Prof. Donna Ankerst, Katharina Selig

June 17th & 18th 2019

Problem 1

The data below show post-treatment gene expressions for four rats randomly assigned a control versus four rats assigned a drug, collected to test the null hypothesis that both groups have the same expression.

Control	Drug
12	23
17	18
14	26
9	21

(a) Calculate by hand the mean, median, min, and max of each group. Comment on the differences seen between the two groups.

(b) A permutation test is desired to test differences between the 2 groups. Calculate the number of permutations required, showing all work.

(c) Write `R` code to perform an exact permutation test at the 0.05 level of the null hypothesis that both groups have the same mean versus the alternative hypothesis that the means are not the same. Presume you have access to the `gtools`-package and hence the `combinations`-function, where `combinations(x,y)` returns a matrix and each row contains one possible subset of `1:x` of length `y`.

(d) Modify the code to test that both groups have the same median.

`In R:`

(e) Implement the code from (c), (d) in `R`. For both tests, construct a histogram of permutation replicates, indicating the observed values of the test statistics on each. Interpret the hypotheses test results for both the median and mean at the 0.05 level of significance.

Problem 2

Consider the `R chickwts` data set introduced in the lecture on the effect of five diets (casein, horse bean, linseed, meat meal, soybean, sunflower) on weight of chicks. The number of chicks in each group is the following

casein	horse bean	linseed	meat meal	soybean	sunflower
12	10	12	11	14	12

(a) Suppose we are interested in performing an analysis of variance test of

$$H_0 : \text{all 5 diets have the same mean} \quad \text{vs.} \quad H_A : \text{at least two differ}$$

using the F-statistic

$$F = \frac{\frac{1}{k-1}\sum_{i=1}^{k}(\bar{y}_i - \bar{y})^2}{\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2},$$

where k is the number of groups, n_i is the number of chicks in group i, $\bar{y}_i$ is the mean weight of group i, y_{ij} is the j^{th} observation in group i, and $\bar{y}$ is the overall mean weight. We reject H_0 at the $\alpha = 0.05$ level if $F > F_{k-1,n-k,1-\alpha}$. The test is implemented in `R` as the function `aov`. How can you calculate the number of permutations needed for an exact test? Write down the solution and simplify it to a reasonable extent.

In R:

(b) Perform the F-test in R using the aov-function on the chickwts data, which is build in in R. Extract and store the F-statistic for this test.

(c) Perform an approximate permutation test using 1000 replicates, plotting a smoothed histogram of the permutation replicates with the observed F-statistic indicated. Compare p-values from the observed F-test and the permutation test.

Not in R:

(d) The two-sample Kolmogorov-Smirnov test statistic is

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|,$$

where $F_{1,n}(x)$ and $F_{2,m}$ are the ecdf for the first and second sample, respectively. Calculate by hand the ecdfs of weight for the soybean and linseed diets. Summarize the results in a table and then calculate D.

linseed	141	148	169	181	203	213	229	244	257	260	271	309		
soybean	158	171	193	199	230	243	248	248	250	267	271	316	327	329

In R:

(e) Implement D in R using the ecdf and ks.test function. The p-value returned by ks.test is for a test of the null hypothesis that $D = 0$ vs. $H_A : D > 0$. Interpret this p-value for a test at the 0.05 level of significance. Plot the ecdfs on the same graph and explain the value of D in terms of this graph. Indicate D with a red line on the graph.

(f) Perform an approximate permutation test with 1000 replicates under the null hypothesis that chicks with the soybean and linseed diet have the same weights at the 0.05 level. Report the ASL. Plot a histogram of the replicates and overlay the observed value of D.

Problem 3

The following data contains LSAT (average score on law school admission test score) and GPA (average undergraduate grade-point average) for 15 law schools.

LSAT	576	635	558	578	666	580	555	661	651	605	653	575	545	572	594
GPA	339	330	281	303	344	307	300	343	336	313	312	274	276	288	296

(a) Develop and write down an algorithm to compute the jackknife estimate of bias and standard error of Pearson's correlation statistic.

(b) Write R code to implement the algorithm of (a).

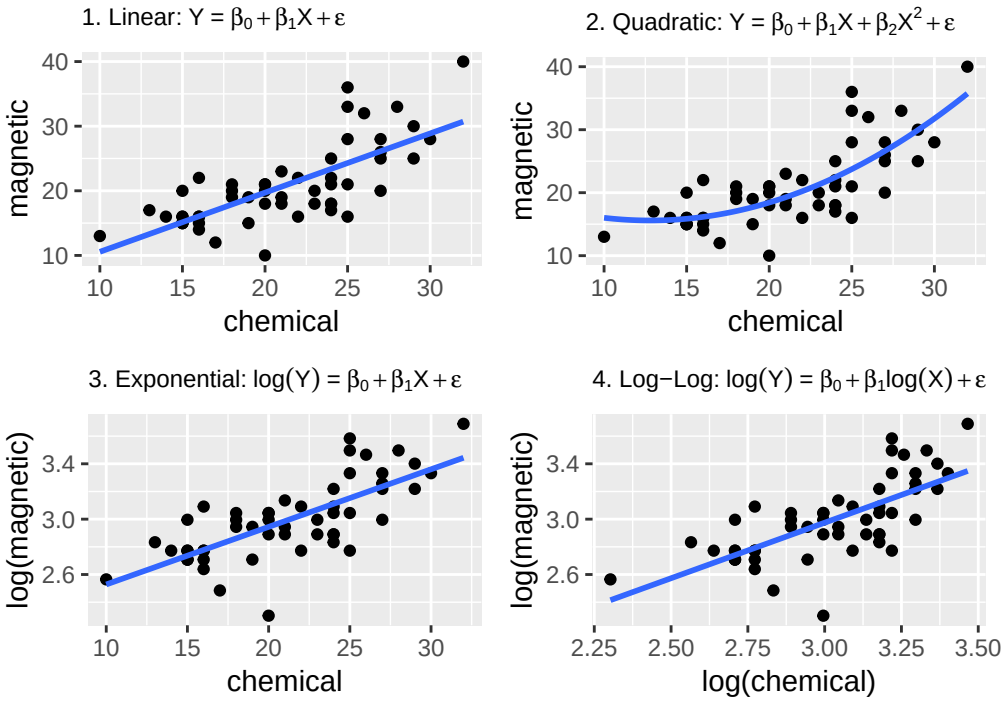
In R:

(c) Implement the code in R showing a histogram of the jackknife replicates with the observed value and jackknife mean overlaid. You can read in the data using

```
LSAT <- c(576, 635, 558, 578, 666, 580, 555, 661, 651, 605, 653, 575, 545, 572, 594)
GPA <- c(339, 330, 281, 303, 344, 307, 300, 343, 336, 313, 312, 274, 276, 288, 296)
```

Problem 4

Recall the `ironslag` data from the `DAAG` package, which contains 53 measurements of iron content by two methods, `chemical` and `magnetic`. Four models are proposed for predicting the magnetic measurement (Y) from chemical measurement (X):



(a) Write down an algorithm to perform 26-fold cross-validation whereby each held out group has 2 observations, except for one that has 3. Take the average of the squared error terms for each test as the metric.

(b) Write Rcode to implement the algorithm of (a) for the linear model. You can use `predict` to calculate the estimates for the test data for a model. `predict(model, newdata = test_set)` returns the fitted reponse values using the `model` with the `newdata`.

`In R:`

(c) Implement the code in (b) for all four regression models. Summarize the prediction errors over the 26 test sets by side-by-side boxplots, means and standard errors. Comment on which model is preferred based on 26-fold cross-validation.