Lecture 17

Topics:

I) Bivariate Normal prediction given/predicting z-score, raw scores or percentiles.

II) start estimation.

Since raw scores, percentiles & z-scores are all essentially equivalent ways of specifying the same thing prediction for BV (X,Y) can apply to all three.

e.g. Midterm 1 scores: X

$E(X) = 10.62$, $sd(X) = 1.46$

Midterm 2 scores: Y

$E(Y) = 8.54$, $sd(Y) = 2.08$

Correlation btwn X & Y

$\rho = 0.155$.

Question: Suppose you got $X=8$ on midterm 1. what is your predicted percentile for Y among

(a) the whole class

(b) the subgroup of students who also got $X=8$ on midterm 1.

Answer: First compute the predicted z-score for Y (among the whole class).

$$(\text{predicted z-score}) = \rho(\text{observed z-score})$$

$$= 0.155\left(\frac{8 - E(X)}{sd(X)}\right)$$

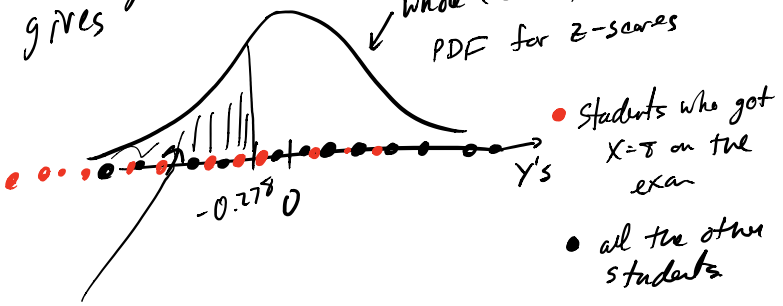
$$= 0.155\left(\frac{8 - 10.62}{1.46}\right)$$

$$= -0.278$$

Since predicted z-score $= \dfrac{E(Y|X=8) - E(Y)}{sd(Y)}$ ← computed from the whole class

-0.278 describe the z-score prediction among the whole class.

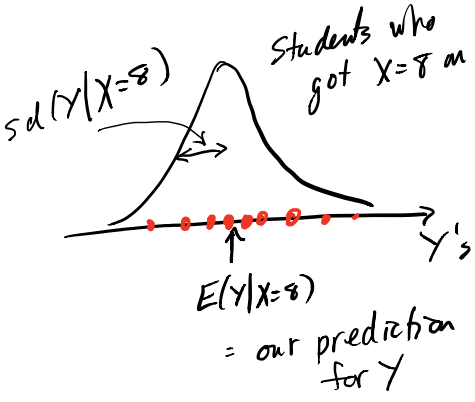
Relating z-scores to percentiles gives



This area $= 0.39$

$\therefore$ for (a) the predicted percentile on midterm 2 (Y) is 39^{th} percentile.

What about (b)?



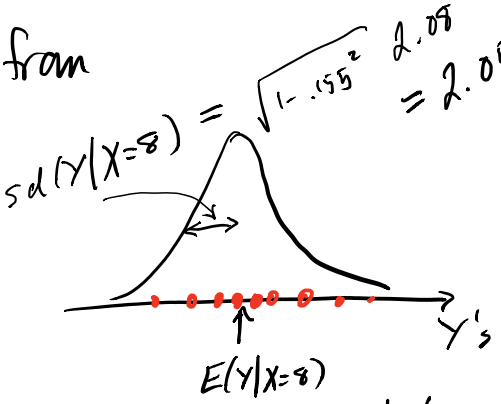
$\therefore$ Among these students the predicted z-score is 50^{th} percentile

e.g. Same problem from last example.

given $X=8$ compute

$$P(Y \geq 10 \mid X=8).$$

Given $X=8$, Y behaves like a random draw from



= our prediction for Y

$$= (-0.278)(2.08) + 8.54$$

↑ predicted z-score ↑ $sd(Y)$ ↑ $E(Y)$

$$= 7.962$$

i.e. given $X=8$, Y behaves like a random draw from

$$N(7.962, 2.05^2).$$

$$\therefore P(Y \geq 10 \mid X=8) = P\left(Z \geq \frac{10-7.962}{2.05}\right)$$

↑ $Z \sim N(0,1)$

$$= P(Z \geq 0.994)$$

$$= 0.16.$$

Basics of estimation

A typical estimation problem goes like this ... I'm interested in the average house square footage of households in California (μ). I sampled 45 homes at random, measured their sq footage (denoted $X_1, \ldots, X_{45}$). Since $\bar{X} = 2857\ ft^2$ I'm estimating μ to be $2857\ ft^2$.

Now, clearly μ isn't exactly $2857\ ft^2$ how far off could it be?

trying to quantify this is the main goal of this section.

Note $\mu - \bar{X}$ = estimation error.

There are two possible sources of est error.

(1) Sampling error (due to the randomness in $\bar{X}$).
Quantified by $\text{var}(\mu - \bar{X})$

(2) systematic error (also called bias)
Quantified by $E(\mu - \bar{X})$.

These combine to give an overall estimate of the typical size of the estimation error:

typical size of the estimation error

$$= \sqrt{E\left[(\mu - \bar{X})^2\right]}$$

called mean squared error

fact

$$= \sqrt{\text{var}(\mu - \bar{X}) + \left(E(\mu - \bar{X})\right)^2}$$

var: quantifies sampling variability; bias

$$= \sqrt{\text{var}(\bar{X}) + \left[E(\mu - \bar{X})\right]^2}$$

Since $E(\bar{X}) = \mu$ &

$$\text{var}(\bar{X}) = \frac{\sigma^2}{45}$$

(σ^2: variance of each X_i)

for this example we have:

the typical estimation error

$$= \sqrt{E\left[(\mu - \bar{X})^2\right]}$$

$$= \sqrt{\text{var}(\bar{X}) + 0^2}$$

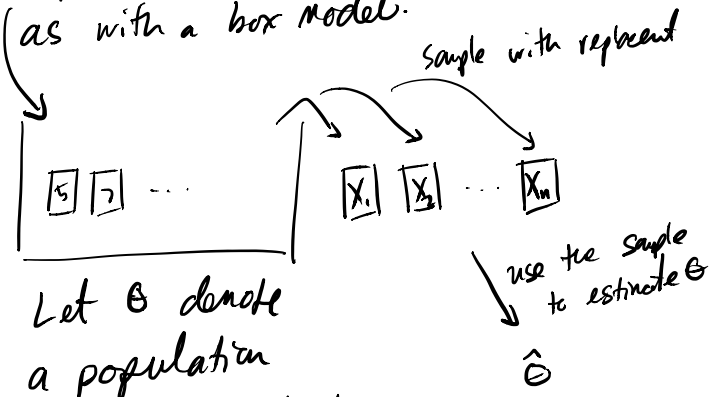
$$= \frac{\sigma}{\sqrt{45}}$$

So if $\sigma = 1000\ ft^2$ then $\mu - \bar{X}$ is typically off by $\pm \frac{1000}{\sqrt{45}} = 149$.

& if $\sigma = 200\ ft^2$ then $\mu - \bar{X}$ is typically off by $\pm \frac{200}{\sqrt{45}} = 29.8$

The above example was special... lets now outline the general case

A population can be represented as with a box model.



Let θ denote a population parameter, which is computed from these tickets, that you want to estimate.

$\theta - \hat{\theta}$ = estimation error

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

= systematic error

$\text{var}(\hat{\theta})$ = sampling uncertainty.

$MSE(\hat{\theta}) = E\left[(\theta - \hat{\theta})^2\right]$

= mean squared estimation error

typical estimation error

$$= \sqrt{MSE(\hat{\theta})}$$

Fact:

$$MSE(\hat{\theta}) = var(\hat{\theta}) + (bias(\hat{\theta}))^2$$

combines the two possible sources of estimation error to compute "typical estimation error".