STAT 231

Tutorial.

Quick Review for TQ 1

(i) Given a data set, do we know how to calculate numerical data summary measures? (PROPERTIES OF THESE MEASURES)

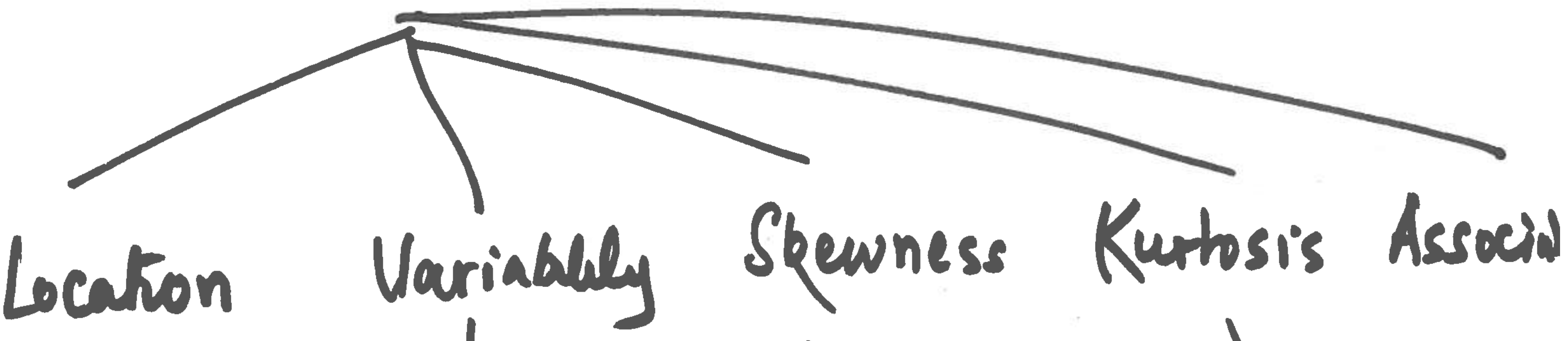
(ii) Given a data, can we draw the different graphical measures and/or interpret them if they are given?

(iii) Are we aware of the various terminologies that were introduced?

(iv) R (commands that we learned in Assignment 1)

(v) STAT 230

Numerical Measures



Range = Max - Min

$IQR = Q_3 - Q_1$

s^2, s

$$\frac{\text{Mean-Median}}{}$$

$> 0 \Rightarrow$ Right Skew

$< 0 \Rightarrow$ Left

$0 \Rightarrow$ Symm.

K = 3 for Gaussian.

Mean $\bar{y} = \frac{1}{n}\sum y_i$

Median $y_{(n/2)}$ = middle most obs

Mode = Max. frequency

$$s^2 = \frac{1}{n-1}\sum (y_i - \bar{y})^2 = \text{Sample Variance}$$

$$s = +\sqrt{s^2} = \text{Sample s.d.}$$

$K > 3 \Rightarrow$ More extreme/peaked compared to normal

$< 3 \Rightarrow$ Less frequency. of extreme. compared to normal.

<u>Two variables</u> (Response / Explanatory)

Relative Risk ↓ Categorical

Sample Correlation Coefficient. ↓ Numerical

Properties

20,000

$$\text{(i)}\quad s^2 = \frac{1}{n-1}\sum (y_i - \bar{y})^2$$

20

$$= \frac{1}{n-1}\left[\sum y_i^2 - n\bar{y}^2\right]$$

79

(ii) Add/subtract an obs. How to recalculate all the measures

(iii) If we transform the data (Linear Transformation) how do the measures change?

Example

$$\{y_1, \ldots \ldots y_{79}\}$$

$$y_1 \leq y_2 \ldots \ldots \leq y_{79}$$

$$\sum y_i = 1580 \qquad \sum y_i^2 = 200{,}000$$

- Find $\bar{y}$, s
- Find Q_1, Q_3

$$\left.\begin{array}{l} \cdot \text{ Find } \bar{y},\ s \\ \cdot \text{ Find } Q_1,\ Q_3 \end{array}\right\}$$

$$\bar{y} = \sum y_i / n = 1580/79 = 20$$

s^2 = use the 2nd formula

$$= \frac{1}{78}\left[200{,}000 - 79 \times 20^2\right]$$

$n = 79$

$Q_1 = ?$

p = Corresponding percentile

$\underline{m = (n+1) \times p}$

$= (79+1) \times 0.25 = 20$

Integer

Q_3

$m = (79+1) \times 0.75 = 60$

If not an integer, take the Average of the nearest two integers.

IQR = 60th obs — 20th obs.

Suppose we add one more observation $y_{80} = 25$

New mean $\quad \bar{y}_{new} = \dfrac{79 \times 20 + 25}{80}$

New variance. $\quad =$

$$\Downarrow$$

$$\sum {y_i^2}_{new} = \text{Old Sum of Squares} + 25^2$$

$$= 200{,}000 + 625$$

$$= 200{,}625$$

Use the 2nd formula to find the new variance.

We can calculate the new IQR, new range, after adding this point

---

TRANSFORMATION

$x_1, \ldots x_n.$

$y_1, \ldots y_n$ $\quad \boxed{y_i = a + b x_i.}$

$a, b \neq 0$

$b > 0.$

$$\boxed{y = 2 + 3x}$$

$$\left.\begin{array}{l}\bar{x} = 40 \\ \bar{y} = 122\end{array}\right\}$$

$\bar{y} = 2 + 3 \times 40$

For any measure of location, the formula is applied directly

MEASURES OF VARIABILITY

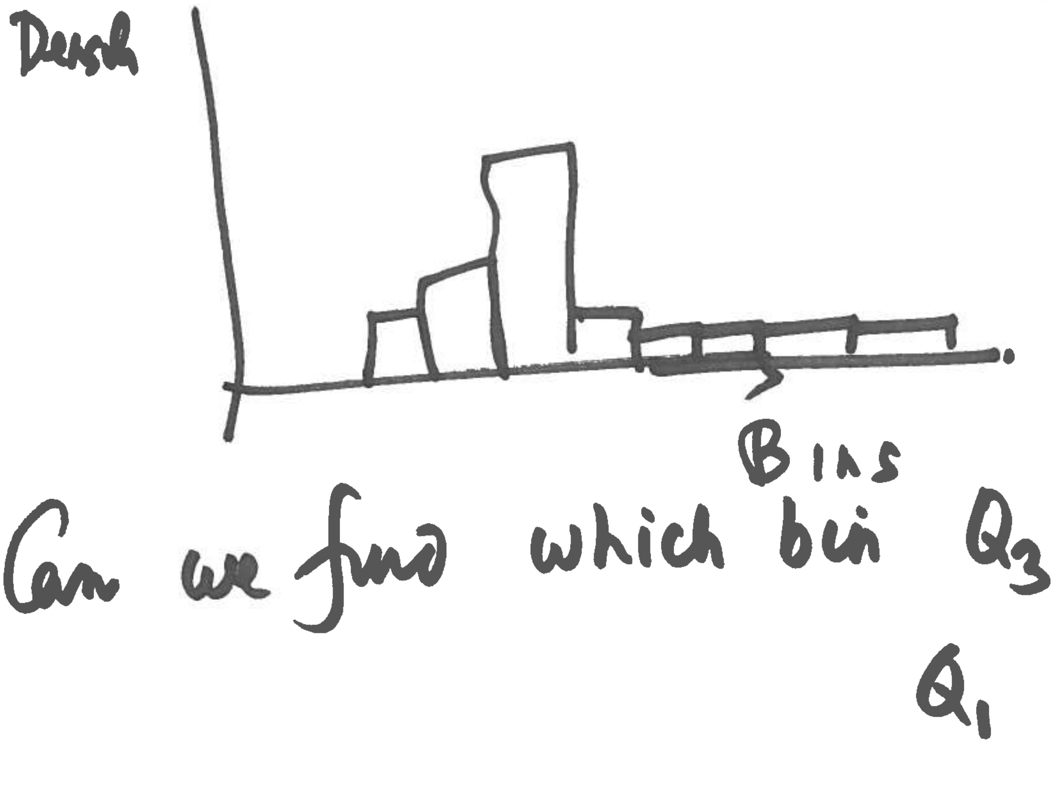
New Range = b × Old Range.

New Variance = b^2 × Old Variance.

New s.d = $|b|$ × Old s.d.

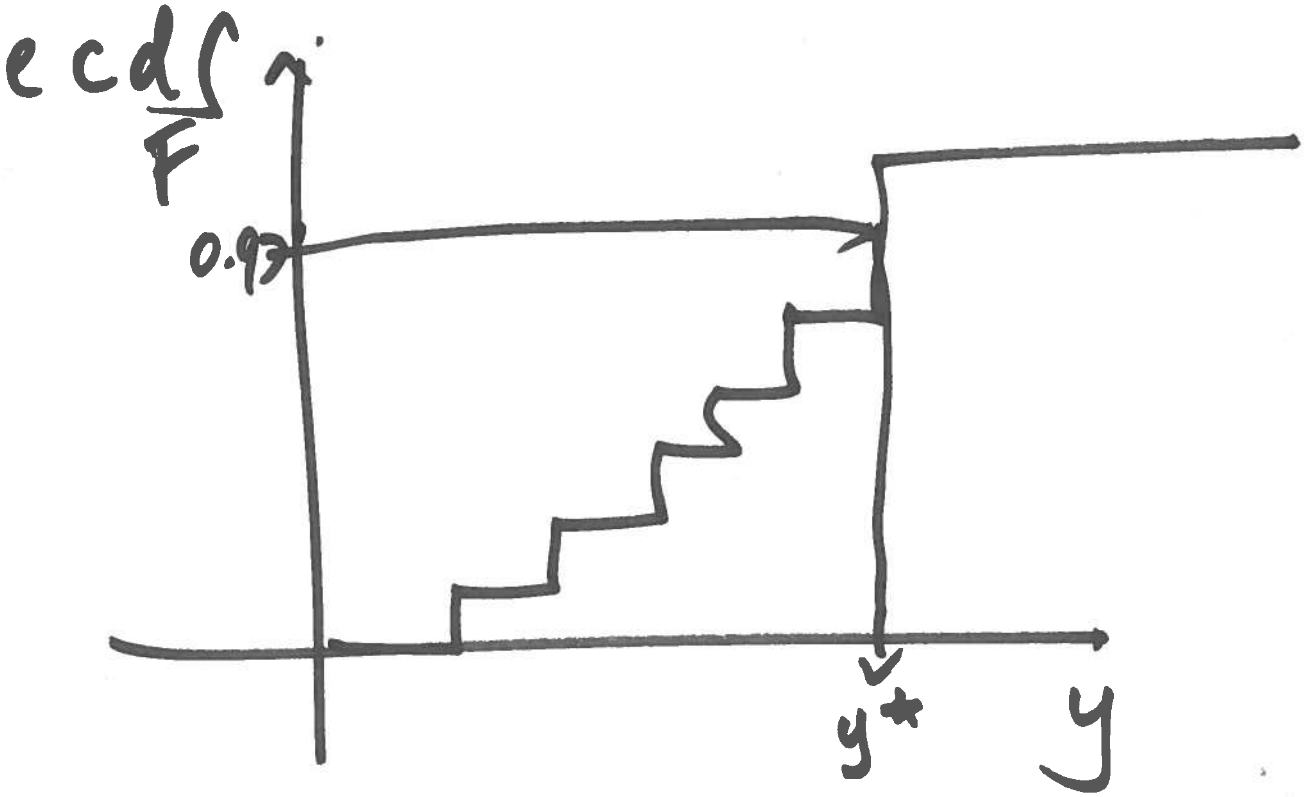
Graphical measures

- Relatvie Frequency Histogram
- Empirical cdf
- Box - Plot
- Scatter plot



Can we find which bin Q_3 belongs to?

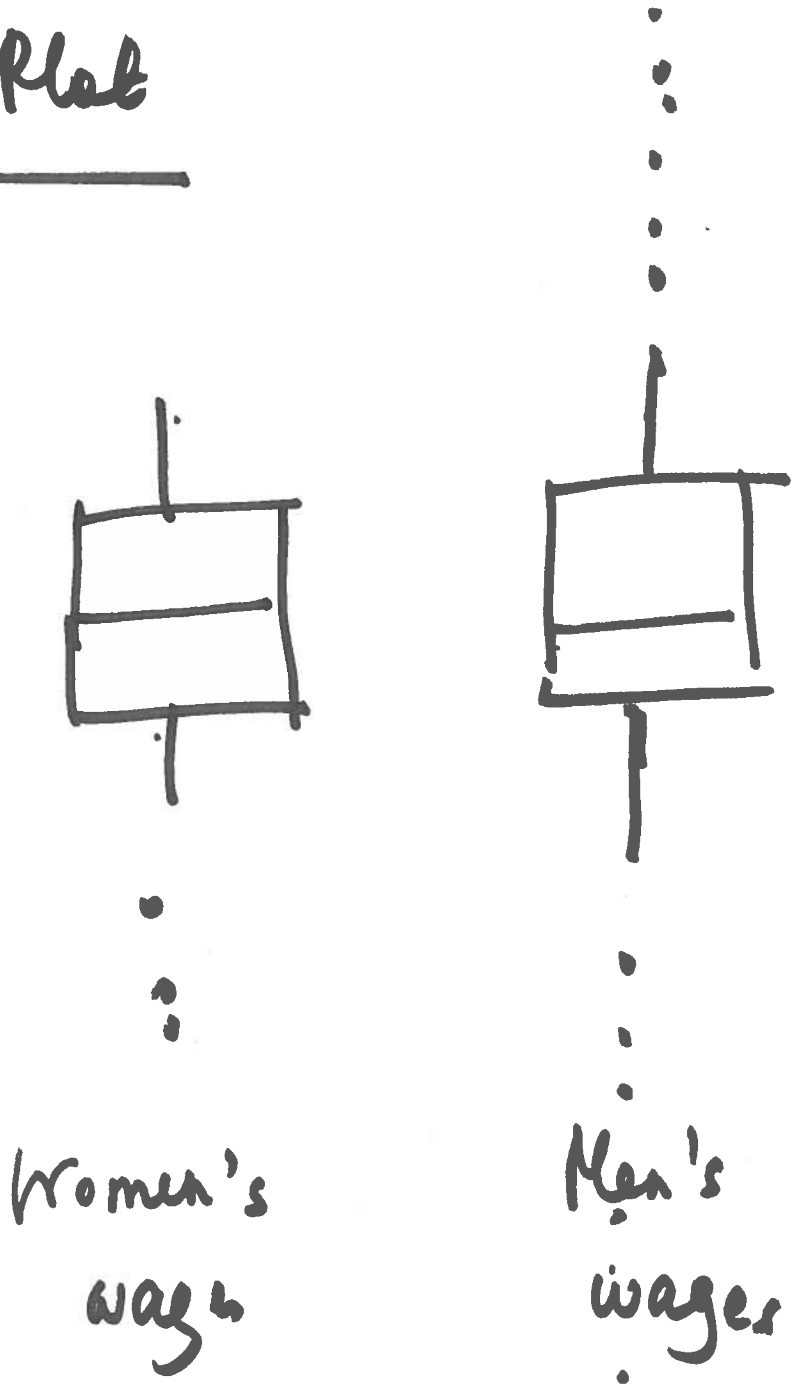
Q_1? median?



$$F(y) = \frac{\#\text{ of obs } \le y}{n}$$

n → Sample Sui

Box-Plot



If the median is half-way between the Q_3 and Q_1 and the two whiskers same size $\Rightarrow$ symmetric.

Scatter plots help identify associations

$$R.R = \frac{y_{11}/y_{11}+y_{12}}{y_{21}/y_{21}+y_{22}}.$$

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\left[\sum (x_i - \bar{x})^2\right]^{1/2}\left[\sum (y_i - \bar{y})^2\right]^{1/2}}$$

$RR \approx 1 \Rightarrow$ no evidence of association

$|r| \approx 1 \Rightarrow$ strong evidence of linear association/correlation

Types of data, inference, response, explanatory variable, etc.

$Y_1, \ldots Y_n \sim G(\mu, \sigma)$

indep

$\bar{Y} \sim G(\mu, \sigma/\sqrt{n})$