Stat 512
Emerson, Fall 15
Homework #4
October 23 2015

Written solutions to the homework problems are due on Friday October 30, 2015 at the beginning of class.

The homework problems are divided into "regular" and "more involved" problems. In order to facilitate multiple graders, you should hand in these categories of problems separately. That is, hand in one paper that contains only the "regular" problems, and another paper that contains only the "more involved" problems.
As noted on the syllabus and discussed during the first class, copying of homework solutions is not allowed and, when detected, will be investigated as an infraction of the academy integrity policy of the University of Washington. While it is permissible to discuss problems with other students, TAs, or the instructor in order to learn how to solve a problem, your written solutions must be prepared without directly referencing any notes or solutions derived from other students or sources found on the internet.

REGULAR PROBLEMS

1. In this problem we consider alternative approaches to handle initial measurements on subjects used in a randomized experiment.

   - We consider a problem in which we randomly select n independent pairs of two independent individuals from a population, so all individuals are totally independent. (In terms of statistical design, the pairs constitute "blocks".)
   - On each individual, we make initial measurements of some random variable X_{ik} having mean μ and variance σ^2 for $i = 1, 2$ and $1 \leq k \leq n$.
   - We then apply treatment A to the first chosen individual of each pair, and make again measure the random variable which for notational convenience we label Y_{1k} having mean $\mu + \delta_A$ and variance σ^2. (So X and Y represent the same scientific quantity measured at different times.) We apply treatment B to the second chosen individual, and make measurement Y_{2k} having mean $\mu + \delta_B$ and variance σ^2. (In statistical design, we would randomize which individual of each block receives treatment A and which receives treatment B.)
   - We consider that repeat measurements are being made on the same individual are potentially correlated, but measurements made on different individuals are independent, with $\text{corr}(X_{ik}, Y_{i'k'}) = \rho \mathbf{1}_{[i=i' \,\&\, k=k']}$ for $i = 1, 2$.

Ultimately, we are interested in the difference in treatment effects as measured by $\theta = \delta_A - \delta_B$.

(a) Write down the mean and variance of random vector $W_k = (X_{1k}, Y_{1k}, X_{2k}, Y_{2k})^T$.

(b) Define $Z_k^{(0)} = Y_{1k} - Y_{2k}$. Express $Z_k^{(0)}$ as a linear transformation of W_k for some vector $\vec{c}_{(0)}$ (so $Z_k^{(0)} = \vec{c}_{(0)}^T W_k$), and provide the mean and variance of $Z_k^{(0)}$. (This approach completely ignores the initial measurements.)

(c) Define differences $D_{ik} = Y_{ik} - X_{ik}$ for $i = 1, 2$ and $1 \leq k \leq n$. Find the mean and variance for D_{ik}. Now define $Z_k^{(1)} = D_{1k} - D_{2k}$. Express $Z_k^{(1)}$ as a linear transformation of W_k for some vector $\vec{c}_{(1)}$ (so $Z_k^{(1)} = \vec{c}_{(1)}^T W_k$), and provide the mean and variance of $Z_k^{(1)}$. (This approach considers changes in the measurements.)

(d) For some specified a, define transformations $G_{ik} = Y_{ik} - aX_{ik}$ for $i = 1, 2$ and $1 \leq k \leq n$. Find the mean and variance for G_{ik}. Now define $Z_k^{(a)} = G_{1k} - G_{2k}$. Express $Z_k^{(a)}$ as a linear transformation of W_k for some vector $\vec{c}_{(a)}$ (so $Z_k^{(a)} = \vec{c}_{(a)}^T W_k$), and provide the mean and variance of $Z_k^{(a)}$. (This approach completely considers arbitrary linear handling of the initial measurements.) In what sense is this result a generalization of the previous two approaches?

(e) Now consider how averages across blocks might serve as estimators of θ. That is, consider the distributions of

$$\overline{Z}^{(*)} = \frac{1}{n} \sum_{k=1}^{n} Z_k^{(*)}$$

in terms of their means and variances. Find the value of a such that $\overline{Z}^{(a)}$ would have the greatest precision in terms of variance. Find the value of a such that $\overline{Z}^{(a)}$ would have the lowest mean squared error (MSE) as an estimator of θ, where

$$MSE_\theta = E\left[(\overline{Z}^{(a)} - \theta)^2\right].$$

2. For each of the following distributions derive the mean, variance, skewness, and kurtosis.

(a) Poisson: $X \sim \mathcal{P}(\lambda)$

(b) Exponential: $X \sim \mathcal{E}(\lambda)$

(c) Normal: $X \sim \mathcal{N}(\mu, \sigma^2)$

3. Let X_1, X_2 be independent, identically distributed random variables. Find the density for random variable Y for the following combinations of distributions and transformations.

(a) X_i has an exponential distribution with mean μ, and $Y = X_1^2 + X_2^2$

(b) $X_i \sim \mathcal{N}(\mu, \sigma^2)$, and $Y = X_1^2 + X_2^2$

(c) $X_i \sim \mathcal{U}(0, \theta)$, , and $Y = \log(X_1 + X_2)$

4. For each of the hierarchical models, find the density, mean, and variance of Y.

(a) $X \sim \mathcal{N}(\nu, \tau^2)$ and $Y|X = x \sim \mathcal{N}(x, \sigma^2)$

(b) $X \sim \mathcal{U}(0, 1)$ and $Y|X = x \sim \mathcal{B}(n, x)$

(c) $X \sim \mathcal{P}(\lambda)$ and $Y|X = x \sim \mathcal{B}(x, p)$

(d) $X \sim \mathcal{E}(\lambda)$ and $Y|X = x \sim \mathcal{E}(x)$ (use the hazard parameterization)

MORE INVOLVED PROBLEMS

5. We consider a sequential experiment in which we have potential observations X_1 and X_2 which are independent and identically distributed $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Our sequential sampling plan is as follows: We observe X_1, and if, for some prespecified $a < b$, $X_1 \leq a$ or $X_1 \geq b$, we stop. Otherwise we continue sampling to observe X_2. At the end of our experiment, we have the bivariate sequential test statistic

$$(M, S) = \begin{cases} (1, X_1) & X_1 \leq a \text{or} X_1 \geq b \\ (2, X_1 + X_2) & otherwise. \end{cases}$$

In order to estimate the unknown mean μ, we use the observed sample mean $\hat{\mu} = S/M$.

(a) Find the density for (M, S). (This cannot be solved in closed form, so it is sufficient to write down the integral you would use to find it.)

(b) Suppose $\mu = 0$, $\sigma^2 = 1$, $a = 0$, $b = 2.7897$.

i. Find $Pr[M = 1, S \leq a]$.

ii. Find $Pr[M = 1, S \geq b]$.

iii. Find $Pr[M = 2, S \leq 0]$.

iv. Find the value of c such that $Pr[M = 1, S \geq b] + Pr[M = 2, S \geq c] = 0.025$

(c) Derive a formula for the expected value for $\hat{\mu}$ in the general case. Under what conditions will $E[\hat{\mu}] = \mu$? If an estimator $\hat{\mu}$ satisfies $E[\hat{\mu}] = \mu$, $\forall \mu$, we call that estimator unbiased. Under what conditions on our sampling plan is $\hat{\mu}$ unbiased?

6. Consider a sample of independent, identically distributed continuous random variables $X_i \sim F_X$, $i = 1, \ldots, n$, and a sample of independent, identically distributed continuous random variables $Y_i \sim F_Y$, $i = 1, \ldots, m$, where the X's and the Y's are also totally independent. We are interested in whether the X_i's have the same distribution as the Y_i's, and we choose consider estimating the probability θ that a randomly chosen X will be greater than a randomly chosen Y.

(a) Provide a formula for θ in terms of the joint distribution of (X_i, Y_j).

(b) Let $U_{ij} = 1_{[X_i \geq Y_j]}$. Find the probability distribution for U_{ij}.

(c) What is the mean and variance for U_{ij}?

(d) What is the covariance $Cov(U_{ij}, U_{k\ell})$ for arbitrary i, j, k, ℓ? (Be sure to consider cases when $i = k$ and/or $j = \ell$.)

(e) Are the random variables $\{U_{ij} : 1 \leq i \leq n; 1 \leq j \leq m\}$ identically distributed? Are they totally independent?

(f) Now consider the statistic

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} 1_{[X_i \geq Y_j]}.$$

Find the expectation and variance of U when $F_X(u) = F_Y(u),\ \forall u \in \mathcal{R}^1$

7. Consider again the setting of the previous problem, but that we transform all of the random variables from the scale they were originally measured on to their ranks in the combined sample

$$R_i = rank(X_i) = \sum_{j=1}^{n} 1_{[X_j \leq X_i]} + \sum_{j=1}^{m} 1_{[Y_j \leq X_i]}$$
$$S_i = rank(Y_i) = \sum_{j=1}^{n} 1_{[X_j \leq Y_i]} + \sum_{j=1}^{m} 1_{[Y_j \leq Y_i]}$$

and define

$$R = \sum_{i=1}^{n} R_i.$$

(a) Find the mean and variance of R when $F_X(u) = F_Y(u),\ \forall u \in \mathcal{R}^1$. (Hint: Under the null hypothesis, R has the same distribution as the sum of m numbers randomly chosen without replacement from the integers $\{1, \ldots, m + n\}$.

(b) Find the correlation between U from the previous problem and R as defined in this problem.