Topic: Exploratory Data Analysis (EDA)

Presentation of Univariate Data

Part B: In Graphs

School of Mathematics and Applied Statistics

UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Presentation of Data in Graphs

Graphs

- represent the data *visually*
- help in understanding the nature or *distribution of the data*
- are used to illustrate *relationships* between variables

Presentation of Data in Graphs

Graphs

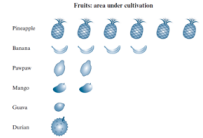
- represent the data *visually*
- help in understanding the nature or *distribution of the data*
- are used to illustrate *relationships* between variables

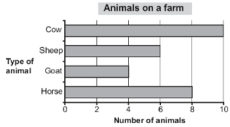
There are many different types of graphs, some are

- Pictograms
- Pie graphs
- Bar / column graphs
- Dot plots
- Histograms
- Stem-and-leaf plots

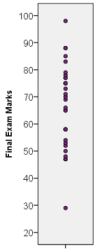
The type of data variable will determine the types of graphs that are suitable.

Appropriate Graphics - Examples

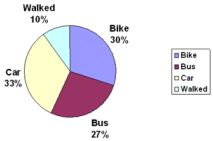


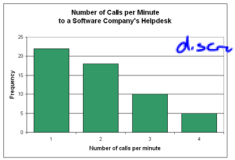


bar graph



dot plots





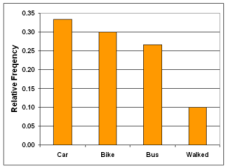
discrete

Examination result Stem-and-Leaf Plot

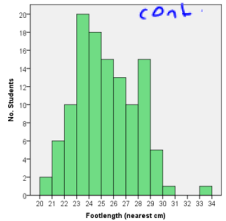
Frequency	Stem &	Leaf
1.00	2 .	9
.00	3 .	
3.00	4 .	778
7.00	5 .	0234888
4.00	6 .	5569
10.00	7 .	0135557789
4.00	8 .	3588
1.00	9 .	8

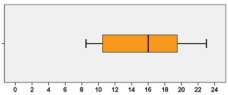
Stem width: 10
Each leaf: 1 case(s)

column "bar graph"



cont.





boxplot.

Different plots for different data types . . .

For one qualitative variable:

- pictograms
- pie charts
- bar graphs

Different plots for different data types . . .

For one qualitative variable:

- pictograms
- pie charts
- bar graphs

For one quantitative variable:

- dot plots
- bar graphs
  (a small number of discrete values)
- histograms
  (grouped discrete or continuous data)
- stem-and-leaf plots

Different plots for different data types . . .

For one qualitative variable:

- pictograms
- pie charts
- bar graphs

For one quantitative variable:

- dot plots
- bar graphs
  (a small number of discrete values)
- histograms
  (grouped discrete or continuous data)
- stem-and-leaf plots

For two qualitative variables:

- stacked bar graphs
- clustered bar graphs

Different plots for different data types . . .

For one qualitative variable:

- pictograms
- pie charts
- bar graphs

For one quantitative variable:

- dot plots
- bar graphs
  (a small number of discrete values)
- histograms
  (grouped discrete or continuous data)
- stem-and-leaf plots

For two qualitative variables:

- stacked bar graphs
- clustered bar graphs

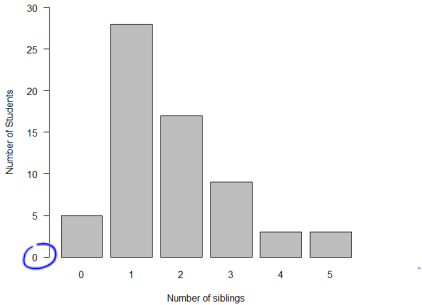
For two quantitative variable/s:

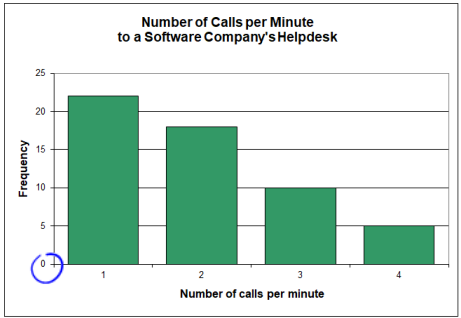
- scatterplots
- line plots (against time)

Bar chart structure

In a bar chart

- the bars are separated - they do not touch
- the width of the bars should be the same for each category
- the height (or length) of each bar represents a quantity, whereas its width means nothing
- the frequency scale MUST start at zero

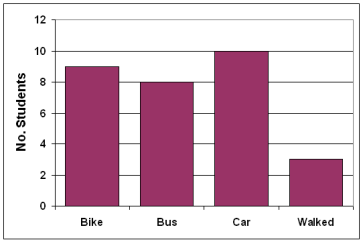




Bar Charts: Examples

Example: Bar Chart for qualitative data: order of bars can be rearranged

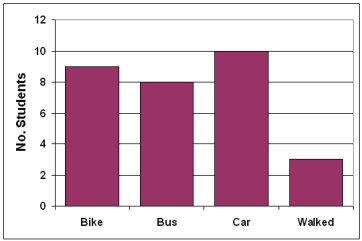
Vertical scale
-can show frequencies



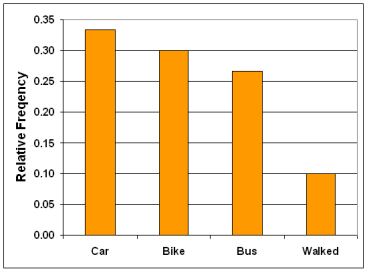
Bar Charts: Examples

Example: Bar Chart for qualitative data: order of bars can be rearranged

Vertical scale
-can show frequencies



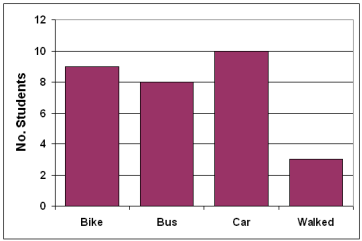
- or relative frequencies



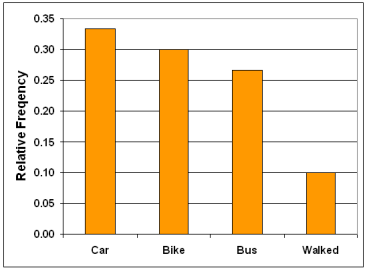
Bar Charts: Examples

Example: Bar Chart for qualitative data: order of bars can be rearranged

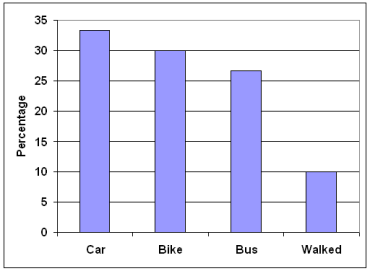
Vertical scale
-can show frequencies



- or relative frequencies



- or percentages



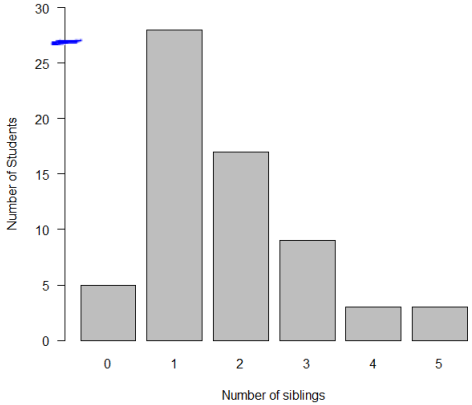
In R: Bar Charts

R code:

```
Siblings <- c(M100data$Siblings)
Siblingfreq <- table(Siblings)
Siblingfreq
```

```
Siblings
0  1  2  3  4  5
5 28 17  9  3  3
```

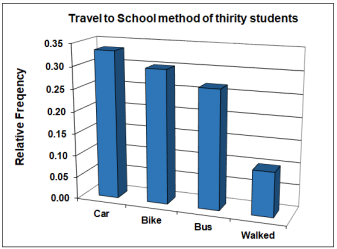
```
barplot(Siblingfreq,
xlab = "Number of siblings",
ylab="Number of Students",
las=1, ylim =c(0, 30))
```

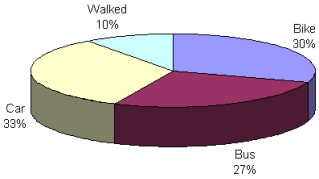


Inappropriate Graphics - Examples

Use the fewest dimensions possible:

- using 3D is volume which can distract
- don't use unless necessary - i.e. not for univariate data





- avoid pie charts
- a simple bar chart is often more effective

Histograms- Quantitative Data

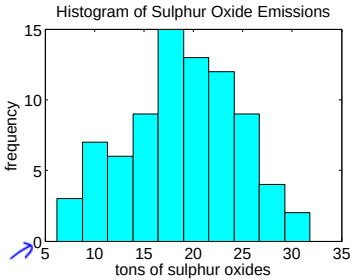
Histograms are used to represent

- **continuous:** interval or ratio data
- **discrete:** grouped data (too many unique values for a bar chart)

Histograms- Quantitative Data

Histograms are used to represent

- **continuous:** interval or ratio data
- **discrete:** grouped data (too many unique values for a bar chart)



- Real number scale on horizontal axis, **no gaps** between bars (bins).
- Observations are **grouped** into bins (classes), not necessarily of constant width.
- Vertical scale must **start at zero**
- Frequency (count) or rel. freq. is represented by **area** of bar.

Area of histogram bars

- Area $=$ height $\times$ width, so vertical axis of histogram should ideally display **density** (relative frequency $\div$ width).
- For **constant bin width**, area is proportional to height, so vertical axis can display frequency if preferred.
- For **non-constant bin width,** vertical axis **must** display **density**.

Histogram with constant bin width

Example: Sulphur emission data

Histograms: describing shape

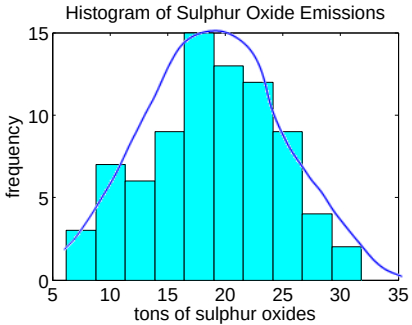
- Reasonably symmetric
- Unimodal - single hump

Histogram with constant bin width

Example: Sulphur emission data

Histograms: describing shape

- Reasonably symmetric
- Unimodal - single hump



Histogram with Non-constant bin width

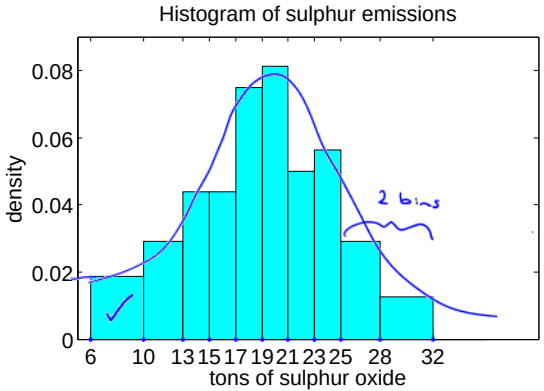
If 6 out of 80 observations satisfy $6 \leq x < 10$, then for first bar

density= $\frac{\text{rel. freq.}}{\text{width}} = \frac{6/80}{(10-6)} = \frac{6^3}{80} \times \frac{1}{4_2}$

$= \frac{3}{160.}$

$= 0.01875$

Total area of bars $= 1$



Purpose of histogram

To display **overall shape and interesting features**, including

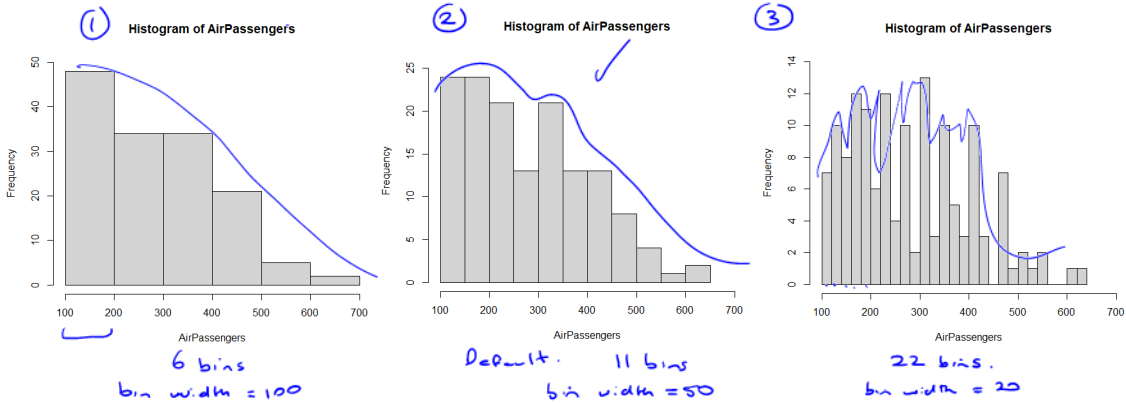
- Outliers?
- Long or short tails?
- Symmetry or skewness?
- Bell shape, U shape, uniform, . . . ?
- Unimodal/bimodal (1 or 2 humps)?

Appearance varies according to number and choice of bins;

- Rule of thumb: use number of bins as $\approx \sqrt{n}$
- avoid too few (uninformative)
- or too many (bumpy plot)

Histogram: Choice of interval length is important

Number of monthly air passengers (1949-1960): same data plotted left to right with 6, 11, and 22 bins respectively.



In R: Histogram

R code:

```
#1. Default uses 11 bins (see middle plot)
hist(AirPassengers,xlim=c(100, 700), ylim=c(0, 26))

#2. Use 6 bins (see left plot)
hist(AirPassengers, breaks=6, xlim=c(100, 700), ylim=c(0, 50))

#3. Use 22 bins (see right plot)
hist(AirPassengers, breaks=22, xlim=c(100, 700), ylim=c(0, 14))
```