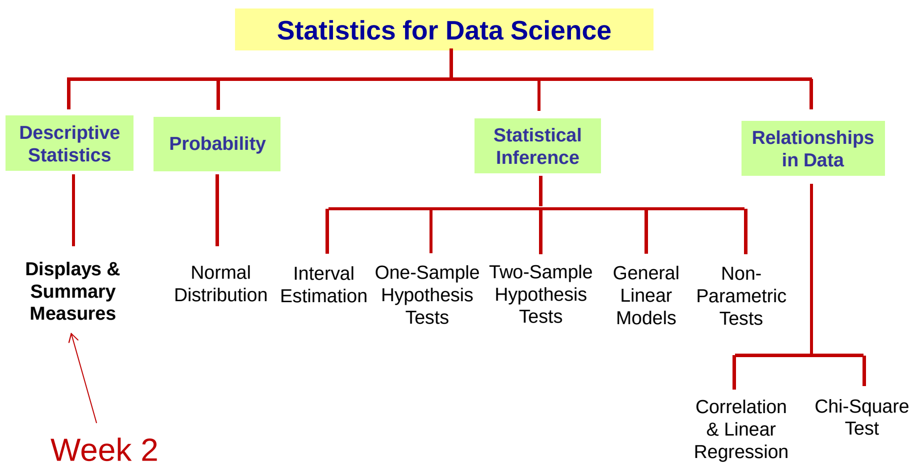
MATH 4044
Statistics for Data Science

Descriptive statistics

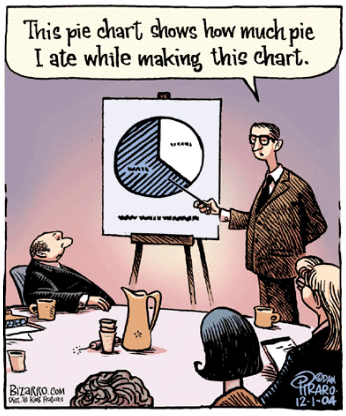
Course outline



Field, A & Miles, J, *Discovering Statistics Using SAS*, Chapters 3 & 4

Topics to be covered

- Graphical displays:
  - Frequency tables, bar charts, histograms, boxplots, scatterplots.
- Numerical summaries for quantitative variables:
  - Measures of centre
  - Measures of spread or dispersion
  - Five-number summary
- Exploratory Data Analysis



Descriptive Statistics

- A quantitative (numerical) summary of a sample.
- Meaningful presentation of data such that the sample characteristics can be effectively observed.
- Graphical display, table, summary measure (e.g. *average*).

A good data display should…

- Serve a clear purpose:
  - Show the data.
  - Be the correct kind of display for the data type.
  - Simplify complex information.
  - Highlight particular figures and situations.
- Focus on substance:
  - Avoid “chart junk” (e.g. distracting designs).
  - Avoid distortion - reveal true nature of the data.

Which data display to use?

Data Displays

Categorical data

- Tables of counts or percentages
- Pie charts
- Bar/Column and Pareto charts

Numerical data

- Frequency distributions
- Histograms
- Boxplots

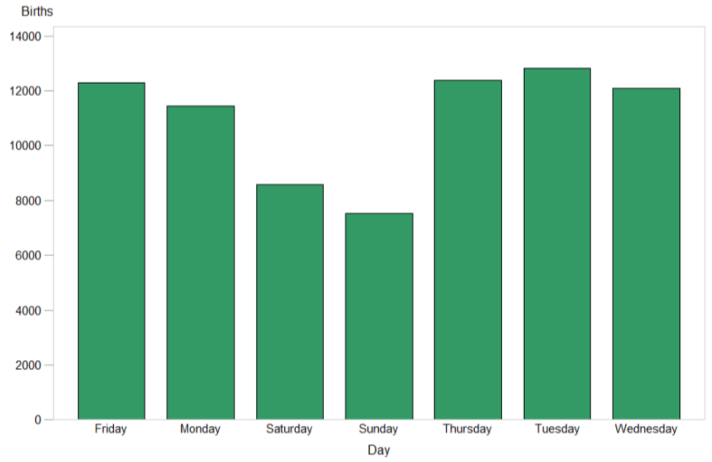
Data displays: Categorical data

Never on a Sunday?

- Births are not, as you might think, evenly distributed across the days of the week.
- In the table are the average numbers of babies born on each day of a particular week in 2002.

Day	Births
Sunday	7,526
Monday	11,453
Tuesday	12,823
Wednesday	12,083
Thursday	12,366
Friday	12,285
Saturday	8,573

Never on a Sunday – Bar chart



Tasks > Graph > Bar Chart > Simple Vertical Bar

'Column to Chart' is Day, 'Sum of' is Births

Case Study: Sit or Run?

- Regular running offers many health benefits and is an exercise approx. one in five Australians try at some stage in their lives.
- However running demands considerable effort from the heart and lungs and requires high levels of fitness.
- To understand the effects running would have on a person who is fit versus a person is not fit, an experiment was conducted on a group of University students.

Case Study: Sit or Run?

- The students were asked to run for a 1-minute period after which their pulse rates were measured.
- *The following data was collected:*
  - Height (cm) continuous
  - Weight (kg) continuous
  - Age (years) discrete (rounded)
  - Gender (Male or Female) nominal (binary)
  - Smokes (Yes or No) nominal (binary)
  - Drinks Alcohol (Yes or No) nominal (binary)
  - Exercise Frequency (High, Moderate and Low) ordinal
  - Pulse rate (bpm) discrete

Case Study: Sit or Run?

- Some questions of interest:
  - What are the characteristics of the pulse rates measured?
  - What is a typical pulse rate?
  - Is there a relationship between pulse rate and smoking, gender or exercise frequency?

Data displays: Categorical data

- Categorical (qualitative) data can be organised as either a table of counts/percentages or as a frequency distribution table.

The FREQ Procedure

Frequency / Percent / Row Pct / Col Pct

Table of Gender by Exercise

Gender	Exercise 1	Exercise 2	Exercise 3	Total
1	11	31	17	59
	10.00	28.18	15.45	53.64
	18.64	52.54	28.81	
	78.57	52.54	45.95	
2	3	28	20	51
	2.73	25.45	18.18	46.36
	5.88	54.90	39.22	
	21.43	47.46	54.05	
Total	14	59	37	110
	12.73	53.64	33.64	100.00

Data displays: Frequency table code

- You can either use ***Tasks > Describe > Table Analysis*** or run the following SAS program:

Data to be analysed

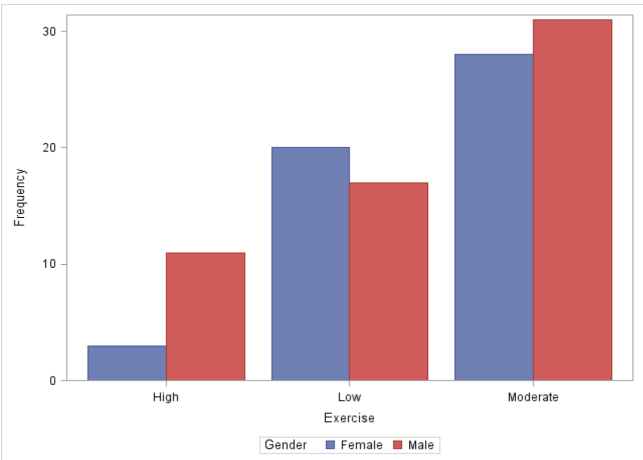
```
proc freq data=work.pulse_rates;
     tables Gender * Exercise;
run;
```

Tabulate data with *Gender* as the row variable (because it is listed first) and *Exercise* as the column variable. '*' means 'by'.

In this case, the data was in a file called `pulse_rates` stored in SAS temporary library `Work`.

Data display: Bar Chart

Clustered Bar Chart



These types of charts are useful for broad comparisons.

Data display: Bar chart code

- The code to produce this diagram is as follows:

```
proc sgplot data=work.pulse_rates;
    vbar Exercise / group=Gender groupdisplay=cluster;
run;
quit;
```

Variable to be plotted

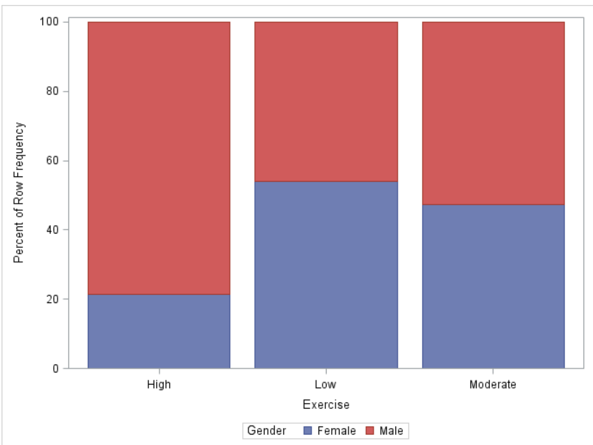
Selected options following a slash '/'

Produces vertical bars; use `hbar` to obtain horizontal bars

Notice the QUIT statement in this program.
Certain procedures in SAS have something called RUN-group processing. This kind of processing keeps the procedure in memory, even if it encounters a RUN statement. Because the procedure is still in memory, you can request additional charts or models. The QUIT statement ends the procedure.

Data display: Bar Chart

100% Stacked Bar Chart



These types of charts are useful for broad comparisons.

Data display: Bar chart code

- The code to produce this diagram is as follows:

```
proc freq data=work.pulse_rates;
    table Exercise * Gender / out=freq outpct;
run;

proc sgplot data=freq;
    vbar Exercise / response=pct_row
/* response= means SGPLOT will plot the summed values */
        group=Gender;
run;
quit;
```

New data file consisting of percent frequencies

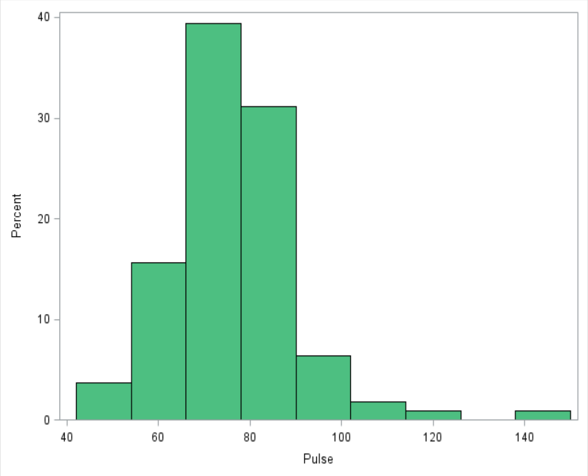
Commented text

The trick to making relative frequencies add up to 100% is to pre-calculate the relative frequencies using PROC FREQ.

Data displays: Numerical data

- Quantitative (numerical) data can be organised as either a frequency distribution table, a histogram or a boxplot.
- Data needs to be sorted in ascending order first (ordered arrays).
- Ultimately we are interested the distribution of the data.
- Why? Quantitative variables often take many values; distributions tells us what value a variable takes and how often it takes these values.
- The most common distribution display is a histogram.

Data display: Histogram



- This distribution is skewed right.
- It appears to have one peak between 70 and 80.
- The spread is from 45 to 150 beats per minute.
- There appears to be potentially two outliers, a pulse rate of approx. 120 and another one above 140.

What does this distribution suggest about the *typical* pulse rate?

Data display: Histogram code

- One way to produce a histogram is to submit the following code:

```
proc sgplot
data=work.pulse_rates;
     histogram Pulse / fillattrs=fill (color=big);
run; quit;
```

Option to specify box fill colour other than light grey SAS uses by default. [A few different colours are available.]

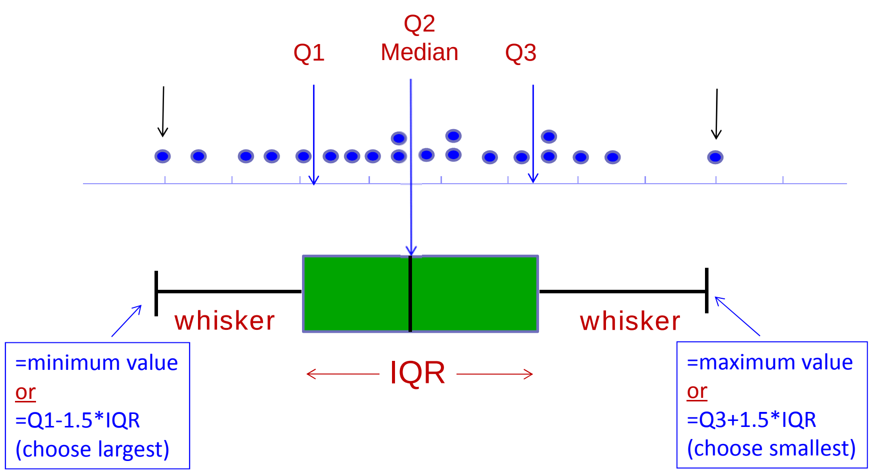
The five-number summary

- The five-number summary consists of:
  - ☐ Minimum
  - ☐ First quartile Q_1
  - ☐ Median
  - ☐ Third quartile Q_3
  - ☐ Maximum

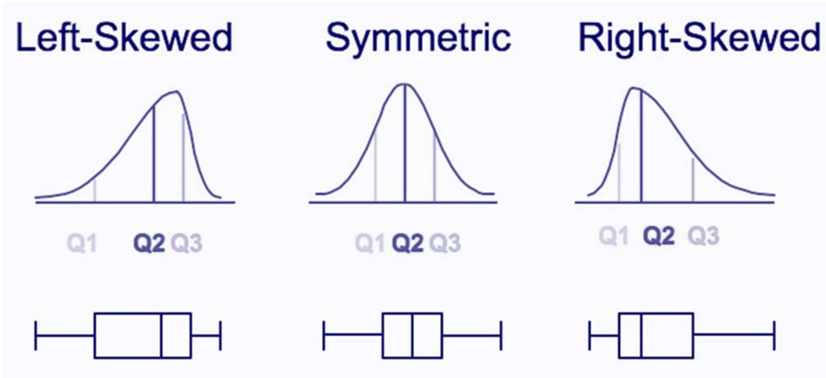
- Represented graphically with a boxplot.

The Boxplot

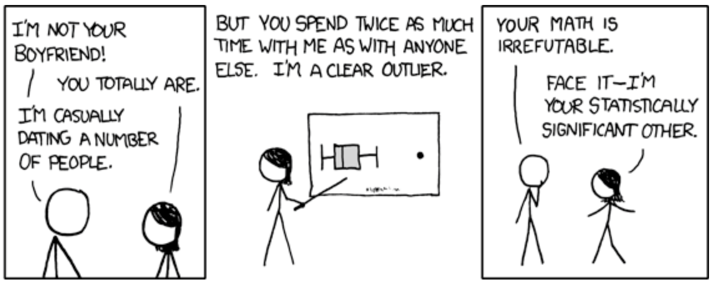
- Use the boxplot if using median, IQR.



Boxplots and distribution shapes



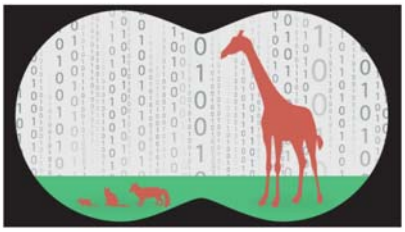
Outliers



- Sample values that lie far away from the vast majority of the other sample values.
- 'Extreme' or 'unusual' observations.
- What is 'extreme' or 'unusual'?

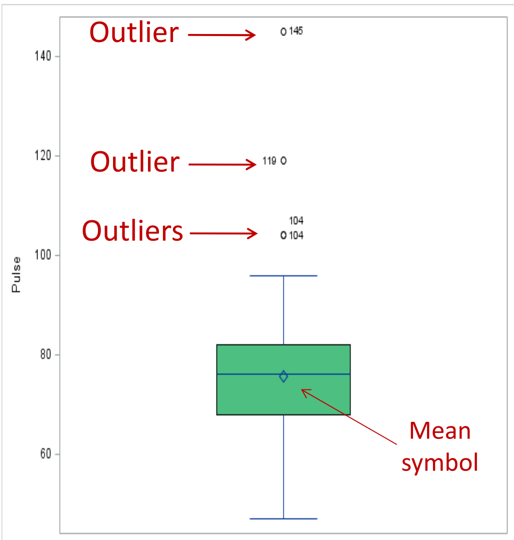
Why look for outliers?

- Examination of data for possible outliers serves many useful purposes, including:
  - Identifying strong skew in the distribution.
  - Identifying data collection or entry errors.
  - Providing insight into interesting properties of data.



Beware the giraffes in your data!

Data display: Boxplot



Min = 47, Max = 145, Q1 = 68, Q3 = 82, IQR = 14

Q3 + 1.5xIQR = 103

The maximum value is larger than 103, so the whisker stops at 103 and values beyond this point are outliers.

Q1 – 1.5xIQR = 47

This value exactly the minimum so the lower whisker stops at 47 and there are no outliers at the lower end of the data.

Data display: Boxplot code

- One way to produce a boxplot is to submit the following code:

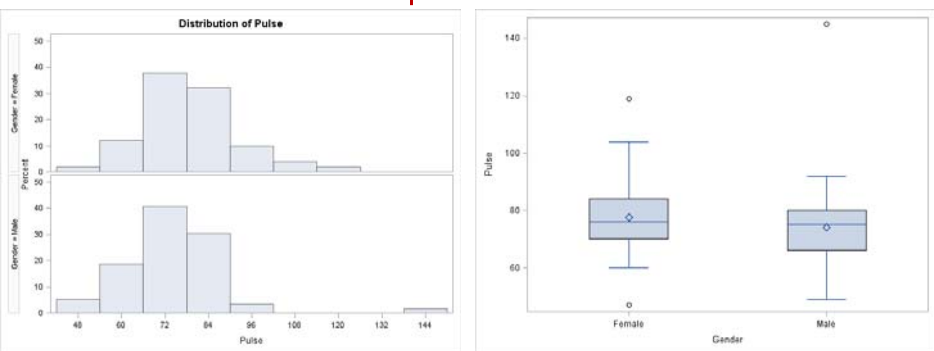
```
proc sgplot data=work.pulse_rates;
      vbox Pulse / datalabel=Pulse fillattrs=fill (color=big);
run;
quit;
```

Option to change box fill colour

The statement VBOX tells SAS to produce a vertical box. DATALABEL option was added after a slash '/' to identify outliers by their *Pulse* values. Alternatively, outliers could have been identified by their subject number in the data file.

Distribution of *Pulse* rate by *Gender*

Is there a relationship between *Pulse* and *Gender*?



Tasks > Describe > Summary Statistics with *Pulse* as 'Analysis variable' and *Gender* as 'Classification variable'. Go to 'Plots' tab and select 'Histogram' and 'Boxplot'.
The code produced by SAS for this task can be modified to change the colour of histogram bars and boxplots or add labels for outliers.

Measures of centrality and dispersion

Mean (centre)
Standard Deviation (dispersion)

Coefficient of variation

Use for (nearly) symmetric distributions without outliers

Median (centre)
IQR (dispersion)

Use for skewed distributions or distributions with outliers

The standard deviation

- The most common measure of dispersion.
  - Same units as the data.
- Measures the 'average deviation' of observations from the *mean*.

Sample standard deviation

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

almost computing the "mean"

where n = the number of observations

Calculating standard deviation

x	$x-\bar{x}$	$(x-\bar{x})^2$
0	-3	9
2	-1	1
3	0	0
4	1	1
6	3	9
$\sum x = 15$	$\sum(x-\bar{x}) = 0$	$\sum(x-\bar{x})^2 = 20$

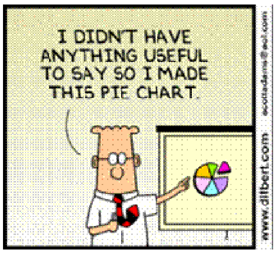
$\bar{x} = 3$

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\frac{20}{5-1}} = \sqrt{\frac{20}{4}} = \sqrt{5} = 2.24$$

On average, observations are 2.24 units below or above the mean.

Exploratory Data Analysis (EDA)

- The process of using statistical tools to investigate data sets in order to understand their important characteristics.
- Statistical tools:
  - Graphs;
  - Measures of centre;
  - Measures of dispersion.



Example: Pulse and Exercise

- Describe the shape of the distribution of *Pulse* by *Exercise*.
- Nominate and interpret values of appropriate measures of centre and spread.
- Compare and contrast the distributions.
- Is there a relationship between *Pulse* and *Exercise*?

Analysis Variable : Pulse										
Exercise	N Obs	N	N Miss	Mean	Std Dev	Minimum	Maximum	Median	Lower Quartile	Upper Quartile
High	14	14	0	68.643	12.689	49.000	96.000	68.500	60.000	76.000
Low	37	37	0	78.351	11.458	52.000	119.000	78.000	71.000	85.000
Moderate	59	58	1	75.690	14.093	47.000	145.000	75.000	68.000	80.000

Example: Descriptive statistics code

- Using the MEANS procedure:

```
proc means data=work.pulse_rates n nmiss mean std
min max median maxdec=3 q1 q3;
      var Pulse;
      class Exercise;
run;
```

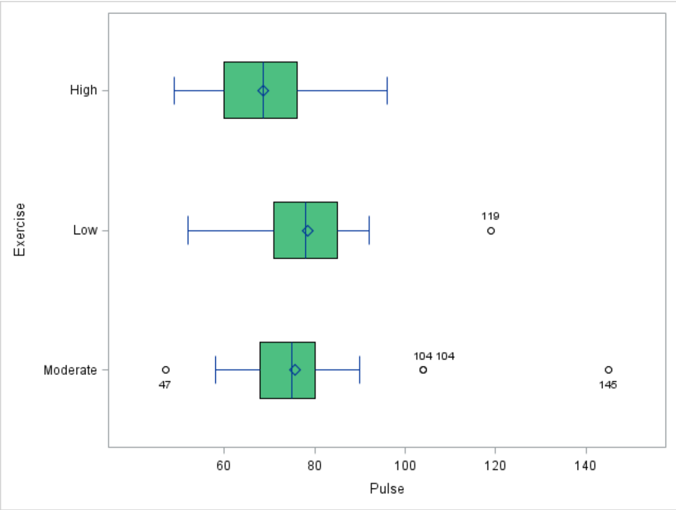
PROC MEANS is a popular SAS procedure that produces a number of useful statistics.
If no options are specified, only the number of non-missing observations, the mean, standard deviation, the min and the max value are printed. The option MAXDEC=n specifies how many decimal places we want.

The VAR statement specifies the variable to be analysed.

The CLASS statement tells the procedure to produce selected statistics for each value of categorical variable *Exercise*.

Example: Boplots

Distribution of *Pulse* by level of *Exercise*



Example: Boxplots code

Using the SGPLOT procedure:

```
title "Boxplots of Pulse for each value of Exercise";
proc sgplot data=work.pulse_rates;
      hbox Pulse / category=Exercise datalabel=Pulse
      fillattrs=fill (color=big);
run;
quit;
```

The statement HBOX tells SAS to produce a horizontal box. To see a boxplot for each value of the categorical variable *Exercise,* CATEGORY option was added.

Example: Discussion

- Distribution of pulse rates is reasonably symmetric for all three levels of exercise.
- Boxplots indicate outliers for low and moderate levels of exercise.
  - There is one subject with unusually high pulse rate of 119 bpm in the low exercise frequency group.
  - In the moderate exercise frequency group, there is one subject with unusually low pulse rate of 47, and subjects with relatively high pulse rates of 104 (two subjects) and 145 (one subject).
  - We could examine the data file to identify these subjects further.

Example: Discussion

- As there are outliers present, median and IQR will be used to describe centre and spread of the distribution of pulse rates by level of exercise.
  - For the low exercise group, Median = 78 bpm and IQR = 85 – 71 = 14 bpm.
  - For the moderate exercise group, Median = 75 bpm and IQR = 80 – 68 = 12 bpm.
  - For the high exercise group, Median = 68.5 bpm and IQR = 76 – 60 = 16 bpm.
- Typical pulse rate appears to be lowest for the high exercise group and highest for low exercise group.
- The difference in typical pulse rate between low and moderate exercise groups is less pronounced.
- Variability in pulse rates, as measured by IQR, appears to be similar for the three groups.
- Formal statistical tests can be performed to determine whether observed differences among groups are statistically significant.

Coefficient of variation

- A measure of relative variability used to:
  - Measure changes that have occurred in a population over time.
  - Compare variability of two populations that are expressed in different units of measurement.
- Expressed as a percentage rather than in units of the particular data:

Standard deviation

$$CV = \left(\frac{S}{\bar{X}}\right) \times 100\%$$

Mean

Variability
relative to
typical value

Example: Comparing variation

- Descriptive Statistics for Age (years) and salary including bonuses (thousands of dollars) of CEOs:

Variable	Mean	Std Dev	Minimum	Maximum	N
Age	51.4666667	8.9223898	32.0000000	74.0000000	60
Salary	404.1694915	220.5335343	21.0000000	1103.00	59

- Difference in units and magnitude make it not appropriate to compare standard deviations directly.
- Coefficient of variation should be used instead.

Forbes, November 8, 1993, 'America's Best Small Companies'. Small companies were defined as those with annual sales greater than five and less than $350 million and ranked according to 5-year average return on investment. This data covers the first 60 ranked firms.

Example: Comparing variation

- Age:

$$CV = \frac{s}{\bar{x}} \times 100\% = \frac{8.92}{51.47} \times 100\% = 17.34\%$$

- Salary:

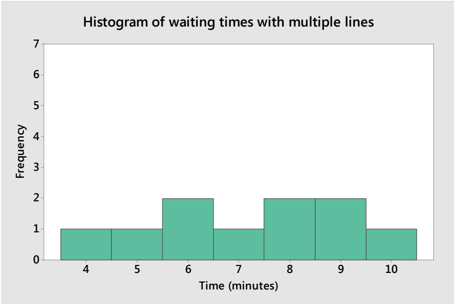
$$CV = \frac{s}{\bar{x}} \times 100\% = \frac{220.53}{404.17} \times 100\% = 54.56\%$$

- We can see that CEO age has considerably less variation than CEO salary.

Example: How long do I have to wait in line?

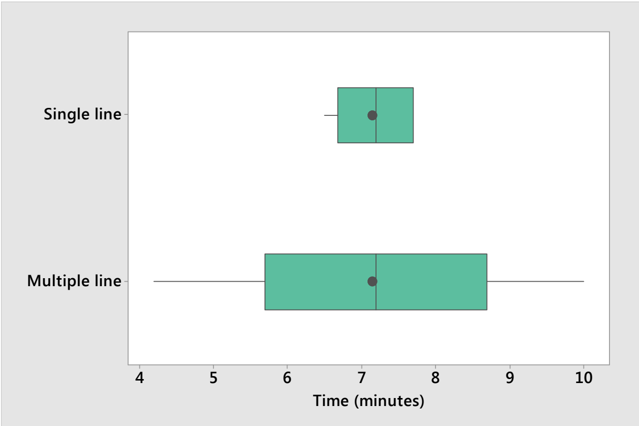
- A shop experiments with two different configurations for serving customers:
  - A single waiting line for three different checkouts;
  - Individual lines at three different checkouts.
- Which is the better configuration?

Example: Which is the better configuration?



Variable	Mean	Std Dev	Minimum	Maximum	N	Lower Quartile	Median	Upper Quartile
Single	7.150	0.477	6.500	7.700	10	6.700	7.200	7.700
Multiple	7.150	1.822	4.200	10.000	10	5.800	7.200	8.500

Example: Which is the better configuration?



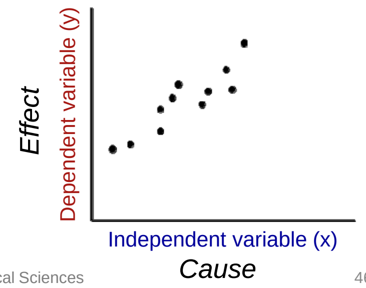
Does it make sense to use mean and standard deviation to describe centre and spread of the distribution of waiting times?

Example: Which is the better configuration?

- With either configuration, both the mean and median waiting time is 7.15 minutes.
- However, there is considerably less variability with a single waiting line:
  - Standard deviation of approximately 0.5 minutes, compared to approximately 1.8 minutes with multiple lines.
- The single line configuration seems to work better.

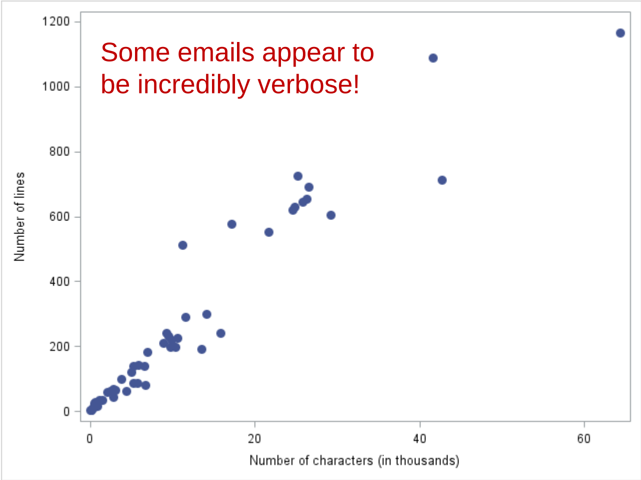
Data displays: Relationships in numerical data

- Visual impression of whether a relationship or association exists between numerical variables can be formed using a simple scatterplot.
  - Case by case view of data for two numerical variables.
- We are typically interested in cause and effect relationship between variables:
  - Dependent variable is the variable to be predicted.
  - Independent variable is the variable used to make predictions.

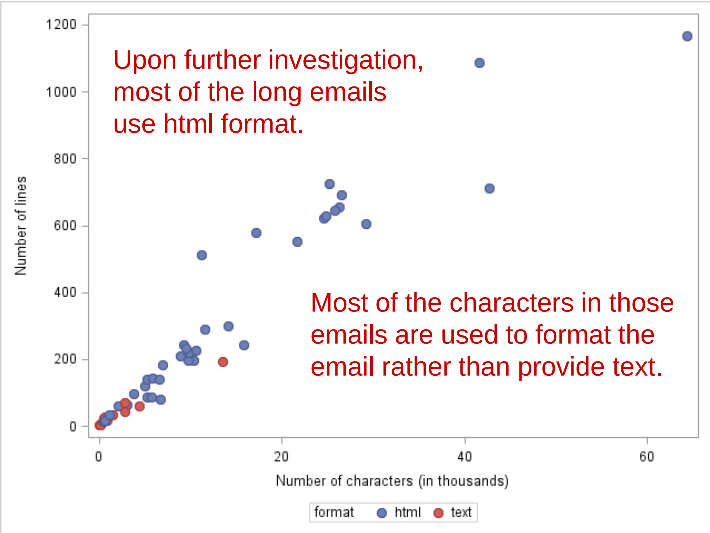


Data display: Scatterplot

- Number of line breaks vs number of characters in a sample of 50 emails:



Data display: Scatterplot grouped



Data display: Scatterplots

- Use tasks or write your own code, e.g.:

Simple scatter

```
proc sgplot data=work.email_sample;
       scatter x=num_char y=line_breaks /
       markerattrs=graphdata1(symbol=circlefilled size=8pt);
       label line_breaks = 'Number of lines';
       label num_char = 'Number of characters (in thousands)';
run; quit;
```

Grouped scatter

```
proc sgplot data=work.email_sample;
       scatter x=num_char y=line_breaks / group=format
       markerattrs=graphdata1(symbol=circlefilled size=8pt);
       label line_breaks = 'Number of lines';
       label num_char = 'Number of characters (in thousands)';
run; quit;
```