STAT 3119

Week2: 9/5/2019 @GWU

Today's class

Blackboard is now available and will be updated each week for homework assignment and submission.

Class websites: https://github.com/npmldabook/Stat3119

Email: xtian@gwu.edu

TA: Xiang Shen (xiangshen@gwu.edu)

Outline Week 2 (Chapter 16)

In the following 14 weeks, we will start to go over different types of designs, model and analysis of the data.

- A1. Design of single factor experimental studies
- A2. Single factor ANOVA model, estimation and testing
- A3. Alternative formulation of the ANOVA model
- A4. Data analysis example with R

  - B1. Relationship: ANOVA vs. Regression (Ch 16.8)
  - B2. Power and sampple size analysis
  - B3. Walkthrough of R examples.
  - B4. Questions for Chapter 16 and homework assignments.
  - Chapter 16.9 and Chapter 17 next Tuesday.

One factor ANOVA

We studies in the last lesson about the one factor (one-way) ANOVA model (**Cell means model**) :

$$Y_{ij} = \mu_i + \epsilon_{ij} \qquad (16.2)$$

And **the factor effects** model:

$$Y_{ij} = \mu_. + \tau_i + \epsilon_{ij} \qquad (16.62)$$

- For estimation: we went over how to estimate the model parameters such as μ_i (factor mean response), $\mu_.$ (overall mean response), τ_i (deviation from factor mean) based on the observed data (Y_{ij}).

- For testing, we construct a F-test to test for equality of the factor levels means.
- We illustrate how to perform an ANOVA analysis with R.

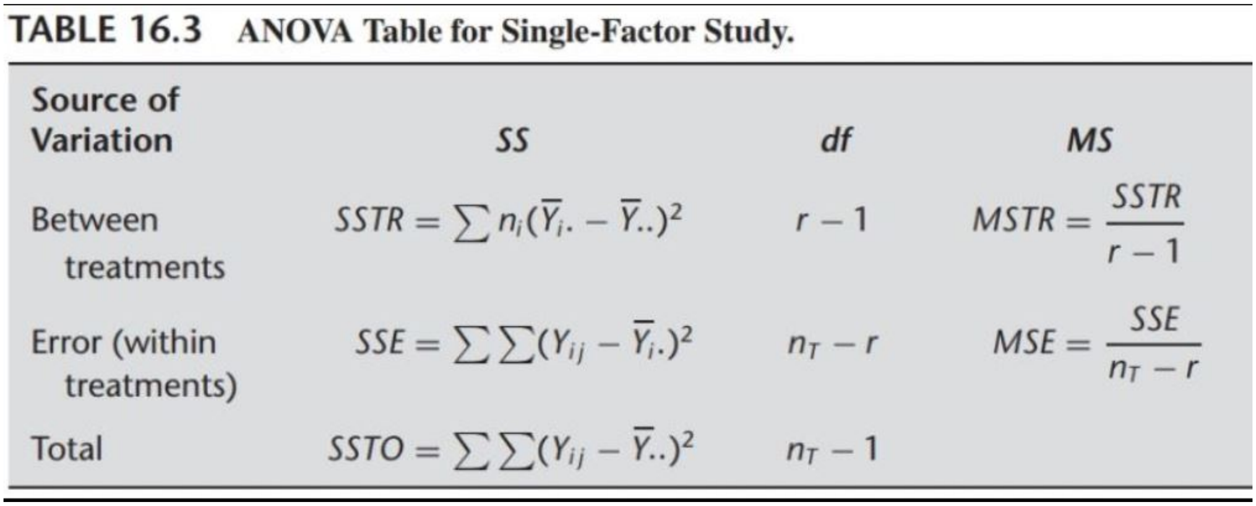


TABLE 16.3 ANOVA Table for Single-Factor Study.

Source of Variation	SS	df	MS
Between treatments	$SSTR = \sum n_i(\overline{Y}_i. - \overline{Y}..)^2$	$r - 1$	$MSTR = \frac{SSTR}{r-1}$
Error (within treatments)	$SSE = \sum\sum(Y_{ij} - \overline{Y}_i.)^2$	$n_T - r$	$MSE = \frac{SSE}{n_T - r}$
Total	$SSTO = \sum\sum(Y_{ij} - \overline{Y}..)^2$	$n_T - 1$	

The test statistic

$$\text{Test statistic}\, F^* = MSTR/MSE$$

When H_0 holds. F^* is distributed as $F(r-1, n_T - r)$ where r= number of factor levels, n_T the totla sample size.

Regression Approach to ANOVA (ch 16.1, 16.8)

Both the means model and the effects model are **linear statistical models**; that is, the response variable Y_{ij} is a linear function of the model **parameters** (Here: the 'linear' is not relative to the X).

- When indicator variables are used in regression models, the regression results will be identical to those obtained with analysis of variance models.
- There is no fundamental choice between regression and analysis of variance models when the predictor variables are qualitative, because the models are equivalent, leading to the same findings and conclusion.
- The textbook shows the regression approach to single-factor analysis of variance for three alternative models: (1) the factor effects model with unweighted mean, (2) the factor effects model with weighted mean, and (3) the cell means model. You will find the ANOVA model setup is easier than the regression approaches.

It is important to emphasize that the choice of model affects the definition of the model parameters, i.e. how to set the **design matrix** $X = (X_1, X_2, ..)$ in the regression model to obtain the coefficient vector $\beta = (X'X)^{-1}X'Y$, but not affecting the results of test for equality of the factor means.

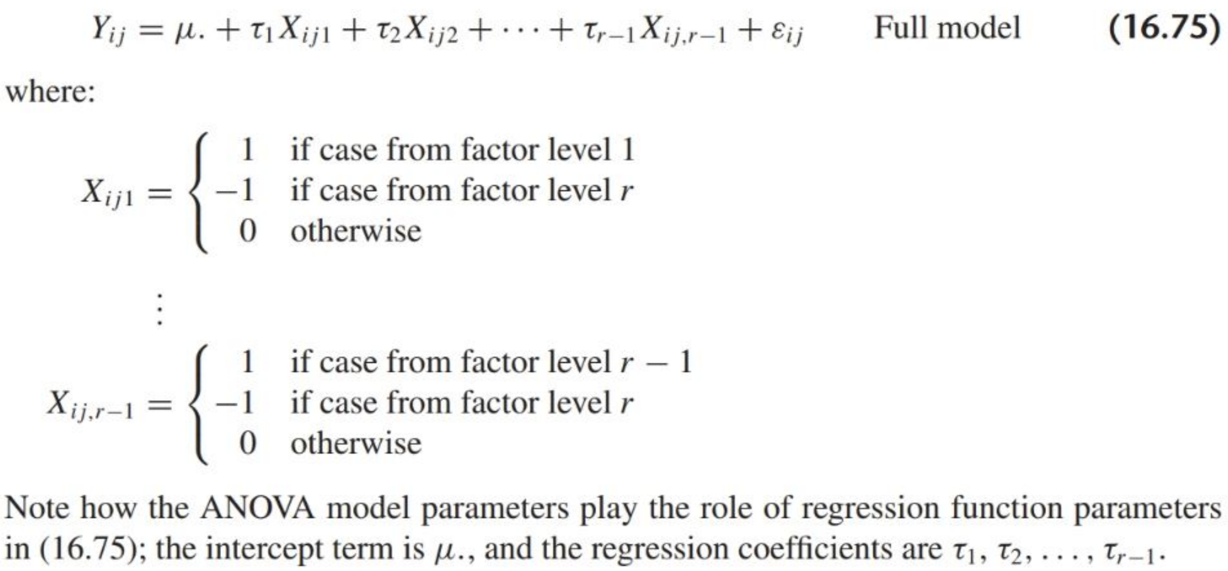
A. Regression for factor models with unweighted mean (Ch 16.8, p.705)

In this case, the ANOVA model is $Y_{ij} = \mu_. + \tau_i + \epsilon_{ij}$, where the overall mean $\mu_. = \sum \mu_i / r$, which leads to the constraints in $\tau_i = \mu_i - \mu_.$,

$$\sum_{i=1}^{r} \tau_i = 0$$

Then we show that the best estimators for μ_i is $\hat{\mu}_i = \overline{Y_{i.}}$ (sample mean in factor level i), and $\hat{\mu}_. = \sum_i \overline{Y_{i.}}/r$, and $\hat{\tau}_i = \hat{\mu}_i - \hat{\mu}_.$.

Now we can have an equivalent model using regression:



$$Y_{ij} = \mu. + \tau_1 X_{ij1} + \tau_2 X_{ij2} + \cdots + \tau_{r-1} X_{ij,r-1} + \varepsilon_{ij} \qquad \text{Full model} \qquad \textbf{(16.75)}$$

where:

$$X_{ij1} = \begin{cases} 1 & \text{if case from factor level 1} \\ -1 & \text{if case from factor level } r \\ 0 & \text{otherwise} \end{cases}$$

$$\vdots$$

$$X_{ij,r-1} = \begin{cases} 1 & \text{if case from factor level } r-1 \\ -1 & \text{if case from factor level } r \\ 0 & \text{otherwise} \end{cases}$$

Note how the ANOVA model parameters play the role of regression function parameters in (16.75); the intercept term is $\mu.$, and the regression coefficients are $\tau_1, \tau_2, \ldots, \tau_{r-1}$.

Example for fitting the regression model (R):

- read the data as in the last class:

```
Ex16 <- read.table("C:/Users/tianx/Documents/GitHub/Stat3119/week2/CH16_TA01.txt", header=F) # find the
#Ex16 <- read.table(url("https://raw.githubusercontent.com/npmldabook/Stat3119/master/Week2/CH16_TA01.t
names(Ex16)<- c("sales", "package", "stores")
Ex16$package<- as.factor(Ex16$package)
```

- Define the indicator variables for factor $1, ..., r-1$.

```
Factor1<-  (Ex16$package==1)*1+(Ex16$package==4)*(-1)
Factor2<-  (Ex16$package==2)*1+(Ex16$package==4)*(-1)
Factor3<-  (Ex16$package==3)*1+(Ex16$package==4)*(-1)
(DesignX <- cbind(1, Factor1, Factor2, Factor3))
```

```
##         Factor1 Factor2 Factor3
##  [1,] 1       1       0       0
##  [2,] 1       1       0       0
##  [3,] 1       1       0       0
##  [4,] 1       1       0       0
##  [5,] 1       1       0       0
##  [6,] 1       0       1       0
##  [7,] 1       0       1       0
##  [8,] 1       0       1       0
##  [9,] 1       0       1       0
## [10,] 1       0       1       0
## [11,] 1       0       0       1
## [12,] 1       0       0       1
## [13,] 1       0       0       1
## [14,] 1       0       0       1
## [15,] 1      -1      -1      -1
## [16,] 1      -1      -1      -1
## [17,] 1      -1      -1      -1
## [18,] 1      -1      -1      -1
## [19,] 1      -1      -1      -1
```

- use **lm()** in R to a linear regresion model

```
LMfit<-lm( sales~ Factor1+Factor2+ Factor3, data=Ex16 )
summary(LMfit)
```

```
##
## Call:
## lm(formula = sales ~ Factor1 + Factor2 + Factor3, data = Ex16)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5.20  -1.95  -0.20   1.50   5.80
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.6750     0.7485  24.949 1.25e-13 ***
## Factor1      -4.0750     1.2708  -3.207 0.005884 **
## Factor2      -5.2750     1.2708  -4.151 0.000854 ***
## Factor3       0.8250     1.3706   0.602 0.556221
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.248 on 15 degrees of freedom
## Multiple R-squared:  0.7881, Adjusted R-squared:  0.7457
## F-statistic: 18.59 on 3 and 15 DF,  p-value: 2.585e-05
```

So that the fitted regression model would be

$$\hat{Y} = 18.675 - 4.075X_1 - 5.275X_2 + 0.825X_3 \,.$$

Note: As in Table 16.4, the coefficents are exactly the same as the estimators $\hat{\mu}_.$ and $\hat{\tau}_i, i = 1, 2, 3$ as in the ANOVA fit. And the final $\hat{\tau}_4$ is the negative sum of the $\tau_i, i = 1, 2, 3$.

```
(tau4<- -sum(LMfit$coef[2:4]))
```

```
## [1] 8.525
```

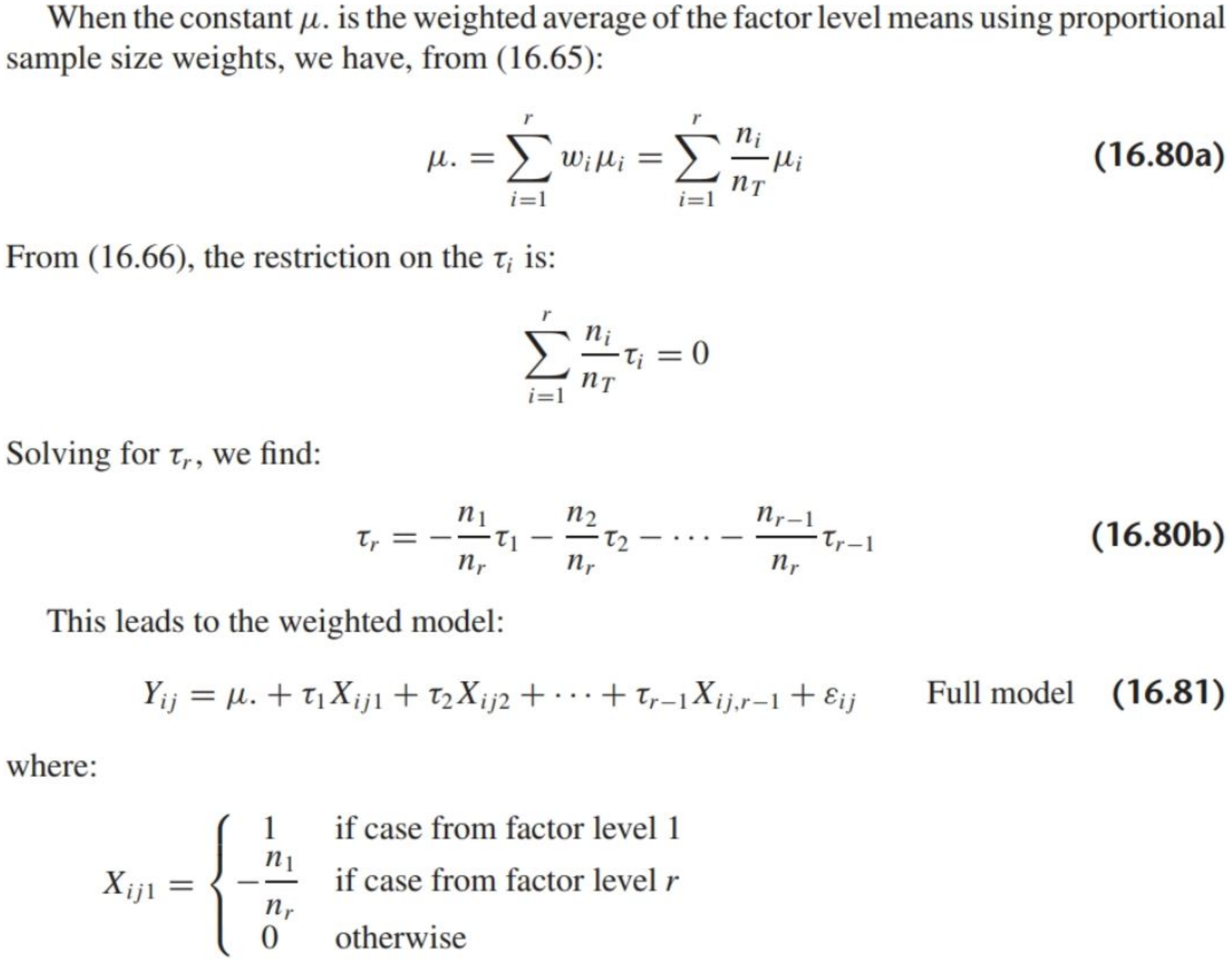
B. Regression for factor models with weighted mean (ch 16.8, p.708)

In this case, the ANOVA model is $Y_{ij} = \mu_. + \tau_i + \epsilon_{ij}$, where the overall mean $\mu_. = \sum w_i \mu_i$, which leads to the constraints in $\tau_i = \mu_i - \mu_.$,

$$\sum_{i=1}^{r} w_i \tau_i = 0$$

Then since the best estimators for μ_i is $\hat{\mu}_i = \overline{Y_{i.}}$ (sample mean in factor level i), and $\hat{\mu}_. = \sum_i w_i \overline{Y_{i.}}$, and $\hat{\tau}_i = \hat{\mu}_i - \hat{\mu}_.$.

Now we can have an equivalent model using regression for the case when the weights $w_i = n_i/n_T$:



When the constant $\mu.$ is the weighted average of the factor level means using proportional sample size weights, we have, from (16.65):

$$\mu. = \sum_{i=1}^{r} w_i \mu_i = \sum_{i=1}^{r} \frac{n_i}{n_T} \mu_i \tag{16.80a}$$

From (16.66), the restriction on the τ_i is:

$$\sum_{i=1}^{r} \frac{n_i}{n_T} \tau_i = 0$$

Solving for τ_r, we find:

$$\tau_r = -\frac{n_1}{n_r}\tau_1 - \frac{n_2}{n_r}\tau_2 - \cdots - \frac{n_{r-1}}{n_r}\tau_{r-1} \tag{16.80b}$$

This leads to the weighted model:

$$Y_{ij} = \mu. + \tau_1 X_{ij1} + \tau_2 X_{ij2} + \cdots + \tau_{r-1} X_{ij,r-1} + \varepsilon_{ij} \qquad \text{Full model} \tag{16.81}$$

where:

$$X_{ij1} = \begin{cases} 1 & \text{if case from factor level 1} \\ -\dfrac{n_1}{n_r} & \text{if case from factor level } r \\ 0 & \text{otherwise} \end{cases}$$

Example for fitting the regression model (R):

- Define weights in (16.80b) (note: they are not the same as the weights in 16.80a):

```
(Ni <- table(Ex16$package))
```

```
##
## 1 2 3 4
## 5 5 4 5
```

```
(weights<- Ni[1:3]/Ni[4])
```

```
##
##   1   2   3
## 1.0 1.0 0.8
```

- Define the indicator variables for factor $1, ..., r-1$. Note the difference in the matrix for factor $r = n_i/n_r$ may not be 1.

```
Factor1<-  (Ex16$package==1)*1+(Ex16$package==4)*(-weights[1])
Factor2<-  (Ex16$package==2)*1+(Ex16$package==4)*(-weights[2])
Factor3<-  (Ex16$package==3)*1+(Ex16$package==4)*(-weights[3])
(DesignX <- cbind(1, Factor1, Factor2, Factor3))
```

```
##         Factor1 Factor2 Factor3
##  [1,] 1       1       0     0.0
##  [2,] 1       1       0     0.0
##  [3,] 1       1       0     0.0
##  [4,] 1       1       0     0.0
##  [5,] 1       1       0     0.0
##  [6,] 1       0       1     0.0
##  [7,] 1       0       1     0.0
##  [8,] 1       0       1     0.0
##  [9,] 1       0       1     0.0
## [10,] 1       0       1     0.0
## [11,] 1       0       0     1.0
## [12,] 1       0       0     1.0
## [13,] 1       0       0     1.0
## [14,] 1       0       0     1.0
## [15,] 1      -1      -1    -0.8
## [16,] 1      -1      -1    -0.8
## [17,] 1      -1      -1    -0.8
## [18,] 1      -1      -1    -0.8
## [19,] 1      -1      -1    -0.8
```

- Then we fit the regression model and use formula (16.80b) to get $\hat{\tau}_4$ (same as results in p.710).

```
LMfit<-lm( sales~ Factor1+Factor2+ Factor3, data=Ex16 )
tau4<-  -sum(weights*LMfit$coefficients[2:4])
cat("The regression coefficent is \n")
```

```
## The regression coefficent is
```

```
LMfit$coefficients
```

```
## (Intercept)     Factor1     Factor2     Factor3
##  18.6315789  -4.0315789  -5.2315789   0.8684211
```

```
cat("\n The estimate of tau4 is ", tau4 )
```

```
##
##  The estimate of tau4 is  8.568421
```

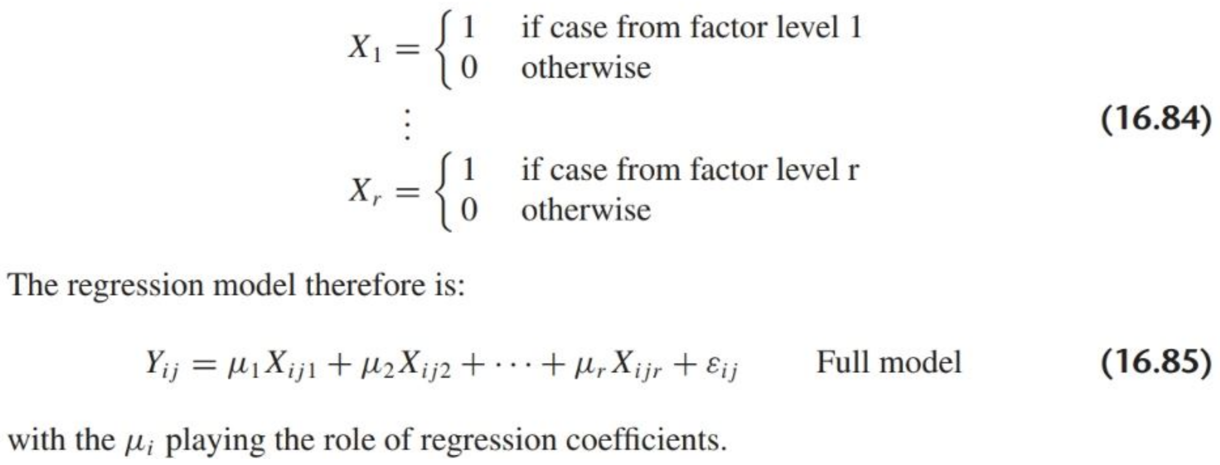
So that the fitted regression model equivalent to this ANOVA model would be

$$\hat{Y} = 18.63 - 4.03X_1 - 5.23X_2 + 0.87X_3\,.$$

C. Regression for factor models with cell means model (Ch 16.8, p.710)

In this case, the ANOVA model is $Y_{ij} = \mu_i + \epsilon_{ij}$, then the best estimators for μ_i is $\hat{\mu_i} = \overline{Y_{i.}}$ (sample mean in factor level i).

Now we can have an equivalent model using regression:



$$X_1 = \begin{cases} 1 & \text{if case from factor level 1} \\ 0 & \text{otherwise} \end{cases}$$
$$\vdots \tag{16.84}$$
$$X_r = \begin{cases} 1 & \text{if case from factor level r} \\ 0 & \text{otherwise} \end{cases}$$

The regression model therefore is:

$$Y_{ij} = \mu_1 X_{ij1} + \mu_2 X_{ij2} + \cdots + \mu_r X_{ijr} + \varepsilon_{ij} \qquad \text{Full model} \tag{16.85}$$

with the μ_i playing the role of regression coefficients.

Example for fitting the regression model (R):

1. We use indicator vector to indicate the factor levels

```
Factor1<-  (Ex16$package==1)*1
Factor2<-  (Ex16$package==2)*1
Factor3<-  (Ex16$package==3)*1
Factor4<-  (Ex16$package==4)*1
(DesignX <- cbind( Factor1, Factor2, Factor3, Factor4))
```

```
##       Factor1 Factor2 Factor3 Factor4
##  [1,]       1       0       0       0
##  [2,]       1       0       0       0
##  [3,]       1       0       0       0
##  [4,]       1       0       0       0
##  [5,]       1       0       0       0
##  [6,]       0       1       0       0
##  [7,]       0       1       0       0
##  [8,]       0       1       0       0
##  [9,]       0       1       0       0
## [10,]       0       1       0       0
## [11,]       0       0       1       0
## [12,]       0       0       1       0
## [13,]       0       0       1       0
## [14,]       0       0       1       0
## [15,]       0       0       0       1
## [16,]       0       0       0       1
## [17,]       0       0       0       1
## [18,]       0       0       0       1
## [19,]       0       0       0       1
```

2. You can use the matrix operations as what you have learned in the regression class (Chapter 6.3, not in the e-book version): $\beta = (X'X)^{-1}X'Y$.

```
beta <-  solve(t(DesignX) %*% DesignX) %*% t(DesignX) %*% Ex16$sales
t(beta)
```

```
##      Factor1 Factor2 Factor3 Factor4
## [1,]    14.6    13.4    19.5    27.2
```

3. Or fit the linear regression software with no intercept (need to add 0+ in the formula to remove the default intercept)

```
LMfit<-lm( sales~ 0+Factor1+Factor2+ Factor3+Factor4, data=Ex16 )
LMfit$coefficients
```

```
## Factor1 Factor2 Factor3 Factor4
##    14.6    13.4    19.5    27.2
```

Therefore, if the predictors are indicators of the factor levels, the fitted regression model would be

$$\hat{Y} = 14.6X_1 + 13.4X_2 + 19.5X_3 + 27.2X_4 \,.$$

Note: Although equivalent, for simple ANOVA model, it is easier to directly apply the ANOVA fitting rather than setting the design matrix. We will use regression in the more complex setting in later chapter. Except for setting the design matrix differently, the regression approach for the weighted mean is very similar.

The key concepts 1) ANOVA and regression approaches are equivalent if setting up correctly 2) they lead to the same conclusion.

Sample size planning and power analysis for single factor studies (ch 16.10)

Planning of sample sizes is an integral part of the design of a study (both observational and experimental):

- It is important to plan the sample sizes so that needed protection against both Type I and Type II errors can be obtained, or so that the estimates of interest have sufficient precision to be useful. This planning is to ensure that the sample sizes are large enough to detect important differences with high probability. Otherwise, if the study power is low, it would be waste of money if the results are not significant, but we can only state that the study results are inconclusive.

- At the same time, the sample sizes should not be so large that the cost of the study becomes excessive. A power that is greater than 85-90% are often adquate. There is no need to go for 99% power, but we also don't design a study with lower than 80% power (chance of 1 out 5 to miss the 'real treatment effect').

Note. Calculating power or sample size (for a given power) is like a "thought experiment". We may not need the data but a specification of the parameter setting under the alternative that we believe in ("what would happen if …") or use some reasonable estimate from historic or published studies.

Sample size/power analysis (2)

- We usually assme all treatments are to have equal sample sizes (balanced or $n_i = n_r/r$). First, the test statistic is relatively insensitive to small departures from the assumption of the ANOVA model if the sample sizes are equal. This is not the case for unequal sample sizes. Second, the power of the test is maximized if the samples are of equal size.
- Planning of sample sizes can be approached in terms of (1) controlling the risks of making Type I and Type II errors, (2) controlling the widths of desired confidence intervals for the certain paramater , e.g. mean difference.
- We consider the planning of sample sizes with the given power here in this Chapter for one-factor studies.
- Under the alternative hypothesis, the test statistic follows a non-central F-distribution with non-centrality parameter ϕ. The textbook listed a limited cases with different degrees of freedom, ϕ and α. This is one approach that people used in the past. The modern statistical software such as R, SAS, Minitab provide much easier solutions, than using probability tables. We will study how to plan the sample size of one-factor study with a R function.

Sample size/power analysis (3)

We use the Food Company example in the textbook (page 717): We need the design assumptions of the cell means and variance (SD).

1. The analyst wishes to determine the power of the decision rule in the example on page 699 when there are substantial differences between the factor level means. Specifically, the analyst wishes to consider the case when 1 = 12.5, 2 = 13, 3 = 18, and 4 = 21.
2. We still need to know , the standard deviation of the error terms. Suppose that from past experience it is known that = 3.5 cases approximately

Then, we can use R function **power.anova.test()** to calculate the power given these design assumptions and given sample size, or verse versa, we can calculate the sample size with a chosen power level. By default it uses a 5% significance level. You can use help() function to understand the meaning of different arguments for this function.

A. First, from sample size to power: If the company set sample size per group is set $n_i = 5$, then what would be the study power?

```r
mu     <- c(12.5, 13, 18, 21)
sigma2 <- 3.5^2
power.anova.test(groups = length(mu), n = 5, between.var = var(mu), within.var = sigma2)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##          groups = 4
##               n = 5
##     between.var = 16.72917
##      within.var = 12.25
##       sig.level = 0.05
##           power = 0.9330586
##
## NOTE: n is number in each group
```

B. Second, if the company set power to be 90%, then what would be the required sample size per group?

```
mu     <- c(12.5, 13, 18, 21)
sigma2 <- 3.5^2
power.anova.test(groups = length(mu),  between.var = var(mu), within.var = sigma2 , power = 0.9)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##          groups = 4
##               n = 4.576923
##     between.var = 16.72917
##      within.var = 12.25
##       sig.level = 0.05
##           power = 0.9
##
## NOTE: n is number in each group
```

In this case, $n = 4.6$ is needed per group, so we will round to the next integer to have 5 cases per group for this study.

If we want to choose a small type I error, $\alpha = 0.01$, we can also specify that in the function by setting sig.level $= 0.01$. Then more cases will the needed given the same power and design assumption.

```
power.anova.test(groups = length(mu),  between.var = var(mu), within.var = sigma2 , power = 0.9,
                 sig.level = 0.01)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##          groups = 4
##               n = 6.224297
##     between.var = 16.72917
##      within.var = 12.25
##       sig.level = 0.01
##           power = 0.9
##
## NOTE: n is number in each group
```

Snow tires example (the textbook page 719):

Define: Minimum range of factor level means for which it is important to detect differences between the μ_i is $\Delta = max(\mu_i) - min(\mu_i)$. We note that the between group variance does not depends on the mean values of each group, but their difference from each other; also it follows from the (16.92) that the SSTR is mimimized by $\Delta^2/2$. Then we first divide SSTR by $r - 1$ to get the MSTR as the between group variance.

- Case 1: A company owning a large fleet of trucks wishes to determine whether or not four different brands of snow tires have the same mean tread life (in thousands of miles). It is important to conclude that the four brands of snow tires have different mean tread lives when the difference between the means of the best and worst brands is 3 (thousand miles) or more. Thus, the minimum range specification is $= 3$. Thus we can calculate the sample size for $\Delta = 3$ and $\sigma = 2$ as follows. The resulting sample size is $n = 14$ per group.

```
Delta = 3
sigma = 2
r = 4
power.anova.test(groups = r,  between.var =  Delta^2/(2*(r-1)), within.var = sigma^2, power = 0.9)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##          groups = 4
##               n = 13.61847
##     between.var = 1.5
##      within.var = 4
##       sig.level = 0.05
##           power = 0.9
##
## NOTE: n is number in each group
```

- Case 2: In particular, the sample size is only determined by the relative difference,$\Delta/\sigma = k$ (also called the 'effect size'). Thus, can calculate the sample size for $k = 2$ as follows by setting within.var $= 1$:

```
# specify the effect size
 k=2
 r = 4
## calculate the sample size
power.anova.test(groups = r,  between.var = k^2/(2*(r-1)), within.var = 1, power = 0.9)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##          groups = 4
##               n = 8.139055
##     between.var = 0.6666667
##      within.var = 1
##       sig.level = 0.05
##           power = 0.9
##
## NOTE: n is number in each group
```

Note 1: The sample of the test is a function of k (the standardize difference), and the sample is decreased with a larger k for the given power. As we can see, in the first case of Snow tires example, the sample size per group is 14 (rounded up) for $k = 1.5$, which was decreased to 9 for $k = 2$.

Note 2: As we can see if it is very convenient to use the statistical software to caculate the sample size/power with one-click, rather than using the probability table. We can vary the value of k and the number factor level r to generate the sample size for different senarios on page 720.

Sample size and power analysis (4)

Because at the design stage, the researchers may not have very good knowledge of the effect size or a good estimate of the variability, it is generally desirable to investigate the needed sample sizes for a range of likely values of parameters before deciding on the sample sizes to be employed.

In practice, in designing a randomized clinical trial, we often set a range of possible treatment effects and various power (0.80, 0.85, 0.90), then calculate a sample size table to discuss with the investigators to choose the most reasonable sample size.

Summary this week and Homework

(* to be reviewed and tested in Quiz for basic concepts)

- *One factor (one-way) ANOVA model
- *Estimation and hypothesis testing in two formulations
- Relationship with regression and ANOVA
- *power and sample size analysis
- Data fitting with R
- Reading: Chapter 16.1-16.8, and 16.10, Appendix A on F-dstribution.
- Homework for week2 (R can be used in the analysis for plotting, ANOVA table, power calculation, no need to search the probability tables in the Appendix).

Problems 16.7, 16.10, 16.25 (assume $n_i = 9$ for all levels), 16.27, 16.29 at the end of Chapter 16 (Due 9/12/19, extended to 9/17/19 for problems for online submission).