ISyE 6404 – Nonparametric Statistics

Lecture #1 (08/21/18) – Order Statistics (Textbook Chapter 5)

1. Introduction/Motivating Examples (page 76)

Suppose a researcher is sent out to Leech Lake, Minnesota, to ascertain the average weight of Walleye fish caught from that lake. She obtains her data by stopping the fishermen as they are returning to the dock after a day of fishing. In the time the researcher waited at the dock, three fishermen arrived, each with their daily limit of three Walleye. Because of limited time, she only has time to make one measurement with each fisherman, so at the end of her field study, she will get three measurements.

McIntyre (1952) discovered that with this forced limitation on measurements, there is an efficient way of getting information about the population mean. We might assume the researcher selected the fish to be measured randomly for each of the three fishermen that were returning to shore. McIntyre found that if she instead inspected the fish visually and selected them non-randomly, the data could beget a better estimator for the mean. Specifically, suppose the researcher examines the three Walleye from the first fisherman and selects the smallest one for measurement. She measures the second smallest from the next batch, and the largest from the third batch.

Opposed to a **simple random sample (SRS)**, this **ranked set sample (RSS)** consists of independent **order statistics** which we will denote by X[1:3], X[2:3], X[3:3]. If X-bar is the sample mean from a SRS of size n, and XRSS-bar is the mean of a ranked set sample X[1:n],...,X[n:n], it is easy to show that like X-bar, XRSS-bar is an **unbiased estimator** of the population mean. Moreover, it has **smaller variance**. That is, Var(XRSS-bar) ≤ Var(X-bar). The key is that variances for order statistics are generally smaller than the variance of the i.i.d. measurements. If you think about the SRS estimator as a linear combination of order statistics, it differs from the linear combination of order statistics from a RSS by its **covariance structure**.

The sampling aspect of RSS has received the most attention. Estimators of other parameters can be constructed to be more efficient than SRS estimators, including nonparametric estimators of the CDF (Stokes and Sager, 1988).

This chapter presents distributions of single or multiple order-statistics. Our lecture will emphasize on how to apply the results given below (copied from the textbook) to solve real-life problems. Examples of Applications of Order-statistics include:

a) Estimation of Population Distribution Parameters Based on Type-II Censored Data:

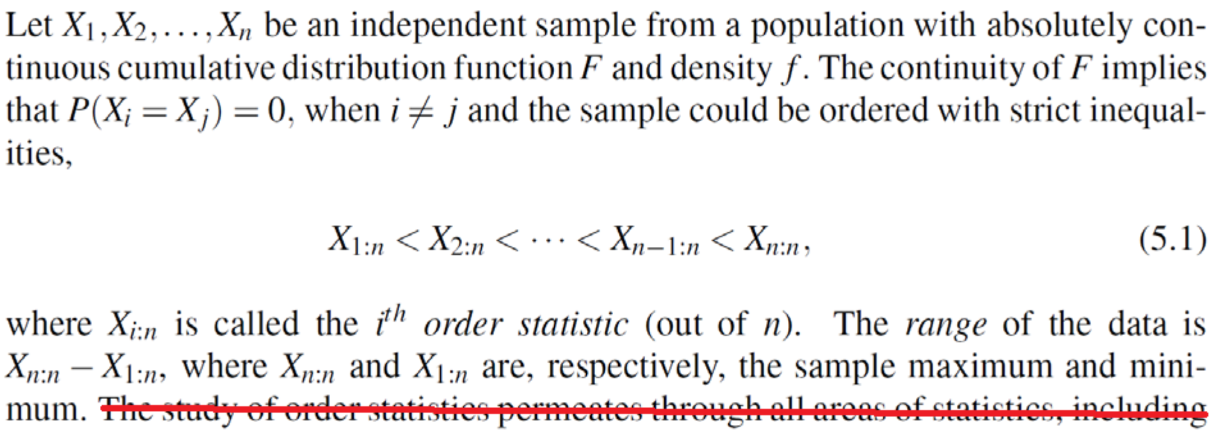
In Sarhan and Greenberg (1956) 10 individual patients were taken systolic blood pressures (SBP). For some technical reasons 20% of smallest and 30% of largest observations are censored leaving only five observations. How to estimate mean and standard deviation of the population? Can one perform statistical inference including hypothesis-testing and confidence interval for the population parameters?

Remark #1 (Accelerated Life Tests): Due to limited budget of time or cost, in testing of high-reliability products (called life- or reliability-testing), Type-II censored are commonly seen. The data were collected to provide statistical inference for "certifying" product-reliability of comparing new against old products for a significant improvement.

Remark #2 (Potential Enrichment Project): Type-II censored studies have been done very long time ago. Students might want to work on an enrichment project in exploring whether there are recent applications of order-statistics. Do they need any information other than the distribution results described below?

Remark #3 (potential computing/simulation project): Imagine that some practitioners do not have the ability to comprehend or derive the distributional results for order-statistics. Will simulation be possible to construct the needed distribution for statistical inference? What assumption(s) is needed? What are the general steps for making the simulation possible? Note that when *certain assumptions in the derivation of distributional results for order-statistics might be violated, simulation might be the only way to get the results*. Thus, this might be a potential computing/simulation project.

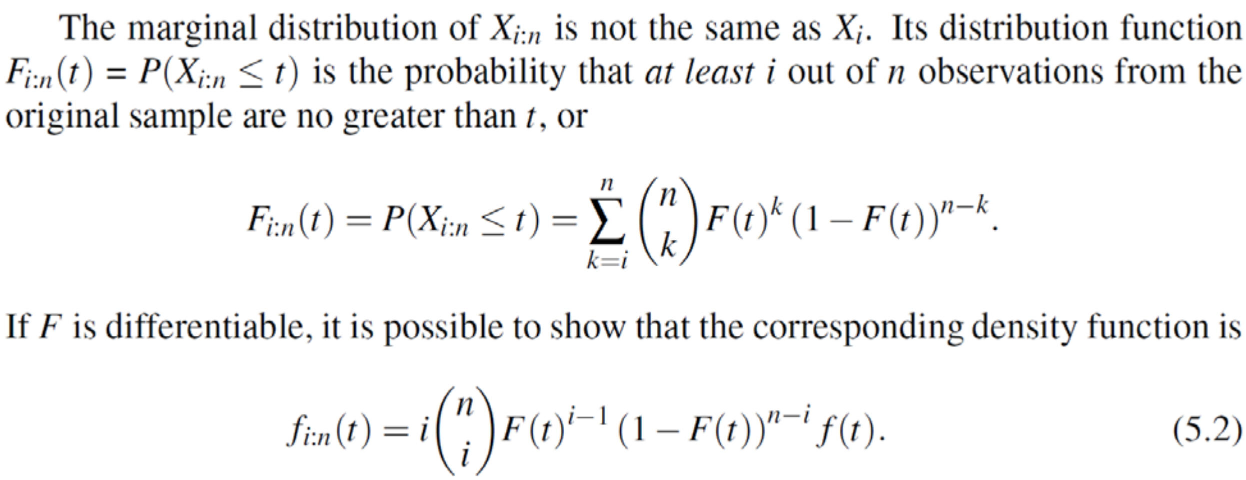
1. Order-Statistics: Random-Variables and Their Joint-Distributions



Let $X_1, X_2, \ldots, X_n$ be an independent sample from a population with absolutely continuous cumulative distribution function F and density f. The continuity of F implies that $P(X_i = X_j) = 0$, when $i \neq j$ and the sample could be ordered with strict inequalities,

$$X_{1:n} < X_{2:n} < \cdots < X_{n-1:n} < X_{n:n}, \tag{5.1}$$

where $X_{i:n}$ is called the i^{th} *order statistic* (out of n). The *range* of the data is $X_{n:n} - X_{1:n}$, where $X_{n:n}$ and $X_{1:n}$ are, respectively, the sample maximum and minimum. ~~The study of order statistics permeates through all areas of statistics, including~~

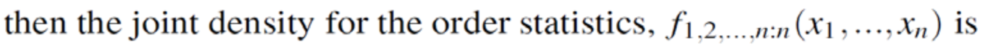


The marginal distribution of $X_{i:n}$ is not the same as X_i. Its distribution function $F_{i:n}(t) = P(X_{i:n} \le t)$ is the probability that *at least* i out of n observations from the original sample are no greater than t, or

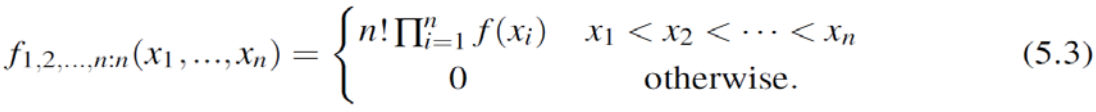
$$F_{i:n}(t) = P(X_{i:n} \le t) = \sum_{k=i}^{n} \binom{n}{k} F(t)^k (1 - F(t))^{n-k}.$$

If F is differentiable, it is possible to show that the corresponding density function is

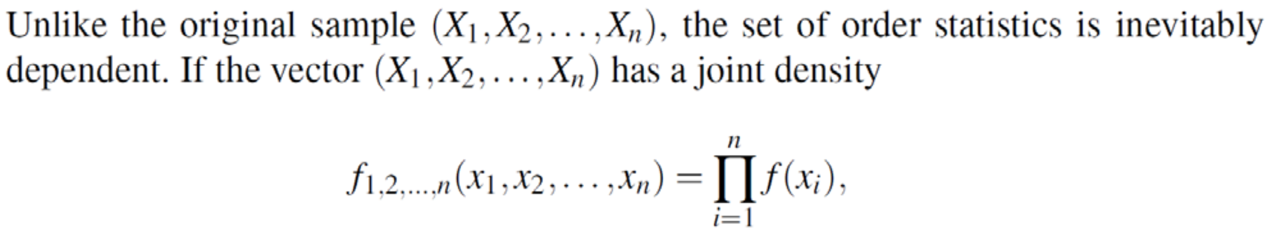
$$f_{i:n}(t) = i \binom{n}{i} F(t)^{i-1} (1 - F(t))^{n-i} f(t). \tag{5.2}$$



then the joint density for the order statistics, $f_{1,2,\ldots,n:n}(x_1, \ldots, x_n)$ is



$$f_{1,2,\ldots,n:n}(x_1, \ldots, x_n) = \begin{cases} n! \prod_{i=1}^{n} f(x_i) & x_1 < x_2 < \cdots < x_n \\ 0 & \text{otherwise.} \end{cases} \tag{5.3}$$



Unlike the original sample $(X_1, X_2, \ldots, X_n)$, the set of order statistics is inevitably dependent. If the vector $(X_1, X_2, \ldots, X_n)$ has a joint density

$$f_{1,2,\ldots,n}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i),$$

From (5.3) we can obtain the distribution of any subset of order statistics. The joint distribution of $X_{r:n}, X_{s:n}$, $1 \leq r < s \leq n$ is defined as

$$F_{r,s:n}(x_r, x_s) = P(X_{r:n} \leq x_r, X_{s:n} \leq x_s),$$

which is the probability that at least r out of n observations are at most x_r, *and* at least s of n observations are at most x_s. The probability that *exactly* i observations are at most x_{r} and j are at most x_s is

$$\frac{n!}{(i-1)!(j-i)!(n-j)!} F(x_r)^i \left(F(x_s) - F(x_r)\right)^{j-i} \left(1 - F(x_s)\right)^{n-j},$$

where $-\infty < x_r < x_s < \infty$; hence

$$\begin{aligned} F_{r,s:n}(x_r, x_s) \quad &= \quad \sum_{j=s}^{n} \sum_{i=r}^{s} \frac{n!}{(i-1)!(j-i)!(n-j)!} \times \\ &\quad F(x_r)^i \left(F(x_s) - F(x_r)\right)^{j-i} \left(1 - F(x_s)\right)^{n-j}. \qquad (5.4) \end{aligned}$$

If F is differentiable, it is possible to formulate the joint density of two order statistics as

$$\begin{aligned} f_{r,s:n}(x_r, x_s) \quad &= \quad \frac{n!}{(r-1)!(s-r-1)!(n-s)!} \times \qquad (5.5) \\ &\quad F(x_r)^{r-1} \left(F(x_s) - F(x_r)\right)^{s-r-1} \left(1 - F(x_s)\right)^{n-s} f(x_r) f(x_s). \end{aligned}$$

Sample Range. The range of the sample, R, defined before as $X_{n:n} - X_{1:n}$, has density

$$f_R(u) = \int_{-\infty}^{\infty} n(n-1)[F(v) - F(v-u)]^{n-2} f(v-u) f(v) dv. \qquad (5.6)$$

5.2 Sample Quantiles

Recall that for a distribution F, the p^{th} quantile (x_p) is the value x such that $F(x) = p$, if the distribution is continuous, and more generally, such that $F(x) \geq p$ and $P(X \geq x) \geq 1-p$, if the distribution is arbitrary. For example, if the distribution F is discrete, there may not be any value x for which $F(x) = p$.

Analogously, if $X_1, \ldots, X_n$ represents a sample from F, the p^{th} *sample quantile* $(\hat{x}_p)$ is a value of x such that $100p\%$ of the sample is smaller than x. This is also called the $100p\%$ *sample percentile*. With large samples, there is a number $1 \leq r \leq n$ such that $X_{r:n} \approx x_p$. Specifically, if n is large enough so that $p(n+1) = r$ for some $r \in \mathbb{Z}$, then $\hat{x}_p = X_{r:n}$ because there would be $r-1$ values smaller than $\hat{x}_p$ in the sample, and $n-r$ larger than it.

5.4 Asymptotic Distributions of Order Statistics

Let $X_{r:n}$ be r^{th} order statistic in a sample of size n from a population with an absolutely continuous distribution function F having a density f. Let $r/n \to p$, when $n \to \infty$. Then

$$\sqrt{\frac{n}{p(1-p)}} f(x_p)(X_{r:n} - x_p) \Longrightarrow \mathcal{N}(0,1),$$

where x_p is p^{th} quantile of F, i.e., $F(x_p) = p$.

Let $X_{r:n}$ and $X_{s:n}$ be r^{th} and s^{th} order statistics $(r < s)$ in the sample of size n. Let $r/n \to p_1$ and $s/n \to p_2$, when $n \to \infty$. Then, for large n,

$$\begin{pmatrix} X_{r:n} \\ X_{s:n} \end{pmatrix} \overset{\text{appr}}{\sim} \mathcal{N}\left(\begin{bmatrix} x_{p_1} \\ x_{p_2} \end{bmatrix}, \Sigma \right),$$

where

$$\Sigma = \begin{bmatrix} p_1(1-p_1)[f(x_{p_1})]^{-2}/n & p_1(1-p_2)/[nf(x_{p_1})f(x_{p_2})]^{-1} \\ p_1(1-p_2)/[nf(x_{p_1})f(x_{p_2})]^{-1} & p_2(1-p_2)[f(x_{p_2})]^{-2}/n \end{bmatrix}$$

and x_{p_i} is p_i^{th} quantile of F.

EXAMPLE 5.8

Let $r = n/2$ so we are estimating the population median with $\hat{x}_{.50} = x_{(n/2):n}$. If $f(x) = \theta \exp(-\theta x)$, for $x > 0$, then $x_{0.50} = \ln(2)/\theta$ and

$$\sqrt{n}\left(\hat{x}_{0.50} - x_{0.50}\right) \Longrightarrow \mathcal{N}\left(0, \theta^{-2}\right).$$

5.5 Extreme Value Theory

Earlier we equated a series system lifetime (of n i.i.d. components) with the sample minimum $X_{1:n}$. The limiting distribution of the minima or maxima are not so interesting, e.g., if X has distribution function F, $X_{1:n} \to x_0$, where $x_0 = \inf_x\{x : F(x) > 0\}$. However, the *standardized limit* is more interesting. For an example involving sample maxima, with $X_1, ..., X_n$ from an exponential distribution with mean 1, consider the asymptotic distribution of $X_{n:n} - \log(n)$:

$$\begin{aligned} P(X_{n:n} - \log(n) \le t)) &= P(X_{n:n} \le t + \log(n)) = [1 - \exp\{-t - \log(n)\}]^n \\ &= [1 - e^{-t} n^{-1}]^n \to \exp\{-e^{-t}\}. \end{aligned}$$

This is because $(1 + \alpha/n)^n \to e^{\alpha}$ as $n \to \infty$. This distribution, a special form of the Gumbel distribution, is also called the *extreme-value distribution*.

Extreme value theory states that the standardized series system lifetime converges to one of the three following distribution types F^* (not including scale and location transformation) as the number of components increases to infinity:

$$\begin{array}{lll} \text{Gumbel} & F^*(x) = & \exp(-\exp(-x)), \quad -\infty < x < \infty \\ \\ \text{Fréchet} & F^*(x) = & \begin{cases} \exp(-x^{-a}), & x > 0,\ a > 0 \\ 0, & x \le 0 \end{cases} \\ \\ \text{Negative Weibull} & F^*(x) = & \begin{cases} \exp(-(-x)^{a}), & x < 0,\ a > 0 \\ 0, & x \ge 0 \end{cases} \end{array}$$