1/12/2010 36-402/608 ADA-II H. Seltman

Breakout #1 Comments

1. Look at the Breakout #1 Results and summarize the EDA.

   In addition to bird number (whose statistics are meaningless), the "demographics" of gender (female indicator variable) and age status (adult vs. juvenile) and the survival outcome status, we have nine measurements on each bird. E.g., the Humerus length ranges from 0.659 to 0.780 inches with a median of 0.733 and a mean of 0.732. The similarity of the median and mean suggests a symmetric distribution.

   Adult vs. juvenile status is missing for all female birds. Among the females 28 out of 49 died. Among the juvenile males 12 of 28 died, and for adult males 24 of 59 died.

   The boxplots of humerus length by survival show symmetric distributions (but you can't see whether or not they are normal on a boxplot) with a slightly higher median for surviving birds compared to perished birds. The boxplots of humerus length by survival and demographic group show some degree of unequal spread (we can see IQRs) though probably not more that can be expected by chance under the condition of equal variances in the population (see simulated plot). All three groups show a higher median humerus length for surviving birds. The pattern of length medians is highest for juvenile males, and lowest for females.

2. Look at the code and come up with explanations or questions about anything that is unfamiliar or questionable.

   The form `table(..., useNA="ifany")` is much safer than the "naked" table command.

   This code uses `with(my.dataframe, someExpression)` rather than `attach(my.dataframe); someExpression` as a method to avoid typing `my.dataframe$` many times. The advantage of this approach is that you are free to modify the data.frame as needed, and those changes will be saved (if you quit R with a yes answer to the "save workspace" question). Remember that many commands, such as `lm()` have a `data=` argument that requires neither "with" nor "attach" to use.

3. Explain what a statistical model is, what model applies to the given results, and what the null and alternative hypotheses are.

   Means model: Surviving and perishing are groups with humerus length outcomes with means assigned some parameters, say, μ_S and μ_P

   Error model: Errors (ϵ) are independent and normally distributed with a common parametric variance, say, σ^2.

   Full model: $Y_i = \mu_{ij} + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2),\ \epsilon_i \perp\!\!\!\perp \epsilon_i$

   where $\mu_{ij} = \mu_S$ if subject i is in the survived group and $\mu_{ij} = \mu_P$ if subject i is in the perished group.

   As must always be the case, the model "allows" both $H_0 : \mu_S = \mu_P$ and $H_A : \mu_S \neq \mu_P$.

4. Give a clear statement of the meaning of the p-value and confidence intervals.

   The p-value is the probability over many repeat experiements (i.e., samples from the population) of seeing the observed statistic, i.e., t=-1.777, or one more un-null-like, i.e., $t < -1.777$ or $t > +1.777$, **if the null hypothesis is true *and* the model assumptions are met**. The common statement that the p-value is "the probability that the results are due to chance" is commonly misunderstood, and only true when taken as a shortcut for the exact above statement. In fact all results we see are "due to chance".

   The 95% confidence interval is a **random** interval (in both width and center) across repeated experiments, that has the characteristic that 95% of them include the true, fixed parameter, and 5% miss the parameter completely. This is only true if the model assumptions are correct, but is true under both the null and alternative hypotheses.

5. Explain how the model assumptions might be justified and/or checked. Explain what might happen if the assumptions are incorrect.

   Normality is justified by the central limit theorem as applied to all of the many small genetic, environmental, and measurement factors that can be thought of as small "plus or minus" effects around some mean effect. Independent errors can be justified by lack of interaction between subjects. Equal variance can be justified by supposing that the "plus or minus" effects are similar for both groups.

   We can check equal variance with side-by-side boxplots or a comparison of observed group variances. We can check normality with a quantile-normal plot of residuals. We can check for serial correlation with a Durbin-Watson test. Within-group correlation can be checked for known groups using two-way ANOVA. Correlation due to unknown collaborations are not likely to be detected.

6. Define power. Explain how it applies to this problem.

   Power is the chance (over repeat experiments) that you will reject the null hypothesis. When $\alpha = 0.05$ and H_0 is true, any appropriate test has 5% power. For each combination of sample size, error variance and true effect size (e.g., population mean difference), a power value exists, and is at least 5%. Without further calculation you cannot tell whether power is closer to 5% or 100%. If power is low, and you obtain $p > 0.05$ for your experiment, you have no way of making a preference between the null hypothesis being true and making a type 2 error. But if the power is high, you can be reasonably sure that the null hypothesis is true.

7. If time permits: think about how the t-test is constructed.

   For any statistic, G, designed to estimate parameter γ, if G is known to be normally distributed, a test of $H : \gamma = \gamma_0$ can be made by constructing

$$t = \frac{G - \gamma_0}{\mathrm{SE}(G)}$$

   where SE=estimate standard error (standard deviation of the sampling distribution, using an estimate of σ^2.

   With the approprate degrees of freedom, df, for the problem, the t-statistic will follow the t(df) distribution.