Last time:

Sec 9.3 Exponential

Sec 10.1-2 Normal

Today:

finish Normal

Bias / Unbiased Estimator.

Sec 10.4 (Continued)

CI & test for differences between population means ($\mu_1 - \mu_2$)
proportions ($p_1 - p_2$)

CI

	Size	Sample proportion	Population proportion
setup: Cty X	$n = 400$	37%	p_1
Cty Y	$m = 600$	28%	p_2

For Bern(p), mean $= p$, Var $= pq$, sd $= \sqrt{pq}$.

sd of sample mean $= \frac{1}{\sqrt{n}}$ sd $= \sqrt{\frac{pq}{n}}$

By CLT, $\bar{x} \sim N(p_1, \frac{p_1 q_1}{n})$

$\bar{y} \sim N(p_2, \frac{p_2 q_2}{m})$ (37%)²

$\bar{x} - \bar{y} \sim N(p_1 - p_2, \frac{p_1 q_1}{n} + \frac{p_2 q_2}{m})$

$\bar{x} - \bar{y} = 37\% - 28\% = 9\%$

$sd(\bar{x} - \bar{y}) = \sqrt{\frac{37\% \cdot 63\%}{400} + \frac{28\% \cdot 72\%}{600}} = 3\%$

95% CI for $p_1 - p_2$ is

$9\% \pm 2 \times 3\% = [3\%, 15\%]$

test:

H_0: City X and City Y has the same proportion
H_a: City X has a higher proportion.

What is the difference between a 95% CI and a test with significant level 0.95?

H_0: $p_1 = p_2$, $p_1 - p_2 = 0$
H_a: $p_1 > p_2$, $p_1 - p_2 > 0$

test statistic $= \bar{x} - \bar{y}$
distribution under H_0: $N(0, 3.2\%)$
observed value: $\bar{x} - \bar{y} = 9\% \approx +3SD$
Conclusion: reject H_0

Bias

"estimator"
T is an estimator for θ (fixed unknown parameter)
↑ computed from data

Why are we using these estimators?
What makes a good estimator?

Error $= T - \theta$

$x = \theta$
$a = T$ (from different runs of data)

① bias / unbiased
② variance

θ - fixed unknown parameter (fixed)
T - estimator (R.V.)

$Bias: B_\theta(T) = E(T - \theta) = E T - \theta$

unbiased: $B_\theta(T) = 0$ or $E T = \theta$

Example:

$X_1, X_2, X_3, \dots, X_n$ i.i.d with mean μ

X_1 is an unbiased estimator of μ
Since $E X_1 = \mu$

$\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$ is also an unbiased estimator of μ

$E \bar{X} = \frac{1}{n} \sum_{i=1}^{n} E X_i = \frac{1}{n} \cdot n\mu = \mu$

$MSE_\mu(X_1) = Var(X_1)$
$MSE_\mu(\bar{X}) = Var(\bar{X}) = \frac{1}{n} Var(X_1)$

In particular for Bern(p) random
sample proportion is an unbiased estimator of population proportion

Example 2

uniform $\{1, 2, \dots, N\}$

$X_1, X_2, \dots, X_n$

$E \bar{X} = \frac{N+1}{2}$

$N = 2E\bar{X} - 1$

$E(2\bar{X} - 1) = N$

$\Rightarrow 2\bar{X} - 1$ is an unbiased estimator of N

WW2 German tanks

n $X_1, X_2, \dots, X_n$ SRS from $1, 2, \dots, N$

$\bar{x} = 94$

$n = 5$

$x_1 = 1$, $x_2 = 2$ $x_5 = 999$

$\bar{x} = \frac{1}{5}(1 + 2 + 999) = 334$

$2\bar{x} - 1 = 667$ not reasonable since $999 > 667$!

$T_2 = M = \max_{\text{across } x_i} x_i \leq N$

$E T_2 < N \Rightarrow$ biased!

Sec 11.1

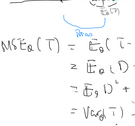
Mean Squared Error

$MSE_\theta(T) = E_\theta(T - \theta)^2$ (just means this is a function of θ)

lower MSE ⇒ good

Decomposition of Error

$Bias: B_\theta(T) = E_\theta(T) - \theta$ — constant
$Deviation: D_\theta(T) = T - E_\theta(T)$ — random
$Error: T - \theta = D_\theta(T) + B_\theta(T)$.



$MSE_\theta(T) = E_\theta(T - \theta)^2$
$= E_\theta(D + B)^2$
$= E_\theta D^2 + 2E_\theta(DB) + B^2$ (constant)
$= Var_\theta(T) + B_\theta^2(T)$

$E_\theta D = E_\theta(T - E_\theta T)$
$= E_\theta T - E_\theta T$
$= 0$

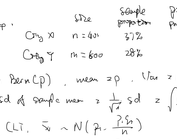
MSE is the sum of variance of the estimator and the squared bias.

Notes: In terms of MSE, the only aspects that we take into account are $Var_\theta(T)$ and $B_\theta(T)$.

Example Germany tank problem

$T_1 = 2\bar{X} - 1$

$T_2 = M$



$M - N = -G_N$
$B_N(M) = E(M - N) = -\frac{N}{n+1}$

G_1, G_2, G_3 & G_N has the same distr.

$E G_1 = \dots = E G_N \quad (= N - M)$
$G_1 + G_2 + G_3 + G_N = 15 - 5$
$E G_N = \frac{1}{4}(15 - 5) = 3$

Generally, $E G_{n+1} = \frac{1}{n+1}(N - n)$

$B_N(M) = -\frac{N-n}{n+1}$
$E M = N - \frac{N-n}{n+1} = \frac{n(N+1)}{n+1}$

$T_3 = \frac{n+1}{n} M - 1$ — Augmented Maximum

$E T_3 = N$

Compare $MSE_\theta(T_2)$ & $MSE_\theta(T_3)$

$MSE_\theta(T_3) = Var(T_3) + B_\theta^2(T_3)$
$MSE_\theta(T_2) = Var(T_2) + B_\theta^2(T_2)$
$Var(T_3) = \left(\frac{n+1}{n}\right)^2 Var(T_2)$ ($= \left(\frac{n+1}{n}\right)^2$)
$> Var(T_2)$ but very close if n is large

T_3 has slightly larger variance but is unbiased