

Chapter 1 — Statistical Inference

- *causal inference* — To infer causality, you need a randomized experiment (or a huge observational study and lots of outside information).
- *inference to populations* — Generalizations to populations are justified by statistics alone under random sampling. Otherwise, generalizations may be speculative and can only be justified by subjective judgment.
- *null and alternative hypotheses* — A null hypothesis is usually a simple statement that there is no effect or no difference between groups, and is formally an equation about a parameter. We might find data to be consistent with a null hypothesis, but we do not prove a null hypothesis to be true. We may fail to reject a null hypothesis, but we do not accept it based on data. An alternative hypothesis is what we accept if we are able to reject a null hypothesis. Formally, it is usually an inequality about one or more parameters.
- *test statistic* — a test statistic is something that can be calculated from sample data. Statistical inferences are based on the null distribution of the test statistic, the random distribution the statistic would have if we were able to take several samples of the same size and if the null hypothesis were true. In an experiment where individuals are randomly allocated, we can also consider the randomization distribution of the test statistic.
- *p-value* — A p-value is the probability that the test statistic would take on a value at least as extreme (in reference to the alternative hypothesis) as that from the actual data, assuming that the null hypothesis is true. Small p-values indicate evidence against the null hypothesis. P-values are probabilities about the values of test statistics, but are not probabilities of hypotheses directly.
- *permutation and randomization tests* — In a randomization test, we ask how unusual the results of an experiment are compared to all possible randomizations. A permutation test is identical in practice, but differs in interpretation because the groups are observed instead of randomly assigned.
- *confounding* — Two variables are confounded when we cannot distinguish between their effects.

Chapter 2 — *t*-Distribution Methods

- *standard error* — The standard error of a statistic is the estimated standard deviation of its sampling distribution.
- *Z-ratio and t-ratio* — The Z-ratio (z-score) of a statistic that is an estimator for a parameter is the $(\text{estimate} - \text{parameter}) / (\text{standard deviation})$. In many situations the Z-ratio will have an (approximate) standard normal distribution. The t-ratio is $(\text{estimate} - \text{parameter}) / (\text{standard error})$. The difference is that the t-ratio has variability in both the numerator (estimate) and the denominator (standard error), so the t-ratio is a more variable statistic than the Z-ratio. In many situations the t-ratio will have a t distribution with a number of degrees of freedom that is a function of the sample size.
- *paired samples versus two independent samples* — In a paired setting, pairs of observations are sampled together. It is appropriate to take differences within the pair and to base inferences on the distribution of these differences. When there are two independent samples, inferences are based on the distribution of the difference of sample means, which has a different standard error than the

mean of paired differences. In a paired design, the effects of confounding factors are often reduced more efficiently than by randomization.

- *pooled standard deviation* — If the model assumes that there is a common standard deviation, then it makes sense to pool information from all samples to estimate it. This is an assumption of ANOVA and regression as well. However, if in a two-sample setting there is a large discrepancy in variability in the two samples, it is best not to pool.
- *confidence intervals* — A 95% confidence interval is made from a procedure that will contain the parameter for 95% of all samples. The values in a 95% confidence interval are precisely those for which the corresponding two-sided hypothesis test would have a p-value larger than 0.05.

Chapter 3 — Assumptions

- *robustness* — A statistical procedure is robust if it is still valid when assumptions are not met.
- *resistant* — A statistical procedure is resistant when it does not change much when part of the data changes, perhaps drastically.
- *outlier* — An outlier is an observation that is far away from the rest of the group.
- *transformations* — Transformed variables (logs are a common choice) often come closer to fitting model assumptions.

Chapter 4 — Alternative Methods

- *ranks* — Transforming observations to ranks is an alternative to t tools. This transformation may be especially useful with *censored data*, where an observation may not be known exactly, but its value may be known to be larger than some value. (For example, if the variable being measured is survival time and the individual has not died by the end of the study, the survival time is censored.)
- *permutation test* — In a permutation test, the p-value is a proportion of regroupings.
- *sign test* — For paired data, we can consider randomizing the sign of each paired difference (or permuting the groups within each pair) to find a p-value.
- *Normal approximation* — P-values from permutation tests can be found by a normal approximation.

Chapter 5 — One-way ANOVA

- *pooled estimates of SD* — If there are two or more samples and we assume constant variance in the populations, the best estimate of the common variance is the pooled variance, s_p^2 , which is a weighted average of the sample variances, weighted by their degrees of freedom.
- *degrees of freedom* — For a random sample, the degrees of freedom are one less than the sample size, $n - 1$.
- *One-way ANOVA* — In a one-way ANalysis Of VAriance, there is a single categorical explanatory variable and a quantitative response variable. The null hypothesis is that all population means are equal. To test this hypothesis, we use an *F*-test. The ANOVA table is an accounting method for computing the *F* test statistic.

- *F-distribution* — An F distribution has both numerator and denominator degrees of freedom. The mean is near one. P-values in F -tests are areas to the right.
- *Extra-sum-of-squares F-test* — An extra sum of squares F -test is for comparing nested models that differ by one or more parameters having values fixed at zero.
- *residual plots* — A plot of residuals versus fitted values (or group in an ANOVA) can be useful for detecting deviations from constant variance assumptions or skewness.
- *random effects* — In a random effects model, the groups are thought to be sampled from some larger population of interest about which inferences are desired. In a fixed effects model, the groups in the study are the only ones of interest.

Chapter 6 — Multiple Comparisons

- *linear combination* — A linear combination of variables is a sum of the variables where each variable may be multiplied by a coefficient.
- *simultaneous inference* — Simultaneous inference is when more than one inferences are made at the same time.
- *familywise confidence level* — The familywise confidence level is the proportion of samples for which a set of more than one confidence intervals will all contain the true parameter values.
- *multiple comparisons* — When there are more than two groups, it is often of interest to make multiple comparisons. To control for multiple comparisons, it is best to increase the width of standard confidence intervals by using a larger multiplier than the standard t -distribution one to be confident that all intervals are correct.
- *Tukey-Kramer* — The Tukey-Kramer procedure assumes that sample sizes are equal, and is based on the Studentized range statistic, a normalized difference between the maximum and minimum sample mean. Tukey-Kramer is only valid for comparing sample means.
- *Scheffé* — The Scheffé procedure uses a multiplier based on the F distribution and is based on the distribution of the most extreme linear contrast. This procedure is very conservative, but valid for searching for comparisons after the fact.
- *Bonferroni* — The Bonferroni procedure is valid for a predetermined number of comparisons of any type. The multiplier comes from a t distribution, but with increased confidence level based on the number of comparisons.

Chapter 7 — Simple Linear Regression

- *simple linear regression* — In simple linear regression, there is a single quantitative explanatory variable and a quantitative response variable.
- *interpolation* — Interpolation is making predictions or estimates of the mean response within the range of the explanatory variable.
- *extrapolation* — Extrapolation is making predictions or estimates of the mean response outside the range of the explanatory variable.

- *model assumptions* — the four model assumptions are (1) normality of the responses around their means; (2) linearity of the means as a function of the explanatory variable; (3) constant variance for each value of the explanatory variable; and (4) independence among all observations.
- *fitted values* — A fitted value is the estimated mean for a set of explanatory variables.
- *residual* — The residual is the difference between the actual value and the fitted value.
- *least squares estimates* — The principle of least squares says that the best estimated regression coefficients minimize the sum of squared residuals. (In simple linear regression and in regular multiple regression, these are also maximum likelihood estimates.)
- *degrees of freedom* — the degrees of freedom for a set of residuals are the number of observations minus the number of parameters for the mean.
- *estimate of σ* — The standard estimate of σ is the square root of the residual sum of squares over the square root of the degrees of freedom. The square of this is an unbiased estimate of the variance. (The maximum likelihood estimate divides by n instead of $n - p$ where there are n observations and p parameters.)
- *standard error* — The standard error associated with an estimate is an estimate of the standard deviation of the distribution of the estimate.
- *the regression effect* — the regression effect, or *regression to the mean* is the tendency for future observations to be more average than the previous observation, on average. This is a necessary result of correlations being less than 1 (in situations where there is not a perfect linear relationship).
- *correlation* — The correlation coefficient measures the strength of the linear relationship between two quantitative variables. Graphs are important: r near 1 or -1 does not imply that a linear fit is better than a nonlinear fit, just that the linear fit is better than a fit of a horizontal line. An r near 0 does not imply no relationship between the variables; there could be a very strong nonlinear relationship.

Chapter 8 — Regression Assumptions

- *testing linearity* — A curve in the plot of residuals versus fitted values can indicate nonlinearity.
- *testing constant variance* — A wedge-shaped pattern of residuals is evidence that variance is large for larger means than for smaller means.
- *testing normality* — Skewness in residual plots can indicate lack of normality, as can a lack of straightness in a normal quantile plot of residuals.
- R^2 The R^2 test statistic can be interpreted as the proportion of variance in the response variable explained by the regression.

Chapter 9 — Multiple Regression

- *multiple regression* — In multiple regression, there is a single quantitative response variable and at least two explanatory variables.

- *dummy variables* — categorical variables can be included in a regression analysis by coding the presence of each level but one in a separate variable.
- *interaction* — two variables display interaction if the size of the effect of one depends on the value of the other. This can be modeled in multiple regression by including the product of the variables in the model.
- *coded scatterplots* — It is often useful to use different plotting symbols (or colors) for each different population.
- *interpreting coefficients* — A coefficient in a regression model only has meaning in the context of all of the other coefficients from the same model. In a strict mathematical sense, a regression coefficient is a measure of how the response changes per unit change in the explanatory variable keeping other variables fixed. This interpretation is often faulty, because it may not be possible or plausible in the context to keep all other variables fixed. This is especially obvious when, for example, X and X^2 are two variables, but can also occur if X_1 and X_2 are strongly correlated.

Chapter 10 — Inference in Multiple Regression

- *tests of single coefficients* — Tests of single coefficients in a regression model are made using a t -test, and confidence intervals are based on multipliers from the t distribution.
- *tests of linear combinations* — Tests or confidence intervals for linear combinations of coefficients requires first finding the standard error of the estimate (the linear combination of coefficients). This will be the square root of a linear combination of variances and covariances of the coefficients.
- *covariance* — The covariance of two estimates is a measure of how they vary together.

Chapter 11 — Model Checking

- *dealing with outliers* — If an observation is an outlier, it is worth fitting the model with and without the observation. If the inferences of main interest are the same, include the observation. If the inferences of main interest are different, you have a choice. You can include the observation, but indicate in the summary of the analysis that the results are strongly dependent on the inclusion of an influential observation. Or, you can drop the observation and limit the analysis to a portion of the population of interest.
- *influence* — Cook's distance is one way to measure the influence of an observation. It is a scaled measure of how all of the fitted values change when a single observation is left out of a fit. A very rough guide is that a Cook's distance near one or larger indicates an influential observation. Cook's distance only measures influence one observation at a time, so it is not a perfect measure. For example, a pair of observations could be highly influential, but each might individually be not so influential.
- *leverage* — Leverage is a measure of the explanatory variables alone that measures how far the explanatory variable values are from the rest. Observations with high leverage have the potential to be influential, but may not be. The leverage values are the diagonals of the “hat matrix” from the matrix approach to regression.

Chapter 12 — Model Selection Strategies

- *The curse of dimensionality* — I have not used this phrase in the course, but the general idea is that to maintain a set level of precision, the number of data points needs to increase exponentially with the number of parameters. What this means in a multiple regression setting is that, even if there are a lot of important variables and we want to include polynomial terms and interactions to better model what is really happening, a lot of data is necessary to model all of the parameters precisely. Bias and variance are in conflict — adding more parameters to a model decreases bias but adds variability to the estimates of the parameters from data, and the amount of data new data necessary to maintain variability increases exponentially. A solution is to accept a certain amount of bias and fit models that are simpler, but that provide more reliable estimates.
- *variable selection* — There are several methods for deciding which variables to include in a model. Different methods need not result in the same model. This is okay as long as we realize that it is rare for there to be a single model that is much better than all others. Variable selection strategies can be thought of as means to find useful models.
- *forward selection* — In forward selection, variables are added one at a time, adding the best available variable at each step, until there are no more variables for which the improvement in fit is worth the penalty for adding the variable. In the forward selection procedure, once a variable is in, it cannot be removed.
- *backward elimination* — In backward elimination, a model with many parameters is fit first and then variables are eliminated one at a time, always removing the least beneficial variable until removing any additional variables results in a worse model by the objective. Once a variable is removed, it is gone for good.
- *stepwise regression* — Stepwise regression alternates between forward and backward steps.
- *all subsets regression* — If there are p variables, there are 2^p possible linear models with these variables. In all subsets regression, the model that optimizes some objective criterion from all of the 2^p models is selected as the best. In practice, this is only feasible for relatively small p , which is why sequential procedures (forward selection, backward elimination, stepwise regression) are used as alternative procedures.
- C_p — The C_p statistic picks models for which the total mean square error is estimated to be small.
- AIC — AIC is based on a likelihood model and is the deviance plus a penalty of 2 per parameter.
- BIC — BIC is also based on likelihood models and is the deviance plus a penalty of $\log n$ per parameter. Because $\log n > 2$ for all but the tiniest of data sets, BIC is less likely than AIC to include too many variables.

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 it 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.

Chapter 21 — Binomial Regression

- *binomial distribution* — The binomial distribution counts the number of successes in a fixed number of independent trials when there is the same success probability for each trial.
- *binomial regression* — In a logistic regression setting where there are multiple observations for each set of values of the explanatory variables, binomial regression can be appropriate. The estimated regression coefficients are identical. The deviance will be off by a constant factor because in binomial regression, the order of the observations does not matter, but it does for logistic regression. (The probability that the first two of five observations are successes is different than the probability that two of five are successes.)
- *deviance residuals* — Deviance residuals have a functional form that is based on the likelihood function. The sum of squared deviance residuals is the deviance. If many deviance residuals are larger than 2 in absolute value, this is an indication that extra-binomial variability may be present.
- *maximum likelihood* — The principle of maximum likelihood says to estimate parameters with values that make the probability (likelihood) of the observed data to be as high as possible.

Chapter 22 — Poisson Regression

- *poisson distribution* — The Poisson distribution is a probability distribution on the nonnegative integers. It may be derived as a limit of binomial distributions for which n tends to infinity but the mean np tends to a constant μ . It also results from a generally plausible assumption on a distribution of points where: (1) the mean is proportional to the length (or area or volume) of a region; (2) the counts in disjoint regions are independent; and (3) the probability of two or more points in a small region gets small quickly enough. The mean and variance of the Poisson distribution are the same.
- *comparison between Poisson and binomial regression* — Poisson regression and binomial regression each have a nonnegative count as the response. The main difference is that in binomial regression, it is necessary for the response in each case to be modeled as a sum of a known number of independent 0/1 random variables, so there is a known maximum value. In Poisson regression, there is no theoretical maximum value.

- *link function* — In a log-linear model for Poisson counts, the conventional choice of link function is the log.
- *extra-Poisson variation* — In practice, it is often the case that observed variability is larger than that predicted by the Poisson distribution. In this case, a more general model that assumes that the variance is a scalar multiple of the mean can be fit. This is called a quasiliikelihood model.