Q : sample v/s population meaning?

Variability (spread / dispertion)

- shows how 'spread out' data is
- measures of variability are :
  → range
  → interquartile range
  → variance
  → standard deviation

Range

range = (highest score - lowest score)

Interquartile Range

- IQR = (75th% - 25th%) ≈ H-spread

Semi-IQR

- SIQR = (IQR)/2
- in a symmetric distribution, the median +/- the SIQR contains half the scores in the distribution

Variance

- $\sigma^2 = \frac{\sum (X - \mu)^2}{N}$ mean of deviation squared

  X : data

  μ : mean

  variance in a population

- $s^2 = \frac{\sum (X - M)^2}{N - 1}$

  Variance est. from a sample

Standard Deviation

- sd (σ) = $\sqrt{\sigma^2}$ (population standard deviation)
- sd (s) = $\sqrt{s^2}$ (sample standard deviation)

Normal Distributions

- can calculate the proportion of the distribution within a given number of SD from the mean

68% w/in 1 SD (+/-)

95% w/in 2 SD (+/-)

one that is not skewed.

Central Tendency

- for symmetric distributions : mean, median, trimean, trimmed mean are equal
  (mode is also equal except in bimodal dist)
  in bell shaped normal dist

- in positive skew, mean is usually higher that the median and geometric mean is lower than all measures except the mode.

- with skewed distributions, the geometric mean, trimean and trimmed mean are between the median and the mean.

- what report? : mean, median and trimean / mean trimmed 50%.

media reports the median for skewed distributions

trimean

- weighted average of the $25^{th}\%$, $50^{th}\%$ and $75^{th}\%$
- trimean $= (P_{25} + 2P_{50} + P_{75}) / 4$

geometric mean

- geometric mean : $(\prod x)^{1/N}$ — multiply all numbers and take the n^{th} root

  related to logs

trimmed mean

- mean calced after removing some of the higher and lower scores
- trimmed x% ⇒ $\frac{x}{2}$% scores from the bottom
  and
  $\frac{x}{2}$% scores from the top are removed

  less influenced by extreme scores

median is basically mean trimmed 98+%

measure of CT ①

mean

- balance point of the distribution
- minimises sum of squared deviation

  $\sum (x - \bar{x})^2$

  (basically means that if deviation is calced and squared wrt mean, twill be the smallest sum)

- $\mu = \frac{\sum X}{N}$

median

- minimises the sum of absolute deviations

  ↓

  Same as value that minimises avg absolute deviation.

- midpoint of dist (same no of scores above & below)
  ↑ 50th percentile value

mode

- most frequently occuring value
- for continuous data, the mode is computed as the midpoint of the most frequent interval

measure of CT 2

- smallest absolute difference (when calced wrt a mid value)

???

measure of CT 3

Variance of the sum of 2 variables

- select a number each from 2 populations, add them. repeat that.
  what is the variance of the sums?

↓

$\sigma^2_{sum} = \sigma_{p_1}^2 + \sigma'_{p_2}$ (or diff)

mean(sum) = mean(p_1) + mean(p_2)

↖ regardless of sign

variance sum law : $\sigma^2_{x \pm y} = \sigma^2_x + \sigma^2_y \pm 2\rho\sigma_x\sigma_y$ ← for independent this term is = 0

ONLY WHEN X, Y are independent variables

(randomly paired ⇒ independence)
(for 2 people or something)

effects of linear transformation

- mean → changed (same)
- std dev → changed (coeff)
- variance → changed (coeff^2)
- corr → unchanged.

if X has mean μ_x
std dev σ_x
var σ^2_x

$Y = bX + A$

$\mu_y = b\mu_x + A$
$\sigma_y = b\sigma_x$
$\sigma^2_y = b^2\sigma^2_x$

measuring distribution skew

- pearson's measure of skew : $\frac{3(\text{mean} - \text{median})}{\sigma}$

- third moment measure of skew about the mean (more common) : $\sum \left\{ \frac{(X-\mu)^3}{N\sigma^3} \right\}$

- estimating the skew of samples : $\frac{n}{(n-1)(n-2)} \sum \frac{(X-M)^3}{s^3}$

  sample size — n; sample mean — M; sample std dev — s

- measure of kurtosis : $\sum \frac{(X-\mu)^4}{N\sigma^4} - \underline{\underline{3}}$

  measure of kurtosis — fourth moment about the mean

  3 — kurtosis of a normal dist

- by hand : $\frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \frac{(X-M)^4}{s^4} - \frac{3(n-1)^2}{(n-2)(n-3)}$

lecture notes

eg 2   1   3   4   4   4   5   5   7   8   9   31

$$\bar{x} = \frac{\sum x}{n} = \frac{81}{11} = 7.36\ldots$$

$$\text{rank} = \frac{1}{N}(n^{th}\% + 1)$$

median (50th %) = 5

$\%^{th}$ = data at rank_{IR} +

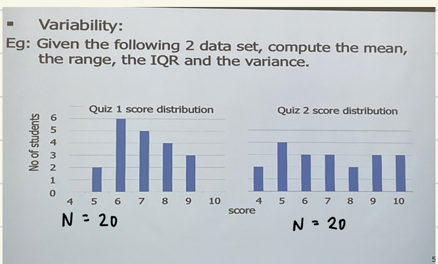
mode = 5

$$\text{trimean} = \frac{4 + 2*5 + 8}{4} = 5.5$$

$$\text{geometric mean} = (1 \times 3 \times 4 \ldots\ldots \times 31)^{1/11} = 5.2$$

mean trimmed 18.2% (remove 1, 31) $= \frac{49}{9} = 5.4$

eg 2



μ	7	7
range	9 - 5 = 4	10 - 4 = 6
IQR	8 - 6 = 2	9 - 5 = 4
σ^2	$\frac{\sum(x-\mu)^2}{N} = 1.5$	3.9

Population

mean $\mu = E[X] = \frac{\sum X}{N}$

variance $\sigma^2 = E[(X-\mu)]^2 = \frac{\sum(X-\mu)^2}{N} = \frac{\sum X^2}{N} - \frac{(\sum X)^2}{N^2}$

$E(Y^2) = \sigma_y^2 + \mu_y^2$

$= E[X^2] - (E[X])^2$

$= E[X^2] - \mu^2$

Sample

mean $\bar{x} = \frac{\sum X}{n}$

variance $s^2 = \frac{\sum(X-\bar{X})^2}{n-1}$ or $\frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1}$

only for linear, so if graph is curved, nope!

Pearson Correlation

population

$$\rho = \frac{\overbrace{E[(X-\mu_x)(Y-\mu_y)]}^{cov(X,Y)}}{\sigma_x \sigma_y} = \frac{E[XY] - \mu_x\mu_y}{\sqrt{E[(X-\mu_x)^2]\,E[(Y-\mu_y)^2]}}$$

$$= \frac{\frac{\Sigma XY}{N} - \frac{\Sigma X \Sigma Y}{N^2}}{\sqrt{E[X^2]-\mu_x^2}\sqrt{E[Y^2]-\mu_y^2}} = \frac{\Sigma XY - \frac{\Sigma X \Sigma Y}{N}}{\sqrt{\left(\Sigma X^2 - \frac{(\Sigma X)^2}{N}\right)\left(\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}\right)}}$$

sample

sample!!

$$cov(X,Y) = \frac{1}{n-1}\sum (X-\bar{X})(Y-\bar{Y})$$

$$r = \frac{E[(X-\bar{X})(Y-\bar{Y})]}{S_x S_y}$$

$$= \frac{\Sigma XY - \frac{\Sigma X \Sigma Y}{n}}{\sqrt{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}\sqrt{\Sigma Y^2 - \frac{(\Sigma Y)^2}{n}}}$$

Chpt 4

Bivariate Data

dataset w/ pair of variables which may be correlated to each other

eg. ice cream sales + temperature

Pearson correlation ρ

$$\rho = \frac{E[(X-\mu_x)(Y-\mu_y)]}{\sigma_x \sigma_y}$$

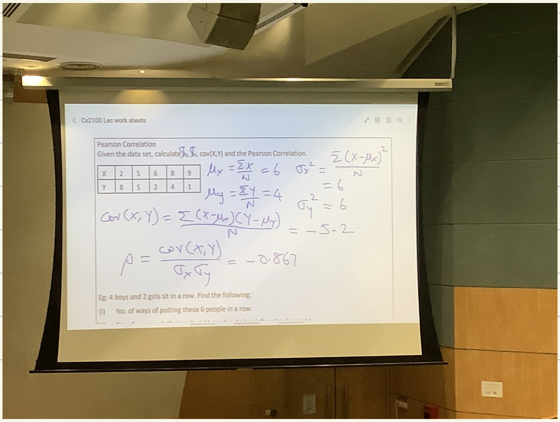
Co-variance of X and Y denoted as cov(XY)

$$= \frac{E[XY] - \mu_x \mu_y}{\sqrt{(E[X^2] - (\mu_x)^2)(E[Y^2] - (\mu_y)^2}}$$

$$\rho = \frac{\sum XY - (\sum X \sum Y)/N}{\sqrt{\sum X^2 - \frac{(\sum X)^2}{N}} \sqrt{\sum Y^2 - \frac{(\sum Y)^2}{N}}}$$

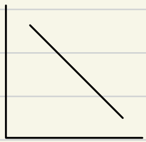
$\mu_x = \mu_y = 0$

$$\rho = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}$$

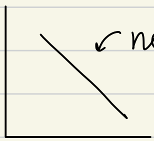


$$cov(yx) = \frac{1}{n-1} \sum (x-\bar{x})(y-\bar{y})$$

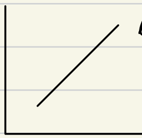
sample



points along a line : linear relationship



negative association



positive association

Pearson Product-Moment correlation

- Strength of linear relationship b/w 2 variables
- valid only for linear relationships
- $\rho \rightarrow$ population
  $r \rightarrow$ sample
- linear transformation of a variable does not change its correlation with other variable
- $r = \dfrac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$
- variance sum law for dependent variables :

$$\sigma^2_{x \pm y} = \sigma_x^2 + \sigma_y^2 \pm 2\rho\sigma_x\sigma_y$$

- estimate mean and variance of a population of size N

population mean $\mu = E[X] = \frac{\sum X}{N}$

population variance $\sigma^2 = E[(X-\mu)^2] = \frac{\sum (X-\mu)^2}{N}$

or $\frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N}$

$\approx \sigma^2 = E[X^2] - \mu^2$

$\therefore E[X^2] = \sigma^2 + \mu^2$