Probability Theory
Lecture 4

Mattias Villani

Division of Statistics
Dept. of Computer and Information Science
Linköping University

Overview Lecture 4

- Order statistics
- Probability in data mining

Order statistics

- Finding the distribution of **extremes**:
  - $\min(X_1, X_2, ..., X_n)$
  - $\max(X_1, X_2, ..., X_n)$.
- **Median**: $Pr(X \leq m) = 1/2$.
- **Range**: $R = \max(X_1, X_2, ..., X_n) - \min(X_1, X_2, ..., X_n)$.

DEF The **kth order variable**

$$X_{(k)} = \text{the } k\text{th smallest of } X_1, X_2, ..., X_n$$

DEF The order statistics: $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$.

- Even if the original sample $X_1, X_2, ..., X_n$ are independent, their order statistics $X_{(1)}, X_{(2)}, ..., X_{(n)}$ are not clearly not.

Distribution of the maximum

TH The distribution of the maximum $X_{(n)}$

$$F_{X_{(n)}}(x) = P\left(X_1 \leq x, X_2 \leq x, ..., X_n \leq x\right) = \prod_{i=1}^{n} P(X_i \leq x) = [F(x)]^n .$$

The density of the maximum $X_{(n)}$

$$f_{X_{(n)}}(x) = n\,[F(x)]^{n-1} f(x)$$

✎ Let $X_1, ..., X_n \sim L(a)$. Find $F_{X_{(n)}}(x)$. Solution: If $X \sim L(a)$ then

$$F(x) = \begin{cases} \frac{1}{2}\exp\left(\frac{x}{b}\right) & \text{if } x < 0 \\ 1 - \frac{1}{2}\exp\left(-\frac{x}{b}\right) & \text{if } x \geq 0 \end{cases}$$

so

$$F_{X_{(n)}}(x) = [F(x)]^n = \begin{cases} \frac{1}{2^n}\exp\left(\frac{nx}{b}\right) & \text{if } x < 0 \\ \left[1 - \frac{1}{2}\exp\left(-\frac{x}{b}\right)\right]^n & \text{if } x \geq 0 \end{cases}$$

Distribution of the minimum

TH The distribution of the minimum $X_{(1)}$

$$F_{X_{(1)}}(x) = 1 - P\left(X_{(1)} > x\right)$$
$$= 1 - P\left(X_1 > x, X_2 > x, \ldots, X_n > x\right)$$
$$= 1 - \prod_{i=1}^{n} P(X_i > x) = 1 - \left[1 - F(x)\right]^n.$$

The density of the minimum $X_{(n)}$

$$f_{X_{(1)}}(x) = n\left[1 - F(x)\right]^{n-1} f(x)$$

✏ Let $X_1, \ldots, X_n \sim Exp(a)$. What is $f_{X_{(1)}}(x)$ and $E(X_{(1)})$?
Solution: We have

$$F(x) = 1 - e^{-ax}$$

so

$$f_{X_{(1)}}(x) = n\left[e^{-ax}\right]^{n-1} ae^{-ax} = ane^{-anx}$$

so $X_{(1)} \sim Exp(an)$ and $E(X_{(1)}) = \frac{1}{an}$. [Serial electric circuits]

Marginal distribution of $X_{(k)}$

TH The distribution of the kth order variable $X_{(k)}$ from a random sample from $F(x)$:

$$F_{X_{(k)}}(x) = F_{\beta(k,n+1-k)}\left[F(x)\right]$$

where $F_{\beta(k,n+1-k)}(\cdot)$ is the cdf of a $Beta(k, n+1-k)$ variable.

✏ Let the individual jumps of n athletes in a long jump tournament be independently $U(a,b)$ distributed. What is the probability that the recorded score of the silver medalist is longer than c meters?
Solution: First, calculate the distribution of Y_i = longest jump out of three jumps for the ith athlete, for $i = 1, \ldots, n$:

$$F_{Y_i}(y) = \left[F(y)\right]^3 = \left(\frac{y-a}{b-a}\right)^3$$

Then derive $Y_{(n-1)}$

$$F_{Y_{(n-1)}}(y) = F_{\beta(n-1,2)}\left(\left(\frac{y-a}{b-a}\right)^3\right)$$

Joint distribution of the extremes and range

- So far: only *marginal* distributions of order statistics.

TH The joint density of $X_{(1)}$ and $X_{(n)}$

$$f_{X_{(1)},X_{(n)}}(x,y) = \begin{cases} n(n-1)\left(F(y) - F(x)\right)^{n-2} f(y)f(x) & \text{if } x < y \\ 0 & \text{otherwise} \end{cases}$$

- From $f_{X_{(1)},X_{(n)}}(x,y)$ we can derive the distribution of the Range $R_n = X_{(n)} - X_{(1)}$ by the transformation theorem.

TH The distribution of the **Range** $R_n = X_{(n)} - X_{(1)}$ is

$$f_{R_n}(r) = n(n-1)\int_{-\infty}^{\infty} \left(F(u+r) - F(u)\right)^{n-2} f(u+r)f(u)du$$

Joint distribution of order statistics

TH The joint density of the order statistics is

$$f_{X_{(1)},\ldots,X_{(n)}}(y_1,\ldots,y_n) = \begin{cases} n!\prod_{k=1}^{n} f(y_k) & \text{if } y_1 < y_2 < \cdots < y_n \\ 0 & \text{otherwise} \end{cases}$$

- The marginal densities of any order variable can be derived by integrating $f_{X_{(1)},\ldots,X_{(n)}}(y_1,\ldots,y_n)$ in the usual fashion.