PROBLEM SET 1, 18.465

Due Friday, Feb. 13, 2015

Venables and Ripley, p. 108, table 5.1, give a list of probability distributions available in R. There is a standard or basic (default) form of each distribution, such as mean 0 and variance 1 for the normal, but one can also specify parameters. The basic names can be preceded by "p" to give the probability distribution function, "q" to give quantiles, and "r" to generate independent random variables with the given distribution. Also, the prefix "d" gives the density for a continuous distribution and the probability mass function $\Pr(x = x)$ for a discrete distribution. Specific examples will be spelled out in some problems.

1. Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be i.i.d. samples, also independent of each other, from two distribution functions F and G. Let x be the vector $(X_1, ..., X_m)$ and $y = (Y_1, ..., Y_n)$. A test of $F = G$, sensitive to location alternatives such as $G(x) \equiv F(x - \theta)$ for some $\theta \neq 0$, is the Mann–Whitney–Wilcoxon two-sample test.performed in R by

wilcox.test(x,y).

(a) If one considers samples from distributions on disjoint half-lines, so that for some c, $F(c) = 1 = 1 - G(c)$ or $G(c) = 1 = 1 - F(c)$ then they will be as different as possible both from the Mann–Whitney–Wilcoxon and Kolmogorov–Smirnov points of view: the supremum $D_{m,n}$ of $|F_m - G_n|$ will be 1 and the Mann–Whitney statistic W will be either as small as possible (0) or as large as possible (mn). Find the probability, under the null hypothesis H_0 that the X_i and Y_j are all i.i.d. with the same continuous distribution function F, of getting such samples, as a function of m and n, for a 2-sided test where we want to find the probability that either all Y's are smaller than all X's or all Y's are larger than all X's.

(b) For $m = n$, what is the smallest n for which the probability is less than 0.05, and then what is it?

2. (a) Now consider samples x = runif(n) [giving U[0,1] variables] and y = runif(n,0.5,1.5), or equivalently y = runif(n) + 0.5 because if in R you add a number to a vector, the number is added to all components of the vector. Anyhow, the two intervals now overlap in just half the length of each. For n only moderately large, the Wilcoxon test tends

Date: 18.465, Feb. 2, 2015.

to reject equality of the distributions at the 0.05 level more often than the Kolmogorov–Smirnov test does. I tried $n = 10$ 4 times and the Wilcoxon test rejected equality of the distributions every time and the Kolmogorov–Smirnov test did only once (but came close twice), anyhow the Wilcoxon test gave a smaller p-value in all 4 cases. Try such an experiment for 5 pairs of samples x and y.
(b) Now, however, consider comparing a sample x = runif(n) with a sample y =
runif(n,-1,2), so we have two uniform distributions with the same mean (namely 1/2). The Wilcoxon test has a poor chance of distinguishing these distributions, even for n large, since they don't differ in location but rather in scale. But as n gets large, the Kolmogorov–Smirnov test can easily distinguish them. What is $\sup_x |(F - G)(x)|$ for the two distribution functions F and G?
(c) For $n = 40$, do the two tests, comparing 5 independent pairs (x,y) of samples as in part (b).

3. (a) For $n = 30$, 35, and 40, consider:
(i) The 0.95 quantile ($\alpha = 0.05$) of the 1-sample Kolmogorov statistic as given accurate to 3 decimal places in
www-math.mit.edu/∼rmd/465/onesamplequants
(ii) the approximation $1.36/\sqrt{n}$ proposed by Hollander and Wolfe, also rounded to 3 places;
(iii) the approximation $1.358/\sqrt{n}-0.144/n$ obtained by regression with fixed intercept in subsection 1.1 of the handout "Kolmogorov–Smirnov and Mann–Whitney–Wilcoxon tests," also rounded to 3 places.

As (iii) is more complicated that (ii) one might expect it to give a better approximation to (i) than (ii) does; is that right? Further, does (iii) agree with (i) to the number of significant digits given in (i)?

(b) Find constants A and B for an approximation $An^{-1/2} + B/n$ for the 0.99 quantile of D_n by the same regression method as for the 0.95 quantile. The quantiles are in the last "onepctq" column of the table "onesamplequants" on www-math.mit.edu/∼rmd/465. Use all 6 values of n in the table, as in the case $q = 0.95$. See the handout "Suggestions for using R," www-math.mit.edu/∼rmd/465/rathena
Answer the analogous questions to those in part (a).

4. For $m = 19$ and $n = 20$, as mentioned in the next to last paragraph of Section 2 of the handout "EDF Tests...," there is a value of $D_{m,n}$, namely 152/380, where $380 = 19 \cdot 20$ is the least common

multiple of 19 and 20 because they are relatively prime, such that under $H_0: \ F = G$ continuous, the probability of observing this large a value of $D_{m,n}$ or larger is $\doteq 0.0503$ (given in Table A.10 of Hollander and Wolfe, although that is not accessible to most of us). So one can perform the test in this case at very close to the 0.05 level, unlike the $m = n = 20$ case. But, how close is the probability to the limiting distribution given in (3) of the handout? For that we'd take $M = \sqrt{mn/(m+n)}D_{m,n}$ and evaluate $2\sum_{j=1}^{\infty}(-1)^{j-1}\exp(-2j^2M^2)$. A few terms should suffice to approximate the sum. R source code is provided in www-math.mit.edu/∼rmd/465/supabsbt which you can use as follows: download the code into your R directory. Then get into R and upload the code by typing

source("supabsbt")

Then

supabsbt(M,k)

will compute the kth partial sum of the above series. See what you get for $k = 1, 2, 3$ (or more if necessary).

(a) If the partial sums stop changing when you add another term, the series has converged and you have your answer. What is it for the given M?

(b) How close is it to 0.05?

(c) In a "correction for continuity" we could replace 152 by 152.5. How would this affect the result?

5. Does the (Dvoretzky–Kiefer–Wolfowitz)–Massart inequality, given as (2) of the handout, also apply to the two-sample case? In other words, if $F = G$ is a continuous distribution function and F_m and G_n are independent empirical distribution functions based on $m + n$ total random variables i.i.d. (F), is it true for all $M > 0$ that

$\Pr\left(\sup_x \sqrt{\frac{mn}{m+n}}|(F_m - G_n)(x)| \geq M\right) \leq 2\exp(-2M^2)$?

For very small values of m and n it's easy to find the exact distribution of the statistic.

(a) Find values of M such that the inequality fails:

(i) For $m = n = 1$;

(ii) For $m = 1$ and $n = 2$;

(iii) For $m = n = 2$.

(b) Show that for $m = 1$ and $n = 4$ the inequality always holds.

6. *Extra credit.* For any positive integers m and n let $\mathrm{lcm}(m,n)$ be the least common multiple of m and n. An integer-valued form of two-sample Kolmogorov–Smirnov statistic is $J := \mathrm{lcm}(m,n)\sup_x |(F_m -$

$G_n)(x)|$. In Hollander and Wolfe Table A.10 for the Kolmogorov–Smirnov 2-sample statistic, for $m = 19$ and $n = 20$, $P_0\{J \geq x\}$ is given for some integer values of x but not for some intermediate integer values. Specifically, values are not given for $153 \leq x \leq 159$, $172 \leq x \leq 179$, and some other ranges of x between values that are included. Show that under H_0: $F = G$ continuous, $\Pr(J = x) = 0$ for $153 \leq x \leq 159$ although not for $x = 152$ or 160. Some number theory may be needed.