Tutorial: 6 pm: Check the UW schedule

Review video before TQ1 → Posted on Friday

Roadmap

- Measures of Association

Graphical Data Summaries

- Relative Frequency Histogram
- Empirical c.d.f.
- Box-Plot
- Scatter plot
- Q-Q plot

Association

We have bivariate data

$(x_1, y_1), \ldots (x_n, y_n)$

Objective: To check whether the two variables are correlated

- Sign
- Strength:

x = # of beers drunk/week.
y = STAT 231 score

Sample Correlation Coefficient

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

r_{xy} measures the degree of linear association between x and y

If the relationship is - ve.

If $x > \bar{x}$, likely that $y < \bar{y}$

$$(x - \bar{x})(y - \bar{y}) < 0$$

The sign of the numerator gives us the direction of the relationship

Properties of r_{xy}

(i) $0 \leq |r_{xy}| \leq 1$

The sample c.c. will lie between -1 and 1.

The closer the value of $|r|$ is to 1

$\Rightarrow$ more evidence we have of linear association between the two variables

$r_{xy} \approx 0 \Rightarrow$ little evidence of linear association between x and y.

(ii) If $y_i = a + b x_i \quad \forall i = 1, \ldots n$

$$r = \begin{cases} 1 & \text{if } b > 0 \\ -1 & \text{if } b < 0 \end{cases}$$

Proof: ?

(iii) Drawback: r captures only the linear part of the relationship between x and y

X	Y
-1	1
0	0
1	1

$Y = X^2$

What is r?

X	Y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
-1	1	-1	$1/3$	$-1/3$
0	0	0	$-2/3$	0
1	1	1	$1/3$	$1/3$
				0

$\bar{x} = 0$

$\bar{y} = 2/3$

$$\boxed{r_{xy} = 0}$$

Implication : $r = 0 \not\Rightarrow$ There is no association

$\Rightarrow$ There is no evidence of linear association

CATEGORICAL VARIABLES

A = smoker / non-smoker

B: under 30 / over 30

	A: Smoker	A^c: Non-Smoker
B: Under 30	y_{11}	y_{12}
B^c: Over 30	y_{21}	y_{22}

Relative Risk

R.R =

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

$P(A|B)$

$P(A|B^c)$

The closer the Relative Risk is to 1, the stronger is the evidence of no association.

If A and B are independent.

$$P(A|B) = P(A) = P(A|B^c)$$

The ratio $\frac{P(A|B)}{P(A|B^c)} = 1$

If R.R $>> 1$ or $<< 1$, $\Rightarrow$ evidence of association

Unanswered question(s)

- What is the cut-off value ?

GRAPHICAL DATA SUMMARIES

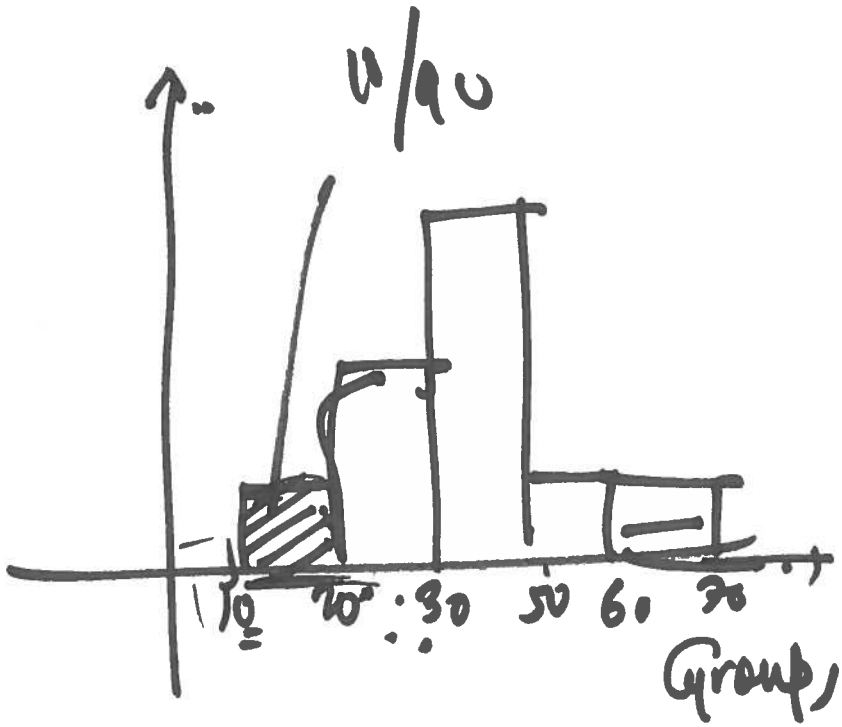
Objective: To identify the "statistical model" appropriate for my sample.

RELATIVE FREQUENCY HISTOGRAM

Density Histogram.

It is applied to grouped data.

Groups	Frequency
10-20	10
20-30	20
30-50	50
50-60	5
60-70	5



RELATIVE FREQUENCY HISTOGRAM.

x-axis → groups

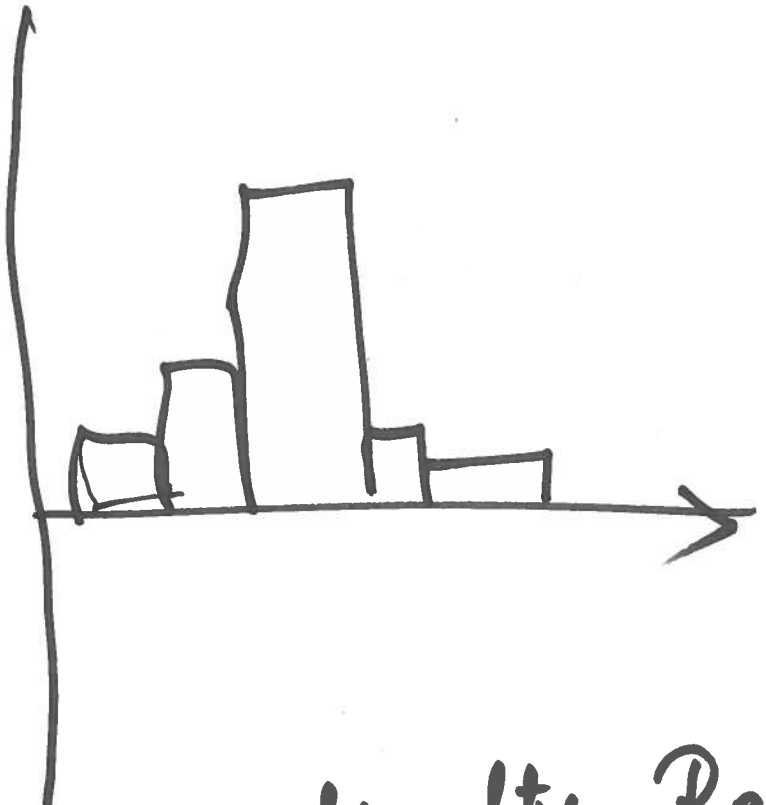
y-axis → The height of each group is chosen so that the relative frequency of the group = area of the rectangle.

Group	Frea	R.F
10-20	10	10/90
20-30	20	20/90
30-50	50	50/90
50-60	5	5/90
60-70	5	5/90.

$\frac{50/90}{20}$

Relative Frequency = $\frac{\text{Frequency}}{\text{Total sample size}}$

The height of $[10, 20]$ = $\frac{10/90 \rightarrow \text{R.F}}{10 \rightarrow \text{Length of group}}$



The area under the Relative Frequency Histogram = 1

⇒ Helps us to compare the data set with known p.d.f's from STAT 230.