Chapter 6. Point Estimation
Chapter 7. Interval Estimation

Math 3215 Summer 2023

Georgia Institute of Technology

Section 6.3.
Order Statistics

$X_1, X_2, \cdots, X_n$ i.i.d. ____

$u(X_1, X_2, \cdots, X_n)$: a statistic.

Example: $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$, $\dfrac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$,

$\min\{X_1, \cdots, X_n\}$

$\max\{X_1, \cdots, X_n\}$

median $\{ \text{——} \}$, quartiles, percentiles.

Order Statistics

Order statistics are the observations of the random sample, arranged, or ordered, in magnitude from the smallest to the largest.

Example

The values $x_1 = 0.62$, $x_2 = 0.98$, $x_3 = 0.31$, $x_4 = 0.81$, and $x_5 = 0.53$ are the $n = 5$ observed values of five independent trials of an experiment with pdf $f(x) = 2x$ for $0 < x < 1$.

The observed order statistics are $y_1 = 0.31$, $y_2 = 0.53$, $y_3 = 0.62$, $y_4 = 0.81$, $y_5 = 0.98$

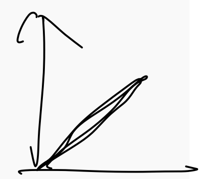
The sample median is $y_3 = 0.62$

The sample range is $y_5 - y_1 = \underline{0.98 - 0.31}$.

$[y_1, y_5]$

Order Statistics

$$F(x) = \begin{cases} 0 & ; \quad x < 0 \\ x^2 & ; \quad 0 < x < 1 \\ 1 & ; \quad x > 1 \end{cases}$$



Let X_1, X_2, X_3, X_4, X_5 be independent random variables with density $f(x) = 2x$ for $0 < x < 1$.

$$\mathbb{P}(X_1 = X_2) = 0 = \iint_A f(x_1) \cdot f(x_2)\, dx_1 dx_2 = 0$$

Let $Y_1 \leq Y_2 \leq Y_3 \leq Y_4 \leq Y_5$ be the order statistics of $X_i, i = 1, 2, 3, 4, 5.$

What is the distribution of Y_1?

$$Y_1 = \min\{X_i : i = 1, 2, \dots, 5\}$$

$$F_{Y_1}(y) = \mathbb{P}(Y_1 \leq y) = \mathbb{P}(\min\{X_i : i = 1, \dots, 5\} \leq y)$$

$$= 1 - \mathbb{P}(\min\{X_i\} > y)$$

$$= 1 - \mathbb{P}(X_1 > y \text{ and } X_2 > y \text{ and } \cdots)$$

$$= 1 - \underbrace{\mathbb{P}(X_1 > y)}_{(1 - F(y))} \underbrace{\mathbb{P}(X_2 > y)}_{(1 - F(y))} \cdots \underbrace{\mathbb{P}(X_5 > y)}_{(1 - F(y))}$$

$$= 1 - (1 - F(y))^5$$

$$f_{Y_1}(y) = -5(1 - F(y))^4 \cdot (-f(y))$$

$$= 5 \cdot (1 - F(y))^4 \cdot f(y).$$

Order Statistics

Let X_1, X_2, X_3, X_4, X_5 be independent random variables with density $f_X(x) = 2x$ for $0 < x < 1$.

Let $Y_1 \leq Y_2 \leq Y_3 \leq Y_4 \leq Y_5$ be the order statistics of $X_i, i = 1, 2, 3, 4, 5$.

What is the distribution of Y_5?

$$F_{Y_5}(y) = \mathbb{P}(Y_5 \leq y) = \mathbb{P}(\max\{X_i\} \leq y)$$

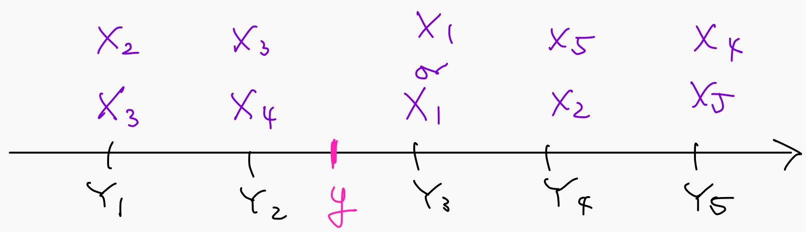
$$= \mathbb{P}(X_1 \leq y \text{ and } X_2 \leq y \text{ and } \cdots X_5 \leq y)$$

$$= \mathbb{P}(X_1 \leq y)\,\mathbb{P}(X_2 \leq y) \cdots \mathbb{P}(X_5 \leq y)$$

$$= \underbrace{F(y) \cdot F(y) \cdots F(y)}_{5 \text{ times}} = F(y)^5$$

$$f_{Y_5}(y) = 5 \cdot F_X(y)^4 \cdot f_X(y).$$

Order Statistics



Let X_1, X_2, X_3, X_4, X_5 be independent random variables with density $f(x) = 2x$ for $0 < x < 1$.

Let $Y_1 \le Y_2 \le Y_3 \le Y_4 \le Y_5$ be the order statistics of $X_i, i = 1, 2, 3, 4, 5$.

What is the distribution of Y_2?

$$F_{Y_2}(y) = \mathbb{P}(Y_2 \le y) = \mathbb{P}(Y_1 \le y,\ Y_2 \le y,\ Y_3 > y,\ Y_4 > y,\ Y_5 > y)$$

$$= \mathbb{P}\big(X_3 \le y,\ X_4 \le y,\ X_1 > y,\ X_2 > y,\ X_5 > y$$

or $\cup$

$$X_2 \le y,\ X_3 \le y,\ X_1 > y,\ X_5 > y,\ X_4 > y$$

or $\cup$

$$\vdots \big)$$

$$= \mathbb{P}(\underline{\qquad}) + \mathbb{P}(\underline{\qquad}) + \cdots +$$

$$= \mathbb{P}(X_3 \le y)\cdot\mathbb{P}(X_4 \le y)\cdots\mathbb{P}(X_5 > y) + \cdots$$

$$= \binom{5}{2} F(y)\cdot F(y)\,(1-F(y))(1-F(y))(1-F(y)) = \binom{5}{2} F(y)^2 (1-F(y))^3$$

$$F_{Y_2}(y) = \mathbb{P}(Y_2 \le y)$$

$$= \mathbb{P}(Y_2 \le y,\ Y_3 > y) + \mathbb{P}(Y_3 \le y,\ Y_4 > y)$$

$$+ \mathbb{P}(Y_4 \le y,\ Y_5 > y)$$

$$+ \mathbb{P}(Y_5 \le y)$$

$$= \binom{5}{2} F(y)^2 (1 - F(y))^3 + \binom{5}{3} F(y)^3 (1 - F(y))^2$$

$$+ \binom{5}{4} F(y)^4 (1 - F(y)) + \binom{5}{5} F(y)^5$$

$$= \sum_{k=2}^{5} \binom{5}{k} F(y)^k (1 - F(y))^{5-k}$$

Order Statistics

In general, if $Y_1 < Y_2 < \cdots < Y_n$ are the order statistics of independent random variables X_i for $i = 1, 2, \cdots, n$ with pdf f,

then the distribution of Y_i is

$$F_{Y_i}(y) = \sum_{k=i}^{n} \binom{n}{k} F(y)^{k} \left(1 - F(y)\right)^{n-k}$$

$$f_{Y_i}(y) = \frac{d}{dy} F_{Y_i}(y)$$

$$= \sum_{k=i}^{n} \overbrace{\binom{n}{k} \cdot k}^{= \binom{n-1}{k-1} \cdot n} \cdot F(y)^{k-1} \cdot f(y) \left(1 - F(y)\right)^{n-k}$$

$$+ \sum_{k=i}^{n} \underbrace{\binom{n}{k} F(y)^{k} (n-k)} \left(1 - F(y)\right)^{n-k-1} \cdot (-f(y))$$

$$\frac{n!}{k!\,(n-k)!} \cdot (n-k) = \frac{n!}{k!\,(n-k-1)!} = n \binom{n-1}{k}$$

$$= \frac{n!}{(i-1)!\,(n-i)!} \left(F(y)\right)^{i-1} \left(1 - F(y)\right)^{n-i} \cdot f(y).$$

Order Statistics

Example

Let X_1, X_2, X_3, X_4, X_5 be independent random variables with density $f(x) = 2x$ for $0 < x < 1$.

Let $Y_1 \leq Y_2 \leq Y_3 \leq Y_4 \leq Y_5$ be the order statistics of $X_i, i = 1, 2, 3, 4, 5$.

Find $\mathbb{P}(Y_3 \leq y)$.

Find the density of Y_3.

$$\mathbb{P}(Y_3 \leq y) = \begin{cases} \sum_{k=3}^{5} \binom{5}{k} (y^2)^k \cdot (1-y^2)^{5-k}, & 0 \leq y \leq 1 \\ 0, & y < 0 \\ 1, & y > 1 \end{cases}$$

$$r\binom{n}{r}$$

$$3 \cdot \binom{5}{3}$$

$$f_{Y_3}(y) = \frac{5!}{(3-1)!\,(5-3)!} (y^2)^{3-1} (1-y^2)^{5-3} \cdot (2y)$$

$$= 30 \cdot y^4 (1-y^2)^2 \cdot 2 \cdot y \qquad 0 \leq y \leq 1$$

Exercise

$$\begin{cases} F(x) = 1 - e^{-\frac{x}{\theta}} \\ f(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}} \end{cases}, \quad \mathbb{P}(X \geq y) = e^{-y/\theta}$$

Let $Y_1 < Y_2 < \cdots < Y_{19}$ be the order statistics of $n = 19$ independent observations from the exponential distribution with mean θ.

Find the pdf of Y_1.

Find $\mathbb{E}[F(Y_1)]$ where F is the cdf of the exponential distribution. with mean θ.

$$F_{Y_1}(y) = \mathbb{P}(Y_1 \leq y) = 1 - \mathbb{P}(Y_1 > y) = 1 - \mathbb{P}(X > y)^{19}$$

$$= 1 - \left(e^{-\frac{y}{\theta}}\right)^{19} = 1 - e^{-\frac{19y}{\theta}}$$

$$f_{Y_1}(y) = \frac{19}{\theta} e^{-\frac{19y}{\theta}} \qquad Y_1 \sim \mathrm{Exp}\left(\frac{19}{\theta}\right)$$

$$\mathbb{E}[F(Y_1)] = \int_0^\infty F(y) \frac{19}{\theta} e^{-\frac{19y}{\theta}} \, dy$$

$$= \int_0^\infty \left(1 - e^{-\frac{y}{\theta}}\right) \frac{19}{\theta} e^{-\frac{19y}{\theta}} \, dy$$

$$= \frac{19}{\theta} \cdot \left[\int_0^\infty e^{-\frac{19y}{\theta}} \, dy - \int_0^\infty e^{-\frac{20y}{\theta}} \, dy \right]$$

$$= \frac{19}{\theta} \cdot \left[\frac{\theta}{19} - \frac{\theta}{20} \right] = 1 - \frac{19}{20} = \frac{1}{20}.$$

$W = X + Y$

$$\overline{W} = \frac{1}{n}\big(\overset{W_1}{\overset{\shortparallel}{(X_1 + Y_1)}} + \cdots + \overset{W_n}{\overset{\shortparallel}{(X_n + Y_n)}} \big)$$

$$= \overline{X} + \overline{Y}$$

$$E[\overline{W}] = E\overline{X} + E\overline{Y} = EX + EY = 80$$

$$Var(\overline{W}) = Var(\overline{X} + \overline{Y}) = \frac{144}{25} = \frac{\sigma^2}{n}$$

$$Var(W_1) = Var(X + Y) = \underline{Var(X)} + \underline{Var(Y)} + 2\underbrace{Cov(X, Y)}$$

$$= 52 + 64 + 2 \cdot 14$$

$$= 144 = \sigma^2$$

$$\frac{\overline{W} - 80}{\sqrt{144/25}} \Rightarrow N(0, 1)$$

<u>CLT</u> :

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{\overline{X} - E[\overline{X}]}{\sqrt{Var(\overline{X})}}$$

$$\sigma/\sqrt{n} = \sqrt{Var(\overline{X})}$$

$$\sigma = \sqrt{Var(X)}$$

Section 6.4.
Maximum Likelihood and Method of Moments Estimation

Maximum Likelihood Estimators
(MLE)

Estimators

We consider an experiment whose outcome follows a certain distribution.

Assume that the distribution is characterized by a pdf $f(x;\theta)$ where θ is an unknown parameter.

Let Ω be the set of all possible para$\overset{\theta}{\text{m}}$eters. We call it the parameter space.

Suppose we repeat the n independent experiment and observe the outcome (sample) $X_1, X_2, \cdots, X_n$.

Our goal is to estimate θ based on the samples.

Example

- $f(x;\theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}}$ $\quad \Omega = (0, \infty)$ or $[1, 2]$
- $f(x;\theta_1,\theta_2) = \frac{1}{\sqrt{2\pi}\,\theta_2^2} e^{-\frac{|x-\theta_1|^2}{2\theta_2^2}}$

Estimators

Definition

For a function $u : \mathbb{R}^n \to \mathbb{R}$, $u(X_1, X_2, \cdots, X_n)$ is called an estimator (or a point estimator) of θ.

Estimators

PMF $\begin{cases} f(1) = \underline{p} \\ f(0) = 1-p \end{cases}$, $\Omega = [0, 1]$

Example

Suppose $X_1, X_2, \cdots, X_n$ are independent Bernoulli RVs with unknown success probability $p \in [0, 1]$.

We want to find an estimator of p which maximizes its possibility.

Find $\underline{u(x_1, \cdots, x_n)} = ?$

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n)$$

$$= \underline{p^{\sum_{i=1}^n x_i} \cdot (1-p)^{n - \sum_{i=1}^n x_i}}$$

a function of $\underline{p}$ parameter.

maximize in p.

$$F(p) = \ln\left(p^{\sum x_i} \cdot (1-p)^{n - \sum x_i}\right)$$

$$= (\ln p) \cdot \left(\sum_{i=1}^n x_i\right) + \ln(1-p) \cdot \left(n - \sum_{i=1}^n x_i\right)$$

$$F'(p) = \frac{1}{p} \cdot \left(\sum x_i\right) - \frac{1}{1-p}\left(n - \sum x_i\right) = 0$$

$$(1-p)\left(\sum x_i\right) - p\left(n - \sum x_i\right) = 0$$

$$\sum x_i - \cancel{p \cdot \sum x_i} - np + \cancel{p \sum x_i} = 0$$

$$\sum x_i = np$$

$$p = \frac{\sum x_i}{n}$$

$$u(x_1, \ldots, x_n) = \frac{\sum x_i}{n} \quad \text{sample mean.}$$

an estimator for p

$1, 0, 0, 1, 1, 1, 0.$

$$\frac{4}{7} = p$$

Maximum Likelihood Estimators

Definition

Let $X_1, X_2, \cdots, X_n$ be a random sample from a distribution $f(x; \theta_1, \cdots, \theta_m)$, $(\theta_1, \cdots, \theta_m) \in \Omega$.

The likelihood function of $\theta_1, \cdots, \theta_m$ is $P(X_1 = x_1, \cdots, X_n = x_n)$

$$= f(x_1; p) \cdot f(x_2; p) \cdots f(x_n; p)$$

The maximum likelihood estimators of $\theta_1, \cdots, \theta_m$ are

$$\begin{cases} \hat{\theta}_{1\,MLE} = u_1(x_1, \cdots, x_n) \\ \vdots \\ \hat{\theta}_{m\,MLE} = u_m(x_1, \cdots, x_n) \end{cases}$$

which maximize the likelihood function.

Example

$$\hat{P}_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Maximum Likelihood Estimators

Example

Suppose $X_1, X_2, \cdots, X_n$ are independent exponential with mean $\theta \in (0, \infty)$.

Likelihood function $= f(x_1;\theta)\, f(x_2;\theta) \cdots f(x_n;\theta)$

$$= \frac{1}{\theta} e^{-\frac{x_1}{\theta}} \cdot \frac{1}{\theta} e^{-\frac{x_2}{\theta}} \cdots \frac{1}{\theta} e^{-\frac{x_n}{\theta}}$$

$$= \underline{\theta^{-n} \cdot e^{-\frac{\sum x_i}{\theta}}}$$

maximize in θ.

$$F(\theta) = \ln\left(\theta^{-n} e^{-\frac{\sum x_i}{\theta}}\right)$$

$$= -n \cdot \ln\theta - \frac{\sum x_i}{\theta}$$

$$F'(\theta) = -\frac{n}{\theta} + \frac{\sum x_i}{\theta^2} = 0$$

$$-n\theta + \sum x_i = 0$$

$$\underbrace{\theta = \frac{\sum x_i}{n}}$$

$$\hat{\theta}_{MLE} = \frac{\sum x_i}{n}.$$

Maximum Likelihood Estimators

$$f(x;\theta_1,\theta_2) = \frac{1}{\sqrt{2\pi\theta_2}} e^{-\frac{|x-\theta_1|^2}{2\theta_2}}$$

Example

Suppose $X_1, X_2, \cdots, X_n$ are normal with mean $\theta_1 \in \mathbb{R}$ and variance $\theta_2 \in (0,\infty)$.

Likelihood function $= f(x_1;\theta_1,\theta_2) \cdots f(x_n;\theta_1,\theta_2)$

$$= \frac{1}{\sqrt{2\pi\theta_2}} e^{-\frac{|x_1-\theta_1|^2}{2\theta_2}} \cdots \frac{1}{\sqrt{2\pi\theta_2}} e^{-\frac{|x_n-\theta_1|^2}{2\theta_2}}$$

$$= (2\pi\theta_2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\theta_2}\sum_{i=1}^{n}|x_i-\theta_1|^2\right)$$

$$F(\theta_1,\theta_2) = \ln\left((2\pi\theta_2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\theta_2}\sum_{i=1}^{n}|x_i-\theta_1|^2\right)\right)$$

$$(\ln\theta_2 + \ln 2\pi)$$

$$= -\frac{n}{2}\log(2\pi\theta_2) - \frac{1}{2\theta_2}\left(\sum_{i=1}^{n}|x_i-\theta_1|^2\right)$$

$$\frac{\partial F}{\partial\theta_1} = -\frac{1}{\cancel{2}\theta_2}\sum_{i=1}^{n}\left(\cancel{2}(x_i-\theta_1)\right) = 0$$

$$\sum_{i=1}^{n}(x_i-\theta_1) = 0$$

$$\sum_{i=1}^{n} x_i - \underbrace{\sum_{i=1}^{n} \theta_1}_{\overset{\shortparallel}{n\theta_1}} = 0 \Rightarrow \theta_1 = \frac{\sum x_i}{n}$$

$$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \shortparallel$$

$$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \hat{\theta}_{1\,MLE}.$$

$$\frac{\partial F}{\partial \theta_2} = -\frac{n}{2\theta_2} + \frac{1}{2\theta_2^2} \sum_{i=1}^{n} |x_i - \theta_1| = 0$$

$$\theta_2 = \frac{1}{n} \sum_{i=1}^{n} \Big| x_i - \underbrace{\frac{\sum x_i}{n}}_{} \overset{=\theta_1}{} \Big|^2 = (\hat{\theta}_2)_{MLE}$$

$\uparrow$

MLE for θ_2

Unbiased estimators

Definition

If an estimator $u(X_1, X_2, \cdots, X_n)$ of θ satisfies $\mathbb{E}[u(X_1, X_2, \cdots, X_n)] = \theta$, we say the estimator is unbiased.

Example $N(\theta_1, \theta_2)$

$$\hat{\theta}_1 = \frac{\sum X_i}{n}$$

$$\mathbb{E}\left[\frac{1}{n}\sum X_i\right] = \frac{1}{n}\sum \mathbb{E}X_i = \frac{1}{n}\cdot n\cdot\theta_1 = \theta_1$$

$\Rightarrow$ $\hat{\theta}_1$ is unbiased.

$$\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n}\left|X_i - \frac{\sum X_j}{n}\right|^2 = \frac{1}{n}\sum_{i=1}^{n}\left|\underbrace{(X_i-\theta_1)}_{\tilde{X}_i} - \frac{\sum \overbrace{(X_j-\theta_1)}^{\tilde{X}_j}}{n}\right|^2$$

$$\mathbb{E}[\hat{\theta}_2] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\tilde{X}_i^2 - 2\cdot\tilde{X}_i\cdot\frac{\sum_{j=1}^{n}\tilde{X}_j}{n} + \frac{\left(\sum_{j=1}^{n}\tilde{X}_j\right)^2}{n^2}\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\underbrace{\mathbb{E}[\tilde{X}_i]^2}_{=\theta_2} - \frac{2}{n}\underbrace{\mathbb{E}\left[\tilde{X}_i\cdot\sum\tilde{X}_j\right]}_{=\mathbb{E}[\tilde{X}_i^2] = \theta_2} + \frac{1}{n}\theta_2\right)$$

$$E[\tilde{X}_1 \cdot \tilde{X}_2] = \underbrace{E[\tilde{X}_1]}_{=0} \underbrace{E[\tilde{X}_2]}_{=0} = 0$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left(\theta_2 - \frac{2}{n}\theta_2 + \frac{1}{n}\theta_2 \right) = \frac{n-1}{n} \theta_2$$

$\hat{\theta}_2$ is biased.

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left| X_i - \frac{\sum X_j}{n} \right|^2$$

degree of freedom: sample variance.

$\Rightarrow$ S^2 is an unbiased estimator for θ_2

Unbiased estimators

Example

Suppose $X_1, X_2, \cdots, X_n$ are normal with mean $\theta_1 \in \mathbb{R}$ and variance $\theta_2 \in (0, \infty)$.

Exercise

A random sample $X_1, X_2, \cdots, X_n$ of size n is taken from a Poisson distribution with a mean of $\theta \in (0, \infty)$.

Find the MLE for θ.

$$\text{Likelihood function} = \left(e^{-\theta} \frac{\theta^{x_1}}{x_1!}\right) \cdot \left(e^{-\theta} \frac{\theta^{x_2}}{x_2!}\right) \cdots \left(e^{-\theta} \cdot \frac{\theta^{x_n}}{x_n!}\right)$$

$$= e^{-n\theta} \frac{\theta^{\sum x_i}}{x_1!\, x_2! \cdots x_n!} = e^{-n\theta} \frac{\theta^{\sum x_i}}{\left(\prod_{i=1}^{n} x_i!\right)}$$

$$F(\theta) = -n\theta + \left(\sum x_i\right) \ln \theta - \ln\left(x_1!\, x_2! \cdots x_n!\right)$$

$$F'(\theta) = -n + \left(\sum x_i\right) \cdot \frac{1}{\theta} = 0 \qquad \therefore \theta = \frac{\sum x_i}{n}.$$

Ex) $X_1, X_2, \cdots, X_n$ i.i.d. Unif on $\cancel{[0, \underline{\theta}]}$

$[\theta_1, \theta_2]$

$\underline{\hat{\theta}_{MLE}} = ?$

Likelihood function $= f(x_1;\theta)\, f(x_2;\theta) \cdots f(x_n;\theta)$

$= \left(\begin{cases} \frac{1}{\theta} \\ 0 \end{cases}\right) \cdot \left(\begin{cases} \frac{1}{\theta} \\ 0 \end{cases}\right) \cdots \left(\begin{cases} \frac{1}{\theta} \\ 0 \end{cases}\right)$

$= \begin{cases} \frac{1}{\theta^n} & \text{if all } x_i \in [0, \theta] \\ 0 & \text{o.w.} \end{cases}$

① $x_i \in [0, \theta]$ $\underline{\underline{\forall i}}$

$0 \leq x_i \leq \theta \Rightarrow \theta \geq \max\{x_1, \cdots, x_n\}$

$\Rightarrow \theta = \underline{\max\{x_1, \cdots, x_n\}} = \hat{\theta}_{MLE}.$

Section 7.1.
Confidence Intervals for Means

Two-sided confidence intervals

Consider a random sample $X_1, X_2, \cdots, X_n$ from a normal distribution $N(\mu, \sigma^2)$.

We are interested in the closeness of the maximum likelihood estimator $\overline{X}$ to the unknown mean μ.

$\hat{\mu}_{MLE} = \overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$

Suppose σ^2 is known.

Our goal is to find an interval $[a, b]$ such that $\mathbb{P}(\overline{X} \in [a, b]) = 1 - \alpha$, where $\alpha > 0$.

$\alpha = 0.05$

Let $z_{\frac{\alpha}{2}} > 0$ be such that $\boxed{\mathbb{P}(Z \geq z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}}$. (Here, Z is standard normal.)

Notation

$= \mathbb{P}(Z > z_{\frac{\alpha}{2}})$

PDF of Z

$$\boxed{\mathbb{P}(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}) = 1 - \alpha}$$

$X_1, \dots, X_n \sim N(\mu, \sigma^2)$

μ unknown, σ^2 known

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \Rightarrow \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

$$1-\alpha = \mathbb{P}(a \le \overline{X} \le b) = \mathbb{P}\left(\underbrace{\frac{a-\mu}{\sigma/\sqrt{n}}}_{= -z_{\frac{\alpha}{2}}} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le \underbrace{\frac{b-\mu}{\sigma/\sqrt{n}}}_{= z_{\frac{\alpha}{2}}}\right)$$

Two-sided confidence intervals

Since $(\overline{X} - \mu)/(\sigma/\sqrt{n})$ is standard normal, we have

$$\mathbb{P}(-z_{\frac{\alpha}{2}} \le \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \le z_{\frac{\alpha}{2}}) = 1-\alpha.$$

Solving for $\overline{X}$, we get

$\mu \le \overline{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$

$$\mathbb{P}\left(\underbrace{\mu - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le \overline{X} \le \mu + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}}_{2 \text{ ineq.}}\right) = 1-\alpha.$$

$\mu \ge \overline{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

$$\mathbb{P}\left(\overline{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$\mu \in \left[\overline{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \ \overline{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

with prob. $1-\alpha$

(for example, $\alpha = 0.05$)

95%

Two-sided confidence intervals

Definition

The interval $\left[\overset{\overline{X}}{\cancel{\mu}} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overset{\overline{X}}{\cancel{\mu}} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$ is called the two-sided confidence interval for unknown μ with confidence coefficient $1-\alpha$ (for $100(1-\alpha)\%$).

Two-sided confidence intervals

Example

Let X equal the length of life of a 60-watt light bulb marketed by a certain manufacturer.

Assume that the distribution of X is $N(\mu, 1296)$.

If a random sample of $n = 27$ bulbs is tested until they burn out, yielding a sample mean of $\overline{X} = 1478$ hours,

then what is a 95% confidence interval for μ?

$$\left[\overline{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \ , \ \overline{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right]$$

$\overline{X} = 1478$

$\sigma = (36) = \sqrt{1296}$

$\sqrt{n} = \sqrt{27} = 3\sqrt{3}$

$z_{\frac{\alpha}{2}} = 1.96$

$\alpha = 0.05$

Two-sided confidence intervals

If we cannot assume that the distribution from which the sample arose is normal, we can still obtain an approximate confidence interval for μ.

By the central limit theorem, provided that n is large enough, the ratio $(\overline{X} - \mu)/(\sigma/\sqrt{n})$ has the approximate normal distribution $N(0, 1)$ when the underlying distribution is not normal.

$$\mu \in \left[\overline{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} , \overline{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right]$$

with $(1-\alpha)$ prob.

Two-sided confidence intervals

Example

Let X equal the amount of orange juice (in grams per day) consumed by an American. Suppose it is known that the standard deviation of X is $\sigma = 96$.

An orange growers' association took a random sample of $n = 576$ Americans and found that they consumed, on average, $\overline{X} = 133$ grams of orange juice per day.

$$\mu \in \left[133 - z_{\frac{\alpha}{2}} \cdot \frac{96}{\sqrt{576}} , \quad 133 + z_{\frac{\alpha}{2}} \cdot \frac{96}{\sqrt{576}} \right]$$

with prob. $1-\alpha$

When σ^2 is unknown

Even though σ^2 is unknown and the sample distribution is not normal, if n is large enough ($n \geq 30$), then

$$\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}, \qquad \boxed{S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

σ replace σ with S.

is approximately the standard normal.

Thus, a $100(1-\alpha)\%$ confidence interval is $\left[\overline{X} - z_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right]$

If the sample distribution is normal, then $\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}}$ has a student's t distribution with degree of freedom $n-1$.

When σ^2 is unknown

Example

Let X equal the amount of butterfat in pounds produced by a typical cow during a 305-day milk production period between her first and second calves.

Assume that the distribution of X is $N(\mu, \sigma^2)$.

To estimate, a farmer measured the butterfat production for $n = 20$ cows and $\overline{X} = 507.50$ and $S = 89.75$.

Exercise

A random sample of size 16 from the normal distri bution $N(\mu, 25)$ yielded $\overline{X} = 73.8$.

Find a 95% confidence interval for μ.

Section 7.2.
Confidence Intervals for the Difference of Two Means

Confidence Intervals for the Difference of Two Means

Let $X_1, X_2, \cdots, X_{n_X}$ and $Y_1, Y_2, \cdots, Y_{n_Y}$ be, respectively, two independent random samples of sizes n_X and n_Y from the two normal distributions $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$.

The distribution of $W = \overline{X} - \overline{Y}$ is

The confidence interval for $\mu_X - \mu_Y$ is

This also works when

Confidence Intervals for the Difference of Two Means

Example

The length of life of brand X light bulbs is assumed to be $N(\mu_X, 784)$.

The length of life of brand Y light bulbs is assumed to be $N(\mu_Y, 627)$ and independent of X.

If a random sample of $n_X = 56$ brand X light bulbs yielded a mean of $\overline{X} = 937.4$ hours and a random sample of size $n_Y = 57$ brand Y light bulbs yielded a mean of $\overline{Y} = 988.9$ hours,

find a 90% confidence interval for $\mu_X - \mu_Y$.

Confidence Intervals for the Difference of Two Means

Suppose σ_X^2 and σ_Y^2 are unknown and the sample sizes are small.

Assume that $\sigma_X^2 = \sigma_Y^2$.

Then,

$$T = \frac{Z}{\sqrt{U/(n_X + n_Y - 2)}}$$

has a student's t distribution of degree of freedom $n_X + n_Y - 2$ where

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\sigma^2/n_X + \sigma^2/n_Y}}, \qquad U = \frac{(n_X - 1)S_X^2}{\sigma^2} + \frac{(n_Y - 1)S_Y^2}{\sigma^2}.$$

After simplication, T can be written as

Confidence Intervals for the Difference of Two Means

Example

Suppose that scores on a standardized test in mathematics taken by students from large and small high schools are $N(\mu_X, \sigma^2)$ and $N(\mu_Y, \sigma^2)$, respectively, where σ unknown.

If a random sample of $n_X = 9$ students from large high schools yielded $\overline{X} = 81.31$ and $S_X^2 = 60.76$, and a random sample of $n_Y = 15$ students from small high schools yielded $\overline{Y} = 78.61$ and $S_Y^2 = 48.24$.

Find a 95% confidence interval for $\mu_X - \mu_Y$.

Section 7.3.
Confidence Intervals for Proportions

Confidence Intervals for Proportions

Example

In a certain political campaign, one candidate has a poll taken at random among the voting population.

The results are that $y = 185$ out of $n = 351$ voters favor this candidate.

Should the candidate feel very confident of winning?

Confidence Intervals for Proportions

In general, suppose Y is a random sample from a binomial distribution with the given number of trials n and an unknown success probability p.

Note that Y/n is an unbiased point estimator for p and

$$\frac{(Y/n) - p}{\sqrt{p(1-p)/n}}$$

has an approximate normal distribution $N(0, 1)$.

One-sided confidence interval

One-sided confidence intervals are sometimes appropriate for p.

For example, we may be interested in an upper bound on the proportion of defectives in manufacturing some item.

For an upper bound, we consider

$$\mathbb{P}\left(-z_\alpha \leq \frac{(Y/n) - p}{\sqrt{(Y/n)(1-(Y/n))/n}}\right) \approx 1 - \alpha.$$

This gives

A lower bound is

One-sided confidence interval

Example

The Wisconsin Department of Natural Resources (DNR) wished to determine the prevalence, p, of chronic wasting disease (a neurological disease similar to mad cow disease) among its whitetail deer population.

In a particular area of the state, 272 deer were legally killed by hunters in a particular season. A tissue sample from each animal was submitted to the DNR.

Laboratory analysis determined that nine of the deer had chronic wasting disease.

Assume that the 272 harvested deer can, to a good approximation, be regarded as a random sample.

Find a one-sided approximate 95% confidence interval that provides an upper bound for p in that area of Wisconsin.

Difference of two proportions

Suppose Y_1 and Y_2 are independent binomial random variables with n_1, n_2 and p_1, p_2.

To make a statistical inference about the difference $p_1 - p_2$, we consider $Y_1/n_1 - Y_2/n_2$.

Its mean is

Its variance is

So the CLT gives

Difference of two proportions

Example

Two detergents were tested for their ability to remove stains of a certain type.

An inspector judged the first one to be successful on 63 out of 91 independent trials and the second one to be successful on 42 out of 79 independent trials.

The respective relative frequencies of success are 0.692 and 0.532.

What is an approximate 90% confidence interval for the difference $p_1 - p_2$ of the two detergents?

Exercise

A machine shop manufactures toggle levers.

A lever is flawed if a standard nut cannot be screwed onto the threads.

Let p equal the proportion of flawed toggle levers that the shop manufactures.

There were 24 flawed levers out of a sample of 642 that were selected randomly from the production line.

Find an approximate 95% confidence interval for p.