14. Comparing More Than Two Means: One-Way Analysis of Variance (ANOVA) and the General Linear Model

Readings: Kleinbaum, Kupper, Nizam, and Rosenberg (KKNR): Ch. 17
Rosner: Ch. 12.4-12.5

SAS: probf, finv, PROC GLM, PROC NPAR1WAY, PROC MULTTEST, Dunn macro

Homework: Homework 6 due by midnight on October 15
Homework 7 due by midnight on October 29

Overview

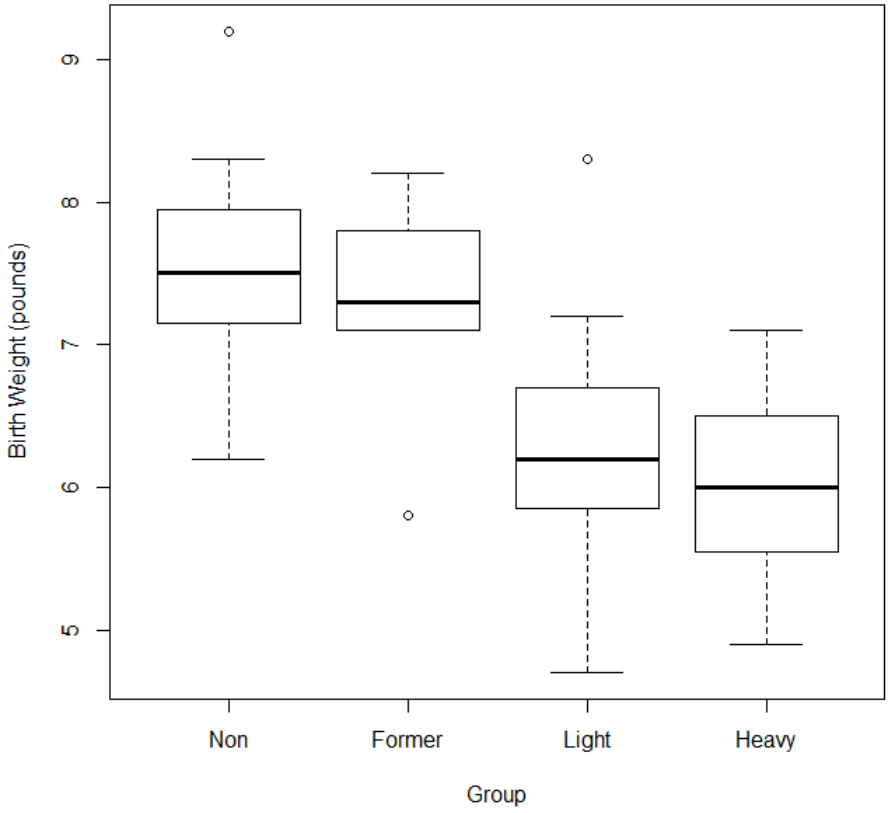
A) One-Way Analysis of Variance (ANOVA)
B) One-Way ANOVA as a General Linear Model (GLM)
C) Extending the One-Way ANOVA
D) ANOVA: Post-Hoc Comparisons
E) Multiple Comparisons
F) Nonparametric ANOVA: The Kruskal-Wallis Test

A. One-Way Analysis of Variance (ANOVA)

One-way ANOVA can be used to compare the means of *J* groups ($J \geq 2$). It can be thought of as a generalization of the independent samples t-test (with equal variances). We'll see that a general linear (regression) model with *J-1 dummy variables* can also be used for this purpose.

Example: Infant birthweight (pounds) and smoking status of mother during the first trimester.

i	*Non Smokers*	*Former Smokers*	*Light Smokers*	*Heavy Smokers*
1	7.50	5.80	5.90	6.20
2	6.20	7.30	6.20	6.80
3	6.90	8.20	5.80	5.70
4	7.40	7.10	4.70	4.90
5	9.20	7.80	8.30	6.20
6	8.30		7.20	7.10
7	7.60		6.20	5.80
8				5.40
$\bar{Y}_j$=	7.586	7.240	6.329	6.013
s_j^2=	0.925	0.833	1.299	0.518



We can conceptualize a one-way ANOVA using the effects model formulation:

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

- Y_{ij} denotes the outcome for the *i*th observation of the *j*th group

- μ is a constant that represents the overall mean of all groups taken together (the "*grand mean*")

- α_j is a constant that represents the difference between the mean of the *j*th group and the grand mean (*between group differences*), where

$$\sum \alpha_j = 0$$

- ε_{ij} is an error term that represents the random errors about the group mean, $\mu + \alpha_j$, for an individual observation from the *j*th group (*within group differences*)

- Each group has n_j observations. In all, we have $N = \sum_j n_j$ total observations.

Note: There is also a means model formulation: $Y_{ij} = \mu_j + \varepsilon_{ij}$, where $\mu_j = \mu + \alpha_j$.

Assumptions for the one-way ANOVA are:

1. **Independence**: The samples are randomly and independently drawn from the respective populations.

2. **Normality**: Each population is normally distributed, and thus the errors follow a normal distribution:
$$\varepsilon_{ij} \sim N(0, \sigma^2) \text{ and } Y_{ij} \sim N(\mu + \alpha_j, \sigma^2)$$

3. **Homoscedasticity**: The variances of the *j* populations are the same. (We'll see that this one is the most problematic with regard to analysis.)

Within and Between Group Variability

If we let $\bar{\bar{Y}}$ denote the grand mean and let $\bar{Y}_j$ denote the mean for the *j*th group, then the deviation of an individual observation from the grand mean can be represented as:

$$Y_{ij} - \bar{\bar{Y}} = \left(Y_{ij} - \bar{Y}_j\right) + (\bar{Y}_j - \bar{\bar{Y}})$$

Where $\left(Y_{ij} - \bar{Y}_j\right)$ represents within-group variability and $(\bar{Y}_j - \bar{\bar{Y}})$ represents between-group variability.

To summarize the total amount of variance we observe and how it breaks down within and between groups, we take the sum of the squared deviations (*sum of squares*) for all Y_{ij}:

$$\begin{aligned}
\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(Y_{ij}-\bar{\bar{Y}}\right)^2 &= \sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(Y_{ij}-\bar{Y}_j\right)^2 + 2\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(Y_{ij}-\bar{Y}_j\right)\left(\bar{Y}_j-\bar{\bar{Y}}\right) + \sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(\bar{Y}_j-\bar{\bar{Y}}\right)^2 \\
&= \sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(Y_{ij}-\bar{Y}_j\right)^2 + 2\sum_{j=1}^{J}\left(n_j\bar{Y}_j-n_j\bar{Y}_j\right)\left(\bar{Y}_j-\bar{\bar{Y}}\right) + \sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(\bar{Y}_j-\bar{\bar{Y}}\right)^2 \\
&= \sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(Y_{ij}-\bar{Y}_j\right)^2 + \sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(\bar{Y}_j-\bar{\bar{Y}}\right)^2 \\
\text{Total SS} &= \text{Within SS} + \text{Between SS} \ (\text{where SS} = \text{sum of squares})
\end{aligned}$$

We see at the final step we have re-written the equation in words to represent the sum of squares for our within variability component and the between variability component.

The ANOVA model determines whether the variability of the data comes mostly from variability within groups or from variability between groups. It does so with a statistical procedure called the *F* test…

Overall *F* Test for the One-Way ANOVA

The null hypothesis of the one-way ANOVA is that all groups have the same mean. Formally,

$H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_J = 0$
H_A: At least one $\alpha_j \neq 0$

or, equivalently,

$H_0: \mu_1 = \mu_2 = \ldots = \mu_J$
H_A: At least one of the means is different

The test of whether there is a difference between groups involves seeing if the between-group variation is significantly larger than the within-group variation. To perform this test, we need to calculate the Mean Square Between ($MS_{Between}$) and the Mean Square Within (MS_{Within}) by dividing the $SS_{Between}$ and SS_{Within} by their respective degrees of freedom:

$$MS_{Between} = SS_{Between} / (J\text{-}1)$$
$$MS_{Within} = SS_{Within} / (N\text{-}J)$$

Under the null hypothesis, the $MS_{Between}$ and MS_{Within} are both unbiased estimators of σ^2 (the common variance among all *J* populations). An *F* statistic can be calculated as the ratio of the two estimates of σ^2.

Summarizing It All: The ANOVA Table

We usually summarize this information in what is called an ANOVA table:

Source	Sum of Squares	Degrees of Freedom	Mean Square	Variance Ratio (F)	p-value
Between	$SS_{Between}$	*J-1*	$MS_{Between}$	*F*	$Pr(F_{J-1,N-J} > F)$
Within	SS_{Within}	*N-J*	$MS_{Within}=\sigma^2_{(Y\|X)}$		
Total	SS_{Total}	*N*-1			

Under the null hypothesis, the *F* statistic will have an *F* distribution with *J*-1 and *N-J* degrees of freedom.

- If $F > F_{J-1,N-J,1-\alpha}$, then reject H_0.
- The exact p-value is given by the area to the right of *F* under an *F* distribution.
- To find probabilities and percentiles in SAS/R see code on the next slide.
- Percentiles of the F-distribution can also be found in Table 9 in the Rosner text.

SAS Code (to calculate the F statistic with df1=J-1, df2=N-J, 1-alpha):

```
data probsqs;
    /* X ~ F(df1=3, df2=23) */
    df1 = 3;
    df2 = 23;
    prob = probf(x, df1, df2);
    /* Computing x so that Pr(X <= x) = 0.95 */
    prob = 0.95;
    df1 = 3;
    df2 = 23;
    x = finv(prob, df1, df2);
run;

proc print data=probsqs;
var prob x;
run;
```

```
Obs prob      x
  1 0.95 3.02800
```

Obs	prob	x
1	0.95	3.02800

R Code:

```
qf(p=0.95, df1=3, df2=23) # determine 95th percentile for F dist.
[1] 3.027998
```

Reading Data into SAS

Similar to R, there are lots of ways we can read data into SAS to use:

- File -> Import Data… -> follow steps in import wizard
- DATA statements (https://stats.idre.ucla.edu/sas/modules/inputting-data-into-sas/)
- PROC IMPORT:

SAS Code:

```
proc import datafile="~/birthweight_smoking_dataset.csv"
     out=BWT /* name for data set for SAS to reference */
     dbms=csv /* identify file as csv */
     replace; /* overwrite BWT if already present */
     getnames=yes; /* take first row as column names from data */
run;
```

Smoking and Birthweight Example Cont.:

SAS Code:

```
proc anova data=BWT;
    class momsmoke;
    model birthwt = momsmoke;
    means momsmoke;
run;
```

The ANOVA Procedure

Dependent Variable: birthwt

Source	R's Name	Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Between	momsmoke	**Model**	3	11.67268915	3.89089638	4.41	0.0137
Within	Residuals	**Error**	23	20.30360714	0.88276553		
Total	NA	**Corrected Total**	26	31.97629630			

Since $p < 0.05$ we reject the null hypothesis and conclude that at least one of the groups has a significantly different birthweight than the rest. (Also, $F = 4.41 > 3.038 = F_{3,23,0.95}$.)

B. One-way ANOVA as a General Linear Model (GLM)

Reference Cell Coding Model

The general linear model is written as $Y = X\beta + \varepsilon$, where *X* is a matrix and β is a vector.

Specifically, *X* is known as a design matrix and has dimensions of *n x p* (n rows and p columns) and β represents our unknown regression coefficients and is a *p x 1* vector.

With this general linear model we know that:
$E(Y) = X\beta$
$V(Y) = \sigma_\varepsilon^2 = \sigma_{Y|X}^2$
$Y|X \sim N(X\beta, \sigma_{Y|X}^2)$ [Note: This does not imply Y *itself* is normally distributed.]

We can draw connections between our ANOVA model by using the intercept model:
$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij} \approx \beta_0 + \beta_F + \beta_L + \beta_H + \varepsilon_{ij}$

β_0, β_F, β_L, and β_H represent the linear effects on infant birthweight of mothers who are non-smokers (β_0), former vs. non-smokers (β_F), light vs. non-smokers (β_L), and heavy vs. non-smokers (β_H).

Categorical Explanatory Variables

Categorical explanatory variables (e.g., sex, smoking status, diabetes status) can be used in a general linear model. We can create an indicator variable (or "dummy variable") that denotes the category. For example:

gender	*smoke*	X_F	X_L	X_H
0 = Female	1 = Non-Smoker	0	0	0
1 = Male	2 = Former Smoker	1	0	0
	3 = Light Smoker	0	1	0
	4 = Heavy Smoker	0	0	1

We can then use these indicator variables in the general linear model:

$$Y = \beta_0 + \beta_F X_F + \beta_L X_L + \beta_H X_H + \varepsilon_{ij}$$

- The "0" category is called the "reference group."
- β_0 is the *mean response* for the reference group: non-smokers.
- $\beta_F, \beta_L, \beta_H$ are the *differences in response* between the reference group of non-smoker mothers and the mothers in each of the respective smoking groups.
- It doesn't matter which category is chosen as the reference (as long as you get the correct interpretation).

Note: For two groups, this approach gives the same result as a two-sample *t* test with *equal* variances.

SAS Code:

```
ODS GRAPHICS ON;
PROC GLM DATA = BWT ORDER = internal PLOT = diagnostics;
    CLASS momsmoke;
    MODEL birthwt = momsmoke/solution;
RUN;
ODS GRAPHICS OFF;
```

Note: The PLOTS=DIAGNOSTICS option in the PROC GLM statement produces a box plot like the one on p. 2 of this document and a gallery of diagnostic plots with which to assess the assumptions we usually make when carrying out an ANOVA.

The GLM procedure results and diagnostic plot gallery for the birthweight and mother's smoking data are found on the next slides. As we review these, consider if the assumptions of the GLM/ANOVA model are met.

The GLM Procedure
Dependent Variable: birthwt

ANOVA Table:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	11.67268915	3.89089638	4.41	0.0137
Error	23	20.30360714	0.88276553		
Corrected Total	26	31.97629630			

R-Square	Coeff Var	Root MSE	birthwt Mean
0.365042	13.96148	0.939556	6.729630

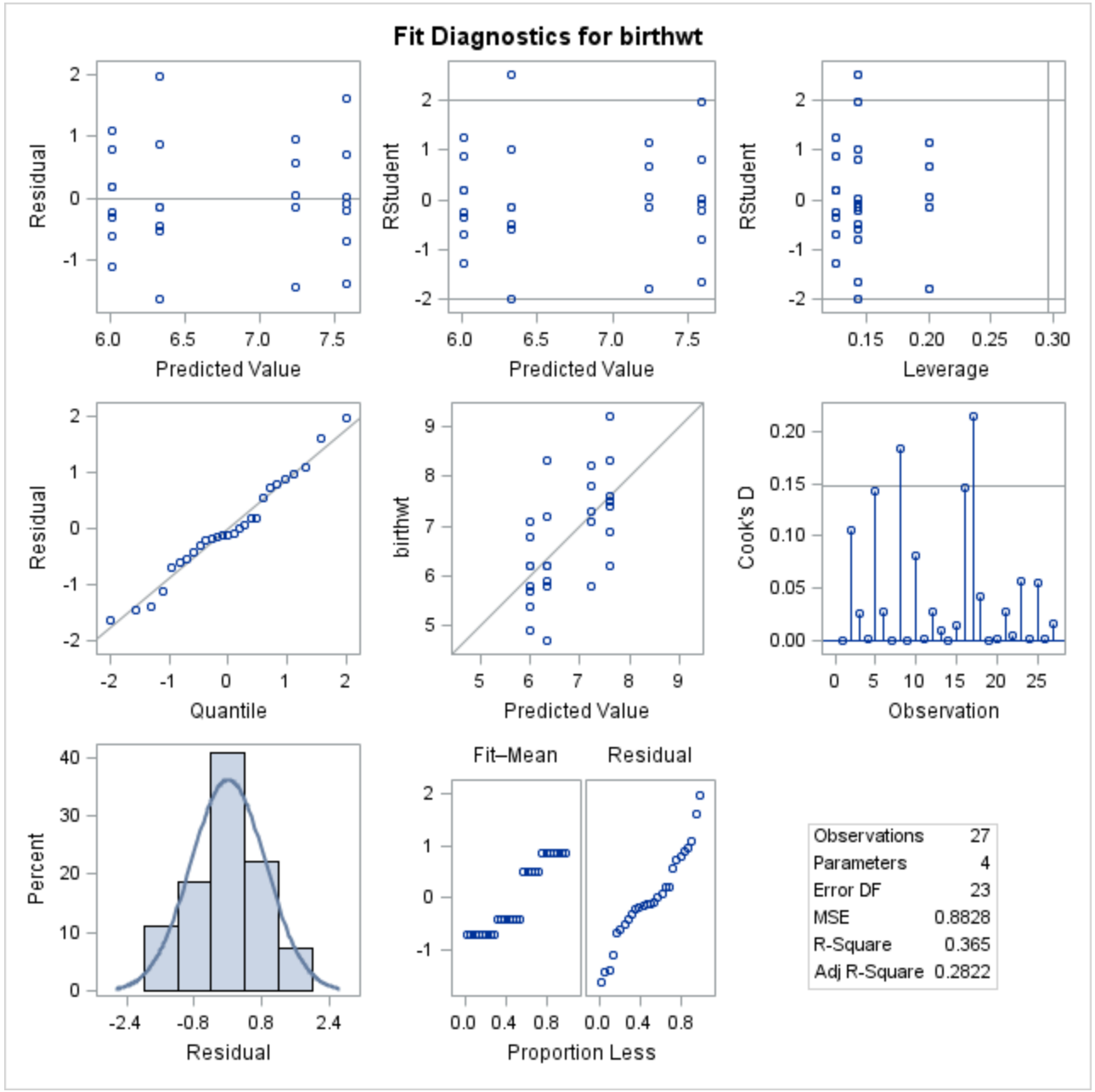
Source	DF	Type I SS	Mean Square	F Value	Pr > F
momsmoke	3	11.67268915	3.89089638	4.41	0.0137

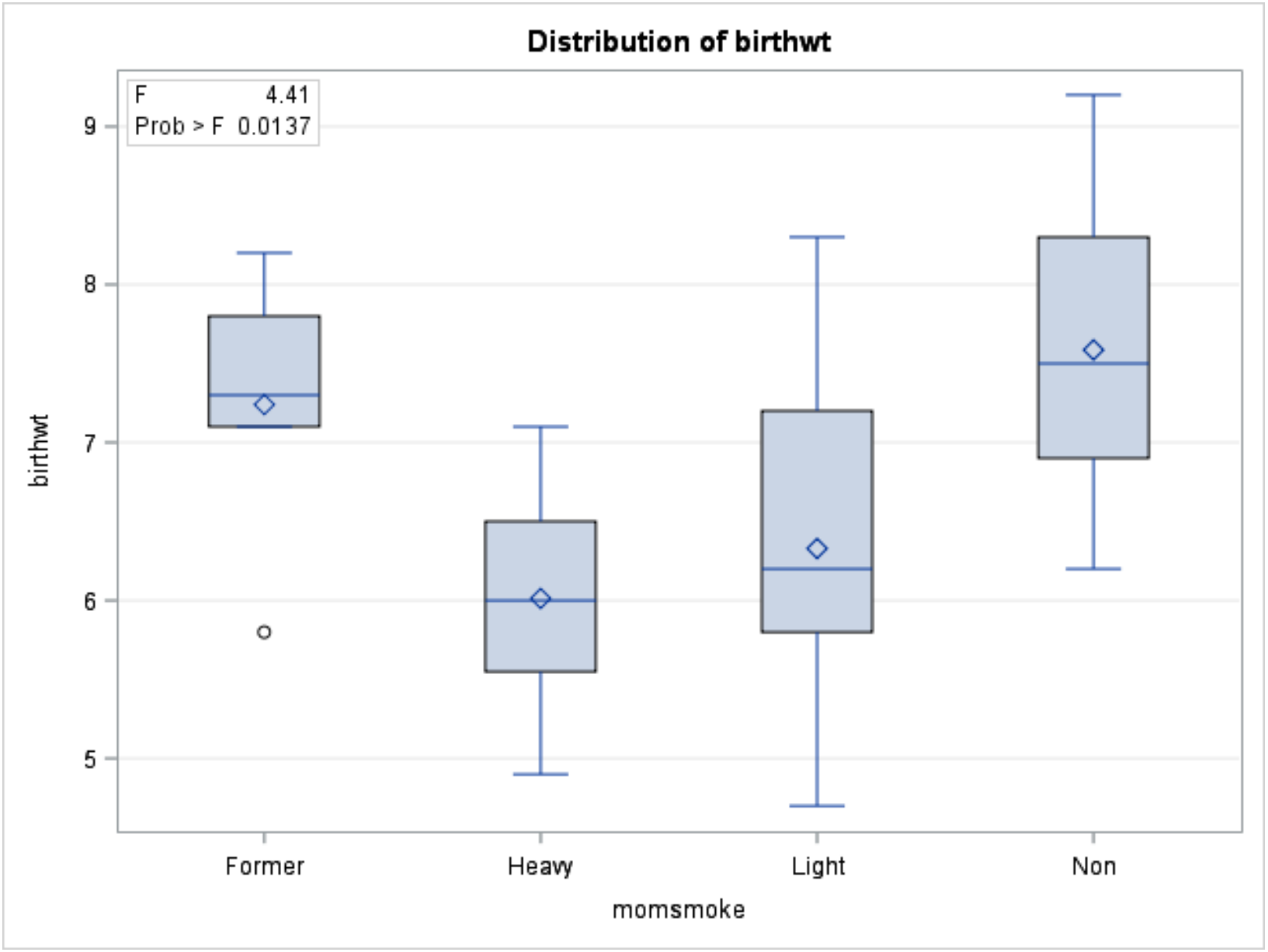
Source	DF	Type III SS	Mean Square	F Value	Pr > F
momsmoke	3	11.67268915	3.89089638	4.41	0.0137

GLM (Regression) Table:

Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	7.585714286	B	0.35511880	21.36	<.0001
momsmoke Former	-0.345714286	B	0.55014768	-0.63	0.5359
momsmoke Heavy	-1.573214286	B	0.48626644	-3.24	0.0037
momsmoke Light	-1.257142857	B	0.50221382	-2.50	0.0199
momsmoke Non	0.000000000	B	.	.	.


Fit Diagnostics for birthwt
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
birthwt
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 27
Parameters 4
Error DF 23
MSE 0.8828
R-Square 0.365
Adj R-Square 0.2822


Distribution of birthwt
F 4.41
Prob > F 0.0137
9
8
7
6
5
birthwt
Former
Heavy
Light
Non
momsmoke

C. Extending the One-Way ANOVA

ANOVA (one-way, two-way and multi-way) can be performed using linear (regression) models. Two-way and multi-way models can include *interactions* as well. (We'll define interactions (*effect modification*) when we talk about categorical data, and you'll learn a lot more about them in a few weeks.)

The *F*-test in the (one-way) ANOVA is a generalization of the two-sample equal variances t-test.

Important issue: Are the variances across the groups equal (i.e. homoscedasticity)?

We can test this assumption of equal variance formally with Bartlett's test or Levene's test for homogeneity of variance. For both, **H_0 is that the variances are equal across all groups**.

What should we do about the one-way ANOVA if we reject H_0?

B.L. Welch[1] proposed an approximate F-test (t-test) that can be implemented when the variances are not equal across the groups. The text from Welch's paper shows how to calculate the approximately F-distributed test statistic:

Our approximate test procedure will, therefore, be:

(i) *Calculate*

$$v^2 = \frac{\sum_t w_t(y_t - \hat{y})^2/(k-1)}{\left[1 + \frac{2(k-2)}{(k^2-1)}\sum_t \frac{1}{f_t}\left(1 - \frac{w_t}{\Sigma w_t}\right)^2\right]}, \qquad (29)$$

$$f_1 = (k-1); \quad f_2 = \left[\frac{3}{(k^2-1)}\sum_t \frac{1}{f_t}\left(1 - \frac{w_t}{\Sigma w_t}\right)^2\right]^{-1}. \qquad (30)$$

(ii) *Refer v^2 to a variance ratio table entered with degrees of freedom f_1 and f_2.*

Notation:

- y_t are the group means
- $\hat{y}$ is the grand mean
- k = the number of groups
- w_t are the group weights: $\frac{n_t}{s_t^2}$, for t = 1, 2, … k

For k = 2 the Welch's test is the version also referred to as the Satterthwaite approximation t-test or Satterthwaite t-test.

1: On the comparison of several mean values: An alternative approach. *Biometrika*, *38*, 330-336, 1951.

In SAS we can perform a Welch's ANOVA test using the following:

SAS Code:

```
PROC GLM DATA = BWT;
    CLASS momsmoke;
    MODEL birthwt = momsmoke;
    MEANS momsmoke/ hovtest=levene(type=abs) hovtest=bartlett
    WELCH;
RUN;
```

In addition to the usual ANOVA table, SAS then provides an ANOVA table based on the approximate Welch's F statistic (continued on next slide):

SAS Results:

Level of momsmoke	N	birthwt	
		Mean	Std Dev
Former	5	7.24000000	0.91268834
Heavy	8	6.01250000	0.71999504
Light	7	6.32857143	1.13975770
Non	7	7.58571429	0.96164542

SAS Results:

Levene's Test for Homogeneity of birthwt Variance ANOVA of Absolute Deviations from Group Means					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
momsmoke	3	0.2412	0.0804	0.23	0.8742
Error	23	8.0205	0.3487		

Bartlett's Test for Homogeneity of birthwt Variance			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
momsmoke	3	1.2656	0.7373

Welch's ANOVA for birthwt			
Source	**DF**	**F Value**	**Pr > F**
momsmoke	3.0000	4.64	0.0235
Error	11.5293		

Adopting a strategy similar to the one suggested by Moser and Stevens for the two-sample case, i.e. always applying the Welch's method for a one-way ANOVA when it isn't known that the variances are equal, is a good strategy. (See "Homogeneity of Variance in the Two-Sample Mean Test" by Moser and Stevens (1992) in the Paper Repository.)

For the smoking and birthweight data the conclusion does not change when we apply Welch's approximate F-test.

Other options if the variances are not equal across the groups:

- Normalizing using variance stabilizing transformations, e.g. log(y).
- Use the unequal variances t-test (in combination with corrections for multiple comparisons – more on this later in the lecture).

We will learn more about transformations later in the semester.

D. ANOVA: Post-Hoc Comparisons

Another issue in ANOVA, if we find that at least one group has a significantly different mean, which groups are different?

To determine which groups are different we could do all possible independent sample t-tests:

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}, \text{ where } s = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$$

For our smoking and birthweight data, this would result in p-values (from t.test in R) for our pairwise comparisons of:

p-values for pairwise comp.	**Former Smoker**	**Light Smoker**	**Heavy Smoker**
Non-Smoker	0.543	0.046	0.005
Former Smoker		0.156	0.038
Light Smoker			0.542

However, if the variances from all *k* groups are assumed to be equal, then a more accurate estimate of σ could be obtained by using all *k* groups. A single estimate of σ is also preferable to using a different estimate for each pair of groups considered.

The pooled estimate of the variance for one-way ANOVA:

$$s^2 = \frac{\sum_{j=1}^{J}(n_j - 1)s_j^2}{\sum_{j=1}^{J}(n_j - 1)}$$

The *t*-test can then be calculated using the pooled estimate of the variance from the one-way ANOVA:

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{N-J}$$

This test is often referred to as the *least significant difference* (LSD) method. Comparisons start with the most extreme means, and then work inward. The procedure stops when the first non-significant result is observed.

Example of LSD Multiple Comparisons for Smoking and Birthweight Data

SAS Code:

```
proc anova data=BWT;
    class momsmoke;
    model birthwt = momsmoke;
    means momsmoke / LSD;
run;
```

SAS Results:

The ANOVA Procedure

t Tests (LSD) for birthwt

Note: This test controls the Type I comparisonwise error rate, not the experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	23
Error Mean Square	0.882766
Critical Value of t	2.06866

LSD results on next slide…

SAS Results cont.:

LSD comparison table:

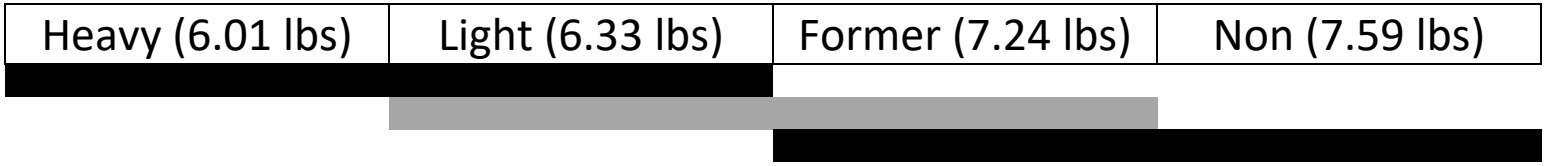
Comparisons significant at the 0.05 level are indicated by ***.				
momsmoke Comparison	Difference Between Means	95% Confidence Limits		
Non - Former	0.3457	-0.7924	1.4838	
Non - Light	1.2571	0.2182	2.2961	***
Non - Heavy	1.5732	0.5673	2.5791	***
Former - Non	-0.3457	-1.4838	0.7924	
Former - Light	0.9114	-0.2266	2.0495	
Former - Heavy	1.2275	0.1195	2.3355	***
Light - Non	-1.2571	-2.2961	-0.2182	***
Light - Former	-0.9114	-2.0495	0.2266	
Light - Heavy	0.3161	-0.6898	1.3220	
Heavy - Non	-1.5732	-2.5791	-0.5673	***
Heavy - Former	-1.2275	-2.3355	-0.1195	***
Heavy - Light	-0.3161	-1.3220	0.6898	

We can summarize this information as the pairwise p-values between smoking groups:
For our non vs. former using R's *pt* function for the probability of a t-distribution (or the pairwise.t.test function, see appendix):

$$p_{\text{(non-form)}} = 2 \times \left(1 - \text{pt}\left(\frac{\bar{Y}_{non} - \bar{Y}_{form}}{s\sqrt{\frac{1}{n_{non}} + \frac{1}{n_{form}}}}, \text{df} \right) \right) = 2 \times \left(1 - \text{pt}\left(\frac{0.3457}{\sqrt{0.882766}\sqrt{\frac{1}{7} + \frac{1}{5}}}, 23 \right) \right) = 0.536$$

p-values for pairwise comp.	**Former Smoker**	**Light Smoker**	**Heavy Smoker**
Non-Smoker	0.536	0.020	0.004
Former Smoker		0.111	0.031
Light Smoker			0.522

One method to convey the relationships between different groups is to order them from smallest to largest mean, and then draw "lines" underneath to connect ones that do not have significant p-values (indicating they are statistically similar):



Notice that light smokers are similar to both heavy and former smokers (p>0.05), ***but*** heavy and former smokers are **NOT** similar to each other (p=0.031)! Identifying which group(s) are different (or similar) can depend a lot on the perspective of which group you are considering.

E. Multiple Comparisons

Problem: When we perform multiple statistical tests, the true overall type I error rate (the experiment-wise error rate) is larger than the type I error rate for the individual tests.

The following table displays the probability of rejecting at least one of the pairwise comparisons using a significance level of 0.05:

# of groups	*# of pairwise comparisons*	*True significance level*
2	1	0.050
3	3	0.143
4	6	0.265
5	10	0.401

We see that as the number of groups, and therefore the total number of pairwise comparisons, increases we are more likely to have a true overall type I error rate (α) that is far greater than the individual test type I error rate.

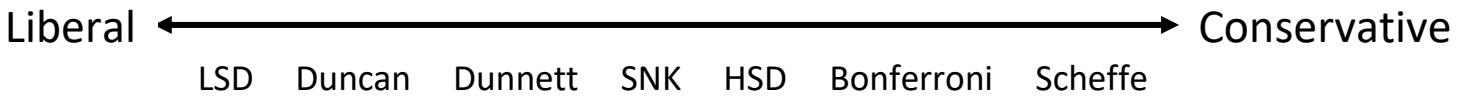
What are some ways we could avoid this issue?

Post-Hoc Comparisons

Many post-hoc comparison procedures have been developed to determine which means differ while adjusting the overall type I error rate. They can be chosen to accommodate the given situation and/or the philosophy of the experimenter (or journal).

- **Bonferroni Adjustment:** Can be used for any *C* independent comparisons. Essentially you conclude that the p-value is significant if it is less than 0.05/*C* instead of 0.05. This is conservative, especially if the tests are not independent.
- **Tukey's Honestly Significant Difference (HSD) Test**: Uses the studentized range distribution to make all pairwise comparisons.
- **Student Newman-Keul's (SNK) Test**: Alternative range test to Tukey's. It adjusts for the number of group means between the two groups being compared.
- **Duncan's Multiple Range Test:** Alternative range test that uses the harmonic mean of the sample size when the sample sizes are unequal.
- **Scheffé's Test**: Can be used for any contrast of interest (not just pairwise comparisons), but can be very conservative.
- **Dunnett's Test:** Used to compare several groups to a single control group; often used in clinical trials.

Comparison of Type I error rates for all pairwise comparisons:



Multiple Comparisons

- There is substantial controversy over the use of post-hoc multiple comparison procedures. Some researchers routinely use these procedures for all ANOVA models, while others never use them.

- Although post-hoc multiple comparison procedures control the overall type I error rates, they inflate type II errors (the probability of accepting the null hypothesis when the alternative is true). With appropriate software, multiple comparisons can be incorporated into sample size and power analyses at the design phase.

- If comparisons of interest are planned in advance (it also helps to limit comparisons to those with the most scientific relevance), then the LSD procedure is appropriate.

- If there are many comparisons to be made, and not all comparisons have been hypothesized in advance, then one of the multiple comparison procedures is appropriate.

SAS Code:

```
PROC GLM DATA = BWT ORDER = internal;
    CLASS momsmoke;
    MODEL birthwt = momsmoke/noint solution;
    MEANS momsmoke/ dunnett('Non') bon tukey;
RUN;
```

Dunnett's t Tests for birthwt

Note: This test controls the Type I experimentwise error for comparisons of all treatments against a control.

Alpha	0.05
Error Degrees of Freedom	23
Error Mean Square	0.882766
Critical Value of Dunnett's t	2.51781

Comparisons significant at the 0.05 level are indicated by *.**

momsmoke Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
Former - Non	-0.3457	-1.7309	1.0395	
Light - Non	-1.2571	-2.5216	0.0073	
Heavy - Non	-1.5732	-2.7975	-0.3489	***

Bonferroni (Dunn) t Tests for birthwt

Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than Tukey's for all pairwise comparison

Alpha	0.05
Error Degrees of Freedom	23
Error Mean Square	0.882766
Critical Value of t	2.88626

Comparisons significant at the 0.05 level are indicated by *.**				
momsmoke Comparison	**Difference Between Means**	**Simultaneous 95% Confidence Limits**		
Non - Former	0.3457	-1.2422	1.9336	
Non - Light	1.2571	-0.1924	2.7067	
Non - Heavy	1.5732	0.1697	2.9767	***
Former - Non	-0.3457	-1.9336	1.2422	
Former - Light	0.9114	-0.6764	2.4993	
Former - Heavy	1.2275	-0.3185	2.7735	
Light - Non	-1.2571	-2.7067	0.1924	
Light - Former	-0.9114	-2.4993	0.6764	
Light - Heavy	0.3161	-1.0874	1.7196	
Heavy - Non	-1.5732	-2.9767	-0.1697	***
Heavy - Former	-1.2275	-2.7735	0.3185	
Heavy - Light	-0.3161	-1.7196	1.0874	

Tukey's Studentized Range (HSD) Test for birthwt

Note: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	23
Error Mean Square	0.882766
Critical Value of Studentized Range	3.91345

Comparisons significant at the 0.05 level are indicated by *.**

momsmoke Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
Non - Former	0.3457	-1.1767	1.8681	
Non - Light	1.2571	-0.1326	2.6469	
Non - Heavy	1.5732	0.2276	2.9188	***
Former - Non	-0.3457	-1.8681	1.1767	
Former - Light	0.9114	-0.6110	2.4338	
Former - Heavy	1.2275	-0.2547	2.7097	
Light - Non	-1.2571	-2.6469	0.1326	
Light - Former	-0.9114	-2.4338	0.6110	
Light - Heavy	0.3161	-1.0295	1.6617	
Heavy - Non	-1.5732	-2.9188	-0.2276	***
Heavy - Former	-1.2275	-2.7097	0.2547	
Heavy - Light	-0.3161	-1.6617	1.0295	

Summary

Bonferroni	Heavy	Light	Former	Non
Tukey (HSD)	Heavy	Light	Former	Non
Dunnett	Heavy	Light	Former	Non

Note: Multiple comparisons tests exist for the case of unequal variances. They are available in SAS PROC MIXED, an advanced procedure you will learn about in BIOS 6612.

Multiple Testing (vs. Multiple Comparisons) and the False Discovery Rate

In some study settings several distinct hypothesis tests may be performed (e.g. genomic and proteomic experiments). The Bonferroni correction is sometimes used in these settings but can be overly conservative when 100s or 1000s of tests are performed.

One commonly employed, but not necessarily optimal, solution is to apply the False Discovery Rate (FDR) correction (latter part of section 12.4 in Rosner). The idea is to limit the number of falsely positive results to a reasonable level (e.g. 5% or 10%).

Letting *k* be the number of tests, we can calculate the FDR by following this algorithm:

1. Calculate the p-values for all comparisons/tests.
2. Rank the comparisons by p-values from smallest to largest.
3. Calculate $q = kp/rank$ for each test.
4. The FDR value for each test is the *minimum* of the *q* values for that test and all tests ranked higher than it.
5. The null hypothesis is rejected for all FDR values that are less than the pre-specified acceptable level (e.g. 5%, 10%).

See the example on Slide 39 for an example of applying this algorithm.

The FDR is still conservative when the tests are not independent.

Many other solutions to the multiple testing problem have been and are being developed.

Permutations and bootstrap sampling of the original data to find adjusted p-values are good improvements over the FDR method since they work with the inherent dependence of tests in the original data.

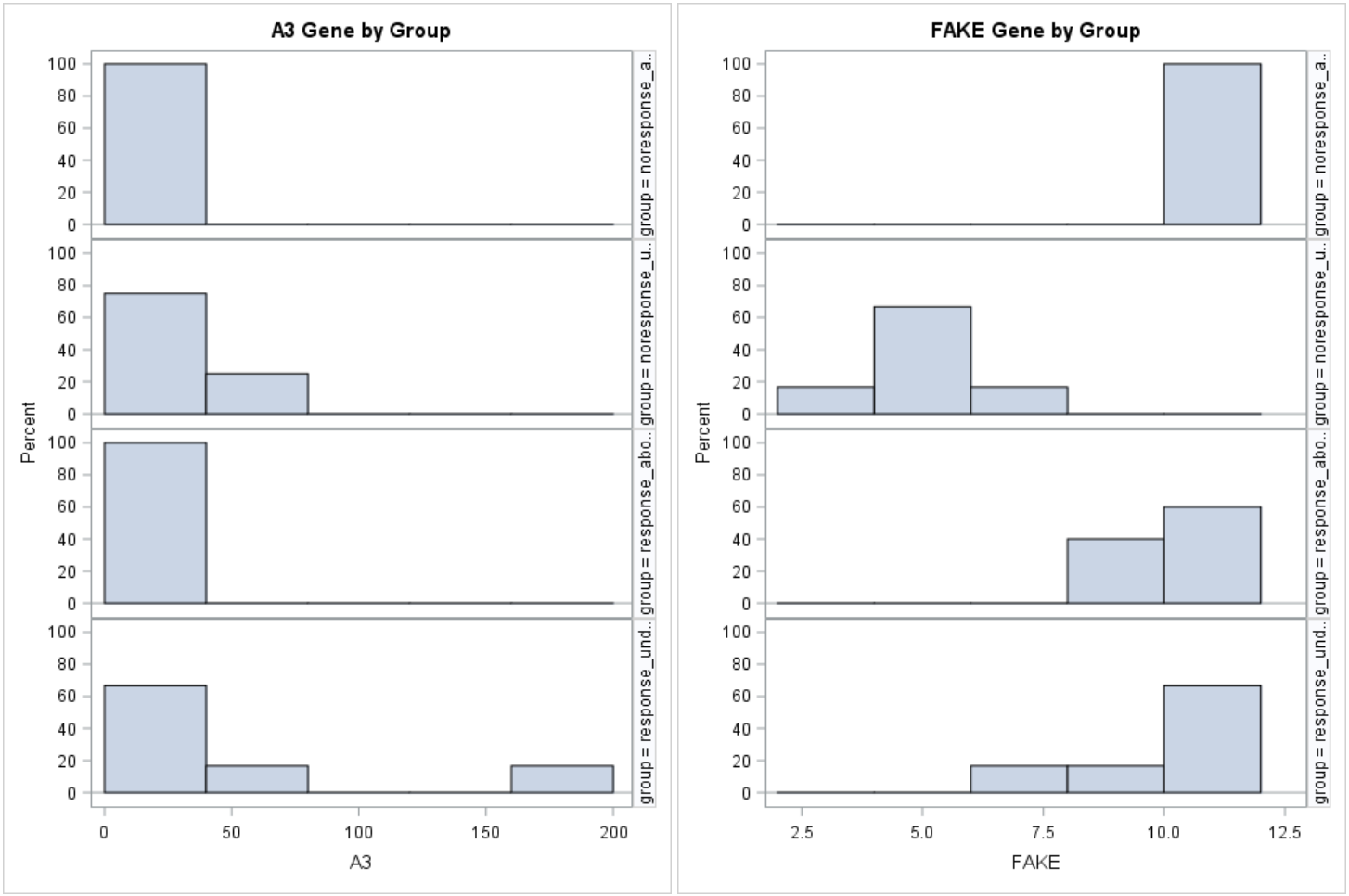
For more information see:

- Benjamini, Y., Hochberg, Y. (1995). Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing, *Journal of the Royal Statistical Society B,* 57: 289-300.

- Westfall, P. H., Young, S. S. (1993), *Resampling based multiple testing,* Wiley, New York.

- Westfall, P.H., Tobias, R.D., Rom, D., Wolfinger, R.D., Hochberg, Y. (1999). *Multiple Comparisons and Multiple Tests Using the SAS System*. Cary, NC: SAS Institute Inc.

Example: In a study of biomarkers for acute myeloid leukemia (AML), several genes in the HOX family were examined as possible predictors of response to therapy. In Lecture 9 we focused on one gene for illustration of the ROC curve. Here we will focus on all the genes and also create four groups based on response (yes/no) and age (60 or under, 61 or older).



SAS Code:

```
proc import datafile="~/HOXgenes.csv"
     out=hox /* name for data set for SAS to reference */
     dbms=csv /* identify file as csv */
     replace; /* overwrite hox if already present */
     getnames=yes; /* take first row as column names from data */
run;


/* Create histograms between groups for example */
title "A3 Gene by Group";
proc sgpanel data=hox;
    panelby group / rows=4 layout=rowlattice;
    histogram A3;
run;

title "FAKE Gene by Group";
proc sgpanel data=hox;
    panelby group / rows=4 layout=rowlattice;
    histogram FAKE;
run;
```

SAS Code:

```
/* Conduct Welch's ANOVA over all genes */
ods output welch=welchtab ; /* tell SAS to save the Welch's ANOVA table as an
    object we can manipulate */
PROC glm DATA = hox NOPRINT ; /* specify NOPRINT to save us from producing
    lots of tables/plots when we really only want Welch's ANOVA p-value */
    CLASS group;
    MODEL A3 A4 HOXA5 A7 HOXA9 A10 B3 B6 B9 MEIS1 MEIS2 PBX2 PBX3 FAKE =
        group; /* tell SAS we want to generate an ANOVA for each gene */
    MEANS group / welch; /* Specify that we want Welch's ANVOA */
RUN;
ods output close; /* tells SAS to stop writing to the function */


/* Extract p-values from Welch's ANOVAs using a data step */
data one;
    set welchtab;
    if Source="group" ;
    keep ProbF;
    rename ProbF=Raw_P; /* Rename ProbF match formatting for PROC MULTTEST */
run;


/* Apply FDR correction */
PROC MULTTEST INPVALUES=one FDR;
RUN;
```

The Multtest Procedure

P-Value Adjustment Information	
P-Value Adjustment	False Discovery Rate

p-Values			
Test	**Raw**	**False Discovery Rate**	**Gene**
1	0.4883	0.8950	A3
2	0.3169	0.8950	A4
3	0.4156	0.8950	HOXA5
4	0.2971	0.8950	A7
5	0.6393	0.8950	HOXA9
6	0.5606	0.8950	A10
7	0.5842	0.8950	B3
8	0.9442	0.9901	B6
9	0.4741	0.8950	B9
10	0.7937	0.9259	MEIS1
11	0.2450	0.8950	MEIS2
12	0.7554	0.9259	PBX2
13	0.9901	0.9901	PBX3
14	<.0001	0.0001	FAKE

Some manipulation in Excel based on ordered Raw p-values from smallest to largest:

Test	**Raw p-Value**	**FDR**	**Gene**	**Rank**	**q (= kp/ Rank)**	**FDR: MIN(q for rank or higher)**
14	0.0001	0.0001	FAKE	**1**	0.0014	0.0014
11	0.245	0.895	MEIS2	**2**	1.715	0.895
4	0.2971	0.895	A7	**3**	1.38647	0.895
2	0.3169	0.895	A4	**4**	1.10915	0.895
3	0.4156	0.895	HOXA5	**5**	1.16368	0.895
9	0.4741	0.895	B9	**6**	1.10623	0.895
1	0.4883	0.895	A3	**7**	0.9766	0.895
6	0.5606	0.895	A10	**8**	0.98105	0.895
7	0.5842	0.895	B3	**9**	0.90876	0.895
5	0.6393	0.895	HOXA9	**10**	0.89502	0.895
12	0.7554	0.9259	PBX2	**11**	0.96142	0.926
10	0.7937	0.9259	MEIS1	**12**	0.92598	0.926
8	0.9442	0.9901	B6	**13**	1.01683	0.9901
13	0.9901	0.9901	PBX3	**14**	0.9901	0.9901

F. Nonparametric ANOVA – the Kruskal-Wallis Test

The Kruskal-Wallis method is a multiple group extension of the Wilcoxon rank sum test. Dunn's method is a Bonferroni adjustment-like nonparametric multiple comparisons method (see SAS macro file "dunn macro.sas" on the Canvas site in the Code Repository).

SAS Code:

```
Calculate Kruskal-Wallis nonparametric ANOVA and Dunn's posthoc
macro;
PROC NPAR1WAY DATA = BWT WILCOXON ANOVA;
    CLASS momsmoke;
    VAR birthwt;
RUN;


FILENAME DUNN '~/dunn macro.sas';
%INCLUDE DUNN;

%DUNN(BWT,momsmoke,birthwt,0.05);
RUN;
```

The NPAR1WAY Procedure

Analysis of Variance for Variable birthwt Classified by Variable momsmoke		
momsmoke	**N**	**Mean**
Non	7	7.585714
Former	5	7.240000
Light	7	6.328571
Heavy	8	6.012500

One-Way ANOVA Table (same as Slide 13):

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Among	3	11.672689	3.890896	4.4076	0.0137
Within	23	20.303607	0.882766		
Average scores were used for ties.					

Wilcoxon Scores (Rank Sums) for Variable birthwt Classified by Variable momsmoke					
momsmoke	**N**	**Sum of Scores**	**Expected Under H0**	**Std Dev Under H0**	**Mean Score**
Non	7	141.50	98.0	18.002057	20.214286
Former	5	88.50	70.0	15.957118	17.700000
Light	7	80.50	98.0	18.002057	11.500000
Heavy	8	67.50	112.0	18.757714	8.437500
Average scores were used for ties.					

Kruskal-Wallis Test	
Chi-Square	10.0808
DF	3
Pr > Chi-Square	0.0179

Kruskal-Wallis Nonparametric ANOVA

Conclusion:

Similar post-hoc comparison test results with Dunn's method:

For Variable = birthwt

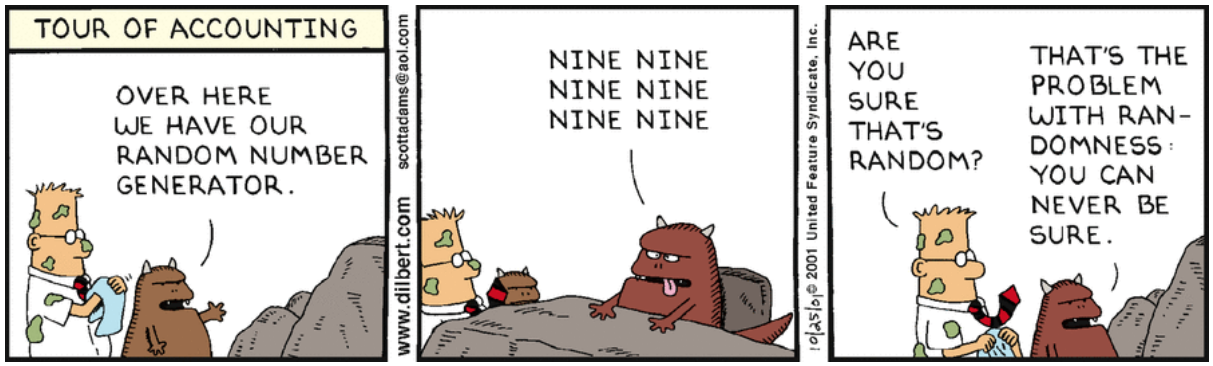
Large sample approximation multiple comparison procedure

designed for unbalanced data

4 groups: Former Heavy Light Non (respective sample sizes: 5 8 7 7)

Alpha = 0.05

Comparison number	Group comparisons	Difference in average ranks	Cutoff at alpha=0.05	Significance difference = **
1	Former-Heavy	9.2625	11.9379	
2	Former-Light	6.2000	12.2615	
3	Former-Non	2.5143	12.2615	
4	Heavy-Light	3.0625	10.8377	
5	Heavy-Non	11.7768	10.8377	**
6	Light-Non	8.7143	11.1932	


TOUR OF ACCOUNTING
OVER HERE WE HAVE OUR RANDOM NUMBER GENERATOR.
www.dilbert.com scottadams@aol.com
NINE NINE NINE NINE NINE NINE
ARE YOU SURE THAT'S RANDOM?
THAT'S THE PROBLEM WITH RAN-DOMNESS: YOU CAN NEVER BE SURE.
© 2001 United Feature Syndicate, Inc.

R Code Appendix

```
#############################################################
### Companion R Code for Lectures 14-15: One-Way ANOVA and Multiple Testing
### Also see Companion SAS Code for examples of implementation in SAS
#############################################################

### Load libraries
library(car) #for ANOVA diagnostic plots, Levene's test
library(MASS)
library(DescTools) #for some types of post-hoc testing
library(ggplot2)

### Read in the two data sets we will be using for examples in Lectures 14/15

## Birth weight and mother's smoking status
BWT <- read.csv("H://Teaching/BIOS 6611/Fall 2018/Lectures/Figure and Example
Code for Lectures/birthweight_smoking_dataset.csv", header=T)

# Tell R the "ordering" for mother's smoking status
BWT$momsmoke <- factor(BWT$momsmoke, c('Non','Former','Light','Heavy'))

## HOX genes csv file
hox <- read.csv("H://Teaching/BIOS 6611/Fall 2018/Lectures/Figure and Example
Code for Lectures/HOXgenes.csv", header=T)
```

```
#############################################################
### Birth weight and smoking related analyses/figures
#############################################################

### Create boxplot of birth weights by group (slides 2 and 16)
boxplot( birthwt ~ momsmoke, data=BWT, xlab='Group', ylab='Birth Weight
(pounds)')


### Based on our data, determine 95th percentile for the F-dist. using qf
function (slide 8)
qf(p=0.95, df1=3, df2=23)


### One-Way ANOVA (slide 10)
aov.pg10 <- aov( birthwt ~ momsmoke , data=BWT)
summary(aov.pg10) #notice that this summary ANOVA table does not include the
"Total" row ("Corrected Total" in SAS-terms)


### One-Way ANOVA as GLM (slides 13-16)
smoke.fit <- lm( birthwt ~ momsmoke, data=BWT )

## Page 14, ANOVA table and GLM regression table
anova(smoke.fit) #the ANOVA table
summary(smoke.fit) #the regression coefficient estimates

## Page 15, diagnostic-type plots to try and represent what SAS provides
# QQ plot to assess normality
qqPlot(smoke.fit)
```

```
# Cook's distance plot
cutoff <- 4/((nrow(BWT) - length(smoke.fit$coefficients) - 2))
plot(smoke.fit, which=4, cook.levels=cutoff)

# Studentized residuals plot, similar to lower-left "Residual" plot in SAS output
on slide 15 of lecture notes
sresid <- studres(smoke.fit)
hist(sresid, freq = FALSE, main = "Distribution of Studentized Residuals")
xfit <- seq(min(sresid), max(sresid), length = 40)
yfit <- dnorm(xfit)
lines(xfit, yfit)

# Spread-Level Plot, similar to top row's first two plots in SAS output on slide
15
spreadLevelPlot(smoke.fit)


### Welch's ANOVA and testing equality of variances (slides 19-20)

leveneTest( birthwt ~ momsmoke, data=BWT, center=mean) #center=mean chosen to
match SAS output, leveneTest documentation notes using center=median is more
robust

bartlett.test( birthwt ~ momsmoke, data=BWT)

oneway.test( birthwt ~ momsmoke, data=BWT, var.equal=FALSE) #Welch's ANOVA
```

```
### Post-hoc testing

## LSD-related functions
PostHocTest( aov.pg10, method=c('lsd') ) #Least Significant Differences (slides
24-25)
pairwise.t.test( x = BWT$birthwt, g = BWT$momsmoke, p.adjust.method='none') #LSD
p-value comparison table on slide 26

## Dunnett test function
DunnettTest( x = BWT$birthwt, g = BWT$momsmoke ) #Dunnett's test (slide 30)

## Bonferroni correction functions
PostHocTest( aov.pg10, method=c('bonferroni') ) #Bonferroni correction (slide 31)
pairwise.t.test( x = BWT$birthwt, g = BWT$momsmoke, p.adjust.method='bonferroni')
#alternatively you can use pairwise.t.test and specify the correction

## Tukey HSD functions
PostHocTest( aov.pg10, method=c('hsd') ) #Tukey's HSD (slide 32)
TukeyHSD( aov.pg10 ) #alternatively you can use TukeyHSD function in the 'stat'
package which is automatically loaded with R


### Kruskal-Wallis non-parametric ANOVA (slide 41)

kruskal.test( birthwt ~ momsmoke, data=BWT) #compared to SAS< it only produces
the Kruskal-Wallis nonparametric ANVOA results (i.e. no parametric ANOVA table or
Wilcoxon scores table)

## Dunn's post-hoc comparison testing (slide 42)

DunnTest(birthwt ~ momsmoke, data=BWT)
```

```
#############################################################
### HOX Genes and False Discovery Rate Example (slides 36-69)
#############################################################

## Histograms by grouping of age/response for A3 and FAKE genes

ggplot(hox, aes(x=A3)) + geom_histogram(breaks=seq(0,200,length=6)) +
facet_grid(group~.)
ggplot(hox, aes(x=FAKE)) + geom_histogram(breaks=seq(2,12,length=6)) +
facet_grid(group~.)

## Calculate p-value for each gene based on Welch's one-way ANOVA

p.vec <- NULL #initialize object to append p-values to
gene.vec <-
c('A3','A4','HOXA5','A7','HOXA9','A10','B3','B6','B9','MEIS1','MEIS2','PBX2','PBX
3','FAKE') #vector of genes to test

for(i in gene.vec){
    waov <- oneway.test( as.formula( paste0( i, ' ~ group')), data=hox,
var.equal=FALSE) #Welch's ANOVA
    p.vec <- c(p.vec, waov$p.value)
}

p.adjust( p.vec, method='fdr' ) #FDR adjustment for p-values

cbind( gene.vec, round(p.adjust(p.vec, method='fdr' ),4) ) #matrix showing genes
and FDR p-values
```