Lecture 11

Confidence Intervals (Interval Estimation)

- A point estimator $\hat{\theta}$ for θ is a single value, and we can talk about its proximity to θ only, and chance of exactly hitting θ with $\hat{\theta}$ is slim, i.e. we cannot attach any "confidence" for $\hat{\theta}$ hitting θ.
- Instead it is more reasonable to estimate θ with a range of "plausible values", e.g. an interval, for which we can attach some level of confidence.
- This "set of plausible values" is called an "interval estimator" or more generally "confidence interval" for θ.
- To construct conf. intervals, we first need to define percentiles or quantiles.

Percentile: 100αth percentile of a distribution is a value x_α so that

$$P(X \leq x_\alpha) \geq \alpha \quad \& \quad P(X < x_\alpha) \leq \alpha$$

So, if X is cts

$$P(X \leq x_\alpha) = P(X < x_\alpha) = \alpha.$$

- $100\alpha^{th}$ percentile of a data set, D, is smallest element $x \in D$ s.t. at least 100α% of the data points are less than or equal to x.

Ex: $D = \{7, 9, 1, 6, 2, 10\}$ 40^{th} percentile? $\hat{x}_{0.40} = ?$

$D_{sort} = \{1, 2, 6, 7, 9, 10\}$

$x=1$ $P(X_i \leq 1) = \frac{1}{6} \approx 0.17 \not\geq 0.40$

$x=2$ $P(X_i \leq 2) = \frac{2}{6} \approx 0.33 \not\geq 0.40$

$x=6$ $P(X_i \leq 6) = \frac{3}{6} = 0.50 \geq 0.40 \Rightarrow \underline{\hat{x}_{0.40} = 6}$

How to find $100\alpha^{th}$ percentile efficiently?

$|D| = n \Rightarrow$ find $\lceil \alpha \times n \rceil^{th}$ smallest value

sample size

"ceiling" $= \begin{cases} \text{round } \underline{\text{up}} \text{ to next integer if decimal} \\ \alpha . n \text{ if integer} \end{cases}$

Ex (above) For 40^{th} percentile, $\alpha = 0.40$, $n = 6$

so $\lceil 6 \times 0.40 \rceil = \lceil 2.4 \rceil = 3$

$\Rightarrow 3^{rd}$ smallest value $= 6 = \hat{x}_{0.40}$

Ex: $\alpha = 0.90$ & $n = 1000 \Rightarrow 90^{th}$ percentile is

$\lceil 1000 \times 0.90 \rceil = \lceil 900 \rceil = 900$

$\Rightarrow 900^{th}$ smallest value.

Quartiles 25^{th} 50^{th} 75^{th} percentiles

25^{th}: first quartile, "lower quartile", Q_1

50^{th}: median, Q_2

75^{th}: 3^{rd} quartile, "upper quartile", Q_3

- Quartiles are useful in finding "outliers"

  " " " " "outlier detection"

- An outlier is very different (i.e. extremely larger or extremely smaller) than (most of the) remaining data.

- IQR = Inter-quartile range = $Q_3 - Q_1$

Tukey's Rule for outlier detection

Elements of D (i.e. data points) are outliers if:

data point > $Q_3 + \alpha \times IQR$

or

data point < $Q_1 - \alpha \times IQR$

where $\alpha = 1.5$

Ex.: $D = \{9, 1, 77, 6, 4, 26\}$ outlier(s)?

$D_{sort} = \{1, 4, 6, 9, 26, 77\}$ $n = 6$

$Q_1 = 25^{th}$ percentile " $\lceil 6 \times 0.25 \rceil = \lceil 1.5 \rceil = 2^{nd}$ value

$\Rightarrow Q_1 = 4$

$Q_3 = 75^{th}$ percentile " $\lceil 6 \times 0.75 \rceil = \lceil 4.5 \rceil = 5^{th}$ value

$\Rightarrow Q_3 = 26$

$IQR = Q_3 - Q_1 = 26 - 4 = 22$

- x_i is an outlier

if $x_i > Q_3 + 1.5\, IQR = 26 + 1.5 \times 22 =$

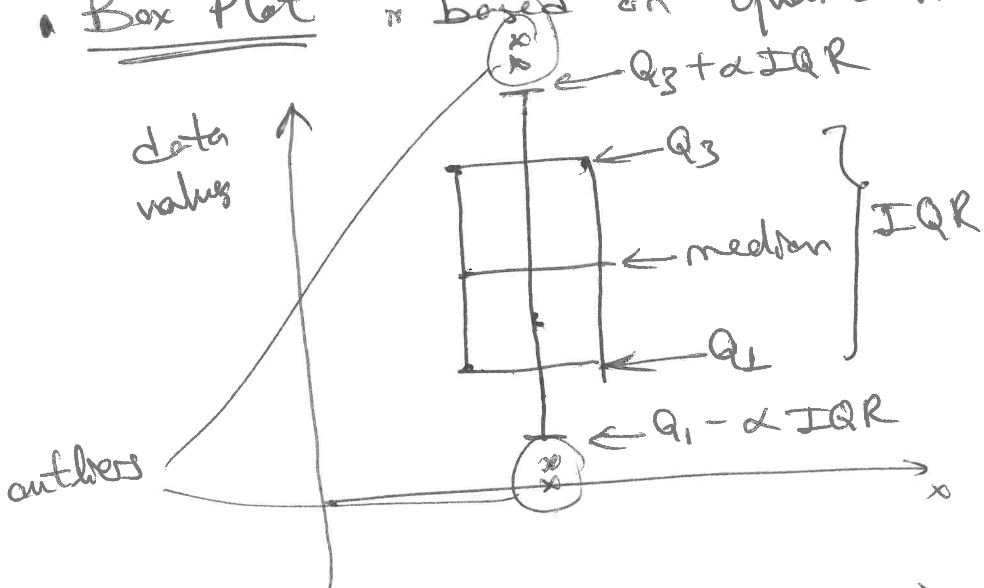
$= 26 + 33 = 59$

or $x_i < Q_1 - 1.5\, IQR = 4 - 33 = -29$

So, **77** is the only outlier in this data set.

- **Box Plot** is based on quartiles.

usually $\alpha = 1.5$



Confidence Intervals (CI)

$D = \{X_1, X_2, \ldots, X_n\}$, $X_i \overset{iid}{\sim} F$

For an unknown parameter θ, and $0 < \alpha < 1$

the $(1-\alpha)$ CI is (A, B) such that

$$P(\theta \in (A, B)) \geq 1 - \alpha,$$

constant RVs. (i.e. here interval is random not θ)

- $(1-\alpha)$ CI $\equiv (1-\alpha)100\%$ CI
- Typically $A = g(X_1, \ldots, X_n)$ & $B = h(X_1, \ldots, X_n)$

- A & B are computed using $\hat{\theta} = T(\underbrace{X_1, \dots, X_n}_{\text{RVs.}})$
  - A: RV, B: RV, $\hat{\theta}$: RV.

<u>Interpretation</u>

$$D_1 = \{X_{1,1}, X_{1,2}, \dots, X_{1,n}\} \longrightarrow \hat{\theta}_1, \ (A_1, B_1)$$

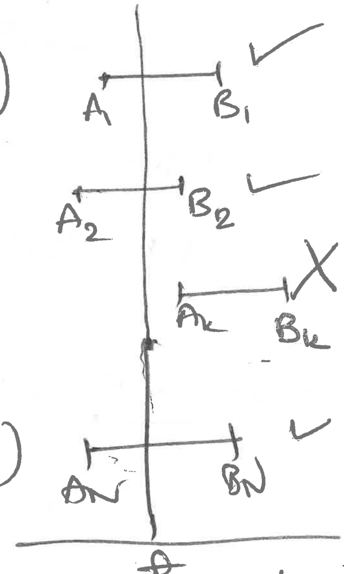
$$D_2 = \{X_{2,1}, X_{2,2}, \dots, X_{2,n}\} \longrightarrow \hat{\theta}_2, \ (A_2, B_2)$$

$$\vdots$$

$$D_k = \{X_{k,1}, X_{k,2}, \dots, X_{k,n}\} \longrightarrow \hat{\theta}_k \ (A_k, B_k)$$

$$\vdots$$

$$D_N = \{X_{N,1}, X_{N,2}, \dots, X_{N,n}\} \longrightarrow \hat{\theta}_N, \ (A_N, B_N)$$



- On average $(1-\alpha)100\%$ of the intervals (e.g. 95% of the intervals) contain θ, & as $N \longrightarrow \infty$, $P(\theta \in (A, B)) \geq 1-\alpha$.
- Usually $(A, B) = (\hat{\theta} - \delta, \hat{\theta} + \delta) = \hat{\theta} \pm \delta$
- <u>Ex:</u> $D = \{X_1, \dots, X_n\}$, $X_i \overset{iid}{\sim} F$ with $E[X_i] = \mu$
  estimator $\hat{\mu}$ for μ.
  95% CI for μ? i.e. find A & B so that
  $$P(\mu \in (A, B)) \geq 0.95.$$
  i.e. $$P(\mu \in (\hat{\mu} - \delta, \hat{\mu} + \delta)) \geq 0.95$$
- Trivially; take $\delta = \infty$,
  so $$P(\mu \in (\hat{\mu} - \infty, \hat{\mu} + \infty)) = P(\mu \in (-\infty, \infty)) = 1 \geq 0.95$$
  but $(-\infty, \infty)$ is <u>a useless interval</u> (to estimate μ).
  (why?)

Case 1: If $\delta = 1$, so 95% CI is $\hat{\mu} \pm 1$ ✓ better (why?)
" 2: If $\delta = 10$, " " " " $\hat{\mu} \pm 10$

- For a given α (i.e. fixed level of CI), the narrower the CI the better (i.e. interval estimator is more precise).

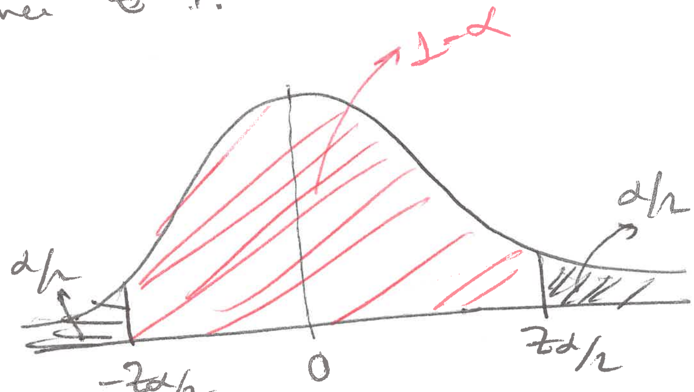
- **Def'n:** Width of CI (A, B) is $B - A$
Usually, " " " $(\hat{\theta} - \delta, \hat{\theta} + \delta)$ is 2δ
where δ is called "margin of error" or "error margin"

- So, a CI with smaller error margin is better at fixed α.

1) Normal-Based CI — $\hat{\theta}$ is an estimator of θ, which is unknown

If $\hat{\theta} \sim N(\theta, se^2)$ then a $(1-\alpha)$ CI for θ
$(\hat{\theta} - z_{\alpha/2} \cdot se, \hat{\theta} + z_{\alpha/2}\, se)$

- This CI is exact. since $\hat{\theta} \sim$ Normal
- width $= 2 z_{\alpha/2} se$
- $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$



- $se = se(\hat{\theta}) = \sqrt{Var(\hat{\theta})}$
$\Leftrightarrow se^2 = Var(\hat{\theta})$

Proof of the Normal-Based CI

Given $\hat{\theta} \sim N(\theta, se^2)$

then $\dfrac{\hat{\theta}-\theta}{se} \sim Z \sim N(0,1)$

So $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha$

$\Rightarrow \underbrace{P\left(-z_{\alpha/2} < \dfrac{\hat{\theta}-\theta}{se} < z_{\alpha/2}\right)} = 1-\alpha$

$= P(-se\, z_{\alpha/2} < \hat{\theta}-\theta < se \cdot z_{\alpha/2})$

$= P(\hat{\theta} - z_{\alpha/2} se < \theta < \hat{\theta} + z_{\alpha/2} se) = 1-\alpha$, (*)

Ex: For 95% CI, $1-\alpha = 0.95 \Rightarrow \alpha = 0.05$

so $z_{\alpha/2} = z_{0.025} = 1.96$

i.e. 95% CI is $(\hat{\theta} - 1.96\, se,\ \hat{\theta} + 1.96\, se)$

that is, $P(\theta \in ($ " , " $)) = 0.95$

Note: If $\hat{\theta} \overset{approx}{\sim} N(\theta, se^2)$, then the norma-based CI is approximal
i.e. prob in (*) above is $\approx 1-\alpha$. (" " " also) is called Wald's CI

Qu: What to do if "se" is not known either?

Ans: Find a consistent estimator $\hat{se}$ of se & replace se with $\hat{se}$.

Then $(1-\alpha)100\%$ CI becomes

$$\underbrace{(\hat{\theta} \pm z_{\alpha/2} \hat{se})}_{(1-\alpha)\text{ Wald CI}} \leftarrow \text{approx CI}$$

Ex: $D = \{X_1, X_2, \ldots, X_n\}$, $X_i \overset{iid}{\sim} F$ with $E[X_i] = \mu$ & $Var(X_i) = \sigma^2$ μ: unknown, σ^2: known

Find a 95% CI for μ (for large n).

Sol'n: $1-\alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025$.

(consider $\hat{\mu} = \bar{X}_n$ for the estimator of μ) by CLT, $\bar{X}_n \overset{approx}{\sim} N(\mu, \sigma^2/n)$

(usually CLT kicks in for $n > 30$)

$z_{\alpha/2} = z_{0.025} = 1.96$

So, a normal-based (or Wald) 95% CI for μ is

$$\hat{\mu} \pm 1.96 \frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \boxed{\bar{X}_n \pm 1.96 \frac{\sigma}{\sqrt{n}}}$$

• What if σ^2 is unknown either?

Replace σ with a consistent estimator $\hat{\sigma}$.

e.g. sample standard deviation

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}$$ is such an estimator

So Normal-Based 95% CI for μ is

$$\boxed{\hat{\mu} \pm 1.96 \frac{\hat{\sigma}}{\sqrt{n}}}$$

• Below are two consistent estimators of σ^2.

(i) $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$ (uncorrected, slightly biased with bias $\to 0$ as $n \to \infty$)

&

(ii) $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ (corrected & unbiased)

are both consistent for σ^2.

• Recall $\hat{\sigma}^2$ is also the "plug-in estimator" for σ^2.

Ex: Recall for $D = \{X_1, \longrightarrow X_n\}$ $X_i \overset{iid}{\sim} P_X$
$\hat{}$ pmf

for any $a \in \mathbb{R}$,

$$\hat{P}_X(a) = \frac{\sum_{i=1}^{n} I(X_i = a)}{n}$$

$\hat{P}_X(a)$ is a "consistent estimator" for $P_X(a)$.

Qu: Can we construct a 95% CI for $P_X(a)$?

Ans: $\hat{P}_X(a) = \dfrac{\sum_{i=1}^{n} I(X_i = a)}{n}$ where $I(X_i = a) \overset{iid}{\sim} \text{Bernoulli}(P_X(a))$

Then, by CLT,

$$\hat{P}_X(a) \overset{approx}{\sim} N\Big(P_X(a), \underbrace{Var(\hat{P}_X(a))}_{= \frac{P_X(a)(1-P_X(a))}{n}}\Big)$$

$\Rightarrow$ 95% CI for $P_X(a)$ is

$$\boxed{\hat{P}_X(a) \pm 1.96\sqrt{\frac{P_X(a)(1-P_X(a))}{n}}}$$

but $P_X(a)$ is unknown!

A consistent estimator of $P_X(a)$ is $\hat{P}_X(a) \equiv$ epmf at a

so 95% CI for $P_X(a)$ becomes

$$\boxed{\hat{P}_X(a) \pm 1.96\sqrt{\frac{\hat{P}_X(a)(1-\hat{P}_X(a))}{n}}}$$

Qu: Can we construct $(1-\alpha)100\%$ CI for the cdf at a, i.e. $F_X(a)$? (for large n)

Ans: Yes, for X discrete or cts.

$$\text{ecdf} = \hat{F}_X(a) = \frac{\sum_{i=1}^{n} I(X_i \le a)}{n} \quad \text{where } I(X_i \le a) \overset{\text{iid}}{\sim} \text{Bernoulli}(F_X(a))$$

$$\text{so CLT} \Rightarrow \hat{F}_X(a) \xrightarrow{\text{approx}} N\left(F_X(a), \frac{F_X(a)(1-F_X(a))}{n}\right)$$

for large n.

So, 95% CI for $F_X(a)$ is

$$\boxed{\hat{F}_X(a) \pm 1.96\sqrt{\frac{F_X(a)(1-F_X(a))}{n}}}$$

unknown.

$\hat{F}_X(a)$ is consistent for $F_X(a)$

so 95% CI for $F_X(a)$ ~~is~~

$$\boxed{\hat{F}_X(a) \pm 1.96\sqrt{\frac{\hat{F}_X(a)(1-\hat{F}_X(a))}{n}}}$$

DKW Confidence Intervals

DKW Inequality:

$$P\left(\max_x |\hat{F}_X(x) - F_X(x)| > \varepsilon\right) \le \frac{2}{e^{2n\varepsilon^2}}$$

$$\Rightarrow \text{for any } a \in \mathbb{R}, \quad P\left(|\hat{F}_X(a) - F_X(a)| > \varepsilon\right) \le \frac{2}{e^{2n\varepsilon^2}}$$

$$\Rightarrow \quad 1 - P\left(|\hat{F}_X(a) - F_X(a)| \leq \varepsilon\right) \leq \frac{2}{e^{2n\varepsilon^2}}$$

$$\Rightarrow \quad \underbrace{P\left(|\hat{F}_X(a) - F_X(a)| \leq \varepsilon\right)}_{\hat{F}_X(a) \in \hat{F}_X(a) \pm \varepsilon} \geq \underbrace{1 - \frac{2}{e^{2n\varepsilon^2}}}_{RHS}$$

So, taking $\varepsilon = \sqrt{\frac{1}{2n}\log(2/\alpha)}$ makes $RHS = 1-\alpha$

$\Rightarrow$ $(1-\alpha)$ CI for $F_X(a)$ is

$$\boxed{\hat{F}_X(a) \pm \sqrt{\frac{1}{2n}\log(2/\alpha)}} \leftarrow \text{DKW CI}$$

<u>Recap</u>

1) Normal-Based CI (ie $(1-\alpha)100\%$ CI) for θ

$$\hat{\theta} \pm z_{\alpha/2}\, se$$

or $\boxed{\hat{\theta} \pm z_{\alpha/2}\, \hat{se}}$ where $\hat{se}$ is consistent for se.

2) <u>Asymptotic Normality:</u> $\hat{\theta}$ is asymptotically normal if $\hat{\theta} \overset{\text{approx}}{\sim} N(\theta, se^2)$

- usually justified with CLT (or MLE theory)
- se is replaced with a consistent estimator $\hat{se}$
- foundation for Normal-Based CI's.

3) DKW CI (see pg 10), used for $F_X(a)$.