Lecture 19

Topics:

I) Estimation continued.

The estimation examples we have been looking at are a bit trivial. Here are a few more interesting ones.

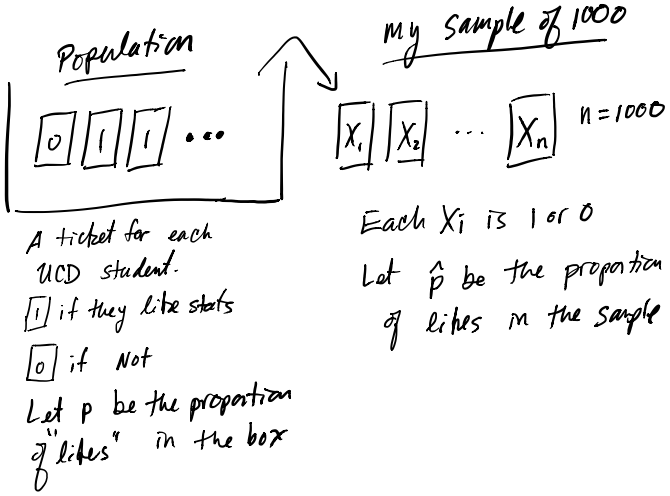
Estimating a population Proportion

e.g. I just interviewed 1000 UC Davis students & only 34% said they liked statistics.

How reliable is this estimate? Could it plausably be off by as much as ±20% (which would mean it could still be true that more than half UC Davis students like statistics :))

Lets analyze the MSE of the estimate.

Here is a box model picture of what we did:



A ticket for each UCD student.
1 if they like stats
0 if not
Let p be the proportion of "likes" in the box

Each X_i is 1 or 0
Let $\hat{p}$ be the proportion of likes in the sample

A couple things to notice.

1) The 34% I obtained is related to $\hat{p}$ by

$$0.34 = \hat{p}.$$

2) Since each X_i is 0 or 1

$$X_1 + \cdots + X_n = 340$$

the number of likes in my sample.

3) $\hat{p} = \frac{X_1 + \cdots + X_n}{n} = \bar{X}.$

Lets find the E & var each X_i:

x	$P_{X_i}(x)$	$x P_{X_i}(x)$	$x^2 P_{X_i}(x)$
1	p	p	p
0	1-p	0	0
		p	p

$\therefore E(X_i) = p$

$\text{var}(X_i) = E(X_i^2) - (E(X_i))^2$

$= p - p^2 = p(1-p).$

Lets find the bias, var & MSE of $\hat{p}$.

$\text{bias}(\hat{p}) = E(\hat{p} - p)$

$\overset{\text{MFI}}{=} E(\hat{p}) - p$

$= E\left(\frac{X_1 + \cdots + X_n}{n}\right) - p$

$= \frac{E(X_1) + \cdots + E(X_n)}{n} - p$

$= \frac{p + \cdots + p}{n} - p$

$= \frac{np}{n} - p = 0$ ↪ zero bias so $\hat{p}$ is accurate "on average".

$$\text{var}(\hat{p}) = \text{var}\left(\frac{X_1 + \cdots + X_n}{n}\right)$$

$$= \frac{\text{var}(X_1) + \cdots + \text{var}(X_n)}{n^2}$$ since the X_i's are indep!

$$= \frac{p(1-p) + \cdots + p(1-p)}{n^2}$$

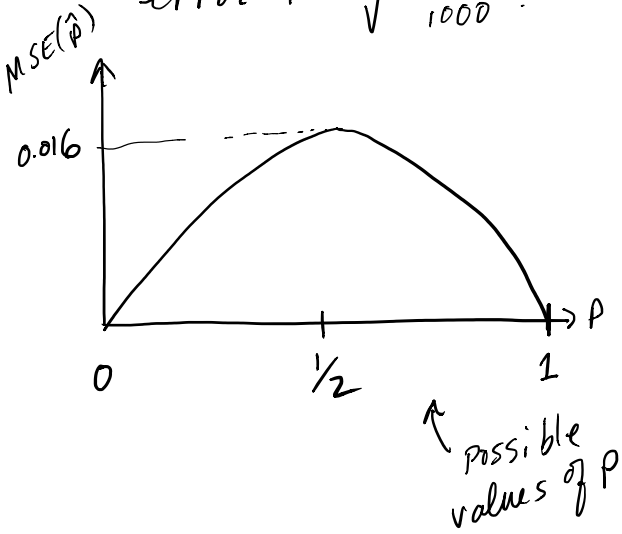
$$= \frac{n\,p(1-p)}{n^2} = \frac{p(1-p)}{n} = \frac{p(1-p)}{1000}.$$

since $n = 1000$ in this case.

$$\therefore \text{MSE}(\hat{p}) = \text{var}(\hat{p}) + (\text{bias}(\hat{p}))^2$$

$$= \frac{p(1-p)}{1000} + 0$$

$\therefore$ the typical estimation error is $\sqrt{\frac{p(1-p)}{1000}}$.



$\therefore$ if 20% lifted stats $\hat{p}$ would be typically wrong by about $\pm\sqrt{\frac{.2(1-.2)}{1000}} = \pm 0.12$

Here are some possible values of p with the corresponding typical $\hat{p}$ estimation errors

Possible true p	typical est error for $\hat{p}$
0.30	0.0145
0.40	0.0155
0.50	0.0158
0.60	0.0155
0.70	0.0145

So if $p = 0.5$ we would ~~expect~~ expect $\hat{p}$ to be around

0.5 ± 0.0158

p — typical discrepancy in $\hat{p}$

our $\hat{p}$ is not like this at all so it is not plausable that $p = 0.5$.

Estimating the difference of two population averages

e.g. Do Trump supporters make more money (on average) than Clinton supporters?

Sample 200 Trump supporters

Got \$41671 as the average income, per year, of these 200

Sample 150 Clinton supporters

Got \$44250 as the average income, per year, of these 150.

The difference

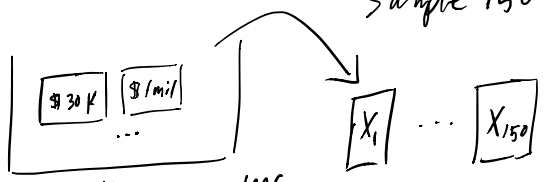
$44250 - 41671 = 2579$

Suggest Clinton supporters make more.

Is this real?

Find the MSE of the estimate 2579.

We need two boxes to describe this setup.



All Clinton supporters
ticket number = yearly income in 2015

$\bar{X} = 44250$

Let μ_1 = Ave ticket #
σ_1 = s.d. of ticket #

All trump supporters
ticket number = yearly income in 2015

$\bar{Y} = 41671$

Let μ_2 = Ave ticket #
σ_2 = s.d. of ticket #

Our estimate of the difference $\mu_1 - \mu_2$ is $\bar{X} - \bar{Y} = 2579$.

Lets find $MSE(\bar{X}-\bar{Y})$ to see how far wrong 2579 could be.

$$\text{bias}(\bar{X}-\bar{Y}) = E\left(\underbrace{\bar{X}-\bar{Y}}_{\text{est}} - \underbrace{(\mu_1-\mu_2)}_{\text{what we are trying to infer}}\right)$$

$$\overset{MF1}{=} E(\bar{X}) - E(\bar{Y}) - (\mu_1-\mu_2)$$

$E(\bar{X}) = \mu_1$, $E(\bar{Y}) = \mu_2$ both by MF1

$$= 0$$

$$\text{var}(\bar{X}-\bar{Y}) \overset{MF2}{=} \text{var}(\bar{X}) + \text{var}(\bar{Y})$$

since $\bar{X}$ & $\bar{Y}$ are indep.

$$\overset{MF2}{=} \frac{\sigma_1^2}{150} + \frac{\sigma_2^2}{200}$$

$$\therefore MSE(\bar{X}-\bar{Y}) = \frac{\sigma_1^2}{150} + \frac{\sigma_2^2}{200}$$

Suppose the Bureau of labor stats says the overall s.d. of income is about $30K.

$\therefore$ if we assume both σ_1 & σ_2 are about $30K then

$$\sqrt{MSE(\bar{X}-\bar{Y})} = \sqrt{\frac{30000^2}{150} + \frac{30000^2}{200}}$$

$$= 3098$$

$\therefore$ we got $\bar{X}-\bar{Y} = 2579$ but it could be wrong by about $\pm$ 3098 (assuming $\sigma_1 = \sigma_2$ = Bureau of labor stats value).

$\therefore$ the fact that we observed a a difference in ave income btwn the two samples this difference could have been due to sampling fluctuation, also called sampling variability.