Statistical Bioinformatics, Makerere
Extreme Values & Order Statistics

Timo Koski

TK

02.08.2018

KTH Matematik

Lecture

The lecture will cover Chapter 1 and 2.11 in Ewens and Grant in some detail.

KTH Matematik

Contents

KTH VETENSKAP OCH KONST
KTH Matematik

Contents

KTH Matematik

Minimum of n I.I.D. R.V's

$X_1, X_2, \ldots, X_n$, I.I.D with the same distribution as X. Let

$$X_{\min} \stackrel{\text{def}}{=} \min\left(X_1, X_2, \ldots, X_n\right).$$

Then

$$P\left(X_{\min} \geq x\right) = \left(P\left(X \geq x\right)\right)^n.$$

KTH Matematik

Proof:

$$P\left(X_{\min} \geq x\right) = P\left(\min\left(X_1, X_2, \ldots, X_n\right) \geq x\right) =$$

$$= P\left(X_1 \geq x, X_2 \geq x, \ldots, X_n \geq x\right) = \prod_{i=1}^{n} P\left(X_i \geq x\right),$$

by independence of $X_1, X_2, \ldots, X_n$. □

KTH Matematik

Minimum of n I.I. Exponential R.V's

Recall that $X_1, X_2, \ldots, X_n$, I.I. and $\in \mathrm{Exp}(\lambda)$

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & 0 < x < \infty \\ 0 & \text{elsewhere.} \end{cases}$$

$$F_X(x) = 1 - e^{-\lambda x}$$

$$P(X \geq x) = 1 - F_X(x) = 1 - \left(1 - e^{-\lambda x}\right) = e^{-\lambda x}.$$

KTH Matematik

Minimum of n I.I. Exponential R.V's

$X_1, X_2, \ldots, X_n$, I.I. and $\mathrm{Exp}(\lambda)$. $X_{\min} \stackrel{\text{def}}{=} \min(X_1, X_2, \ldots, X_n)$.

$$P(X_{\min} \geq x) = (P(X \geq x))^n =$$

$$= \left(e^{-\lambda x}\right)^n = e^{-n\lambda x}$$

Hence $X_{\min} \in \mathrm{Exp}(n\lambda)$.

KTH Matematik

Maximum of n I.I.D. R.V's

$X_1, X_2, \ldots, X_n$, I.I.D with the same distribution as X. Let

$$X_{\max} \overset{\text{def}}{=} \max\left(X_1, X_2, \ldots, X_n\right).$$

$$P\left(X_{\max} \leq x\right) = \left(P\left(X \leq x\right)\right)^n$$

KTH Matematik

Maximum of n I.I.D. R.V's

$$P\left(X_{\max} \leq x\right) = P\left(\max\left(X_1, X_2, \ldots, X_n\right) \leq x\right) =$$

$$= P\left(X_1 \leq x, X_2 \leq x, \ldots, X_n \leq x\right) = \prod_{i=1}^{n} P\left(X_i \leq x\right),$$

by independence of $X_1, X_2, \ldots, X_n$. □

KTH Matematik

Maximum of n I.I.D. Exponential R.V's

$X_1, X_2, \ldots, X_n$, I.I. and $\mathrm{Exp}(\lambda)$. $X_{\max} \overset{\text{def}}{=} \max(X_1, X_2, \ldots, X_n)$.

$$P(X_{\max} \leq x) = (P(X \leq x))^n =$$

$$= \left(1 - e^{-\lambda x}\right)^n.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Mean and Variance of Max. of n Exponential R.V's

Now we shall compute

$$E\left[X_{\max}\right], \operatorname{Var}\left[X_{\max}\right]$$

and an asymptotic distribution of $X_{\max}$. This will be a lengthier story, some details will be skipped. We need the memoryless property of the exponential distributions. We prove this for the geometric distribution.

Memoryless property: Geometric distribution

Recall $X \in Ge(p)$

$$f_X(x) = (1-p) \cdot p^x; x = 0, 1, \ldots,$$

We seek the conditional probability

$$P(X \geq x_1 + x_2 \mid X \geq x_1).$$

By definition

$$P(X \geq x_1 + x_2 \mid X \geq x_1) = \frac{P(X \geq x_1 + x_2)}{P(X \geq x_1)}.$$

We need the distribution function of $X \in \mathrm{Ge}(p)$.

KTH VETENSKAP OCH KONST
KTH Matematik

Distribution function of $X \in \mathrm{Ge}(p)$

$$F_X(x) = P(X \leq x) = \sum_{k=0}^{x} (1-p) \cdot p^k =$$

$$= (1-p) \cdot \sum_{k=0}^{x} p^k = (1-p) \frac{1-p^{x+1}}{(1-p)}$$

$$= 1 - p^{x+1},$$

by summing a geometric sum.

Memoryless property: Geometric distribution

Then

$$P\left(X \geq x_1 + x_2 \mid X \geq x_1\right) = \frac{P\left(X \geq x_1 + x_2\right)}{P\left(X \geq x_1\right)} = \frac{p^{x_1+x_2+1}}{p^{x_1+1}} = p^{x_2},$$

since $P\left(X \leq x\right) = 1 - p^{x+1}$. This is the memoryless property.

KTH VETENSKAP OCH KONST
KTH Matematik

Memoryless property: Exponential distribution

$X \in \mathrm{Exp}(\lambda)$. We find as above the conditional probability

$$P\left(X \geq x_1 + x_2 \mid X \geq x_1\right) = \frac{P\left(X \geq x_1 + x_2\right)}{P\left(X \geq x_1\right)} = e^{-\lambda x_2}$$

KTH Matematik

MGF

$$\frac{d}{ds}\phi(0) = E[X]$$

$X \in \text{Exp}(\lambda)$

$$\phi_X(s) = \frac{\lambda}{\lambda - s}$$

$\Rightarrow$

$$\frac{d}{ds}\phi_X(s) = -\frac{-\lambda}{[(\lambda - s)]^2}$$

Hence $E[X] = \frac{1}{\lambda}$.

KTH VETENSKAP OCH KONST
KTH Matematik

Mean and Variance of Max. of n Exponential R.V's

Now we get to computing

$$E\left[X_{\max}\right], \operatorname{Var}\left[X_{\max}\right]$$

We are going to use the fact that $X_{\min} \in \operatorname{Exp}(n\lambda)$ and the memoryless property.

KTH Matematik

Mean and Variance of Max. of n Exponential R.V's

Suppose, e.g., that n cellular proteins have lifetimes until degradation that are exponentially distributed. Then the expected time to one or other protein dies is the mean of the minimum,

$$E\left(X_{\min}\right) = \frac{1}{n\lambda},$$

since $X_{\min} \in \mathrm{Exp}\left(n\lambda\right)$. By the memoryless property we have that the mean time till next protein dies is the minimum of $\left(n-1\right)$ exponential r.v.'s independent of the first death,

$$\frac{1}{(n-1)\lambda}$$

and so on.

Mean and Variance of Max. of n Exponential R.V's

The expected time till the last of the n proteins dies is the expectation of the maximum lifetime and is thus

$$\frac{1}{\lambda} + \frac{1}{2\lambda} + \cdots + \frac{1}{n\lambda} = E\left[X_{\max}\right].$$

Similarly we can find

$$\operatorname{Var}\left[X_{\max}\right] = \frac{1}{\lambda^2} + \frac{1}{(2\lambda)^2} + \cdots + \frac{1}{(n\lambda)^2}.$$

KTH Matematik

Euler's constant

We write

$$\frac{1}{\lambda} + \frac{1}{2\lambda} + \cdots + \frac{1}{n\lambda} = \frac{1}{\lambda} \sum_{i=1}^{n} \frac{1}{i}$$

It is shown in courses in mathematics that

$$\lim_{n \to \infty} \left(\sum_{i=1}^{n} \frac{1}{i} - \log n \right) = \gamma,$$

where γ is Euler's constant, (see Westergren & Råde: BETA)

KTH VETENSKAP OCH KONST
KTH Matematik

Mean and Variance of Max. of n Exponential R.V's

Hence we approximate

$$E\left[X_{\max}\right] \approx \frac{\gamma + \log n}{\lambda}.$$

log is the natural logarithm.

$$\mathrm{Var}\left[X_{\max}\right] \approx \frac{\pi^2}{6\lambda^2}.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Asymptotic Extreme Value Distribution (EVD) (1.1)

$X_{\max} = \max(X_1, X_2, \ldots, X_n)$, $X_1, X_2, \ldots, X_n$, I.I.D.
Note that $X_1, X_2, \ldots, X_n$ need not be exponentially distributed, but the distribution must satisfy a tail condition (see below). We are going to prove

$$\lim_{n\to\infty} P\left(X_{\max} - \frac{\log n}{\lambda} \leq x\right) = e^{-e^{-\lambda x}}.$$

Here log is the natural logarithm.

KTH VETENSKAP OCH KONST
KTH Matematik

Asymptotic EVD (1.2): Gumbel's distribution

$$F(x) := e^{-e^{-\lambda x}}$$

is a probability distribution function:

- $F(x) \to 1$, as $x \to \infty$,
- $F(x) \to 0$, as $x \to -\infty$,
- $F(x)$ is an increasing function, since $\frac{d}{dx}F(x) = \lambda e^{-\lambda x - e^{-\lambda x}} > 0$

The distribution is known as Gumbel's distribution.

KTH VETENSKAP OCH KONST
KTH Matematik

Asymptotic Distributions

$$F_n(x) = P\left(X_{\max} - \frac{\log n}{\lambda} \leq x\right).$$

We claim thus

$$\lim_{n\to\infty} F_n(x) = F(x).$$

We say that $X_{\max}$ converges asymptotically in distribution to a random variable that has a Gumbel distribution.

KTH VETENSKAP OCH KONST
KTH Matematik

Asymptotic Distributions

In general, if $F_1(x), F_2(x), F_3(x), F_4(x), \ldots$ are distribution functions, and

$$\lim_{n\to\infty} F_n(x) = F(x).$$

where $F(x)$ is a distribution function, we talk about convergence in distribution. Examples: Central Limit Theorem.

KTH Matematik

EVD (2): Proof starts

We know from the preceding that

$$P\left(X_{\max} - \frac{\log n}{\lambda} \leq x\right) = P\left(X_{\max} \leq x + \frac{\log n}{\lambda}\right)$$

$$= \left[P\left(X_1 \leq x + \frac{\log n}{\lambda}\right)\right]^n$$

$$= \left[F\left(x + \frac{\log n}{\lambda}\right)\right]^n.$$

where $F(x) = P(X_1 \leq x)$.

EVD (3)

We write the right hand side as

$$\left[F\left(x+\frac{\log n}{\lambda}\right)\right]^n = e^{n\log F\left(x+\frac{\log n}{\lambda}\right)},$$

EVD (4): Assumption on tails

We assume that

$$F(x) = 1 - V(x)e^{-\lambda x}$$

for large x, where $\lim_{x\to\infty} V(x) = 1$ is assumed for simplicity. In words, the tail of the distribution function

$$1 - F(x) = V(x)e^{-\lambda x}$$

vanishes exponentially. Clearly the exponential distribution satisfies this !

KTH Matematik

EVD (5): Insert the tail

$$e^{n \log F\left(x+\frac{\log n}{\lambda}\right)} = e^{n \log\left[1-V\left(x+\frac{\log n}{\lambda}\right)e^{-x\lambda-\log n}\right]}.$$

EVD (6): Series expansion

We use the series expansion (valid for $-1 \leq x < 1$)

$$\log(1+x) = x - \frac{1}{2}x^2 + \dots$$

with the result

$$n \log\left[1 - V\left(x + \frac{\log n}{\lambda}\right) e^{-x\lambda} e^{\log 1/n}\right] =$$

$$= -n\frac{1}{n} V\left(x + \frac{\log n}{\lambda}\right) e^{-x\lambda} - \frac{1}{2} n \frac{1}{n^2} G(n) + O\left(\frac{1}{n}\right) \dots$$

where $G(n)$ is a bounded function of n and $O\left(\frac{1}{n}\right)$ represents the omitted terms that turn to zero as $n \to +\infty$.

KTH Matematik

EVD (7)

Hence we have

$$\lim_{n\to\infty} n \log\left[1 - V\left(x + \frac{\log n}{\lambda}\right) e^{-x\lambda} e^{\log 1/n}\right]$$

$$= -e^{-x\lambda}.$$

KTH VETENSKAP OCH KONST
KTH Matematik

EVD (8):

In other words we have found that

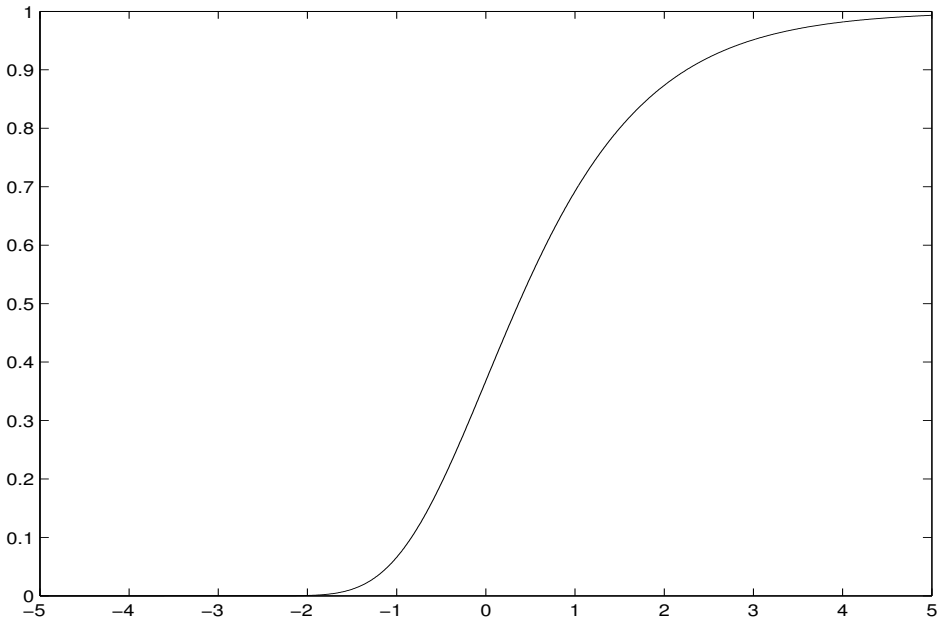
$$\lim_{n\to\infty} P\left(X_{\max} - \frac{\log n}{\lambda} \leq x\right) = e^{-e^{-\lambda x}},$$

as was to be proved.
The rate of convergence may be slow.

KTH VETENSKAP OCH KONST
KTH Matematik

Gumbel's distribution

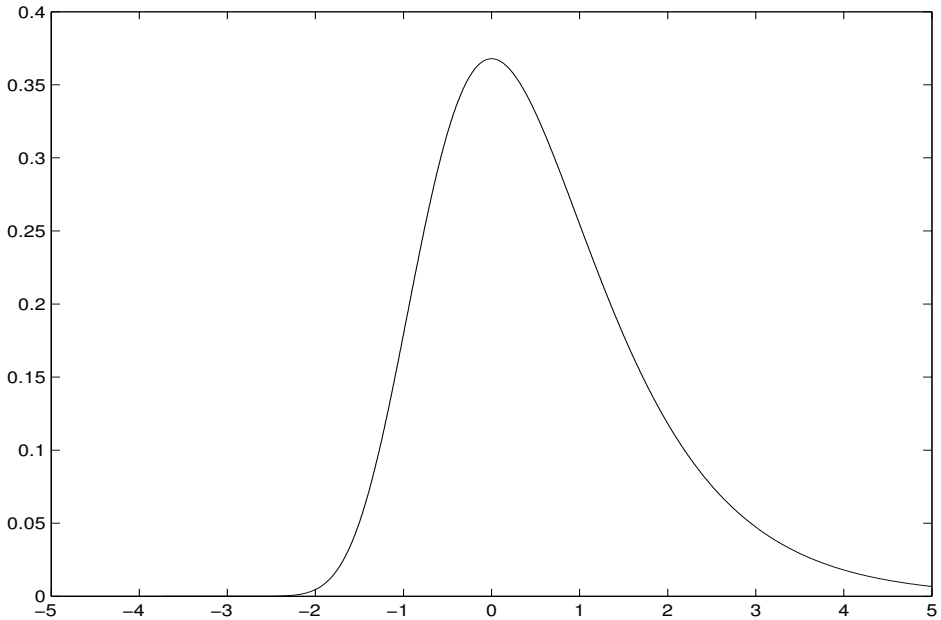
$$P\left(X_{\max} \leq x\right) \approx e^{-e^{-u}}$$



KTH Matematik

Gumbel density

$$e^{-u-e^{-u}}$$



KTH VETENSKAP OCH KONST
KTH Matematik

p-value

The p-value of an observed value x of $X_{\max}$ is the probability $P\left(X_{\max} > x\right)$ under some model for the underlying random variables. If these are exponential, then

$$p\text{−value} \approx 1 - e^{-ne^{-\lambda x}}.$$

The smaller the p-value, the less likely is the hypothesis that the model is the correct one.

KTH VETENSKAP OCH KONST
KTH Matematik

B: Order Statistics: the setting

Let $X_1, \ldots, X_n$ be I.I.D. random variables distributed like X with the distribution function F such that

$$F(x) = \int_{-\infty}^{x} f(u)du.$$

We are for reasons of simplicity of writing going to drop here the subscripts in $F_X(x) = \int_{-\infty}^{x} f_X(u)du$.

KTH VETENSKAP OCH KONST
KTH Matematik

Order Statistics: definition

$X_1, \ldots, X_n$ are I.I.D. random variables distributed like X.

$$X_{(j)} = \text{ j:th order statistic} =$$

$$\text{j:th smallest of } X_1, \ldots, X_n.$$

$X_{(1)} = X_{\min}$, $X_{(n)} = X_{\max}$. Clearly

$$X_{(1)} < X_{(2)} < \ldots < X_{(n-1)} < X_{(n)}$$

KTH VETENSKAP OCH KONST
KTH Matematik

Order Statistics: definition (remark)

$X_1, \ldots, X_n$ are I.I.D. random variables distributed like X. Since X is a continuous r.v. , we have strict equalities in

$$X_{(1)} < X_{(2)} < \ldots < X_{(n-1)} < X_{(n)}$$

For a continuous random variable $P(X = x) = \int_x^x f(u)du = 0$.

KTH VETENSKAP OCH KONST
KTH Matematik

Order Statistics: the theorem on the joint density

$X_1, \ldots, X_n$ are I.I.D. random variables distributed like X with density f. Then

$$f_{X_{(1)},\ldots,X_{(n)}}\left(x_{(1)}, \ldots, x_{(n)}\right)$$

$$= \begin{cases} n! \prod_{i=1}^{n} f\left(x_{(i)}\right) & \text{if } x_{(1)} < \ldots < x_{(n)} \\ 0 & \text{elsewhere.} \end{cases}$$

□

Order Statistics: the theorem on the joint density

$$f_{X_{(1)},\ldots,X_{(n)}}\left(x_{(1)},\ldots,x_{(n)}\right)$$

$$= \begin{cases} n!\prod_{i=1}^{n} f\left(x_{(i)}\right) & \text{if } x_{(1)} < \ldots < x_{(n)} \\ 0 & \text{elsewhere.} \end{cases} \qquad \square$$

The theorem will be explained on the whiteboard.

KTH VETENSKAP OCH KONST
KTH Matematik

A Tool: Multinomial Probability Distribution

The probability

$$P\left(Y_1 = n_1, Y_2 = n_2, \ldots, Y_L = n_L\right) = \frac{n!}{n_1! n_2! \ldots n_L!} \cdot p_1^{n_1} \cdots p_L^{n_L}.$$

is the joint probability of the random variables $Y_1, Y_2, \ldots, Y_L$ as functions of the random variables $X_1, \ldots, X_n$ by

$$Y_l = \text{ the number of times } X_i = x_l,\ i = 1, 2, \ldots, n,$$

with L different outcomes x_l.

KTH VETENSKAP OCH KONST
KTH Matematik

Uniform R.V's

Let $0 < L$ and

$$f_X(x) = \begin{cases} \frac{1}{L} & 0 < x < L \\ 0 & \text{elsewhere.} \end{cases}$$

$X \in \mathrm{U}(0, L)$, uniformly distributed.

KTH VETENSKAP OCH KONST
KTH Matematik

Order Statistics: Uniform R.V's

$X_1, \ldots, X_n$ are I.I.D. random variables $X \in \mathrm{U}(0, L)$

$$f_{X_{(i)}}\left(x_{(i)}\right) = \frac{n!}{(i-1)!(n-i)!} x_{(i)}^{i-1} \left(L - x_{(i)}\right)^{n-i} L^{-n}$$

KTH VETENSKAP OCH KONST
KTH Matematik

Order Statistics: Uniform R.V's

$X_1, \ldots, X_n$ are I.I.D. random variables $X \in \mathrm{U}(0, L)$

$$f_{X_{\min}}(x_{\min}) = f_{X_{(1)}}\left(x_{(1)}\right) = n \cdot (L - x_{\min})^{n-1} L^{-n}.$$

KTH Matematik

Order Statistics: Uniform R.V's

$X \in \mathrm{U}(0,1)$

$$f_{X_{(i)}}\left(x_{(i)}\right) = \frac{n!}{(i-1)!(n-i)!} x_{(i)}^{i-1}\left(1-x_{(i)}\right)^{n-i}.$$

This is $\mathrm{Beta}(i, n-i+1)$.

KTH Matematik

Order Statistics: Uniform R.V's

$X_1, \ldots, X_n$ are I.I.D. random variables $X \in \mathrm{U}(0, L)$

$$f_{X_{(1)},\ldots,X_{(n)}}\left(x_{(1)}, \ldots, x_{(n)}\right) = \begin{cases} \frac{n!}{L^n} & \text{if } x_{(1)} < \ldots < x_{(n)} \\ 0 & \text{elsewhere.} \end{cases}$$

□

KTH Matematik

The R.V's Conditioned by Order Statistics

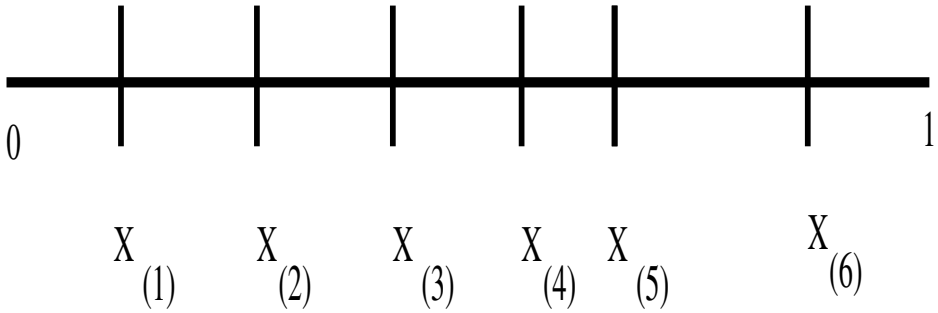
$X_1, \ldots, X_n$ are I.I.D. random variables $X \in \mathrm{U}(0, L)$

$$f_{X_1,\ldots,X_n|X_{(1)},\ldots,X_{(n)}}\left(x_1, \ldots, x_n \mid x_{(1)}, \ldots, x_{(n)}\right)$$

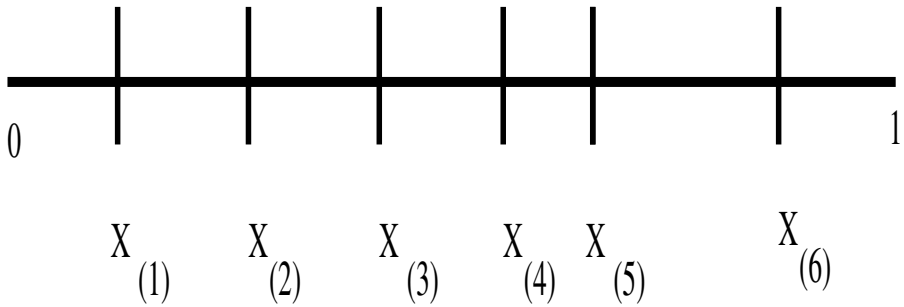
$$= \begin{cases} \frac{1}{n!} & \text{if } x_{(1)} < \ldots < x_{(n)} \\ 0 & \text{elsewhere.} \end{cases} \qquad \square$$

A Stick Broken at Random

$X_1, \ldots, X_n$ are I.I.D. random variables $X \in \mathrm{U}(0,1)$.



A Stick Broken at Random

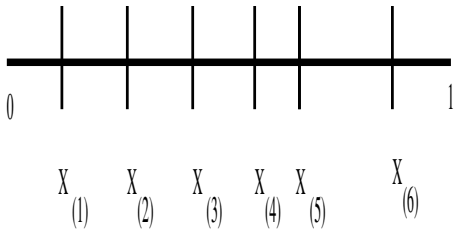


$U_1 = X_{(1)}$ $U_2 = X_{(2)} - X_{(1)}$ $U_3 = X_{(3)} - X_{(2)}$ $U_4 = X_{(4)} - X_{(3)}$

$U_5 = X_{(5)} - X_{(4)}$ $U_6 = X_{(6)} - X_{(5)}$ $U_7 = 1 - X_{(6)}$

$U_1, \ldots, U_{n+1}$ are the lengths of the $n+1$ intervals created by $X_1, \ldots, X_n$ I.I.D. $\in \mathrm{U}(0,1)$.

A Stick Broken at Random



We can show that

$$P\left(U_1 > u, \ldots, U_g > u\right) = (1 - gu)^n$$

for if $1 - ug > 0$. Proof handled during a lektion.

Another result

Let $U_{(n-j)}$ be the $n-j$th order statistic

$$P\left(U_{(n-j)} \leq x\right)$$

$$= \sum_{l=0}^{j} \binom{n+1}{l} \sum_{\nu=0}^{n+1-l} (-1)^{\nu} \binom{n+1-l}{\nu} (1-(l+\nu)x)_{+}$$

where $(a)_{+} = \max(0, a)$.

The exact distribution of $U_{\max}$

Set $j = 0$ in the above

$$P\left(U_{\max} \leq x\right)$$

$$= \sum_{\nu=0}^{n+1} (-1)^{\nu} \binom{n+1}{\nu} (1 - \nu x)_{+}$$

where $(a)_{+} = \max(0, a)$.

Covering the whole genome

The probability of covering a whole genome by n 'markers' of length at most x is also equal to

$$P\left(U_{\max} \leq x\right)$$

$$= \sum_{\nu=0}^{n+1} (-1)^{\nu} \binom{n+1}{\nu} (1-\nu x)_{+}$$

where $(a)_{+} = \max(0, a)$. K. Lange & M. Boehnke (1982): How many polymorphic genes will it take to span the human genome. *American Journal of Human Genetics.* 34, pp. 842−845.

KTH VETENSKAP OCH KONST
KTH Matematik

Asymptotic distribution of $U_{\max}$

$$P\left(nU_{\max} - \ln n \leq x\right) = P\left(nU_n - \ln n \leq x\right) \to e^{-e^{-x}},$$

as $n \to \infty$. Found in (L. Holst 1980).

Number of markers

The number N_α of 'markers' of needed until they are at at most length α:

$$P\left(N_\alpha \leq n\right) = P\left(U_{\max} \leq \alpha\right) \approx e^{-e^{-x}}$$

if $n\alpha - \ln n \to x$, when $\alpha \to 0$, $n \to \infty$.
$e^{-x} = 10^{-c}$ gives $n \geq c \ln 10/\alpha - \ln n/\alpha$.

KTH Matematik

Transformations of Variable (1)

$$U_i = U_i\left(X_1, \ldots, X_n\right), \quad i = 1, \ldots, n.$$

Then

$$f_{U_1,\ldots,U_n}\left(u_1, \ldots, u_n\right) =$$

$$f_{X_1,\ldots,X_n}\left(x_1, \ldots, x_n\right) \cdot |J^{-1}|,$$

J is known as the Jacobian of the transformation.

KTH VETENSKAP OCH KONST
KTH Matematik

Transformations of Variable (1)

$$f_{U_1,\ldots,U_n}(u_1,\ldots,u_n) =$$

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) \cdot |J^{-1}|,$$

where the Jacobian is

$$J = \begin{vmatrix} \frac{\partial U_1}{\partial X_1} & \frac{\partial U_1}{\partial X_2} & \cdots & \frac{\partial U_1}{\partial X_n} \\ \frac{\partial U_2}{\partial X_1} & \frac{\partial U_2}{\partial X_2} & \cdots & \frac{\partial U_2}{\partial X_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial U_n}{\partial X_1} & \frac{\partial U_n}{\partial X_2} & \cdots & \frac{\partial U_n}{\partial X_n} \end{vmatrix}.$$

This and an alternative formula are proved in courses on multivariable integral calculus.

KTH VETENSKAP OCH KONST
KTH Matematik

Transformations of Variable (2)

$$X_i = X_i\left(U_1, \ldots, U_n\right), \quad i = 1, \ldots, n.$$

$$f_{U_1,\ldots,U_n}\left(u_1, \ldots, u_n\right) =$$

$$f_{X_1,\ldots,X_n}\left(x_1, \ldots, x_n\right) \cdot \left|J^*\right|,$$

where

$$J^* = \begin{vmatrix} \frac{\partial X_1}{\partial U_1} & \frac{\partial X_1}{\partial U_2} & \cdots & \frac{\partial X_1}{\partial U_n} \\ \frac{\partial X_2}{\partial U_1} & \frac{\partial X_2}{\partial U_2} & \cdots & \frac{\partial X_2}{\partial U_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial X_n}{\partial U_1} & \frac{\partial X_n}{\partial U_2} & \cdots & \frac{\partial X_n}{\partial U_n} \end{vmatrix}.$$

Proved in courses on multivariable integral calculus.

KTH VETENSKAP OCH KONST
KTH Matematik

End of Lecture

KTH Matematik