

In this handout I will summarize the main ideas from each chapter that we have covered in the course. (It will eventually contain information from all chapters we have covered, but currently focuses on recent chapters.)

Chapter 13 — Two-Way ANOVA

- *two-way ANOVA* — Two-way ANOVA is multiple regression with two categorical explanatory variables (or factors). In general, ANOVA is the special case of regression where there is a quantitative response variable and one or more categorical explanatory variables. The response variable is modeled as varying normally around a mean that is a linear combination of the explanatory variables.
- *additive model* — In two-way ANOVA, an additive model does not have an interaction term between the two explanatory factors. an additive model assumes that the effect of changing levels of one factor is the same at each level of the other factor. If the two factors have I and J levels respectively, an additive model requires $(I - 1) + (J - 1) + 1 = I + J - 1$ parameters for the mean.
- *non-additive model* — In two-way ANOVA, a non-additive model (or saturated model) allows the effects of changing levels in one factor to be different for each level of the other factor. A saturated model would have the same fitted values as a model that gave a separate parameter to each combination of factors (namely the means of the observations for each), but the parameterization is different. In addition to the $I + J - 1$ parameters in the additive model, the saturated model includes an additional $(I - 1) \times (J - 1)$ parameters for the interaction, for a total of $I + J - 1 + (I - 1)(J - 1) = IJ$.
- *F-test* — The ANOVA table takes on special importance in ANOVA as compared to regular multiple regression. In a two-way ANOVA, the extra-sum-of-squares F-test to compare the additive and saturated models is in the line for the interaction term in the ANOVA table.
- *blocks* — Often in a two-way ANOVA, one explanatory variable is of interest (say, an experimental treatment) and the other is a nuisance variable that we need to account for, but are not directly interested in. A blocking variable is a variable that groups similar observations. It is generally a good idea to leave a blocking variable in a model, even if it does not test as statistically significant. In essence, this places a higher priority on controlling for block differences that may exist (even if they are too small to be detected with the data at hand) than in the possible benefit of increased power from a model with fewer parameters.
- *selection of dummy variables* — In setting up an ANOVA, R will take each factor, alphabetize the levels, treat the first of these as the reference, and then use a dummy variable for each other level. These individual dummy variables test specifically the difference between each level and the reference. In some situations where there are specific comparisons you wish to make, it is good to use dummy variables that allow for these comparisons to be made directly.
- *residual plots* — Plots of residuals versus fitted values have the same role in ANOVA as in multiple regression. A difference is that the fitted values can only take on as many values as there are treatment combinations, so you expect the plot to have vertical lines of points, one for each combination of levels. Transformation of the response variable may be helpful to improve departures from model assumptions. I also find it useful and informative to use different plotting symbols for each treatment combination.
- *interaction plots* — The textbook deemphasized these, but plots that plot one categorical variable on the x axis, plot the mean of the response variable for each treatment combination on the y axis, and

then connect the means corresponding to the same level of the other categorical explanatory variable are informative. If the line segments show similar patterns for each level of the second factor, this is consistent with the additive model. If the line segments for each level look quite different, this is evidence of an interaction.

- *alternative parameterization* — It is easier in theory (but messier for computation) to give each level of a factor its own parameter, but to restrict the sum of the parameters to be 0. Under this parameterization, each level is compared to the mean instead of comparing all but one to a reference level. You might see this alternative parameterization in other books about ANOVA.

Chapter 14 — Two-way ANOVA without replication

- *replication* — A design has replication if there are multiple observations for some of the treatment combinations. Without replication, there is only one observation per treatment combination. As a consequence all the data would be used to estimate a saturated model, and there are no degrees of freedom left over for inference. The saturated model will predict fit the data exactly and all residuals would be 0. But there are situations where replication is impossible. Without replication, there is little choice but to adopt an additive model.
- *repeated measures* — In a repeated measures design, the same individual is measured several times. This introduces a special kind of blocking.
- *random effects versus fixed effects* — If a factor contains all of the levels of interest, it is a fixed effect. Examples include sex for which there are no other choices or dosage where an experimenter may select specific levels for the analysis. A factor is a random effect if we want to control for differences among the levels, but see these levels as a sample from some larger population of possible levels of interest. A common example is an animal study where we need to model differences among the animals in the study, but are interested in the effects of the other variables on other animals that may have selected instead for the study. When a factor is included in a model as a random effect, it brings with it another source of variation. Estimates of model parameters for the mean are the same whether or not the factors are considered to be fixed or random, but the resulting inference can be different. In this course we did not explore these differences in analysis.

Chapter 20 — Logistic Regression

- *generalized linear model* — In a generalized linear model, the mean of the response variable is a function of a linear combination of the explanatory variables. Examples in this course are logistic regression and binomial regression where the responses of individuals are 0/1 binary variables (or categorical variables with two levels). In these cases, we use the logit function, the log of the odds ratio, as the function of the mean to model as a linear function.
- *link function* — In a generalized linear model, the link function is the function that has a linear model. You can think of link functions as special transformations so that the linearity assumption of the regular linear model holds. (Constant variance and normality probably will not hold.)
- *maximum likelihood* — Maximum likelihood is a method of parameter estimation in which the parameter values that produce the largest probability of the observed data are the estimates.

- *likelihood function* — The probability density function or probability mass function describes the probability distribution of a random variable. For fixed parameter values and variable possible values of the random variable, these functions integrate (or sum) to one. If we fix the possible values at the observed data and treat the parameters as variables, this is the likelihood function. A likelihood function will usually not integrate (or sum) to one.
- *properties of maximum likelihood estimators* — (1) bias approaches zero as sample size increase, (2) SDs of sampling distribution of estimators can be found analytically (often), (3) maximum likelihood estimators do as well as possible in most situations, (4) sampling distributions of maximum likelihood estimators are approximately normal (which means there are simple approximate statistical inferential tools available).
- *likelihood ratio test* — In a likelihood ratio test, we reject the null hypothesis if the ratio of the maximum likelihood of a reduced model over the maximum likelihood of a full model is too small. In other words, reject when the full model explains the data much better than a reduced model. On the log scale, this is equivalent to a difference of log-likelihoods. It turns out that for large enough samples, twice the difference in log-likelihoods (full - reduced) has an approximate chi-square distribution with degrees of freedom equal to the difference in the number of parameters.
- *deviance* — The deviance is a constant minus twice the log-likelihood. Thus, differences in deviances between models are equivalent to likelihood ratio tests, and differences in deviances can have chi-square distributions if sample sizes are large enough.
- *other types of regression for binary response* — The important qualitative aspect of the logit function is that is an increasing function from $(0, 1)$ to the real line. Other functions, such as the cumulative distribution function of the standard normal curve would work as well.