

CBB 200 Problem set 1

Due Feb 17

In the following problem set solve:

3 from category A

2 from category B

1 from category C

Pr. A.1 The random variable Y has a Poisson distribution with parameter λ . Compute

$$\mathbb{E}[Y(Y-1)].$$

Generalize this result to the r -th factorial moment

$$\mathbb{E}[Y(Y-1)\cdots(Y-r+1)],$$

for $r \leq n$.

Pr. A.2 You are given the integers $(1, \dots, n)$. You then randomly permute this sequence. After this random permutation the probability that an integer i will match its original position is n^{-1} . If A_i is the event that the integer i matches its original position the

$$\begin{aligned}\Pr(A_i) &= n^{-1}, \\ \Pr(A_i A_j) &= (n(n-1))^{-1}, \\ \Pr(A_i A_j A_k) &= (n(n-1)(n-2))^{-1},\end{aligned}$$

etc...

Show that the probability that no number after the permutation is in its original position is

$$1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^n \frac{1}{n!}.$$

Compute the limit of the above probability as $n \rightarrow \infty$.

Pr A.3 Assume Y_1, \dots, Y_n are independent random variables and each Y_i is Poisson distributed with parameter λ_i . Compute the probability distribution of the sum $S_n = \sum_{i=1}^n Y_i$ of these random variables (it should be Poisson with $\lambda = \sum_{i=1}^n \lambda_i$).

Pr A.4 Given a DNA sequence consisting of 10 consecutive nucleotides. Three segments of this sequence are chosen at random: one consisting of 3 consecutive nucleotides, a second consisting of 4 consecutive nucleotides, the third consisting of 5 consecutive nucleotides. By “random” we mean that the segment of 3 nucleotides can be in any of the 8 possible positions with equal probability. Let Y be the number of positions (out of 10) that are in all three segments, so $Y = 0, 1, 2, 3$. What is $\mathbb{E}Y$?

Pr A.5 This is a problem related to linkage analysis. A parent with genetic type Mm has three children. The parent transmits the gene M to each child with probability $\frac{1}{2}$ and the genes that are transmitted to each of the three children are independent. Let $I_1 = 1$ if children 1 and 2 have the same gene transmitted (for example both receive M or m), and $I_1 = 0$ otherwise. Similarly, $I_2 = 1$ if children 1 and 3 have the same gene transmitted and is 0 otherwise, and let $I_3 = 1$ if children 2 and 3 have the same gene transmitted and is 0 otherwise.

1. Show that these three random variables are pairwise independent but not independent.
2. Show that the variance of $I_1 + I_2 + I_3$ is the sum of the variances of I_1 and I_2 and I_3 .
3. Explain the above result in terms of the various covariances between I_1 , I_2 and I_3 and the pairwise independence of I_1 , I_2 , and I_3 .

Pr B.1 Consider the linear shrinkage estimator

$$\hat{\theta}_n = \arg \min_{\theta \in \mathbb{R}^p, \theta_0 \in \mathbb{R}} \left[\sum_{i=1}^n (y_i - x_i \cdot \theta - \theta_0)^2 + \lambda \|\theta\|^2 \right],$$

where $\lambda > 0$ is a parameter of the problem and $\|\theta\|^2 = \sum_{i=1}^p \theta_i^2$. Show that the above minimization is equivalent to the following

$$\hat{\theta}_n^c = \arg \min_{\theta \in \mathbb{R}^p} \left[\sum_{i=1}^n (y_i - (x_i - \bar{x}) \cdot \theta^c)^2 + \lambda \|\theta^c\|^2 \right],$$

where \bar{x} is the p -dimensional vector with $\bar{x}_k = \frac{1}{n} \sum_{j=1}^n x_{kj}$ which is the average value of the k -th element over the n observations. The superscript c stands for centered.

Pr. B.2 Show that the linear shrinkage estimator is the mean and mode of the posterior distribution under a Gaussian prior $\theta \sim N(0, \tau \mathbf{I})$ and a Gaussian sampling model $\mathbf{y} \sim N(\mathbf{X}\theta, \sigma^2 \mathbf{I})$. Find the relationship between λ and the variances τ and σ^2 .

Pr. B.3 Assume that $y_i \sim N(\theta_0 + x_i \cdot \theta, \sigma^2)$ for $i = 1, \dots, n$ and the parameters θ_k for $k = 1, \dots, p$ are each distributed as $N(0, \tau)$ and independent. Assuming that τ, σ^2 are known show that the log-posterior density of θ is proportional to

$$\sum_{i=1}^n (y_i - \theta_0 - x_i \cdot \theta)^2 + \lambda \sum_{k=1}^p \theta_k^2,$$

where $\lambda = \frac{\sigma^2}{\tau}$.

Pr C.1 This problem will be used to illustrate the invariance of rank statistics. It is a programming assignment. We will consider the following distributions

$$\begin{aligned} X &\sim N(0, 1) = F_N \\ X &\sim U[0, 1] = F_U \\ X &\sim \Gamma(\lambda = 1, k = 4) = F_G. \end{aligned}$$

We will look at two sampling methods.

- Sampling with replacement. Define the following procedure $\text{Draw}(dist)$:

$$\begin{aligned} X_1, \dots, X_m &\sim F_{dist} \\ Y_1, \dots, Y_n &\sim F_{dist}, \end{aligned}$$

which corresponds to drawing two sets of m and n observations respectively from one of the three distributions $dist = N, U, G$. We will compute one of two statistics

$$\begin{aligned} d^{KS} &= \sup_x |F_n(x) - F_m(x)| \\ d^{MW} &= \frac{1}{n} \sum_{i=1}^n F_n(x_i) - \frac{1}{m} \sum_{i=1}^m F_m(x_i). \end{aligned}$$

1. $m = 50, n = 20$
 2. For $k = 1, \dots, 10^5$:
 - (a) $\text{Draw}(N)$
 - (b) compute $d_k^{KS,N}$ and $d_k^{MW,N}$
 - (c) $\text{Draw}(U)$
 - (d) compute $d_k^{KS,U}$ and $d_k^{MW,U}$
 - (e) $\text{Draw}(G)$
 - (f) compute $d_k^{KS,G}$ and $d_k^{MW,G}$
 3. Given the six sequences $\{d_k^{KS,N}\}, \{d_k^{MW,N}\}, \{d_k^{KS,U}\}, \{d_k^{MW,U}\}, \{d_k^{KS,G}\},$ and $\{d_k^{MW,G}\}$ plot a histogram for each sequence.
- Sampling without replacement or permuting. Define the following procedure $\text{Draw}(dist)$:

$$\begin{aligned} X_{dist} = X_1, \dots, X_m &\sim F_{dist} \\ Y_{dist} = Y_1, \dots, Y_n &\sim F_{dist}, \end{aligned}$$

which corresponds to drawing two sets of m and n observations respectively from one of the three distributions $dist = N, U, G$. Define the procedure $\text{Permute}(X, Y)$ as randomly permuting the membership of the elements in sets X, Y while preserving the number of elements in each set.

We will compute one of two statistics

$$\begin{aligned} d^{KS} &= \sup_x |F_n(x) - F_m(x)| \\ d^{MW} &= \frac{1}{n} \sum_{i=1}^n F_n(x_i) - \frac{1}{m} \sum_{i=1}^m F_m(x_i). \end{aligned}$$

1. $m = 50, n = 20$
2. $(X_N, Y_N) = \text{Draw}(N)$
3. $(X_U, Y_U) = \text{Draw}(U)$
4. $(X_G, Y_G) = \text{Draw}(G)$
5. For $k = 1, \dots, 10^5$:
 - (a) $\text{Permute}(X_N, Y_N)$
 - (b) compute $d_k^{KS,N}$ and $d_k^{MW,N}$

- (c) $\text{Permute}(X_U, Y_U)$
 - (d) compute $d_k^{KS,U}$ and $d_k^{MW,U}$
 - (e) $\text{Permute}(X_G, Y_G)$
 - (f) compute $d_k^{KS,G}$ and $d_k^{MW,G}$
6. Given the six sequences $\{d_k^{KS,N}\}$, $\{d_k^{MW,N}\}$, $\{d_k^{KS,U}\}$, $\{d_k^{MW,U}\}$, $\{d_k^{KS,G}\}$, and $\{d_k^{MW,G}\}$ plot a histogram for each sequence.

What did we learn about the dependence of the rank statistics with respect to distributions and sampling procedures ?

What does this problem have to do with hypothesis testing ?