TUTORIAL STATISTICS

26/09/2024

Statistics Fall 2024 - TA Session 1
Sampling and Sufficiency

TA: Giacomo Caserta - giacomo.caserta@uniroma2.it
Office Hour: By appointment, office 3D-8, third floor building B

25/09/2024

Problem 1

You pay $1 to play a game in which you roll one standard six-sided die. You lose your dollar if the die is 1, 2, 3 or 4. You get your dollar back if the die is a 5, and if the die is a 6 you get your dollar back plus $2 more (total of $3).

- Calculate expected value and standard deviation for a single toss. (Be sure to include the dollar you pay to play the game.)
- If you play the game 100 times, what are the expected value and standard error for the sampling distribution?
- If you play the game 100 times, what is the probability that your average outcome will be positive? (That is, you walk away with more money than what you had before the game.)
- If you play the game 100 times, your average winnings have a 90 % probability of being below what value?

Problem 2

The waiting time at the post office is distributed as an exponential distribution with mean 20 minutes.

- What is the probability that a client will wait more than 20 minutes?
- What would it be if the distribution were uniform?

You will go to the post office next week every day from Monday to Friday, what is the probability that the maximum time you will wait is between 30 and 32 minutes?

Problem 3

Show that each of the following distributions is a member of the exponential family:

- The Gaussian distribution
- The Bernoulli distribution
- The Poisson distribution
- The Beta distribution

Problem 4

Suppose that $x_1, x_2, ..., x_n$ are i.i.d. Poisson(θ), where $\theta > 0$. Show that:

$$T = T(\mathbf{X}) = \sum_{i=1}^{n} x_i$$

is a sufficient statistic using the Factorization Theorem.

Problem 5

Suppose that $X_1, X_2, ..., X_n$ are i.i.d. $U(\theta, \theta + 1)$, where $-\infty < \theta < \infty$. Show that:

$$\mathbf{T} = \mathbf{T}(\mathbf{X}) = \begin{pmatrix} X_{(1)} \\ X_{(N)} \end{pmatrix}$$

is sufficient.

Problem 6

Consider the linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

For $i = 1, 2, ..., n$ where ϵ_i are iid $N(0, \sigma^2)$ and the x_i's are fixed constants (i.e., not random). In this model, it is easy to show that:

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$$

so that $\theta = (\beta_0, \beta_1, \sigma^2)$. Note that $Y_1, Y_2, ..., Y_n$ are independent random variables (functions of independent random variables are independent); however, Y1, Y2, ..., Yn are not identically distributed because $E(Y_i) = \beta_0 + \beta_1 x_i$ changes as i does.

Use the factorization theorem to find a sufficient statistic $T(X)$. Remember that $dim(T) = dim(\theta) = 3$.

PROBLEM 1

X	-1	0	2
P(x)	2/3	1/6	1/6

$$E(X) = \mu = -1\left(\frac{2}{3}\right) + 0 \cdot \frac{1}{6} + 2\frac{1}{6} = -\frac{1}{3}$$

$$VAR(X) = \left(-1-\left(-\frac{1}{3}\right)\right)^2 \frac{2}{3} + \left(0-\left(-\frac{1}{3}\right)\right)^2 \frac{1}{6} + \left(2-\left(-\frac{1}{3}\right)\right)^2 \frac{1}{6} = \frac{11}{9}$$

$$\sigma = \sqrt{VAR(x)} = \sqrt{\frac{11}{9}} \cong 1,10$$

WE PLAY THE GAME 100 TIMES.

$$SE = \frac{\sqrt{\frac{11}{9}}}{\sqrt{N}} = \frac{\sqrt{\frac{11}{9}}}{\sqrt{100}} = \frac{\sqrt{11}}{30}$$

$$E(x) = \mu = -\frac{1}{3}$$

c) $P(\bar{X} > 0)$

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{N}}} = \frac{0 - \left(-\frac{1}{3}\right)}{\boxed{\frac{\sqrt{11}}{30}}}$$

STANDARD ERROR

$$P(\bar{X} > 0) = P(Z > 3.02) =$$

$$= 1 - P(Z < 3.02)$$

$$= 1 - 0.9987 = 0.0013$$

D) $N = 100$

$$0,8997 = P(Z < 1,28)$$

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{N}}}$$

$$1,28 = \frac{\bar{x} - \left(-\frac{1}{3}\right)}{\frac{\sqrt{11}}{30}}$$

$$1,28 \frac{\sqrt{11}}{30} = \bar{x} + \frac{1}{3}$$

$$\bar{x} \approxeq 0,19$$

PROBLEM 2

A) $$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{IF } x \geq 0 \\ 0 & \text{OTHERWISE} \end{cases}$$

$$\text{CDF} = F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x} & \text{IF } x \geq 0 \\ 0 & \text{OTHERWISE} \end{cases}$$

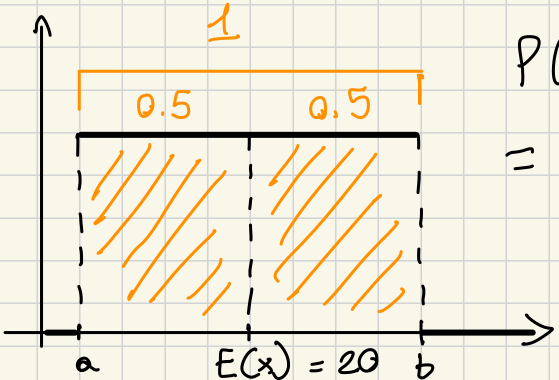
$$E(X) = \frac{1}{\lambda} \Longrightarrow \lambda = \frac{1}{E(X)}$$

$$\lambda = \frac{1}{20} = 0.05$$

$$X \sim EXP(\lambda)$$
$$X \sim EXP(0.05)$$

$$P(X > 20) = 1 - P(X \leq 20)$$
$$= 1 - \left(1 - e^{-0.05 \cdot 20}\right)$$
$$\overset{!}{=} 1 - 1 + e^{-1} = 0.367$$

B) WHAT WOULD IT BE IF THE DISTRIBUTION WERE UNIFORM?



$$P(X \geq 20) = \frac{1}{2}$$
$$= P(X < 20)$$

c) YOU WILL GO TO THE POST OFFICE NEXT WEEK EVERY DAY FROM MONDAY TO FRIDAY...

$P(30 < \underset{\substack{\downarrow \\ \text{MAX}}}{x} < 32)$ CONSIDERING $\underbrace{\text{5 DAYS}}_{\text{5 RVs}}$!!

DISTRIBUTION OF MAX OF N RVs:

$$f_{X_{(n)}}(x) = \underset{\substack{\downarrow \\ 5}}{n}\ \underset{\substack{\downarrow \\ \text{PDF}}}{f(x)}\ \underset{\substack{\downarrow \\ \text{CDF}}}{\left[F(x)\right]}^{N-1}$$

$$f_{X_{(n)}}(x) = 5\lambda e^{-\lambda x}\left(1-e^{-\lambda x}\right)^{4} dx$$

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

$$= \int_{30}^{32} 5 \lambda e^{-\lambda x} \left[\left(1 - e^{-\lambda x} \right) \right]^{4} dx$$

$$= \int_{30}^{32} 5(0.05) e^{-0.05x} \left[\left(1 - e^{-0.05x} \right) \right]^{4} dx$$

$$= \int_{30}^{32} \frac{1}{4} e^{-\frac{1}{20}x} \left(1 - e^{-\frac{1}{20}x} \right)^{4} dx$$

$$= \int_{30}^{32} \frac{\left(1 - e^{-\frac{x}{20}} \right)^{4} e^{-\frac{x}{20}}}{4} dx$$

LET US SOLVE THE INTEGRAL
BY SUBSTITUTION

$$u = 1 - e^{-\frac{x}{20}}$$

$$du = \frac{e^{-\frac{x}{20}}}{20} dx$$

$$20\, du = dx\; e^{-\frac{x}{20}}$$

$$= \int_{1-e^{-\frac{30}{20}}}^{1-e^{-\frac{32}{20}}} \frac{20\, u^4\, du}{4}$$

$$= \int_{1-e^{-\frac{30}{20}}}^{1-e^{-\frac{32}{20}}} 5\, u^4\, du$$

$$= 5\int_{1-e^{-\frac{30}{20}}}^{1-e^{-\frac{32}{20}}} u^4\, du$$

$$= \cancel{5}\,\frac{u^5}{\cancel{5}}\Bigg|_{1-e^{-\frac{30}{20}}}^{1-e^{-\frac{32}{20}}}$$

$$= \Big(\underbrace{1-e^{-\frac{x}{20}}}_{u}\Big)^5\Bigg|_{30}^{32}$$

$$= \left(1-e^{-\frac{32}{20}}\right)^5 - \left(1-e^{-\frac{30}{20}}\right)^5 = 0.0408$$

PROBLEM 3

SHOW THAT EACH OF THE FOLLOWING DISTRIBUTION IS A MEMBER OF THE EXPONENTIAL FAMILY:

DEF A RV X WITH PDF $f(x|\theta)$ IS SAID TO BELONG TO AN EXPONENTIAL FAMILY IF THE PDF TAKES THE FOLLOWING FORM:

$$p(x|\eta) = h(x) \, \mathrm{EXP} \left\{ \eta^T T(x) - A(\eta) \right\}$$

FOR A PARAMETER VECTOR η

FOR GIVEN FUNCTIONS T AND h

$T(x) \Rightarrow$ SUFFICIENT STATISTIC

$A(\eta) \Rightarrow$ CUMULANT FUNCTION

(1) UNIVARIATE GAUSSIAN DISTRIBUTION

$$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} \, \mathrm{EXP} \left\{ -\frac{1}{2\sigma^2} (x-\mu)^2 \right\}$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma} \operatorname{Exp}\left\{ - \frac{(x^2 + \mu^2 - 2x\mu)}{2\sigma^2} \right\}$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma} \operatorname{Exp}\left\{ - \frac{x^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \log\sigma \right\}$$

$$= \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \operatorname{Exp}\left\{ \frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} x^2 - \frac{1}{2\sigma^2} \mu^2 - \ln\sigma \right\}$$

$$\eta = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}$$

$$T(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$$

$$A(\eta) = \frac{\mu^2}{2\sigma^2} + \log\sigma$$

$$h(x) = \frac{1}{\sqrt{2\pi}}$$

(2) BERNOULLI

$$P(x \mid \pi) = \pi^{x} (1-\pi)^{1-x}$$

$$\overset{!}{=} \text{EXP}\left\{ \log\left(\pi^{x} (1-\pi)^{1-x} \right) \right\}$$

$$= \text{EXP}\left\{ x \log \pi + (1-x) \log(1-\pi) \right\}$$

$$= \text{EXP}\left\{ x \log \pi + \log(1-\pi) - x \log(1-\pi) \right\}$$

$$= \text{EXP}\left\{ x \left(\log \pi - \log(1-\pi) \right) + \log(1-\pi) \right\}$$

$$= \text{EXP}\left\{ x \log \frac{\pi}{1-\pi} + \log(1-\pi) \right\}$$

$$\eta = \log\left(\frac{\pi}{1-\pi}\right)$$

$$T(x) = x$$

$$A(\eta) = -\log(1-\pi)$$

$$h(x) = 1$$

(3) POISSON

$$p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

$$= \frac{1}{x!} \text{EXP} \left\{ x \log \lambda - \lambda \right\}$$

$\eta = \lambda$

$T(x) = x$

$A(\eta) = \lambda$

$h(x) = \frac{1}{x!}$

(4) BETA DISTRIBUTION

$$f(x,\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$$

$$= \operatorname{EXP}\left\{(\alpha-1)\log x + (\beta-1)\log(1-x) - \log B(\alpha,\beta)\right\}$$

$$= \operatorname{EXP}\left\{(\alpha-1)\log x + \beta\log(1-x) - \log(1-x) - \log B(\alpha,\beta)\right]$$

$$= \operatorname{EXP}\left\{\alpha\log x - \log x + \beta\log(1-x) - \log(1-x) - \log B(\alpha,\beta)\right\}$$

$$= \frac{1}{x(1-x)}\operatorname{EXP}\left\{\alpha\log(x) + \beta\log(1-x) - \log B(\alpha,\beta)\right\}$$

$$h(x) = \frac{1}{x(1-x)}$$

$$T(x) = \begin{bmatrix} \log(x) \\ \log(1-x) \end{bmatrix}$$

$$\eta = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

$$A(\eta) = \log B(\alpha, \beta)$$

PROBLEM 4

$X_i = 0, 1, 2, \dots, N$

THE PMF IS:

$$f_x(x|\theta) = \prod_{i=1}^{N} \frac{\theta^{x_i} e^{-\theta}}{x_i!}$$

$$= \frac{\theta^{\sum_{i=1}^{N} x_i} e^{-m\theta}}{\prod_{i=1}^{N} x_i!}$$

RECALL THE FACTORIZATION THEOREM

$$f_x(x|\theta) = g(t|\theta)\, h(x)$$

for all support points $x \in \mathcal{X}$

for all $\theta \in \Theta$

$$= \underbrace{\theta^{\sum_{i=1}^{N} x_i} e^{-n\theta}}_{g(t|\theta)} \underbrace{\frac{1}{\prod_{i=1}^{N} x_i!}}_{h(x)}$$

$$T(X) = \sum_{i=1}^{N} X_i$$

$$t(x) = \sum_{i=1}^{N} x_i$$

PROBLEM 5

$X_1, X_2, \dots, X_m$ IID $U(\theta, \theta+1)$

$\dim(\theta) = 1$

$$T = T(X) = \begin{pmatrix} X_{(1)} \\ X_{(N)} \end{pmatrix}$$

$$f_X(X \mid \theta) = \prod_{i=1}^{N} \mathbb{1}(\theta < X_i < \theta + 1)$$

$$= \prod_{i=1}^{N} \mathbb{1}(X_i > \theta) \prod_{i=1}^{N} \mathbb{1}(X_i - 1 < \theta)$$

$$f_X(X \mid \theta) = g(t \mid \theta)\, h(x)$$

$$= \underbrace{I(X_{(1)} > \theta)\, I(X_{(m)} - 1 < \theta)}_{g(t_1, t_2 \mid \theta)} \underbrace{\prod_{i=1}^{N} I(x_i \in \mathbb{R})}_{h(x)}$$

$$X_{(1)} = \text{MIN}_{1 \leq i \leq N} X_i$$

$$X_{(n)} = \text{MAX}_{1 \leq i \leq N} X_i$$

$$t_1 = X_{(1)}$$

$$t_2 = X_{(n)}$$

$$2 = \dim(T) > \dim(\theta) = 1$$

PROBLEM 6

$Y_1, Y_2, \dots, Y_N$ ARE INDEPENDENT RVs

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$\varepsilon_i \sim \text{IID } N(0, \sigma^2)$$

x_i ARE FIXED CONSTANTS

$$\theta = (\beta_0, \beta_1, \sigma^2)$$

σ^2 IS NOT KNOWN

$$E(Y_i) = \beta_0 + \beta_1 x_i$$

$$Y_i \sim N\left(\beta_0 + \beta_1 x_i \,,\, \sigma^2\right)$$

$$f_Y(y \mid \theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma} \, \mathrm{EXP}\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2 \right\}$$

$$= \left(\frac{1}{2\pi \sigma^2} \right)^{\frac{N}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{N} \underbrace{(y_i - \beta_0 - \beta_1 x_i)^2}_{S} \right\}$$

$$S = \sum_{i=1}^{N} y_i^2 - 2 \sum_{i=1}^{N} y_i (\beta_0 + \beta_1 x_i) + \sum_{i=1}^{N} (\beta_0 + \beta_1 x_i)^2$$

$$= \sum_{i=1}^{N} y_i^2 - 2\beta_0 \sum_{i=1}^{N} y_i - 2\beta_1 \sum_{i=1}^{N} y_i x_i + \sum_{i=1}^{N} (\beta_0 + \beta_1 x_i)^2$$

$$f(y|\theta) =$$

$$\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \text{EXP}\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^N y_i^2 - 2\beta_0\sum_{i=1}^N y_i - 2\beta_1\sum_{i=1}^N y_i x_i + \sum_{i=1}^N(\beta_0+\beta_1 x_1)^2\right)\right\}$$

$$= g\left(\sum_{i=1}^N y_i^2, \sum_{i=1}^N y_i, \sum_{i=1}^N y_i x_i, \beta_0, \beta_1, \sigma^2\right) h(x)$$

$$= g(T(y), \theta)\, h(x)$$

WHERE:

$$T(y) = \left(\sum_{i=1}^N y_i^2, \sum_{i=1}^N y_i, \sum_{i=1}^N y_i x_i\right) = (t_1, t_2, t_3)$$

$$g(t,\theta) = (2\pi\sigma^2)^{-\frac{N}{2}} \text{EXP}\left\{-\frac{1}{2\sigma^2}\left(t_1 - 2\beta_0 t_2 - 2\beta_1 t_3 + \sum_{i=1}^N(\beta_0+\beta_1 x_i)^2\right\}\right.$$

AND $h(x) = 1$