AN OUTLIER IS AN OBSERVATION THAT IS MUCH SMALLER OR MUCH BIGGER THAN MOST OF THE OTHERS. SOMETIMES OUTLIERS OCCUR DUE TO DATA ENTRY ERRORS, AND CAN BE CORRECTED OR REMOVED. DETECTION OF OUTLIERS IS IMPORTANT, BECAUSE THEY INVALIDATE SOME PROCEDURES, SUCH AS CALCULATION OF THE MEAN.

## MEASURES OF CENTRE

THE MEAN IS ONE METHOD OF MEASURING THE CENTRE OF A DATA SET. FOR DISCRETE DATA, THE MODE (MOST FREQUENT DATA VALUE) IS ANOTHER. THE MEDIAN (THE VALUE SUCH THAT HALF THE DATA VALUES LIE ABOVE IT AND HALF BELOW IT) IS ANOTHER.

TO FIND THE MEDIAN, SORT THE VALUES. IF $n$ IS ODD, THE MEDIAN IS THE MIDDLE SORTED VALUE. IF $n$ IS EVEN, THE MEDIAN IS THE AVERAGE OF THE TWO MIDDLE VALUES.

EX: THE MEDIAN OF $\{6, 5, 2, 6, 9\}$ IS $6$;

THE MEDIAN OF $\{6, 5, 2, 6\}$ IS $\frac{5+6}{2} = 5.5$.

UNLIKE THE MEAN, THE MEDIAN IS AFFECTED BY OUTLIERS ONLY SLIGHTLY OR NOT AT ALL.

## QUARTILES

OFTEN, A DATA SET IS SEPARATED INTO QUARTILES $Q_1, Q_2, Q_3$. 25% OF THE DATA LIE BELOW $Q_1$, 25% LIE ABOVE $Q_3$, $Q_2$ IS THE MEDIAN.

IN PRACTICE, DIFFERENT BOOKS/SOFTWARE COMPUTE QUARTILES IN DIFFERENT WAYS THAT MAY GIVE SLIGHTLY DIFFERENT ANSWERS.

R CODE:

`quantile(data)` SOMETIMES DISAGREES WITH THE SIMPLE REPEATED MEDIAN METHOD `fivenum(data)`.

REPEATED MEDIAN METHOD

- $Q_1$ = MEDIAN OF LOWER HALF OF SORTED DATA.
- $Q_3$ = MEDIAN OF UPPER HALF OF SORTED DATA.
- FOR ODD $n$, LEAVE $Q_2$ OUT OF EACH HALF.

EX: data = $\{2, 5, 6, 6, 9\}$.

$Q_2 = 6$.

$n$ IS ODD, SO $Q_1 = \frac{2+5}{2} = 3.5$, $Q_3 = \frac{6+9}{2} = 7.5$.

THE RANGE OF THE DATA IS THE MAXIMUM VALUE MINUS THE MINIMUM VALUE, WHICH IS DRASTICALLY AFFECTED BY OUTLIERS.

THE INTERQUARTILE RANGE IS THE DIFFERENCE OF LOWER AND UPPER QUARTILES: $IQR = Q_3 - Q_1$.

THIS IS THE LENGTH OF THE INTERVAL THAT SPANS THE CENTRAL 50% OF THE DATA, A MEASURE OF THE SPREAD OF DATA IN THE MIDDLE REGION. IQR IS NOT AFFECTED BY OUTLIERS.

THE 5-NUMBER SUMMARY PROVIDES A CONCISE DESCRIPTION OF CENTRE AND SPREAD:

MINIMUM VALUE - $Q_1$ - MEDIAN - $Q_3$ - MAXIMUM VALUE

THESE 5 NUMBERS ARE USED TO CONSTRUCT BOX PLOTS FOR EASY COMPARISON OF 2 OR MORE SAMPLES.

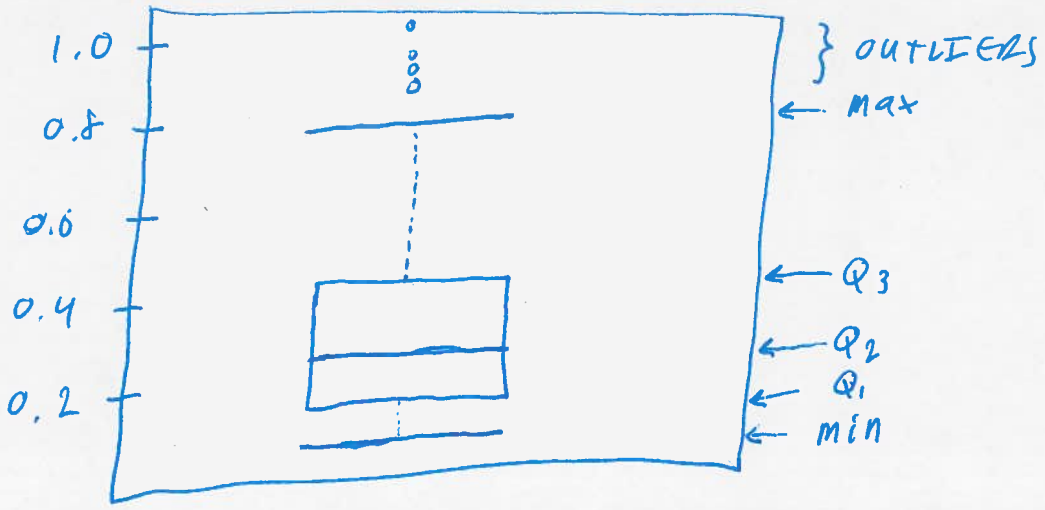
## BOX-AND-WHISKER PLOT

- USE AN AXIS WITH APPROPRIATE SCALE.
- DRAW A BOX FROM $Q_1$ TO $Q_3$, WITH CROSSBAR AT $Q_2$.
- OUTLIERS LOCATED FARTHER THAN 1.5 IQR FROM THE QUARTILES ARE SHOWN AS POINTS.
- DRAW WHISKERS EXTENDING FROM QUARTILES TO MAX/MIN VALUES THAT ARE NOT OUTLIERS.

## R CODE:

`boxplot(x)` DRAWS A SINGLE PLOT OF THE DATA IN x.

`boxplot(x~g)` DRAWS PARALLEL PLOTS, WHERE g IS A CATEGORICAL GROUPING VARIABLE.



## INTERPRETING BOX PLOT

- OUTLIERS, CENTRE AND SPREAD CAN BE SEEN AT A GLANCE.
- HEIGHT OF BOX IS IQR.
- SHOWS WHETHER THE DISTRIBUTION IS ROUGHLY SYMMETRIC (EQUAL WHISKERS, CROSSBAR IN THE MIDDLE OF THE BOX), OR SKEWED.

## MEASURES OF SPREAD

THE SPREAD OF A DATASET DESCRIBES WHETHER OBSERVATIONS VARY A LOT FROM EACH OTHER (HIGH SPREAD) OR ARE SIMILAR TO EACH OTHER (LOW SPREAD). VARIABILITY AND DISPERSION ARE SYNONYMS FOR SPREAD. THE RANGE AND THE IQR ARE MEASURES OF SPREAD.

MULTIPLE MEASURES ARE USEFUL, AS EACH HAS ADVANTAGES AND DISADVANTAGES. RANGE IS USEFUL FOR PLOTTING, BUT IS SENSITIVE TO OUTLIERS. IQR RESISTS OUTLIERS, BUT DESCRIBES ONLY THE MIDDLE OF THE DATASET.

VARIANCE IS ANOTHER MEASURE OF SPREAD, BASED ON SQUARED DISTANCE OF DATAPOINTS FROM THE MEAN:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

VARIANCE IS NONNEGATIVE, AND IS ZERO ONLY WHEN ALL DATA VALUES ARE IDENTICAL.

R CODE:

`var(x)`

WHY IS THE DIVISOR $n-1$ RATHER THAN $n$?

VARIANCE DOESN'T DESCRIBE WHERE THE DATA ARE, JUST HOW CLOSE POINTS ARE TO EACH OTHER. FOR $n$ SORTED POINTS, THERE ARE $n-1$ GAPS, SO THERE ARE $n-1$ GAP LENGTHS THAT PROVIDE INFORMATION ABOUT THE SPREAD.

## STANDARD DEVIATION

STANDARD DEVIATION IS THE SQUARE ROOT OF VARIANCE AND HAS THE SAME UNITS OF MEASUREMENT AS THE DATA.

R CODE:

`sd(x)`

ADDING A CONSTANT TO ALL OBSERVATIONS DOES NOT AFFECT STANDARD DEVIATION, BUT MULTIPLYING DOES.

$$x \mapsto cx \Rightarrow s_{new} = |c| s_{old}.$$

Ex: $X = \{1,2,3\}$, $s_{old} = 1$.

$\tilde{x} = \{11,12,13\} = \{10+1, 10+2, 10+3\} \Rightarrow s_{new} = 1.$

$\tilde{x} = \{-2,-4,-6\} = \{(-2)\cdot 1, (-2)\cdot 2, (-2)\cdot 3\} \Rightarrow s_{new} = |-2| \cdot s_{old} = 2.$

STANDARD DEVIATION IS VERY SENSITIVE TO OUTLIERS, DUE TO SQUARING LARGE DISTANCES.

EX: $X = \{1, 2, 3\} \Rightarrow S = 1$.

$Y = \{1, 2, 3, 10\} \Rightarrow \bar{y} = 4$, so $S^2 = \frac{(-3)^2 + (-2)^2 + (-1)^2 + 6^2}{4-1} = 16.6$.

THE NEW STANDARD DEVIATION IS 4 TIMES GREATER THAN THE OLD ONE.

WE SAW THAT 1.5 IQR BEYOND THE QUARTILES IS ONE WAY OF DETECTING OUTLIERS. ANOTHER IS THE Z-SCORE:

$$z_i = \frac{x_i - \bar{x}}{s}.$$

THIS TELLS YOU HOW MANY STANDARD DEVIATIONS $x_i$ LIES ABOVE OR BELOW THE MEAN. IF $|z_i| > 3$, THEN $x_i$ IS LIKELY AN OUTLIER.

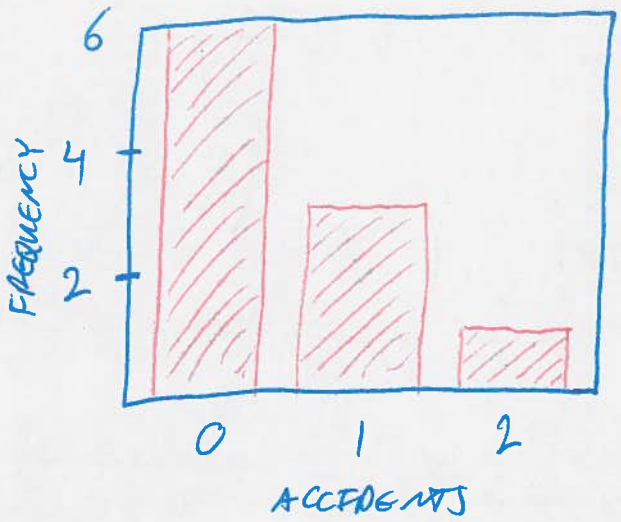
## TABLES AND GRAPHS

A FREQUENCY TABLE IS A METHOD OF ORGANIZING CATEGORICAL OR DISCRETE DATA. IT LISTS ALL POSSIBLE VALUES, AND THE NUMBER OF OBSERVATIONS OF EACH VALUE. THE RELATIVE FREQUENCY ( $\frac{\text{FREQUENCY}}{\text{TOTAL}}$ ) IS OFTEN INCLUDED AS A PERCENTAGE.

EX: NUMBER OF ACCIDENTS REPORTED PER WEEK: 0, 2, 1, 0, 0, 1, 1, 0, 0, 0.

| ACCIDENTS | FREQUENCY | REL. FREQUENCY |
|---|---|---|
| 0 | 6 | 0.6 |
| 1 | 3 | 0.3 |
| 2 | 1 | 0.1 |
| TOTAL | 10 | 1 |

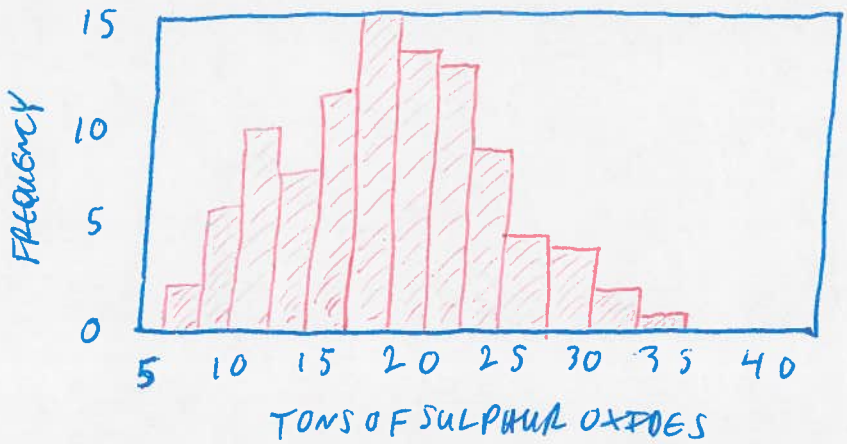
A BAR CHART DISPLAYS FREQUENCY OR RELATIVE FREQUENCY ON VERTICAL AXIS




A HISTOGRAM IS LIKE A BAR CHART, BUT FOR CONTINUOUS INTERVAL OR RATIO DATA.

- HORIZONTAL AXIS HAS REAL NUMBER SCALE, NO GAPS BETWEEN BARS.
- OBSERVATIONS ARE GROUPED INTO INTERVALS, USUALLY OF FIXED WIDTH.
- FREQUENCY IS REPRESENTED BY AREA = HEIGHT × WIDTH OF BAR.

EX: SULPHUR EMISSIONS DATA:




THIS IS A "REASONABLY SYMMETRIC, SINGLE-HUMP" HISTOGRAM.

THE PURPOSE OF A HISTOGRAM IS TO DISPLAY AN OVERALL PATTERN AND INTERESTING FEATURES:

- OUTLIERS
- GAPS
- LONG OR SHORT TAILS
- SYMMETRY OR SKEWNESS
- BELL SHAPE, U SHAPE, UNIFORM
- UNIMODAL/BIMODAL (1 OR 2 HUMPS)

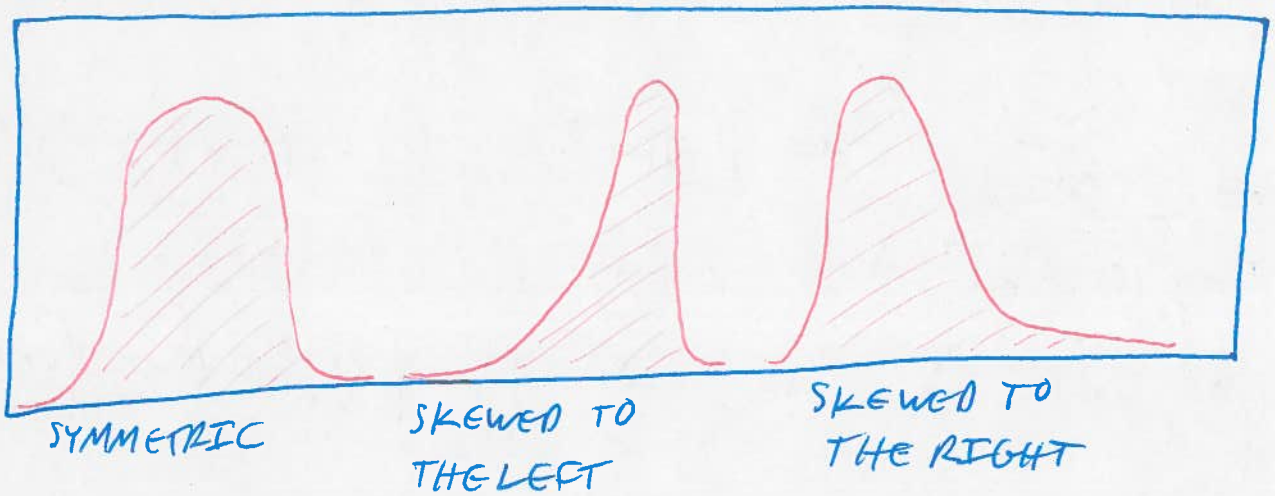
CHOICE OF INTERVAL WIDTH IS IMPORTANT: TOO SMALL GIVES A BUMPY GRAPH, TOO LARGE GIVES AN UNINFORMATIVE GRAPH.

## TAILS

THE LEFT-HAND TAIL IS THE REGION OF LOWEST DATA VALUES.
THE RIGHT-HAND TAIL IS THE REGION OF HIGHEST DATA VALUES.

(DON'T CONFUSE WITH LOW/HIGH FREQUENCY.)

eg. INCOME DISTRIBUTIONS TYPICALLY HAVE LONG RIGHT-HAND TAILS, AS SOME PEOPLE HAVE MUCH HIGHER INCOMES THAN THE MEAN.



IN A DISTRIBUTION SKEWED TO THE LEFT, THE MEAN IS LOWER THAN EXPECTED BECAUSE OF UNUSUALLY SMALL VALUES OR A LARGE NUMBER OF SMALL VALUES, SO MEAN < MEDIAN.

SKEWNESS TO THE RIGHT RESULTS IN AN INFLATED MEAN DUE TO MANY OR UNUSUALLY LARGE VALUES, SO MEAN > MEDIAN.

## STEM-AND-LEAF PLOT

STEM-AND-LEAF PLOTS ARE DISPLAYS OF QUANTITATIVE DATA THAT RETAIN THE NUMERICAL VALUES.

R CODE:

```
stem(x)
```

EX: THE DATA $\{0.4, 1.1, 0.6, 1.9, 2.1, 1.7\}$ IS DISPLAYED

| | |
|---|---|
| 0 | 4 6 |
| 1 | 1 9 7 |
| 2 | 1 |

↑ STEM ↑ LEAF

THE LEFT-HAND DIGIT (OR DIGITS) IS THE STEM; THE REMAINING RIGHT-HAND DIGITS ARE THE LEAF. OBSERVATIONS WITH THE SAME STEM ARE OFTEN SORTED BY LEAF VALUES. EXTRA RIGHT-HAND DIGITS ARE TRUNCATED OR ROUNDED.

IT'S IMPORTANT TO MAKE CLEAR THE STEM UNIT. eg. THE PLOT

2 | 4 CAN REPRESENT 24 OR 2.4, 0.24, 240, ...

UNITS ARE OF THE FORM $10^k$, AND LEAF UNIT = $\frac{\text{STEM UNIT}}{10}$.

Ex: CONSIDER THE SORTED DATASET

$\{24.4, 26.5, 26.5, 26.9, 27.3, 28.0, 28.7, 29.2, 31.8\}$

WITH STEM UNIT = 10 AND TRUNCATION, THE PLOT IS

| | |
|---|---|
| 2 | 46667889 |
| 3 | 1 |

WHICH IS TOO SHORT TO REVEAL ANY USEFUL SHAPE.

WITH STEM UNIT = 1, THE PLOT IS

| | |
|---|---|
| 24 | 4 |
| 25 | |
| 26 | 559 |
| 27 | 3 |
| 28 | 07 |
| 29 | 2 |
| 30 | |
| 31 | 8 |

WHICH IS TOO LONG TO REVEAL ANY USEFUL SHAPE. YOU MIGHT TRY SPLITTING EACH ROW INTO TWO, WITH LEAVES 0-4 IN ONE ROW AND LEAVES 5-9 IN THE NEXT, OR GROUPINGS OF TWO, LEAVES 0-1, 2-3, 4-5, 6-7 AND 8-9 IN 5 ROWS.

| | |
|---|---|
| 2 | 4 |
| 2 | 6667 |
| 2 | 889 |
| 3 | 1 |

THIS IS ABOUT RIGHT FOR THIS SAMPLE SIZE.

- A STEM-AND-LEAF PLOT SHOULD LOOK LIKE A HISTOGRAM ROTATED 90°.
  - THE STEM CORRESPONDS TO THE HORIZONTAL AXIS OF THE HISTOGRAM.
  - ROWS OF LEAVES CORRESPOND TO BARS.
  - LEFT/RIGHT TAILS CORRESPOND TO TOP/BOTTOM REGIONS OF LEAVES.
  - INCLUDE EMPTY ROWS AND ALIGN LEAVES VERTICALLY, IN ORDER TO REVEAL GAPS AND CONCENTRATIONS OF DATA.

EX: WEIGHT IN kg OF GLASS DISCARDED PER WEEK:

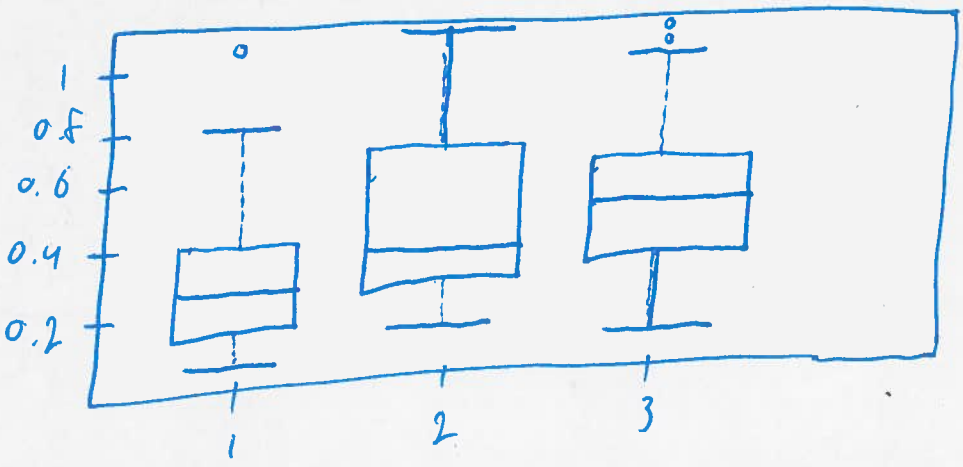
| | |
|---|---|
| 0 | 1 2 4 4 |
| 0 | 5 5 5 6 6 6 7 8 8 8 8 9 9 9 9 9 |
| 1 | 0 1 1 1 2 2 2 3 3 4 4 |
| 1 | 5 6 6 6 6 7 7 8 8 8 |
| 2 | 0 1 2 2 3 4 |
| 2 | 5 6 6 8 9 9 |
| 3 | 1 |
| 3 | |
| 4 | 0 2 |
| 4 | |
| 5 | |
| 5 | 6 6 |
| 6 | |
| 6 | |
| 7 | |
| 7 | |
| 8 | 0 |

THIS PLOT IS SKEWED TO THE RIGHT, SO WE EXPECT MEAN > MEDIAN.

IN FACT, MEAN = 1.702kg, MEDIAN = 1.35kg.

## BIVARIATE DATA

BIVARIATE DATE INVOLVES 2 MEASUREMENTS PER SUBJECT. (eg. HEIGHT AND WEIGHT OF PEOPLE) IF VARIABLE X IS CATEGORICAL AND VARIABLE y IS QUANTITATIVE, THEN X CAN BE USED TO DIVIDE y INTO GROUPS. THEN UNIVARIATE SUMMARIES SUCH AS MEAN AND STANDARD DEVIATION CAN BE COMPUTED FOR EACH GROUP. COMPARISONS OF GROUPS CAN BE DONE WITH PARALLEL PLOTS.

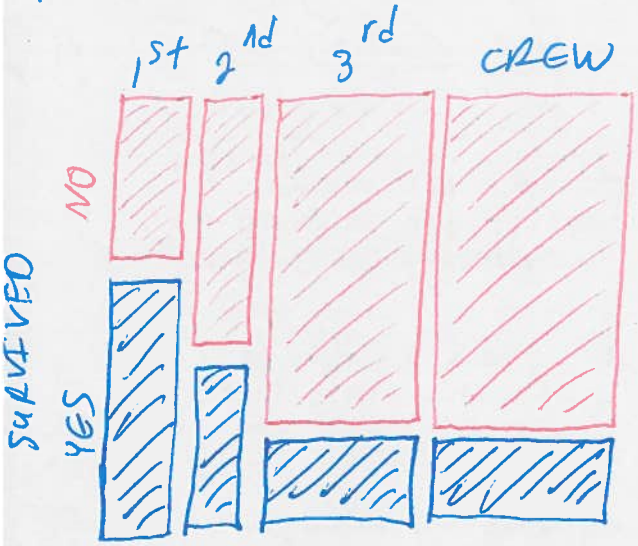



A CONTINGENCY TABLE SUMMARIZES BIVARIATE DATA OF 2 CATEGORICAL VARIABLES.

EX: TITANIC DATA

| SURVIVED | 1st CLASS | 2nd CLASS | 3rd CLASS | CREW |
|---|---|---|---|---|
| NO | 122 | 167 | 528 | 673 |
| YES | 203 | 118 | 178 | 212 |

A MOSAIC PLOT DISPLAYS CATEGORICAL VARIABLES, WITH AREAS PROPORTIONAL TO FREQUENCIES.




BOTH THE CONTINGENCY TABLE AND THE MOSAIC PLOT REVEAL ASSOCIATION BETWEEN THE TWO VARIABLES. A NUMERICAL MEASURE OF CATEGORICAL ASSOCIATION IS PROVIDED BY THE CHI-SQUARED STATISTIC $\chi^2$.

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \text{ WHERE}$$

$O_{ij}$ IS THE OBSERVED FREQUENCY IN ROW $i$, COLUMN $j$, AND

$E_{ij}$ IS THE $\frac{(\text{ROW TOTAL})(\text{COLUMN TOTAL})}{\text{OVERALL TOTAL}}$.

INTERPRETATION: $\chi^2$ IS NONNEGATIVE, SMALL VALUE MEANS WEAK OR NO ASSOCIATION, LARGE VALUE MEANS STRONGER EVIDENCE OF ASSOCIATION.

EX: HERE IS A SIMPLE CONTINGENCY TABLE INCLUDING TOTALS:

$O_{ij}$:

| | | | |
|---|---|---|---|
| 10 | 4 | 6 | 20 |
| 10 | 16 | 4 | 30 |
| 20 | 20 | 10 | 50 |

CALCULATE $E_{ij}$ FOR EACH CELL:

$$E_{11} = \frac{20 \cdot 20}{50} = 8, \text{ ETC.}$$

$E_{ij}$:

| | | | |
|---|---|---|---|
| 8 | 8 | 4 | 20 |
| 12 | 12 | 6 | 30 |
| 20 | 20 | 10 | 50 |

$$\chi^2 = \frac{(10-8)^2}{8} + \frac{(4-8)^2}{8} + \frac{(6-4)^2}{4} + \frac{(10-12)^2}{12} + \frac{(16-12)^2}{12} + \frac{(4-6)^2}{6} \approx 5.83$$

THIS IS A HIGH NUMBER (MORE ON THAT LATER), SO WE SUSPECT THE VARIABLES ARE ASSOCIATED.

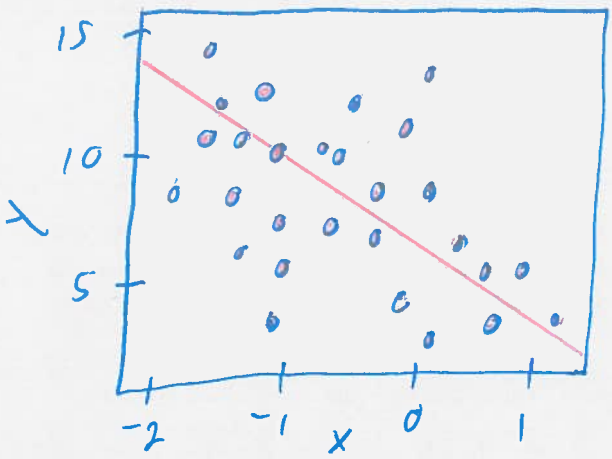
A SCATTERPLOT IS A POINT GRAPH OF TWO QUANTITATIVE VARIABLES.

x-AXIS: EXPLANATORY VARIABLE.

y-AXIS: RESPONSE VARIABLE.

i.e. x IS USED TO PREDICT y.

THE LEAST SQUARES LINE (BEST-FIT LINE) IS OFTEN INCLUDED ON A SCATTERPLOT TO INDICATE TREND. THE LINE MINIMIZES THE SUM OF SQUARED VERTICAL DISTANCES OF ALL POINTS FROM THE LINE.




A SCATTERPLOT REVEALS:

- POSITIVE/NEGATIVE AND WEAK/STRONG ASSOCIATION,
- LINEAR OR CURVED RELATIONSHIP,
- OUTLIERS,
- INFLUENTIAL POINTS THAT IMPACT THE BEST-FIT LINE,
- CLUSTERS AND GAPS.

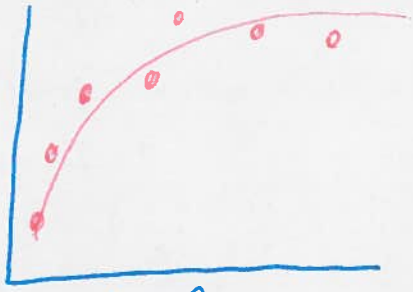
VARIABLES ARE SAID TO HAVE <u>POSITIVE ASSOCIATION</u> WHEN HIGH VALUES OF ONE RESULT IN HIGH VALUES OF THE OTHER. THEY HAVE <u>NEGATIVE ASSOCIATION</u> IF HIGH VALUES OF ONE RESULT IN LOW VALUES OF THE OTHER.

POSITIVE ASSOCIATION

NEGATIVE ASSOCIATION

THE RELATIONSHIP IS SAID TO BE LINEAR IF A STRAIGHT LINE IS A GOOD APPROXIMATION OF MOST POINTS. IT IS CURVED (NONLINEAR) IF A CURVED LINE IS A GOOD APPROXIMATION. IF THE VARIABLES ARE UNRELATED, THE BEST-~~FIT~~ LINE IS HORIZONTAL.

LINEAR



CURVED

UNRELATED

THE CORRELATION COEFFICIENT MEASURES STRENGTH AND DIRECTION OF A LINEAR ASSOCIATION.

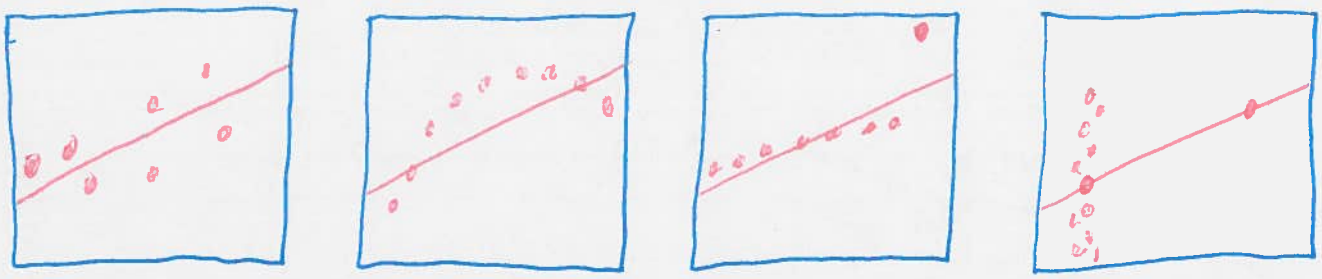
$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} = \frac{1}{n-1} \sum_i \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right).$$

R CODE:

```
cor(x,y)
```

- $-1 \leq r \leq 1$
- THE SIGN INDICATES POSITIVE/NEGATIVE ASSOCIATION.
- $r \approx 0$ INDICATES NO ASSOCIATION. $r = \pm 1$ IS A PERFECT LINEAR ASSOCIATION

BE CAREFUL, $r$ ALONE DOES NOT TELL THE WHOLE STORY.



ALL THESE BEST-FIT LINES ARE THE SAME, BUT THE DATASETS ARE CLEARLY VERY DIFFERENT. $r \approx 0.82$ FOR ALL OF THEM. FIGURE 3 HAS AN EXAMPLE OF AN OUTLIER, AND FIGURE 4 HAS AN EXAMPLE OF A HIGHLY INFLUENCIAL POINT (RIGHTMOST POINT IN BOTH FIGURES).