 # value.name: column name for values in table
                 value.name = "amount"
               )
#names(fat.long) <- c("Row", "type", "amount")
```

```
fat.long

##    Row type amount
## 1    1 fat1    164
## 2    2 fat1    172
## 3    3 fat1    168
## 4    4 fat1    177
## 5    5 fat1    190
## 6    6 fat1    176
## 7    1 fat2    178
## 8    2 fat2    191
## 9    3 fat2    197
## 10   4 fat2    182
## 11   5 fat2    185
## 12   6 fat2    177
## 13   1 fat3    175
## 14   2 fat3    186
## 15   3 fat3    178
## 16   4 fat3    171
## 17   5 fat3    163
## 18   6 fat3    176
## 19   1 fat4    155
## 20   2 fat4    166
```

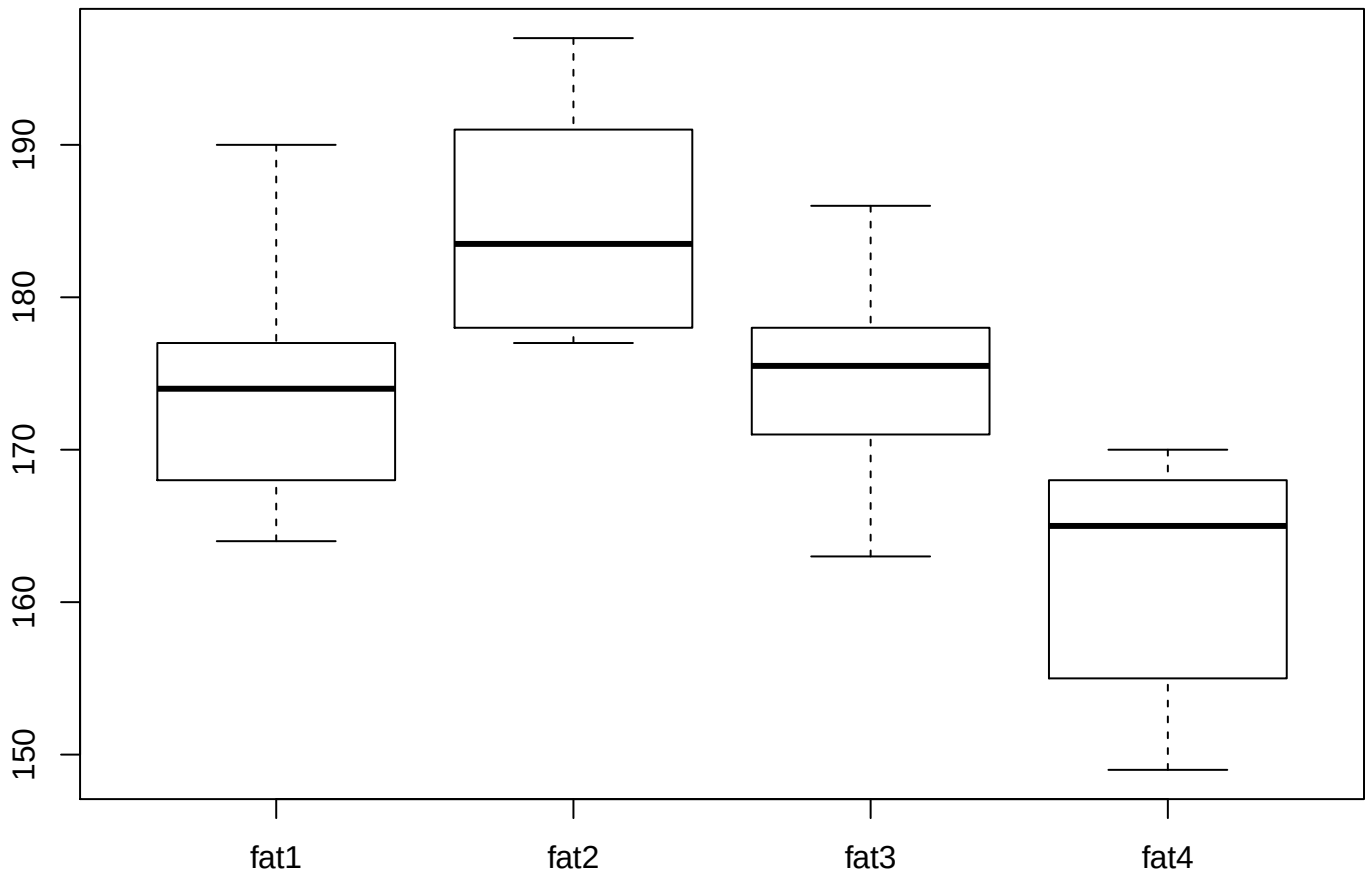
Numerical summaries

```
#### Back to ANOVA
# Calculate the mean, sd, n, and se for the four fats
# The plyr package is an advanced way to apply a function
#to subsets of data, splitting, applying and combining data"
library(plyr)
# ddply "dd" means the input and output are both data.frames
fat.summary <- ddply(fat.long,
                     "type",
                     function(X) {
                       data.frame( m = mean(X$amount),
                                   s = sd(X$amount),
                                   n = length(X$amount)
                                  )})
# standard errors
fat.summary$se   <- fat.summary$s/sqrt(fat.summary$n)
# individual confidence limits
fat.summary$ci.l <- fat.summary$m -
  qt(1-.05/2, df=fat.summary$n-1) * fat.summary$se
fat.summary$ci.u <- fat.summary$m +
  qt(1-.05/2, df=fat.summary$n-1) * fat.summary$se
#fat.summary
```

```
fat.summary

##   type        m        s n       se     ci.l     ci.u
## 1 fat1 174.5000 9.027735 6 3.685557 165.0260 183.9740
## 2 fat2 185.0000 7.771744 6 3.172801 176.8441 193.1559
## 3 fat3 174.8333 7.626707 6 3.113590 166.8296 182.8371
## 4 fat4 162.0000 8.221922 6 3.356586 153.3716 170.6284
```

```
boxplot(amount type,data=fat.long)
```



```
fit.f <- aov(amount ~ type, data = fat.long)
summary(fit.f)

##             Df Sum Sq Mean Sq F value Pr(>F)
## type         3   1596   531.8   7.948 0.0011 **
## Residuals   20   1338    66.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

fit.f

## Call:
##    aov(formula = amount ~ type, data = fat.long)
##
## Terms:
##                     type Residuals
## Sum of Squares  1595.500  1338.333
## Deg. of Freedom        3        20
##
## Residual standard error: 8.180261
## Estimated effects may be unbalanced
```

Findings:

- The pooled standard deviation $s_{\text{pooled}} = 8.18$ is the "Residual standard error".
- $df_M = 4 - 1 = 3, df_E = n - k = 24 - 4 = 20$
- $MSM = SSM/df_M = 1596/3 = 532$,
  $MSE = SSE/df_E = 1338/20 = 66.9$
- $F_s = MSM/MSE = 531.8/66.9 = 7.949178$
- $F_{crit} = 3.098$, $F_s > F_{crit}$, therefore, reject H_0 in favor of H_α
- The p-value for the F-test is 0.001. The scientist would reject H_0 at any of the usual test levels (such as, 0.05 or 0.01).
  ——suggest that the population mean absorption rates differ across fats *in some way*.
  —-The F-test does not say *how* they differ.

Multiple Comparison Methods

- The ANOVA F-test checks whether all the population means are equal.
- Multiple comparisons are often used as a follow-up to a significant ANOVA F-test to determine which population means are different.
  —-Fisher's, Bonferroni's, and Tukey's methods for comparing all pairs of means.

Fisher's least significant difference method **(LSD or FSD)**

Two-step process:

1. Carry out the ANOVA F-test of $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ at the α level. If H_0 is not rejected, stop and conclude that there is insufficient evidence to claim differences among population means. If H_0 is rejected, go to step 2.
2. Compare each pair of means using a pooled two sample t-test at the α level. Use s_{pooled} from the ANOVA table and $df = df_E$ (Residual).

Consider the t-test of $H_0 : \mu_i = \mu_j$ (i.e., populations i and j have same mean).

- The t-statistic is

$$T_s = \frac{\bar{Y}_i - \bar{Y}_j}{S_{\text{pooled}}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}.$$

—-reject H_0 if $|t_s| \geq t_{\text{crit}}$, or equivalently, if

$$|\bar{y}_i - \bar{y}_j| \geq t_{\text{crit}} s_{\text{pooled}} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}.$$

——The minimum absolute difference between $\bar{Y}_i$ and $\bar{Y}_j$ needed to reject H_0 is the LSD, the quantity on the right hand side of this inequality.

- If all the sample sizes are equal $n_1 = n_2 = \cdots = n_k$ then the LSD is the same for each comparison:

$$LSD = t_{\text{crit}} s_{\text{pooled}} \sqrt{\frac{2}{n_1}},$$

where n_1 is the common sample size.

Example: doughnut data, using $\alpha = 0.05$

Recall that: at the first step, you reject the hypothesis that the population mean absorptions are equal because p-value$= 0.001$.
At the second step, compare all pairs of fats at the 5% level.

- $s_{\text{pooled}} = 8.18$ and $t_{\text{crit}} = 2.086$ for a two-sided test based on 20 *df* (the *dfE* for Residual SS).
- Each sample has six observations, so the LSD for each comparison is

$$LSD = 2.086 \times 8.18 \times \sqrt{\frac{2}{6}} = 9.85.$$

- Any two sample means that differ by at least 9.85 in magnitude are **significantly different** at the 5% level.

Another way:

- Order the samples by their sample means.

Fats	Sample Mean
2	185.00
3	174.83
1	174.50
4	162.00

- Two fats are in the same group, if the absolute difference between their sample means is smaller than the LSD $= 9.85$.

Comparison	Absolute difference in means	Exceeds LSD?
Fats 2 and 3	10.17	Yes
2 and 1	10.50	Yes
2 and 4	23.00	Yes
Fats 3 and 1	0.33	No
3 and 4	12.83	Yes
Fats 1 and 4	12.50	Yes

Results of Multiple Comparison

- Three groups for the doughnut data, with no overlap.
  —-Fat 2 is in a group by itself, and so is Fat 4.
  —-Fats 3 and 1 are in a group together.
- This information can be summarized by ordering the samples from lowest to highest average, and then connecting the fats in the same group using an underscore:

```
FAT 4     FAT 1     FAT 3      FAT 2
-----     ----------------     -----
```

- At the 5% level, you have sufficient evidence to conclude that the population mean absorption for Fat 2 and Fat 4 are each different than the other population means.
- However, there is insufficient evidence to conclude that the population mean absorptions for Fats 1 and 3 differ.

Interpreting Groups in Multiple Comparisons

- A group is defined to be a set of populations with sample means that are not significantly different from each other.
- Overlap among groups is common, and occurs when one or more populations appears in two or more groups. Any overlap requires a more careful interpretation of the analysis.

——suppose you obtain two groups in a three sample problem. One group has samples 1 and 3. The other group has samples 3 and 2:

```
   1       3       2
 ----------
         ----------
```

—-this happens when $|\bar{Y}_1 - \bar{Y}_2| \geq LSD$, but both $|\bar{Y}_1 - \bar{Y}_3|$ and $|\bar{Y}_3 - \bar{Y}_2|$ are less than the LSD.

—-The groupings imply that we have sufficient evidence to conclude that population means 1 and 2 are different, but insufficient evidence to conclude that population mean 3 differs from either of the other population means.

FSD Multiple Comparisons in R

pairwise.t.test() with p.adjust.method = "none".

```
#### Multiple Comparisons
# all pairwise comparisons among levels of fat
# Fisher's LSD (FSD) uses "none"
pairwise.t.test(fat.long$amount, fat.long$type,
                pool.sd = TRUE, p.adjust.method = "none")

##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  fat.long$amount and fat.long$type
##
##      fat1  fat2    fat3
## fat2 0.038 -       -
## fat3 0.944 0.044   -
## fat4 0.015 9.3e-05 0.013
##
## P value adjustment method: none
```

Discussion of the FSD Method: family error rate

- Have $c = k(k-1)/2$ pairs of means to compare
- Each comparison is done at the α level, where for a generic comparison of the i^{th} and j^{th} populations

  $\alpha =$ probability of rejecting $H_0 : \mu_i = \mu_j$ when H_0 is true.
- **family error rate** (FER), or the **experimentwise error rate**, is defined to be *the probability of at least one false rejection of a true hypothesis* $H_0 : \mu_i = \mu_j$ *over all comparisons*.
  ——When many comparisons are made, you *may* have a large probability of making one or more false rejections of true null hypotheses.
  ——when all c comparisons of two population means are performed, each at the α level, then

$$\alpha < FER < c\alpha.$$

Example: doughnut problem

- $k = 4$, there are $c = 4(3)/2 = 6$ possible comparisons of pairs of fats.
- Suppose each comparison is carried out at the 5% level, then $0.05 < FER < 0.30$.
  ——At the second step of the FSD method, you could have up to a 30% chance of claiming one or more pairs of population means are different if no differences existed between population means.

Comments:

- The first step of F test of equivalence of the means the FSD method is the ANOVA "screening" test.
- The multiple comparisons are carried out only if the F-test suggests that not all population means are equal.
- FSD method is commonly criticized for being extremely liberal (too many false rejections of true null hypotheses) when some, but not many, differences exist — especially when the number of comparisons is large.
  —-When you do a large number of tests, each, say, at the 5% level, then sampling variation alone will suggest differences in 5% of the comparisons where the H_0 is true. The number of false rejections could be enormous with a large number of comparisons.

Bonferroni Comparisons

Suppose we have two statements: s_1 and s_2

- Statement 1 is correct with probability $1 - \alpha$.
- Statement 2 is correct with probability $1 - \alpha$.
- What is the probability that both statements are simultaneously correct?
  (1) If the statements are independent, then the probability that both are correct is $(1 - \alpha)(1 - \alpha)$.
  (2) But they are not independent. The actual probability is difficult to compute.
- p(s_1 is true and s_2 is true)
  =p(both s_i's are simultaneously true)
  $\geq 1 - 2\alpha$
  —-this gives a lower bound on the probability that both statements are simultaneously true.

- Bonferroni Inequality
  Let $s_1, s_2 \cdots s_c$ be statements with

$$p(s_i \text{ is true}) = 1 - \alpha_i$$

  then
  p(s_1 is true, s_2 is true $\cdots$ and s_c is true)
  =p(all s_i's are simultaneously true)
  $\geq$ 1-$\sum_{i=1}^{c} \alpha_i$
- If α_is are equal, p(s_1 is true, s_2 is true $\cdots$ and s_c is true)
  $\geq$ 1-$c\alpha$ or

$$FER < c\alpha.$$

  Example: Suppose $1 - \alpha_i = .90$, $k = 10$
  $p(\text{All} \quad 10 \quad s_i's \quad \text{true}) \geq 1 - \sum_{i=1}^{10} .10 = 0$
  The Bonferroni inequality works, but might not work very well.

- Example: If β_0 and β_1 both have 95% confidence intervals

$$b_0 \pm t(.975; n-2)s(b_0)$$

and

$$b_1 \pm t(.975; n-2)s(b_1)$$

The joint confidence coefficient using the Bonferroni inequality is greater than or equal to $1 - .05 - .05 = .90$

- To get a joint confidence coefficient of at least $(1-\alpha)$ for β_0 and β_1, we use the confidence intervals

$$b_0 \pm t(1 - \frac{\alpha}{4}; n-2)s(b_0)$$

and

$$b_1 \pm t(1 - \frac{\alpha}{4}; n-2)s(b_1)$$

The confidence coefficient is at least

$$1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha.$$

Genreal Case

To get a joint confidence coefficient of at least $(1-\alpha)$ for c parameters, we construct each interval estimate with statement confidence coefficient $1-\alpha/c$

- The confidence coefficient is at least

$$1 - c * \frac{\alpha}{c} = 1 - \alpha.$$

- The Bonferroni method controls the **family error rate** FER by reducing the individual comparison error rate.
- We have at least $100(1-\alpha)\%$ confidence that all pairwise t-test statements hold simultaneously!

Implementation in R

Bonferroni adjustment in R: p.adjust.method = "bonf"

```
# Bonferroni 95% Individual p-values
# All Pairwise Comparisons among Levels of fat
pairwise.t.test(fat.long$amount, fat.long$type,
                pool.sd = TRUE, p.adjust.method = "bonf")

##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  fat.long$amount and fat.long$type
##
##      fat1    fat2    fat3
## fat2 0.22733 -       -
## fat3 1.00000 0.26241 -
## fat4 0.09286 0.00056 0.07960
##
## P value adjustment method: bonferroni
```

Grouping

We have sufficient evidence to conclude that the population mean absorption for Fat 2 is different than that for Fat 4.

```
FAT 4     FAT 1     FAT 3     FAT 2
----------------------
          ----------------------
```

- The Bonferroni method tends to produce "coarser" groups than the FSD method, because the individual comparisons are conducted at a lower (alpha/error) level.

- Equivalently, the minimum significant difference is inflated for the Bonferroni method.
  —-For example, in the doughnut problem with $FER \leq 0.05$, the critical value for the individual comparisons at the 0.05/6=0.0083 level is $t_{\text{crit}} = 2.929$ with $df = 20$, versus LSD at the 0.05 level with $df = 20$ and $t_{\text{crit}} = 2.086$
  —-The minimum significant difference for the Bonferroni comparisons is

$$LSD = 2.929 \times 8.18 \times \sqrt{\frac{2}{6}} = 13.824$$

  versus an LSD=9.85 for the FSD method.
  —-Recall that the sole comparison where the absolute difference between sample means exceeds 13.824 involves Fats 2 and 4.

Fats	Sample Mean
2	185.00
3	174.83
1	174.50
4	162.00

Example from Koopmans: glabella facial tissue thickness

In an anthropological study of facial tissue thickness for different racial groups,

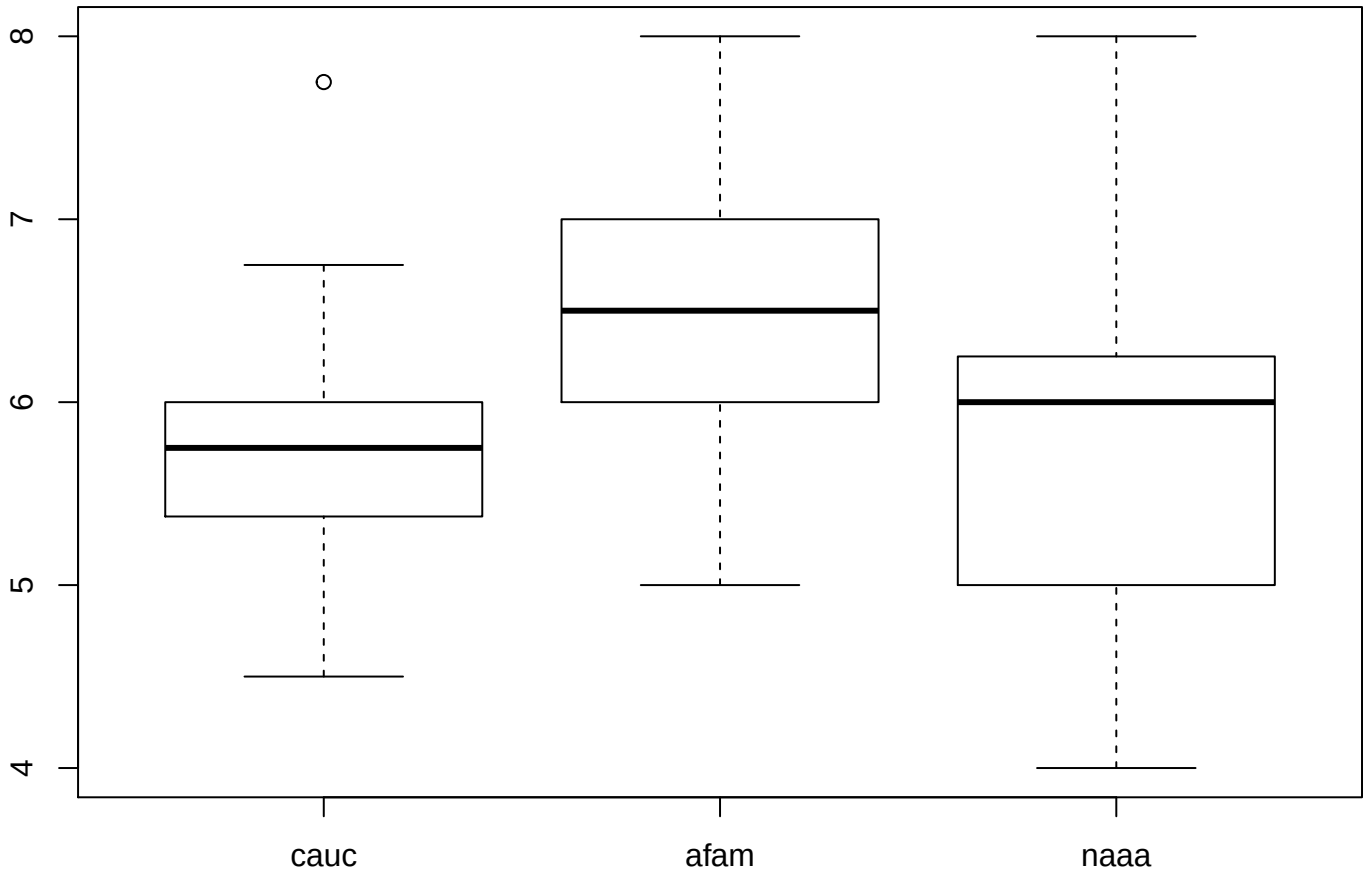
- data were taken during autopsy at several points on the faces of deceased individuals.
- the Glabella measurements taken at the bony ridge for samples of individuals from three racial groups
  —-cauc = Caucasian
  —-afam = African American
  —-naaa = Native American and Asian
- the data values are in mm.

```
#### Example from Koopmans: glabella facial tissue thickness
glabella <- read.table(text="
Row  cauc  afam  naaa
  1  5.75  6.00  8.00
  2  5.50  6.25  7.00
  3  6.75  6.75  6.00
  4  5.75  7.00  6.25
  5  5.00  7.25  5.50
  6  5.75  6.75  4.00
  7  5.75  8.00  5.00
  8  7.75  6.50  6.00
  9  5.75  7.50  7.25
 10  5.25  6.25  6.00
 11  4.50  5.00  6.00
 12  6.25  5.75  4.25
 13    NA  5.00  4.75
 14    NA    NA  6.00
", header=TRUE)
```

```
 glabella.long <- melt(glabella,
                       id.vars=c("Row"),
                       variable.name = "pop",
                       value.name = "thickness",
                       # remove NAs
                       na.rm = TRUE
                     )
 # naming variables manually, the variable.name and value.name not work
 names(glabella.long) <- c("Row", "pop", "thickness")
 # another way to remove NAs:
 #glabella.long <- subset(glabella.long, !is.na(thickness))
glabella.long

##    Row  pop thickness
## 1    1 cauc      5.75
## 2    2 cauc      5.50
## 3    3 cauc      6.75
## 4    4 cauc      5.75
## 5    5 cauc      5.00
## 6    6 cauc      5.75
## 7    7 cauc      5.75
## 8    8 cauc      7.75
## 9    9 cauc      5.75
```

```r
# Plot the data using boxplot
boxplot(thickness~pop,data=glabella.long)
```



Bonferroni Pairwise comparisons

- 3 groups with 3 possible pairwise comparisons
- If we want FER of no greater than 0.05, we should do the individual comparisons at the $0.05/3 = 0.0167$ level.
- Except for the mild outlier in the Caucasian sample, the observed distributions are fairly symmetric, with similar spreads. We would expect the standard ANOVA to perform well here.
- Let μ_c = population mean Glabella measurement for Caucasians,
  μ_a = population mean Glabella measurement for African Americans, and
  μ_n = population mean Glabella measurement for Native Americans and Asians.
  —-interest in simultaneous pairwise comparisons of
  $\mu_c - \mu_a = 0$, $\mu_c - \mu_n = 0$,and $\mu_a - \mu_n = 0$

Summary Statistics

```
glabella.summary <- ddply(glabella.long, "pop",
     function(X) { data.frame( m = mean(X$thickness),
                               s = sd(X$thickness),
                               n = length(X$thickness) ) } )
glabella.summary
```

```
##    pop        m         s  n
## 1 cauc 5.812500 0.8334280 12
## 2 afam 6.461538 0.8946959 13
## 3 naaa 5.857143 1.1168047 14
```

Anova fit

```
fit.g <- aov(thickness ~ pop, data = glabella.long)
summary(fit.g)

##             Df Sum Sq Mean Sq F value Pr(>F)
## pop          2   3.40  1.6991   1.828  0.175
## Residuals   36  33.46  0.9295

fit.g

## Call:
##    aov(formula = thickness ~ pop, data = glabella.long)
##
## Terms:
##                       pop Residuals
## Sum of Squares   3.39829  33.46068
## Deg. of Freedom        2        36
##
## Residual standard error: 0.9640868
## Estimated effects may be unbalanced
```

Findings:

- At the 5% level, you would not reject the hypothesis that the population mean Glabella measurements are identical.
  ——That is, you do not have sufficient evidence to conclude that these racial groups differ with respect to their average Glabella measurement.
  —-**This is the end of the analysis!**
- The Bonferroni intervals reinforce this conclusion, all the p-values are greater than 0.05.
  —- If you were to calculate CIs for the difference in population means, each would contain zero.
  —-You can think of the Bonferroni intervals as simultaneous CI. We're (at least) 95% confident that all of the following statements hold simultaneously:
  $-1.62 \leq \mu_c - \mu_a \leq 0.32$, $-0.91 \leq \mu_n - \mu_c \leq 1.00$, and
  $-1.54 \leq \mu_n - \mu_a \leq 0.33$.
  ——The individual CIs have level $100(1 - 0.0167)\% = 98.33\%$.

```
# Bonferroni 95% Individual p-values
# All Pairwise Comparisons among Levels of glabella
pairwise.t.test(glabella.long$thickness, glabella.long$pop,
                pool.sd = TRUE, p.adjust.method = "bonf")

##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  glabella.long$thickness and glabella.long$pop
##
##      cauc afam
## afam 0.30 -
## naaa 1.00 0.34
##
## P value adjustment method: bonferroni
```

Further Discussion of Multiple Comparisons

- The FSD method is most likely to find differences, whether real or due to sampling variation
- Bonferroni is often the most conservative method.
  —-but tends to work well when the number of comparisons is small, say 4 or less.
  —-focus attention only on the comparisons of interest (generated independently of looking at the data!), and ignore the rest.
- You can be reasonably sure that differences suggested by the Bonferroni method will be suggested by almost all other methods, whereas differences not significant under FSD will not be picked up using other approaches.

Tukey's honest significant difference method (HSD) for multiple comparisons

John Tukey's honest significant difference method is to reject the equality of a pair of means based, not on the t-distribution, but the studentized range distribution.
To implement Tukey's method with a FER of α, reject $H_0 : \mu_i = \mu_j$ when

$$|\bar{Y}_i - \bar{Y}_j| \geq \frac{q_{crit}}{\sqrt{2}} s_{\text{pooled}} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}},$$

where q_{crit} is the α level critical value of the studentized range distribution.

```
#### Tukey's honest significant difference method (HSD)
# Tukey 95% Individual p-values
# All Pairwise Comparisons among Levels of fat
TukeyHSD(fit.f)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = amount ~ type, data = fat.long)
##
## $type
##                   diff        lwr        upr     p adj
## fat2-fat1  10.5000000  -2.719028 23.7190277 0.1510591
## fat3-fat1   0.3333333 -12.885694 13.5523611 0.9998693
## fat4-fat1 -12.5000000 -25.719028  0.7190277 0.0679493
## fat3-fat2 -10.1666667 -23.385694  3.0523611 0.1709831
## fat4-fat2 -23.0000000 -36.219028 -9.7809723 0.0004978
## fat4-fat3 -12.8333333 -26.052361  0.3856944 0.0590077
```

For the doughnut fats, the groupings based on Tukey and Bonferroni comparisons are identical.

```
## Glabella
# Tukey 95% Individual p-values
# All Pairwise Comparisons among Levels of pop
TukeyHSD(fit.g)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = thickness ~ pop, data = glabella.long)
##
## $pop
##                  diff        lwr       upr     p adj
## afam-cauc  0.64903846 -0.2943223 1.5923993 0.2259806
## naaa-cauc  0.04464286 -0.8824050 0.9716907 0.9923923
## naaa-afam -0.60439560 -1.5120412 0.3032500 0.2472838
```

The classical ANOVA assumes

- the populations have normal frequency curves
  —-test the normality assumption using multiple normal QQ-plots and normal scores tests.
  —-An alternative approach that is useful with three or more samples is to make a single normal scores plot for the entire data set.
- the populations have equal variances (or spreads).

One way Anova Model

Consider one way ANOVA model

$$Y_{ij} = \mu_i + \epsilon_{ij} \tag{1}$$

where

- Y_{ij} is the value of the response variable in the jth trial for the ith factor level/sample/group/treatment
- μ_i are parmeters to be estimated
- ϵ_{ij} are independent $N(0, \sigma^2)$, $i = 1, \cdots, k; j = 1, \cdots n_i$

Least square estimators

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

Consider the deviation of Y_{ij} from its expected value $[Y_{ij} - \mu_i]$

- Measure:

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2$$

- Objective: to find estimates μ_i, for which Q is minimum
- $\hat{\mu}_i = \bar{Y}_{i\cdot}$

Residuals

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

with LS estimators $\hat{\mu}_i = \bar{Y}_{i\cdot}$.

- Predicted (fitted or mean) value of Y_{ij} is:

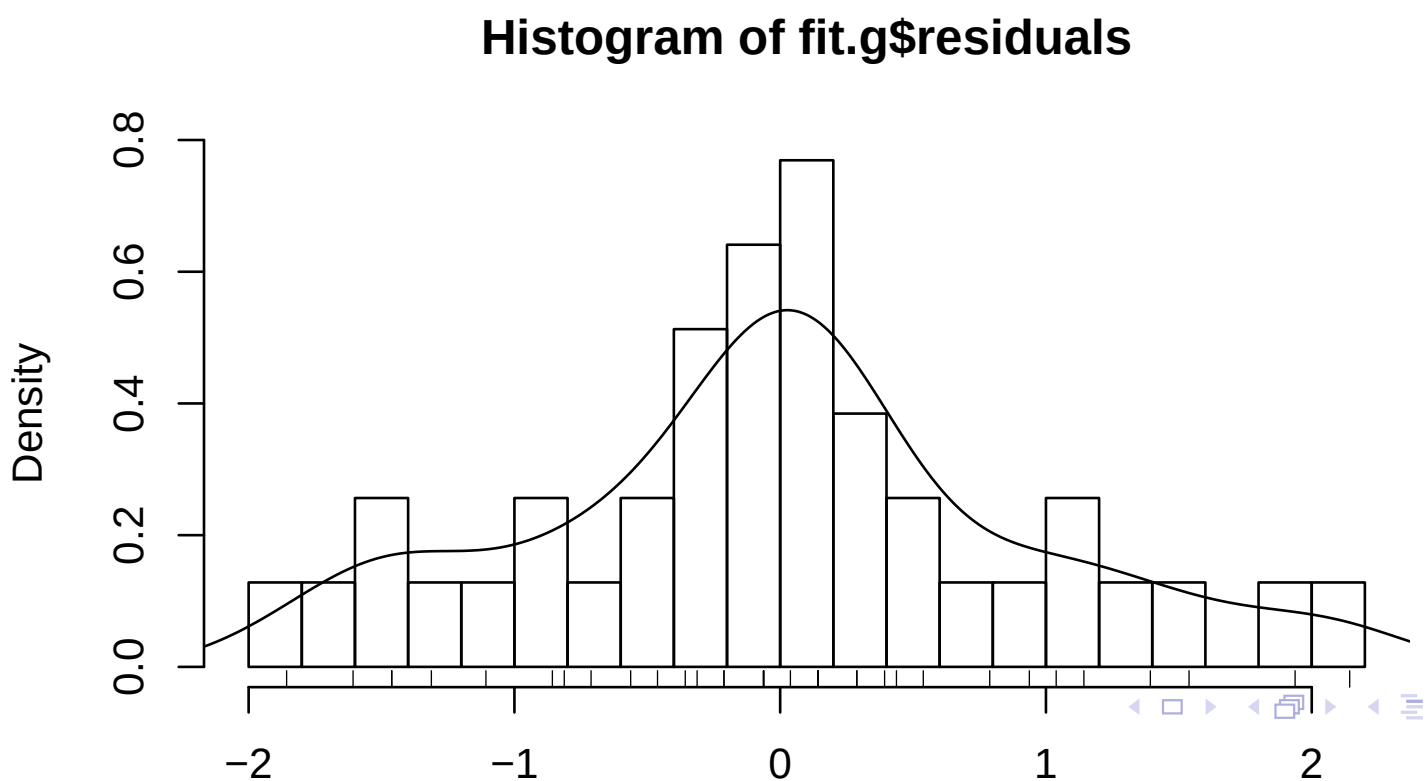
$$\hat{Y}_{ij} = \bar{Y}_{i\cdot}$$

—the fitted value $\hat{Y}_{ij}$ is not the same as Y_{ij}
—Y_{ij} is the observed value and $\hat{Y}_{ij}$ is the predicted value

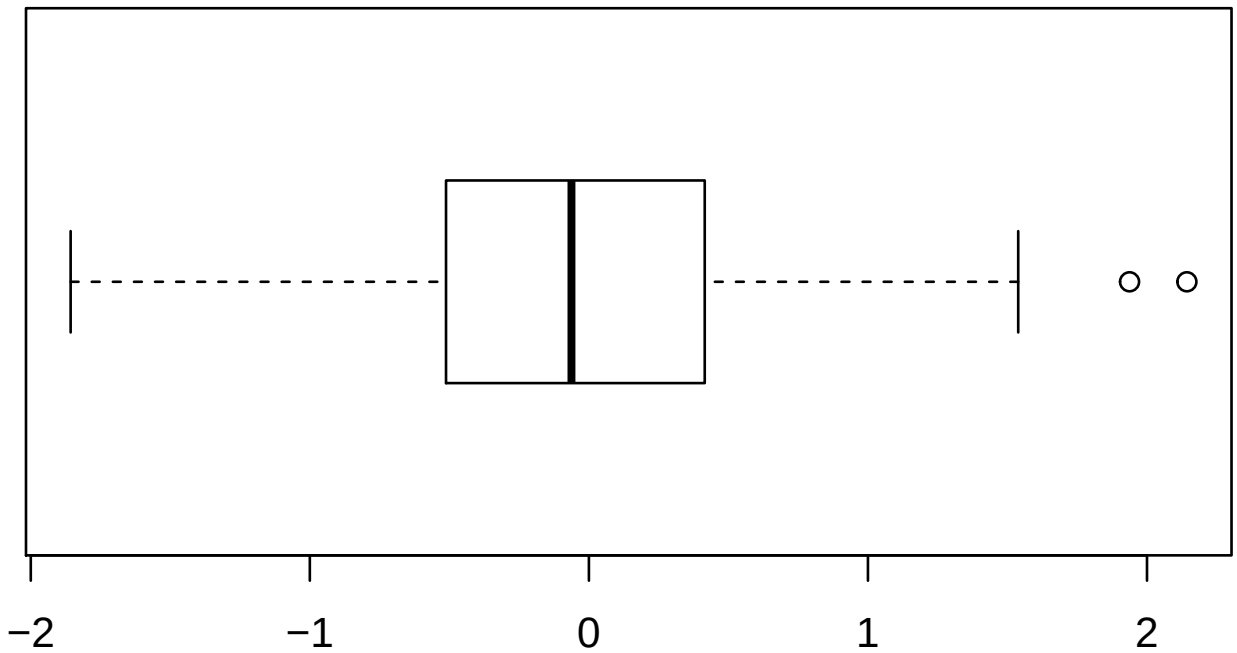
- Residual $e_{ij} = Y_{ij} - \hat{Y}_{ij}$: vertical deviation between Y_{ij} and the estimated μ_i
- Error term $\epsilon_{ij} = Y_{ij} - \mu_i$: vertical deviation between Y_{ij} and the true group mean μ_i
- Residual e_{ij} is a prediction of ϵ_{ij}
—- $e_{ij} \neq \epsilon_{ij}$
A normal scores plot or histogram of the residuals should resemble a sample from a population.

Glabella example diagnostics

```
#### Checking Assumptions in ANOVA Problems
# plot of data
# Histogram overlaid with kernel density curve
hist(fit.g$residuals, freq = FALSE, breaks = 20)
points(density(fit.g$residuals), type = "l")
rug(fit.g$residuals)
```

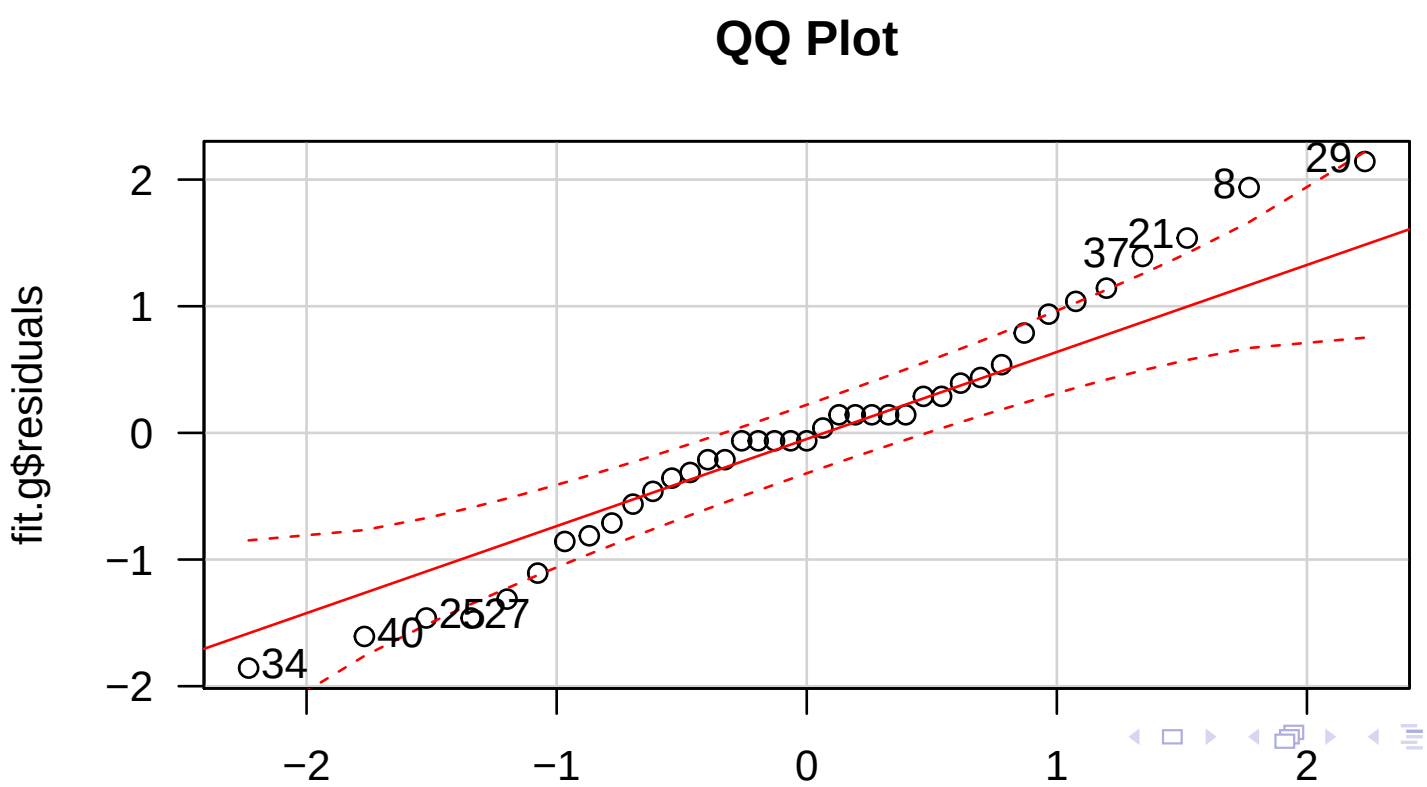


```r
# boxplot
boxplot(fit.g$residuals, horizontal=TRUE)
```



```r
# QQ plot
par(mfrow=c(1,1))
library(car)
qqPlot(fit.g$residuals, las = 1, id.n = 8, id.cex = 1, lwd = 1,
       main="QQ Plot")

## 29  8 34 40 21 25 27 37
## 39 38  1  2 37  3  4 36
```



```
shapiro.test(fit.g$residuals)

##
##  Shapiro-Wilk normality test
##
## data:  fit.g$residuals
## W = 0.97693, p-value = 0.5927

library(nortest)
ad.test(fit.g$residuals)

##
##  Anderson-Darling normality test
##
## data:  fit.g$residuals
## A = 0.37731, p-value = 0.3926
```

```
cvm.test(fit.g$residuals)

##
##  Cramer-von Mises normality test
##
## data:  fit.g$residuals
## W = 0.070918, p-value = 0.2648
```

There are a few observations outside the confidence bands, but the formal normality tests each have $p - values > 0.2$, so there's weak but unconvincing evidence of nonnormality.

Equal variance assumption

- Bartlett Test
- Levene Test

```
## Test equal variance
# Barlett assumes populations are normal
bartlett.test(thickness ~ pop, data = glabella.long)

##
##  Bartlett test of homogeneity of variances
##
## data:  thickness by pop
## Bartlett's K-squared = 1.1314, df = 2, p-value = 0.568
```

Because the p-value > 0.5, we fail to reject the null hypothesis that the population variances are equal. This result is not surprising given how close the sample variances are to each other.

```r
# Levene does not assume normality, requires car package
library(car)
leveneTest(thickness ~ pop, data = glabella.long)

## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  0.5286 0.5939
##       36
```

Levene's tests are consistent with Bartlett's.

Example from the Child Health and Development Study (CHDS)

We consider data from the birth records of 680 live-born white male infants. The infants were born to mothers who reported for pre-natal care to three clinics of the Kaiser hospitals in northern California.

- We will examine whether maternal smoking has an effect on the birth weights of these children.
  —-define 3 groups based on mother's smoking history:
  (1) mother does not currently smoke or never smoked (non smoker, 0 cigs),
  (2) mother smoked less than one pack of cigarettes a day during pregnancy (light smoker, 0-19 cigs)
  (3) mother smoked at least one pack of cigarettes a day during pregnancy (heavey smoker, 20+ cigs)
- Let μ_i = pop mean birth weight (lb) for children in group i, ($i = 1, 2, 3$). We wish to test

$$H_0 : \mu_1 = \mu_2 = \mu_3 \text{ against } H_A : \text{not } H_0.$$

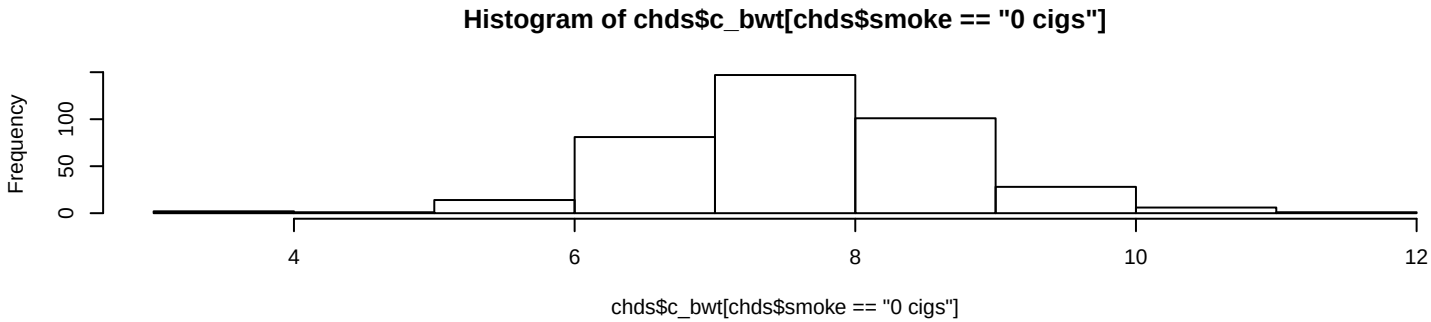
```
#### Example from the Child Health and Development Study (CHDS)
# description at http://statacumen.com/teach/ADA1/ADA1_notes_05-CHDS_de
# read data from website
chds <- read.csv("http://statacumen.com/teach/ADA1/ADA1_notes_05-CHDS.c
  chds$smoke <- rep(NA, nrow(chds));
 # no cigs
 chds[(chds$m_smok == 0), "smoke"] <- "0 cigs";
 # less than 1 pack (20 cigs = 1 pack)
 chds[(chds$m_smok > 0) & (chds$m_smok < 20),"smoke"]<- "1-19 cigs";
 # at least 1 pack (20 cigs = 1 pack)
 chds[(chds$m_smok >= 20),"smoke"] <- "20+ cigs";
 chds$smoke <- factor(chds$smoke)
```

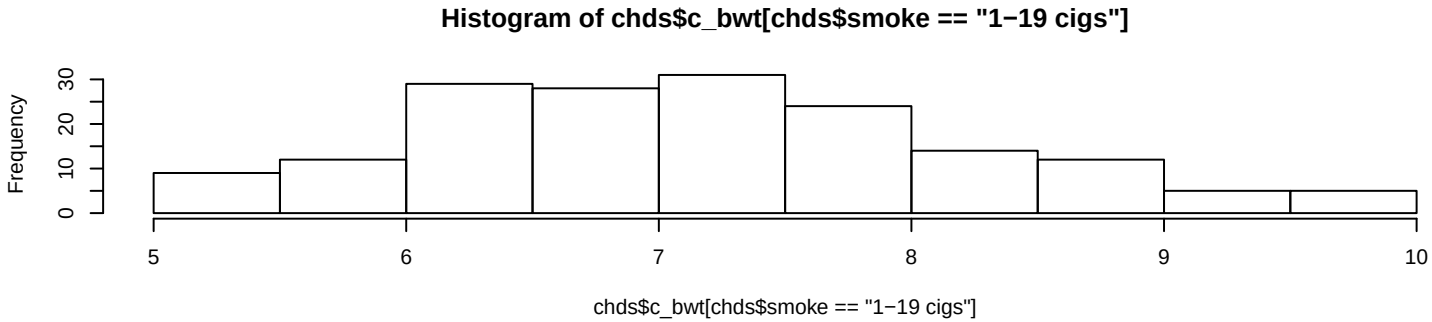
```
head(chds)
  id c_head c_len c_bwt gest m_age m_smok m_ht m_ppwt p_age
1  4     13    20   7.3   37    33     25   66    140    37
2  5     13    21   8.0   41    28      0   63    130    35
3  6     13    21   7.5   39    32      0   61    126    38
4  7     13    20   7.0   39    27      2   68    150    30
5  8     13    19   5.3   37    32     17   67    112    28
6 13     14    20   8.6   43    30      0   63    131    34
p_age p_educ p_smok p_ht  smoke
37      12      25   74    20+ cigs
35      10       7   71    0 cigs
38      12      17   65    0 cigs
30      16       7   73    1-19 cigs
28      10      17   71    1-19 cigs
34      12      17   66    0 cigs
```

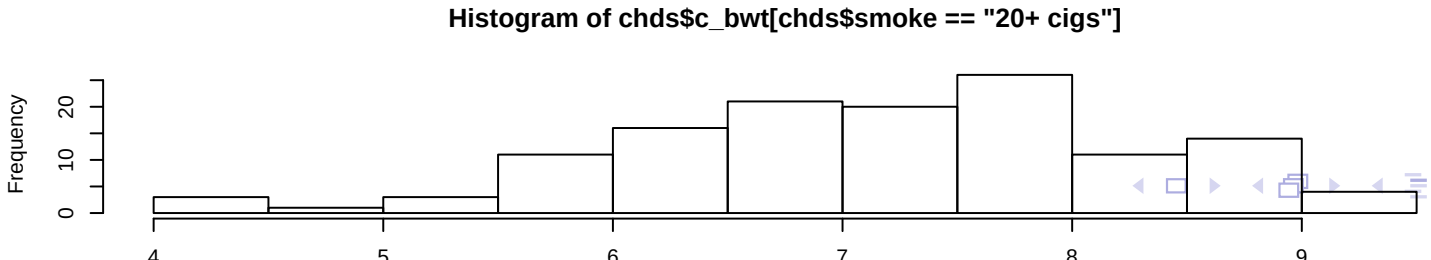
```
# calculate summaries
library(plyr)
chds.summary <-ddply(chds, "smoke",
     function(X) { data.frame( m = mean(X$c_bwt),
                               s = sd(X$c_bwt),
                               n = length(X$c_bwt) ) } )
chds.summary

##       smoke        m        s   n
## 1     0 cigs 7.732808 1.052341 381
## 2 1-19 cigs 7.221302 1.077760 169
## 3  20+ cigs 7.266154 1.090946 130
```

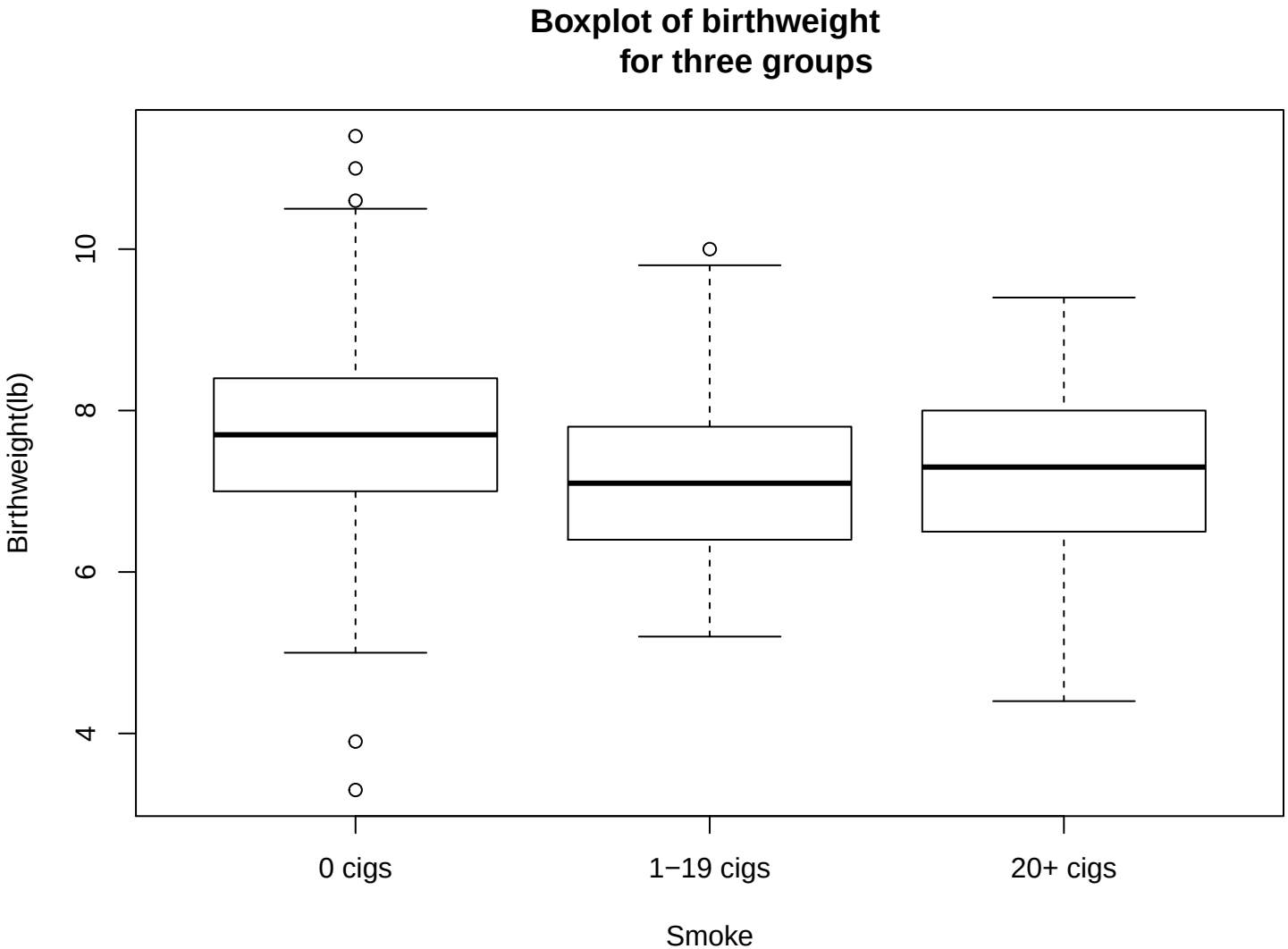
```
# histogram
par(mfrow=c(3,1))
hist(chds$c_bwt[chds$smoke=="0 cigs"])
hist(chds$c_bwt[chds$smoke=="1-19 cigs"])
hist(chds$c_bwt[chds$smoke=="20+ cigs"])
```



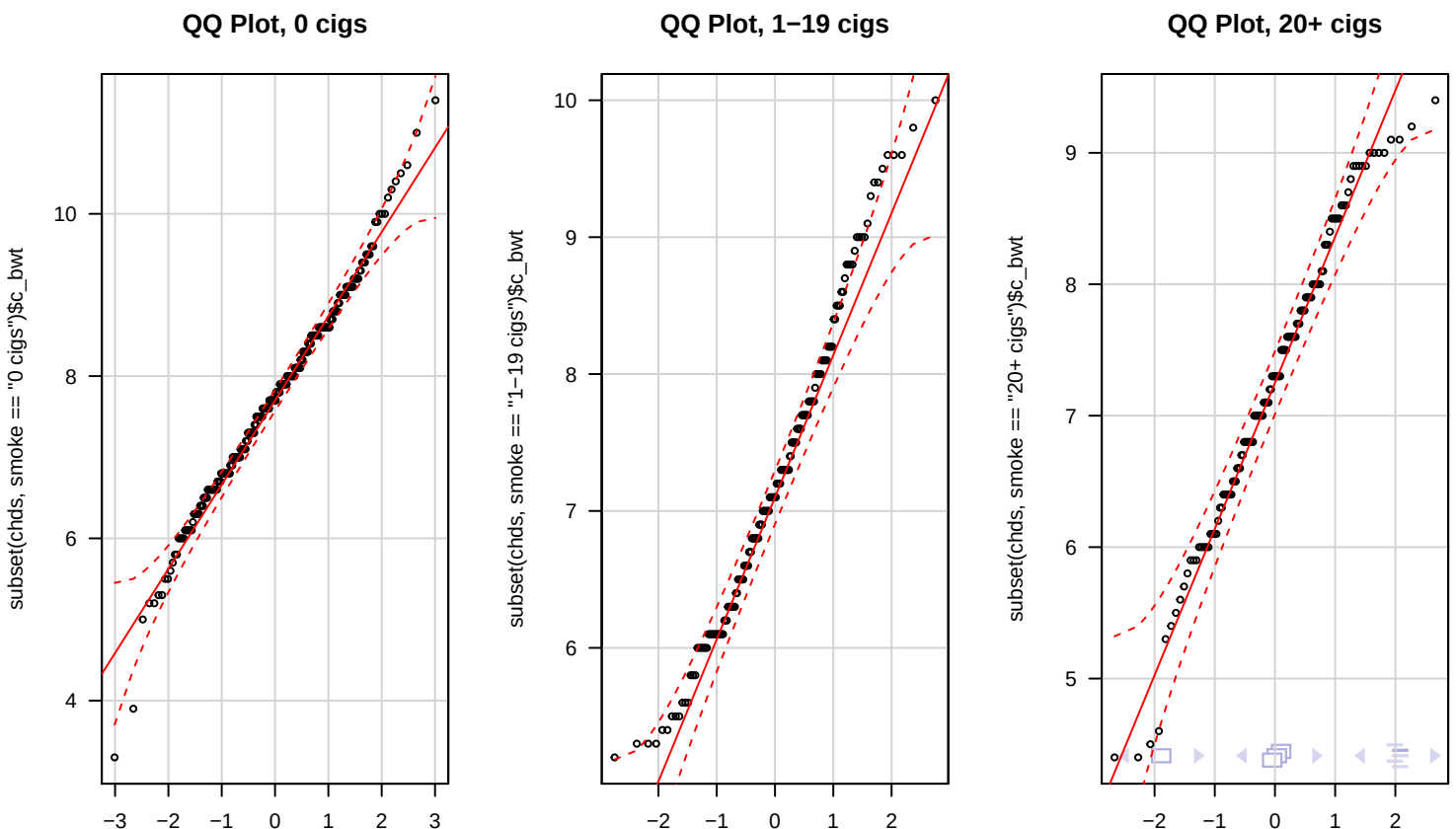




```
boxplot(c_bwt~smoke,data=chds,main="Boxplot of birthweight
        for three groups",xlab="Smoke",ylab="Birthweight(lb)")
```



```
library(car)
par(mfrow=c(1,3))
qqPlot(subset(chds, smoke == "0 cigs"   )$c_bwt, las = 1, id.n = 0,
  id.cex = 1, lwd = 1, main="QQ Plot, 0 cigs")
qqPlot(subset(chds, smoke == "1-19 cigs")$c_bwt, las = 1, id.n = 0,
  id.cex = 1, lwd = 1, main="QQ Plot, 1-19 cigs")
qqPlot(subset(chds, smoke == "20+ cigs" )$c_bwt, las = 1, id.n = 0,
  id.cex = 1, lwd = 1, main="QQ Plot, 20+ cigs")
```



```
shapiro.test(subset(chds, smoke == "0 cigs"   )$c_bwt)

##
##  Shapiro-Wilk normality test
##
## data:  subset(chds, smoke == "0 cigs")$c_bwt
## W = 0.98724, p-value = 0.00199

library(nortest)
ad.test(     subset(chds, smoke == "0 cigs"   )$c_bwt)

##
##  Anderson-Darling normality test
##
## data:  subset(chds, smoke == "0 cigs")$c_bwt
## A = 0.92825, p-value = 0.01831

cvm.test(    subset(chds, smoke == "0 cigs"   )$c_bwt)

##
##  Cramer-von Mises normality test
##
## data:  subset(chds, smoke == "0 cigs")$c_bwt
## W = 0.13844, p-value = 0.03374
```

```
# 1-19 cigs  --------------------
 shapiro.test(subset(chds, smoke == "1-19 cigs")$c_bwt)

##
##  Shapiro-Wilk normality test
##
## data:  subset(chds, smoke == "1-19 cigs")$c_bwt
## W = 0.97847, p-value = 0.009926

 ad.test(     subset(chds, smoke == "1-19 cigs")$c_bwt)

##
##  Anderson-Darling normality test
##
## data:  subset(chds, smoke == "1-19 cigs")$c_bwt
## A = 0.83085, p-value = 0.03149

 cvm.test(    subset(chds, smoke == "1-19 cigs")$c_bwt)

##
##  Cramer-von Mises normality test
##
## data:  subset(chds, smoke == "1-19 cigs")$c_bwt
## W = 0.11332, p-value = 0.07317
```

```
 # 20+ cigs  --------------------
 shapiro.test(subset(chds, smoke == "20+ cigs" )$c_bwt)

##
##  Shapiro-Wilk normality test
##
## data:  subset(chds, smoke == "20+ cigs")$c_bwt
## W = 0.98127, p-value = 0.06962

 ad.test(     subset(chds, smoke == "20+ cigs" )$c_bwt)

##
##  Anderson-Darling normality test
##
## data:  subset(chds, smoke == "20+ cigs")$c_bwt
## A = 0.40008, p-value = 0.3578

 cvm.test(    subset(chds, smoke == "20+ cigs" )$c_bwt)

##
##  Cramer-von Mises normality test
##
## data:  subset(chds, smoke == "20+ cigs")$c_bwt
## W = 0.040522, p-value = 0.6694
```

Observations from plots:

- Looking at the summaries, we see that the sample standard deviations are close.
- Looking at the boxplots, there are outliers in the non smoker group.
- Histogram of the low-smoking and heavy smoking groups show skewness.
- A formal test rejects the hypothesis of normality in the no smoker and low smoker groups.

Fit ANOVA

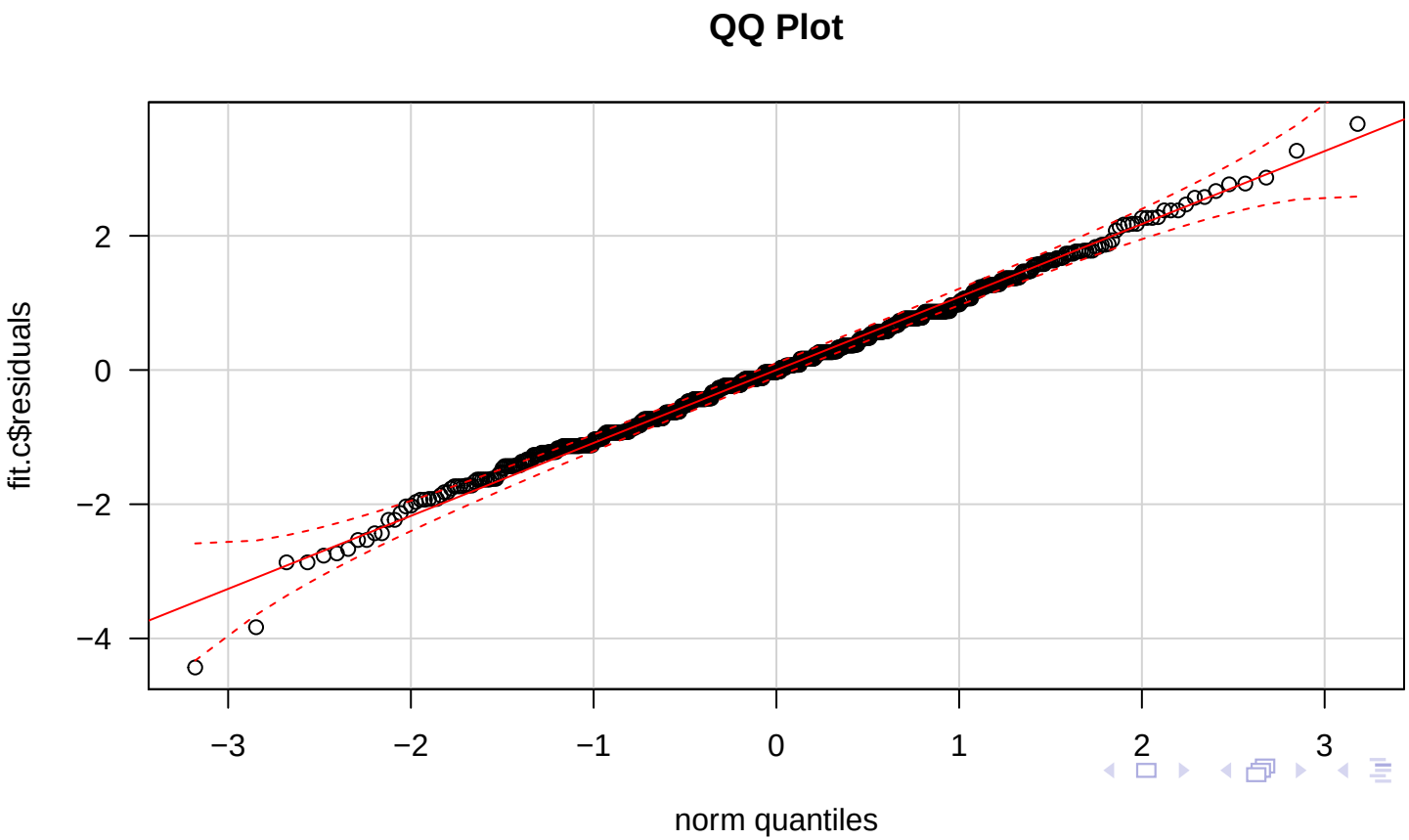
```
fit.c <- aov(c_bwt ~ smoke, data = chds)
summary(fit.c)

##              Df Sum Sq Mean Sq F value   Pr(>F)
## smoke         2   40.7  20.351    17.9 2.65e-08 ***
## Residuals   677  769.5   1.137
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The p-value for the F-test is less than 0.0001. We would reject H_0 at any of the usual test levels (such as 0.05 or 0.01).
- The data suggest that the population mean birth weights differ across smoking status groups.
- We will continue with multiple comparison later.

Test normality by residuals

```
# QQ plot
par(mfrow=c(1,1))
library(car)
qqPlot(fit.c$residuals, las = 1, id.n = 0, id.cex = 1, lwd = 1,
       main="QQ Plot")
```



```
shapiro.test(fit.c$residuals)

##
##  Shapiro-Wilk normality test
##
## data:  fit.c$residuals
## W = 0.99553, p-value = 0.04758

library(nortest)
ad.test(fit.c$residuals)

##
##  Anderson-Darling normality test
##
## data:  fit.c$residuals
## A = 0.62184, p-value = 0.1051

cvm.test(fit.c$residuals)

##
##  Cramer-von Mises normality test
##
## data:  fit.c$residuals
## W = 0.091963, p-value = 0.1449
```

- A formal test of normality on the residuals of the combined sample is marginally significant (SW p-value$=0.047$, others >0.10).
- We are not overly concerned about this since:
  —-in large samples, small deviations from normality are often statistically significant
  —-the small deviations are not likely to impact our conclusions, as inferece is relatively robust to violation of normaility

Checking equal variance assumption

- Summary statistics indicate the variances of the three groups are close to each other
- Formal tests of equal population variances are far from significant.
  —-The p-values for Bartlett's test and Levene's test are greater than 0.4.

Thus, the standard ANOVA appears to be appropriate here.

```r
## Test equal variance
# assumes populations are normal
bartlett.test(c_bwt ~ smoke, data = chds)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  c_bwt by smoke
## Bartlett's K-squared = 0.3055, df = 2, p-value = 0.8583
```

```r
# does not assume normality, requires car package
library(car)
leveneTest(c_bwt ~ smoke, data = chds)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   2  0.7591 0.4685
##       677
```

```r
# nonparametric test
library(car)
fligner.test(c_bwt ~ smoke, data = chds)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  c_bwt by smoke
## Fligner-Killeen:med chi-squared = 2.0927, df = 2, p-value = 0.3512
```

Multiple comparisons

```
## CHDS
# Tukey 95% Individual p-values
TukeyHSD(fit.c)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = c_bwt ~ smoke, data = chds)
##
## $smoke
##                           diff        lwr        upr     p adj
## 1-19 cigs-0 cigs   -0.51150662 -0.7429495 -0.2800637 0.0000008
## 20+ cigs-0 cigs    -0.46665455 -0.7210121 -0.2122970 0.0000558
## 20+ cigs-1-19 cigs  0.04485207 -0.2472865  0.3369907 0.9308357
```

```
## CHDS, multiple comparisons with letters indicating the same group
library(lsmeans)  #tukey comparison

## Loading required package:  estimability

library(multcompView)  #tukey comparison
comp1<-lsmeans(fit.c, "smoke",adjust="tukey")
cld(comp1, alpha=.05,Letters=letters)

##  smoke       lsmean          SE  df lower.CL upper.CL .group
##  1-19 cigs 7.221302 0.08200966 677 7.060278 7.382326  a
##  20+ cigs  7.266154 0.09350540 677 7.082558 7.449749  a
##  0 cigs    7.732808 0.05461927 677 7.625565 7.840052   b
##
## Confidence level used: 0.95
## P value adjustment: tukey method for comparing a family of 3 estimat
## significance level used: alpha = 0.05
```

The Tukey multiple comparisons suggest that the mean birth weights are different (higher) for children born to mothers that did not smoke during pregnancy.