Below you can see a plot of two empirical cumulative distribution functions (eCDF). Each of the eCDFs are plotted using 1 000 realizations of one of two random variables — A and B.

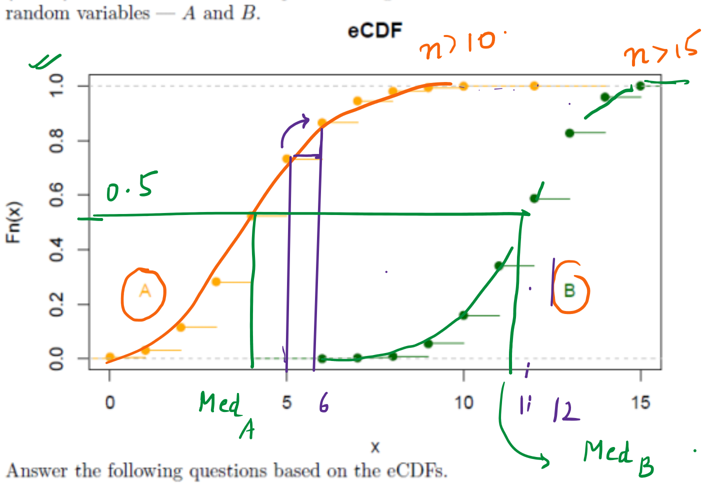
$A \sim Bin(n, p_A)$

$B \sim Bin(n, p_B)$

Eg: A

cdf:

$P[A \leq a] = \sum_{i=0}^{a} f(x)$.



Answer the following questions based on the eCDFs.

i. Suppose both A and B are Binomial distributions with a common n and different ps. What is the value of the common parameter n? Explain in one or two sentences.

Solution

Since both have the same parameter n, I can take the larger value for which there's a jump in the CDF. So $n = 15$.

ii. Which of the 2 variables has a larger median? Explain in one or two sentences.

Solution

B has a larger median since its eCDF crosses 0.5 on the y-axis along at a greater point along the x-axis.

Question 2

A skate rental shop records the time between skates being rented and being returned. Analyzing 100 returns, the average time to return is found to be 2 hours. The shop opens for 4 hours daily and overnight rentals are not allowed.

X : Time for which skates are taken on rent.

$n = 100$

R.S $X_1, X_2, \dots, X_{100}$

$E(X) = 2$, $X \le 4$

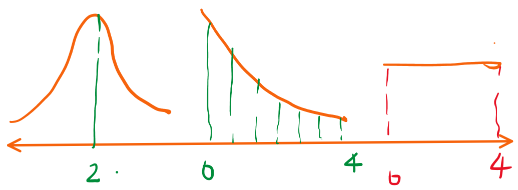
or $P[X > 4] = 0$.

Range of $0 \le X \le 4$

$\Downarrow$

Continuous variable.

$N(\mu, \sigma^2)$ $\quad$ $Exp(\lambda)$ $\quad$ $U[a, b]$.



Part a)

Suggest a reasonable parametric model among the models listed in Table 1 for the rental times assuming they are a random sample. What additional assumptions are you making with the selected distribution? Are any of the model assumptions unrealistic? Explain in a few sentences.

Solution

Given the time until rental return data, an Exponential distribution would be the most suitable choice among the distributions listed.

In addition to the each return time being independent and identical, the model assumes that the rental return is a Poisson process. That is, the likelihood of rental return at any moment is independent and identical.

The rental shop is open for 4 hours per day only. But an exponential random variable can take any positive real number.

Alternatively, you may have assumed a uniform distribution since the rental time is bounded. However, this is not appropriate since not every one rents their skates when the shop opens.

In either case, the indpendence assumption may be unrealistic since skaters rent together with friends and family. They are likely to return at the same time.

Part b)

What is your best guess for the parameter(s) of your selected model based on the given information? State any probability rules that support your guess.

Solution

Let X be the skate return time. If we assume $X \sim \text{Exp}(\lambda)$, $\mathbb{E}(X) = \frac{1}{\lambda}$. The sample mean over the $n = 100$ repairs is $\bar{x}_{50} = 2$ hours. Thus, a reasonable estimate of $\lambda = 1/2 = 0.5$ based on the law of large numbers. (MOM / MLE)

$\therefore$ R.S : $X_1, X_2, \dots, X_n \overset{iid}{\sim} f(x) \Rightarrow X_i \sim f(x)\ \forall i$ and cdf $F(x)$

ordered sample : $X_{(1)} \le X_{(2)} \le \dots \le X_{(n)}$.

$X_{(1)} = \min\{X_1, X_2, \dots, X_n\}$ $\qquad$ $X_{(n)} = \max\{X_1, X_2, \dots, X_n\}$.

$\frac{d}{dx} F(x) = f(x)$

Q. Pdf of $X_{(n)} = \max\{X_1, X_2, \dots, X_n\}$.

c.d.f of $X_{(n)} = P[X_{(n)} \le x]$.

$$= P[\max\{X_1, X_2, \dots, X_n\} \le x].$$

$$= P[X_1 \le x, X_2 \le x, \dots, X_n \le x]$$

$$= P[X_1 \le x] \cdot P[X_2 \le x] \cdots\cdots P[X_n \le x]$$

$= P[X_1 \le x] \cdot P[X_2 \le x] \cdots\cdots P[X_n \le x]$

$P[X_1 \le x] \downarrow$ cdf of r.v X_1 $= F(x)$

$P[X_2 \le x] \downarrow$ cdf of r.v X_2 $= F(x)$

$$= \underbrace{F(x) \cdot F(x) \cdots F(x)}_{n \text{ times}} = \{F(x)\}^n$$

$$\text{pdf of } X_{(n)} = \frac{d}{dx}\{F(x)\}^n = n\{F(x)\}^{n-1} \cdot f(x).$$

Q. pdf of $X_{(1)} = \min\{x_1, x_2, \ldots, x_n\}$

$$\text{cdf of } X_{(1)} = P[X_{(1)} \le x].$$

$$= P[\min\{X_1, X_2, \ldots, X_n\} \le x]$$

$$= 1 - P[\min\{X_1, X_2, \ldots, X_n\} > x].$$

$$= 1 - P[X_1 > x, X_2 > x, \ldots, X_n > x].$$

$$= 1 - P[X_1 > x] \cdot P[X_2 > x] \cdots P[X_n > x].$$

$P[X_1 > x] \Downarrow$ $1 - P[X_1 \le x]$ $= \{1 - F(x)\}$.

$P[X_2 > x] \Downarrow$ $1 - P[X_2 \le x]$. $= \{1 - F(x)\}$.

$$= 1 - \{1 - F(x)\}^n.$$

$$\text{pdf of } X_{(1)} = \frac{d}{dx}\left[1 - \{1 - F(x)\}^n\right]$$

$$= 0 - n\{1 - F(x)\}^{n-1} \cdot (0 - f(x))$$

$$= n \cdot f(x)\{1 - F(x)\}^{n-1}$$

$y = f(x)$

$= \sin^n x$

$= (\sin x)^n$

$\frac{dy}{} = n \cdot \sin^{n-1} x.$

$= n f(x) [1 - F(x)]$

$\frac{dy}{dx} = n \cdot \sin x^{n-1}$

Question 3

Suppose $X_1, X_2, ..., X_n \overset{i.i.d}{\sim} \text{Exp}(\lambda)$. Is the estimator $\hat{\lambda} = X_{(1)} = \min(X_1, X_2, ...X_n)$ an unbiased estimator?

Hint: $\rightarrow$ cdf of $X_{(1)}$

$\boxed{P(X_{(1)} \leq x)} = 1 - P(X_{(1)} > x) = 1 - P((X_1 > x) \cap (X_2 > x) \cap ... \cap (X_n > x))$

Solution:

$$\begin{aligned}
P(X_{(1)} \leq x) &= 1 - P(X_{(1)} > x) = 1 - P((X_1 > x) \cap (X_2 > x) \cap ... \cap (X_n > x)) \\
&= 1 - \prod_{i=1}^{n} [1 - P(X_i \leq x)] \\
&= 1 - \prod_{i=1}^{n} e^{-\lambda x} \\
&= 1 - e^{-n\lambda x} \text{ for } x \geq 0
\end{aligned}$$

$\Rightarrow$ cdf of $X_{(1)} = 1 - e^{-n\lambda x}$.

pdf of $X_{(1)} = \frac{d}{dx}\left[1 - e^{-n\lambda x}\right]$

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

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

$$\implies X_{(1)} \sim \text{Exp}(n\lambda)$$

$$\implies \mathbb{E}(\hat{\lambda}) = \mathbb{E}\left(X_{(1)}\right) = \frac{1}{n\lambda}$$

Therefore $\hat{\lambda}$ is a biased estimator for λ.

$$E[X_{(1)}] = \int_0^\infty x \cdot n\lambda e^{-n\lambda x} dx.$$

Let $n\lambda x = z$. $\Rightarrow$ $x = 0$, $z = 0$

$dx = \frac{dz}{n\lambda}$ $\quad x \to \infty$, $z \to \infty$

$$= \int_0^\infty \cancel{n\lambda} \cdot \left(\frac{z}{\cancel{n\lambda}}\right) \cdot e^{-z} \cdot \frac{dz}{n\lambda}.$$

$$= \frac{1}{n\lambda} \underbrace{\int_0^\infty z e^{-z} dz}_{=1} = \frac{1}{n\lambda}.$$

$$\int z e^{-z} dz = z \int e^{-z} dz - \int - e^{-z} dz$$

$$= z\int e^{-z}\,dz + \int e^{-z}\,dz$$

$$= -ze^{-z} - e^{-z} + c = -e^{-z}(z+1) + c$$

$$\int_0^{\infty} z e^{-z}\,dz = -\left[e^{-z}(z+1)\right]_0^{\infty} = -\left[e^{-\infty}(\infty+1) - e^{-0}(0+1)\right]$$

$$= -[0-1] = 1.$$

Bootstrap :-

$X \sim N(\mu, \sigma^2) \Rightarrow$ popln.

R.S: $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ → Theory.

Sample mean: $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

Sample median: need not follow any particular known distributi

Bootstrap technique is used when the estimators do not a follow a known stat distribution. Hence its properties cannot be analyzed directly from the popln distribution.

Obj: Sampling distribution of the Med:

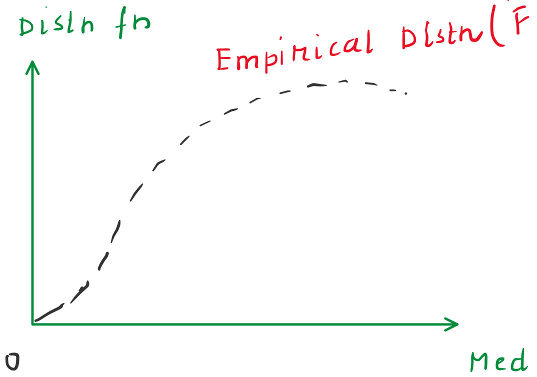
Take $n = 1000$ samples from popln.

$S_1 \Rightarrow Med_1$

$S_2 \Rightarrow Med_2$

$\vdots \qquad \vdots$

$S_{1000} \Rightarrow Med_{1000}$



Ordered set of Med values: $Med_{(1)} \leq Med_{(2)} \leq \cdots \leq Med_{(1000)}$

Q. Var [Sample(Med)] ⇒ obtained from $\hat{F}$.

$$\text{Mean}\left[\text{Sample}(Med)\right] = \bar{x}_{Med} = \frac{1}{\quad}\sum^{1000} Med_{(i)}$$

$$\text{Mean}\left[\text{Sample}(\text{Med})\right] = \bar{x}_{Med} = \frac{1}{1000} \sum_{i=1}^{1000} Med_{(i)}$$

$$\text{Var}\left[\text{Sample}(\text{Med})\right] = \frac{1}{1000-1} \sum_{i=1}^{n} \left(Med_{(i)} - \bar{x}_{Med} \right)^2$$