STAT 391
Homework 7
Out Friday May 22, 2018
Due Friday May 29, 2018

©Marina Meilă
mmp@stat.washington.edu

Reading Lecture notes on Hypothesis testing, part I and II from the `Handouts` web page

Problem 1 – Testing a hypothesis

a. If *positive* integer numbers with at most 3 digits are drawn uniformly, (which means that any number from 1 to 999 are equally likely to appear,) show that the distribution of the first digit is uniform over $S = \{1, \dots 9\}$. The first digit is defined as the highest non-zero digit.

[**b. Not graded**] Prove: If positive integer numbers with at most d digits are drawn uniformly, then the distribution of the first digit is uniform over $S = \{1, \dots 9\}$. (hint: induction.)

c. Consider the event $E_{n,t}$="in a data set of n integers, at least t of them start with 1". Write an expression $p_{n,t} = P_0(E_{n,t})$, the probability that E_n is true given that the highest digits are uniformly distributed over S. This should be a function of t and n.

d. Read the first $n_{\mathcal{D}} = 60$ data from file `hw6_digit.dat` [1], count $t_{\mathcal{D}}$, the number of times 1 appears as the highest digit, and calculate the numerical value of $p_{n_{\mathcal{D}},t_{\mathcal{D}}}$ from these data. *You will find that* $p_{n_{\mathcal{D}},t_{\mathcal{D}}} \ll 1$*!!. In other words, it is extremely unlikely for this data to be generated by* P_0. To convince yourself that not all $p_{n,t}$ values are small, calculate also $p_{n_{\mathcal{D}},t'}$ for $t' = 6$.

[**e. Extra credit**] The formula in **c.** is impractical for large n. Compute the mean and variance of t/n for a given n, as a function of t, then use the CDF, normal approximation, and Φ^{-1} (CDF of standard normal) to obtain an approximate expression $\tilde{p}_{n,t}$ for $p_{n,t}$. Literal answer only for this part.

f. Now read the whole data in `hw6_digit.dat` and compute $\tilde{p}_t$ for the whole data set. Comment on your findings.

g. We again only use the first $n_{\mathcal{D}} = 60$ data from file `hw6_digit.dat`. Now we consider another way of testing. Denote $\theta_i = P(\text{first digit is } i)$. Let Model A be that the first digit follows a uniformly distribution over S, that is $\theta_i = 1/9$, for each $i = 1, \dots, 9$. Let Model B be that the first digit follows a multinomial distribution over S with θ_i's not neccesarily equal. Model A is then a special case of Model B.

Compute the likelihood of the data (only use the first 60 numbers) under Model A: there is no free parameter to estimate. Compute the ML estimates, $\hat{\theta}_i^B$ for $i = 1, \dots, 9$ under Model B, and then use them to obtain the maximum likelihood of the data under Model B. Use these two quantities to obtain the likelihood ratio test statistics value $\lambda_{\mathcal{D}}$.

h. How many (free) parameters d_B are estimated from data in model B? Under model A, $d_A = 0$ parameters are estimated from the data, as we have set them to 1/9. Use the χ^2 table to obtain the probability $Pr[Z_d > -2 \ln \lambda_{\mathcal{D}}]$ where Z_d is a random variable drawn from a χ^2 distribution with $d = d_B - d_A$ degrees of freedom.

i. Now read the whole data in `hw6_digit.dat` and compute the above probability for the whole data set. Comment on your findings.

Problem 2 – Rob at the Flintstone factory

Rob is doing a summer internship in the distant past. He is supposed to be inspecting the products of the ACME Flinstone Factory. This factory produces flintstones, and each flintstone is supposed to be exactly $l_0 = 8in$ long. Rob measures each flintstone with a device whose measurement error X is uniformly distributed between $(-1)in$ and $1in$.

[1] This data contains the population of towns and cities in the US.

a. Let Y denote the length measurement of a flintstone. Under *the model* that all flintstones are exactly l_0 long, what is the distribution of Y? What are $E[Y]$ and $Var(Y)$?

b. Rob has already measured n flintstones, which had lengths $Y_1, \dots Y_n$ and has calculated their mean length $L = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Assume the measurement errors are mutually independent. L is of course a random variables, since all Y_i's are random. Under *the model* above, what is the sample space of this variable? What is $E[L]$? What is $Var(L)$?

c. The actual measurements made by Rob are in `flinstones.dat`. All these flinstones were produced by Fred, and Rob suspects Fred of stealing material and making shorter flintstones. Fred insists that all his flintstones are exactly l_0 and that Rob's observations are due only to measurement error.

Make a plot of the data, also marking clearly the sample size S_Y, the point l_0, and the point $L = l$ the data average.

d. Let us try to help Rob clarify this matter. Rob decides to use Chebyshev's inequality.

$$\boxed{Prob[|z - E[Z]| \geq t] \leq \frac{Var(Z)}{t^2}} \tag{1}$$

Apply this inequality to the variable L; assuming that Fred says the truth, L should have the mean and variance you obtained in **b.** Therefore, (1) will tell you how probable it is for the actual $L = l$ Rob and (you) have calculated from the data to occur. Denote this probability p_{Cheb}

e. Rob now wants to use a more refined tool. He knows about the **Central limit theorem** (see notes Chapter 12.4). He interprets this theorem as follows:

> Under *the model*, the variables $Y_{1:n}$ are independent, identically distributed. So, if n is large (and I am going to assume that it is) the variable $Z = \frac{(Y_1+\dots Y_n)-nl_0}{\sqrt{n \cdot Var(Y)}}$ is $Normal(0,1)$.

Rob is right. Compute the actual value of $Z = z_n$ from the data. Use this fact, and a table or program with Φ the CDF of the Normal, to find the probability p_{CLT} that Z is *no larger* than the value observed.

[f.– not graded] In terms of statistical testing, what are z_n and p?

g. In addition to the probability $p = p_<$ that Z is *no larger* than the value observed, Rob also computed the probability $p_>$ that Z is *no smaller* than the value observed, and the probability $p_{\neq}$ that the absolute value $|Z|$ is *no smaller* than the absolute value $|z_n|$ observed.

Write these quantities as probability statements involving Z and z_n and find the numerical values of $p_<$ and $p_{\neq}$.

Then, write *in words* what are the pairs of hypothesis each of them are used for testing.

h. Help Rob perform a likelihood ratio test for the same question. Let H_0 be Fred's claim that the flintstones are sampled from the true model in **a**. Let H_1 be the alternative that the flintstones are sampled from a uniform distribution with lower mean. Compute the Max Likelihood estimator for the data under the alternative model. Numerical values only.

Notation:a_0, b_0,a_1, b_1 are the parameters for the uniform distribution model under the hypothesis H_0 and the ML model under H_1.

i. Now compute the likelihood ratio λ and the test statistic $t = -2\ln\lambda$. How many free parameters has the model H_1? Let this number be d. For H_0, no parameter is fit to data, so $d_0 = 0$.

What is the p-value, p_{LR} of the probability that a $\chi^2_{d-d_0}$ is more extreme than the observed t?

j. How do you explain the difference between p_{Cheb} and p_{LR}?