Study Design, Sampling and Bias (W1)

Study Design (L1)

What causes lung cancer?

- Number of known cases rose dramatically in second half of 19th century, even more dramatically 1900-1910
- In 1900 only 140 cases documented in medical literature
- In 1910 Dr George Dock invited the two senior classes in medical school at Washington University to witness an autopsy: “the condition was so rare he thought we might never see another case as long as we lived.”
- By the 1920s it was becoming common
- Today it kills about 1.5 million people per year globally
- **What caused the dramatic increase?**

Possible causes

- Few cigarettes before WWI
- Increase in air pollution caused by industry
- Asphalting roads
- Increase in automobile traffic
- Exposure to gas in WWI
- Influenza to gas in WWI
- Influenza epidemic in 1918
- Working with benzene or gasoline
- **How would you find out?**
- **Mueller**, 1939:

	Absolute number		Percent	
	Lung cancer patients	Healthy	Out of 86 lung cancer patients	Out of 86 healthy men
Extreme smoker (daily consumption of 10–15 cigars, more than 35 cigarettes, more than 50 g of pipe tobacco)	25	4	29.07	4.65
Very heavy smoker (7–9 cigars, 26–35 cigarettes, 36–50 g of pipe tobacco)	18	5	20.93	5.81
Heavy smoker (4–6 cigars, 16–25 cigarettes, 21–35 g of pipe tobacco)	13	22	15.12	25.58
Moderate smoker (1–3 cigars, 1–15 cigarettes, 1–20 g of pipe tobacco)	27	41	31.39	47.68
Non smoker	3	14	3.49	16.28
Altogether	86	86	100.00	100.00

Questions

- Is this a good experimental design?
- What do the results show?
- Does this analysis settle the matter?

Criticisms of Mueller, 1939

- Occupation and the flu are discussed for the cases only
- Unexplained assumptions about the smoking habits of 20 cases
- Silent about the sampling, recruitment and interview modes of the “healthy” subjects
- Selection and differential misclassification cannot be ruled out

Statistic is a game changer

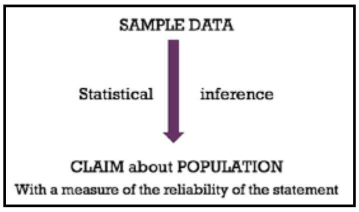
- So many issues in the news can be settled with statistics
  - Does smoking cause lung cancer?
  - Do vaccinations cause autism?
  - Do wind farms cause sickness?
  - Do cell phones cause brain cancer?
  - Does homeopathy work?
- The answers do not have to remain matters of opinions
  - Statistics can answer these questions
  - They remain contentious because people don’t understand how statistics work

Statistics is…

- Important in politics
- Important in agriculture
- Important in medicine
- Important in engineering
- Important in business and finance
- Important in science
- **Statistics is the science of collecting, organising and interpreting numerical facts, which we call data**

Statistics is…
- ‘the science of quantitative reasoning’ — of ways of thinking about and working with numerical facts and ideas
- Is a collection of procedures and principles for gathering data and analysing information in order to help people make decisions when faced with uncertainty
- But the science of statistics ‘has much more in common with philosophy than it does with accounting’

Statistical design of experiments
- Sample data
- Statistical inference of sample data
- Claim about population
  - With a measure of the reliably of the statement (CI)



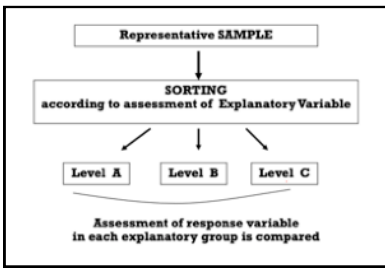
Possible Steps in any study
- **AIM**: exploratory or a specific question?
- What will I measure? What **VARIABLES**
- How will I choose the subjects? - **SAMPLING**
- What is needed for the **DATA COLLECTION**?
- SUMMARY of data and **STATISTICAL ANALYSIS** (INFERENTIAL TESTS)
- Interpretation of data and **CONCLUSION** or DECISION

Type of variable
- **Explanatory (independent)** variable: may explain or cause a change in another variable, may be manipulated or set at a value
- **Response (dependent)** variable: the variable measured to see if it changes in response to another variable
- **Confounding** variable: a variable that is thought/**known** to influence both the explanatory and response and so confuses the interpretation of any relationship between them
- **Lurking** variable: as confound but **NOT known** beforehand

Simpson’s paradox
- Rare cases of misleading information
- In 1973, University of California, Berkeley was sued for **bias against women** who had applied for admission
- Problem is unfair averaging over different groups - better to compare % or average within each level
- Overall % could be misleading… direction of a relationship is reversed within sub-groups compared to the overall total group
- Arises from influence of a Confounding Variable
- Example 4: 47% of men successful, but only 31% of women overall in being admitted to a university BUT women applied in different faculties to men (ones where admission was more difficult) and were more successful than men there!

Types of study - **OBSERVATIONAL**
- **No manipulation** of the factors under investigation
- No random assignment of units to any specific treatment
- Can be very informative and is the only way possible if:
  - Ethical considerations preclude manipulation
  - Want to see what happens in a natural setting, without contrived involvement
- Causation is difficult to establish - too many uncontrolled confounding, lurking variables
- Special type of **observational study**: **case-control study**
- Often seen in medical studies where cases of a disease are compared with others who are known not to have the disease (control group)

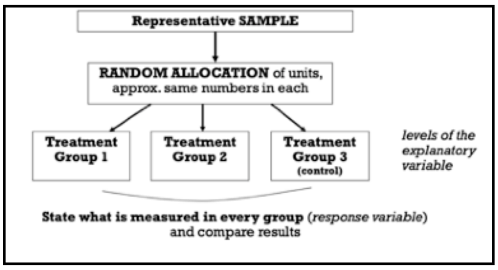
Observational Study Schematic Layout
- Representative **sample**
- **Sorting**: according to assessment of explanatory variable
  - Level A, B and C



- **Assessment** of response variable in each explanatory group is compared

<u>Types of study - **EXPERIMENTAL**</u>
- Do have **active imposition** of a treatment level on the subject
  - Different values of explanatory can one set or controlled
- Do randomly assign units to a specific treatment group
- Measure the responses for each treatment group
- “Cause and effect link” more likely to be established

<u>Simple randomised experiment</u>
- General schematic layout:



- Representative sample
- Random allocation of units, approx same numbers in each
  - Treatment Group 1, 2 and 3
- State what is measured in every group (Response variable) and compare results

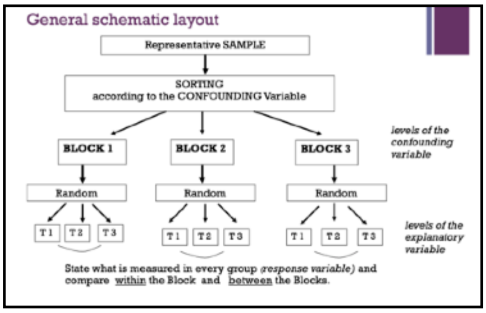
<u>Types of study - **EXPERIMENT**</u>
- Major features of a randomised experiment are:
- **CONTROL** - “absence”, placebo, **blinding**
- **RANDOMISATION** -> all treatment groups have similar background, minimises bias
- **REPLICATION** - sample size sufficient, each treatment group was >1 unit, minimises random error
- **BLOCKING** - according to a confound variable to “control”

<u>Example 3: Health study</u>
- 21,996 male physicians, 2 (different looking) drugs efficacy?
- 2-factorial experiment, random assignment to 4 treatment groups:

Aspirin + Beta carotene	Aspirin + Placebo
Beta carotene + Placebo	Placebo + Placebo

- Response variables: heart attack, rates of cancer after period of time
- Controls?
- Randomisation?
- Replication?

<u>Blocked Random Experiment: General schematic layout:</u>



<u>Special forms of Experimental Design</u>
1. Factorial Experiments:
- Vary more than 1 factor at a time - time and cost considerations
- Can see the interactions between different explanatory variables
2. Repeated-measures designs:
- Blocks = individuals, and
- Units = repeated time periods in which receive varying treatments
3. Matched-Pair designs
- Either two matches individuals or same individual receives each of two treatments
- Important to randomise order of two treatments and use blinding if possible

<u>Summary of Lecture 1</u>
- Types of variables:
  - Explanatory
  - Response
  - Confounding (lurking)
- Types of Study:
  - Observational Study
  - Experiment
- Design Features of Experiments:
  - Control (e.g. placebo)
  - Randomisation
  - Replication
  - Blocking

<u>Sampling and Randomisation (L2)</u>

<u>Shere Hite and female sexuality</u>

- Shere Hite is a sex researcher best known for her research into female sexuality
- Her research method involved distributing huge numbers of surveys (~100,000) to women's organisations and analysing the answers of those who responded
- In one study on marriage satisfaction among women
  - 98% reported dissatisfaction
  - 75% reported extramarital affairs
  - 4% given the survey responded
- Some criticisms:
  - Women who were dissatisfied were more motivated to respond
  - Distribution to women's organisations
  - Long questions and responses bias towards atypical respondents
  - Leading questions
- Some advantages:
  - Responses are more likely to be truthful (?)
  - Non-laboratory setting

<u>Some important terminology</u>

- **Population** -> parameters
  - The larger group of units about which **inferences** are to be made
- **Sample** -> statistics
  - The smaller group of units actually measured
- Even a small sample can be used to make inferences about a much larger group or the population
- If the sample chosen can be considered to be truly **representative** of that population with regard to the questions of interest

<u>Advantages of a Sample Survey over a Census</u>

- **Sometimes** a **census isn't possible**
  - When measurements destroy units
- **Speed**: especially if population is large
- **Accuracy**: devote resources to getting accurate sample results
- **Cost**: less costly and less time than census

<u>Sampling Design</u>

- The sampling process comprises several stages:
  - Defining the population of concern. The entire group of objects or people about which information is wanted is called the **population**
  - Individual members of the population are called **units**
  - A **sample** is a part of the population that is actually observed in order to gather information
- Specifying a **sample frame**, that is an available list that represents the population. A simple one is the electoral roll or a telephone list
- Specifying a **sampling method**, which can be either haphazard and convenience styles to those based on probability and randomness)This will be the major focus of this lecture)
- Determine the **sample size** in order to achieve a desired accuracy (in general the **error** is proportional **$1/\sqrt{n}$** , where n is the sample size)
- Implementing the sampling and data collecting, and
- Applying statistical description and inference to the sample statistics (latter parts of this course)

<u>Types of sampling:</u>

<u>Convenience or haphazard samples (1)</u>

- Problems with bias:
- Selection bias
  - Voluntary response bias
  - Non-response bias
- Undercoverage bias (similar to selection bias)
- Questionnaire wording

Probability samples (2)
- Each member of the population has a known chance of being selected
- This requires effective randomisation of the population list

Simple Random Sampling
- All units in population have the **same chance** of being in the sample (uniform distribution), and
- Every conceivable group of units of the required size has the **same chance** of being the selected sample (a completely randomised sample)
- Usually representative IF population is relatively homogenous

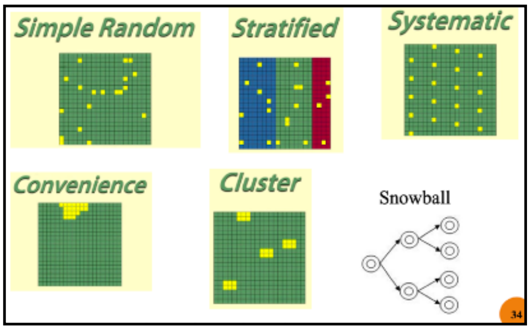
Other probability sampling methods
- Stratified random sampling
- Cluster sampling
- Systemic sampling
- Multistage sampling
- Why not use simple random sampling in all situations?
- Not always practical or best representation of all

Stratified Random Sampling
- Better representation when there are subgroups for the feature of interest that are:
  - Different to each other (heterogeneous), but
  - Similar within each sub-group (homogenous)
- With a simple random sample it is possible to be biased towards one section of population. This is the nature of randomness!

Stratified Random Sampling Procedure
1. Sort population into the sub-groups according to this known confounding variable
2. Randomise each sub-group
3. Select a simple random sample from each sub-group
   - Sample size in each?
     - Simple way: equal amounts in each stratum… not usual
     - Proportional to size of each stratum in population… simplest and reasonably accurate
     - Proportional to the variability within each stratum… most accurate, but not easiest
4. Combine to make the total representative sample

Other probability sampling
- Cluster sampling
  - Separate sample in to clusters for convenience
  - Randomly select 1 or more clusters and sample all in it
- Systematic sampling
  - Randomly select one individual in population and then every nth on list to achieve a total sample size
- Multistage sampling
  - For large national surveys
  - Combination of sampling types, e.g. stratify by religion, then by income level within region… take one area cluster in that region for each income

Summary of Lecture 2
- Population vs Sample
- Types of Sampling
  - Haphazard
  - Simple Random Sampling
  - Stratified Random Sampling
  - Cluster Sampling
  - Systematic Sampling
  - Multistage Sampling
- Survey bias: Selection bias, undercoverage, non-response, response bias, questionnaire wording



Variables and Distributions (L3)

Looking at Data — Variables

- Any set of data contains information about some set of **individuals**
- This information is organised in **variables**

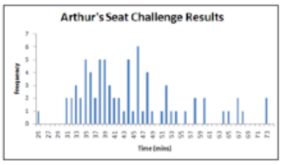
Variables

- **Quantitative variables** have **numerical** values taken on each individual
  - Examples: height, number of siblings, fastest speed driven
  - Averages and other computations make sense
- **Categorical variables** are **group** or **category** names that don't necessarily have logical ordering
  - Examples: gender, eye colour, country of residence
  - Sometimes encoded numerically e.g. Female = 1, Male = 0
  - Categorical variables such as gender can't be averaged
  - **Ordinal** variables: Categorical variables that have a **logical ordering**
    - Examples: t-shirt size (S, M, L, XL), grade achieved (N, P, C, D, HD)
  - **Nominal** variables: Categorical variables that have no logical ordering
    - Examples: hair colour, nationalities

Distributions

- Two sets of numbers
  - Values a variable may take
  - Frequency with which it takes them
- Histogram - displays this **frequency** distribution information

Time (mins)	Frequency	Time (mins)	Frequency	Time (mins)	Frequency	Time (mins)	Frequency	Time (mins)	Frequency
25	1	35	5	45	1	55	0	65	1
26	0	36	4	46	6	56	1	66	0
27	0	37	2	47	1	57	0	67	2
28	0	38	5	48	4	58	2	68	1
29	0	39	5	49	1	59	0	69	0
30	0	40	3	50	0	60	2	70	0
31	2	41	2	51	1	61	0	71	0
32	2	42	2	52	3	62	0	72	0
33	3	43	1	53	1	63	0	73	2
34	2	44	5	54	1	64	1		



Important features of distributions

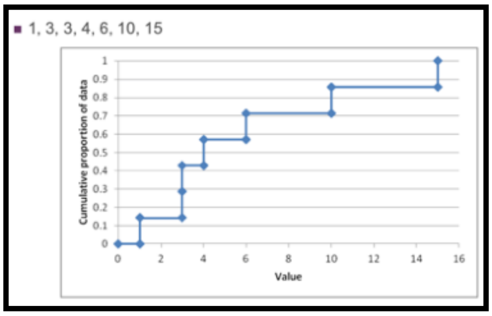
- Location: Around what value are the data located?
  - E.g. centre (median)
- **Spread**: What is the variability among the data values?
  - Are there any deviations? Gaps? Outliers?
- **Shape**: What is the distribution of the data?
  - Unimodal? Multi-modal?
  - Symmetric? Skewed?

Median and Quartiles

- **Median**: The median, M, is the midpoint of a distribution, the number such that half the observations are smaller than it and the other half are larger. It is the 50th percentile
- **First quartile**: The first quartile, Q1, has 25% of the data below it and 75% above it. It is the 25th percentile
- **Third quartile**: The third quartile, Q3, has 75% of the data below it and 25% above it. It is the 75th percentile

Quantiles (Percentiles)

- A **quantile** is a value that is greater than a given **proportion** of the data
- When the **proportion** is expressed as a **percentage**, the value is called a **percentile**



Cumulative distributions:

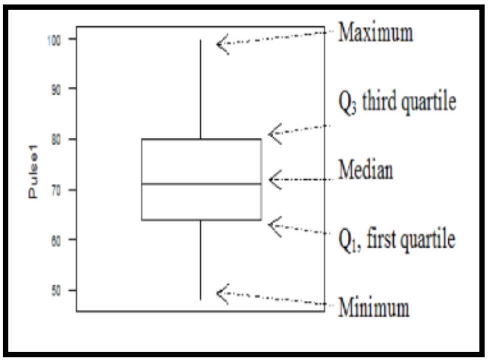
A rule for quartiles

- Calculate **0.25n**, where n is the number of values
- If it's **an integer**, count off that may values in the ordered list. Q1 is halfway between that **value** and **the next**
  - e.g. 1,3,4,4,6,10,15,126 contains n = 8 values
  - N/4 = **2** is an integer, So Q1 is halfway between **second** and **third** values, i.e. 3.5

- If its **not an integer**, **round up** and count off that many values in the ordered list. Q1 is that value
  - E.g. 1,3,3,4,6 contains n=5 values
  - N/4 = **1.25**, so Q1 is the **second** value, i.e. 3

A rule for quartiles
- For Q3, count the same number of values as for Q1 backwards form the largest value
  - E.g. 1,3,4,4,6,10,15,126 contains n = 8 values
  - N/4 = 2 is an integer, so Q3 is halfway between second and third last value, i.e. 12.5
  - E.g. 1,3,3,4,6 contains n = 5 values
  - N/4 = 1.25, so Q3 is the second last value, i.e. 4
- Same rule works for any quantile:
  - Calculate **np**, where p < 0.5 (percentile)
  - If **np** is an **integer**, take the **midpoint** between that **value** and **the next** if not **round up** and take that value

Boxplots (preview):



Measuring Spread
- Interquartile Range **IQR = Q3-Q1**
  - In pulse data: IQR = 80 - 64 = 16
  - **Middle 50% of data has this spread**
- Range = max - min
  - In pulse data: Range = 92 - 48 = 44
  - (Not as useful as IQR)

Quartiles in Excel
- Investigate the built-in function wizard, Fx
  - =QUARTILE(data cell range, x)
    - If x=0 —> minimum)
    - If x=1 -> 1st quartile
    - If x=2 -> 2nd quartile (MEDIAN)
    - If x=3 -> 3rd quartile
    - If x=4 -> maximum
  - e.g. Pulse data set
    - Median Pulse 1 is =QUARTILE(A2:A93,2) = 71

Summary of Lecture 3
- Variables
  - Quantitative
  - Categorical (ordinal)
- Distributions
  - Values
  - Frequencies
- Histograms
- Median, Quartiles, Quantiles
- Interquartile range

Quantitative Data, Linear Transformation & Cat. Data (W2)

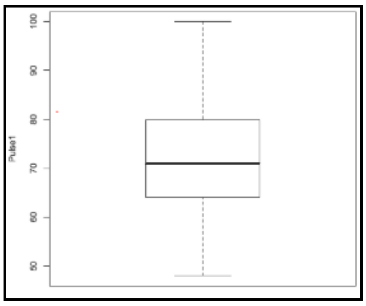
Quantitative Data (L1)

Five-number Summary

- The **five-number summary** of a distribution consists of the median, M, the quartiles Q1 and Q3 and the smallest and largest observations written in the order:
  - **Minimum, Q1, M, Q3, Maximum**
- Example: The five-number summary for our Pulse1 data is: (48, 64, 71, 80, 100)

Box plots - visual representation of a distribution (5-number summary)

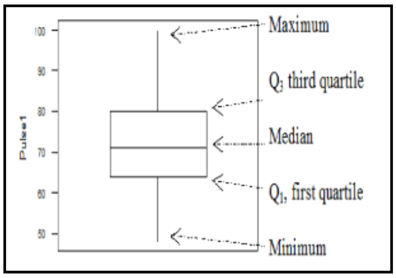
1. Label a vertical (or horizontal) axis with numbered scale from min to max
2. Draw box with lower end at Q1 and upper end at Q3
3. Draw a line through the box at the median
4. Place a dot at each of minimum and maximum
5. Check for outliers: Locate the **lower boundary** at **(Q1-1.5xIQR)** and the **upper boundary** at **(Q3+1.5xIQR)**. All data values outside these are "outliers". Mark each by an asterisk
6. Draw a line from Q1 end of box to smallest data value inside the boundary. Draw a line from Q3 end of box to largest data value inside the boundary



- A central box spans the quartiles, and hence the middle half of the observations lie in this box

Outlier check

- Pulse1 five-number summary: (48, 64, 71, 80, 100)
- Obeservations more than 1.5xIQR outside the central box are plotted individually as possible outliers
- Upper boundary is Q3+1.5xIQR
- Lower boundary is Q1-1.5xIQR
- Any outliers?

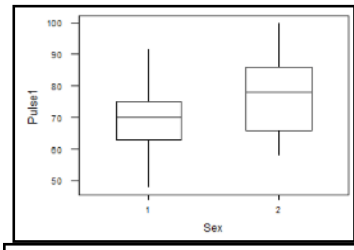


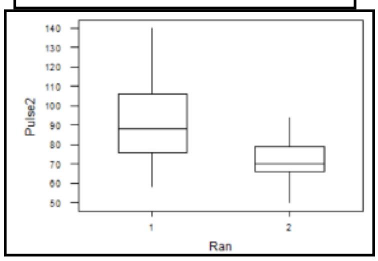
Box plots reveal:

- The centre of the distribution: the median
- The spread of the distribution: the interquartile range (IQR)… the middle 50% of data
- Symmetry or Skewness (How?)
- Outliers (How?)

Comparing Distributions

- Compare make and female resting pulses
- To compute 5-number summary
  - Sort the data according to 'Sex' (and optionally by 'Pulse1' too)
  - Use =QUARTILE(cell ranges) for the pulse1 of each sample: male(1) and female(2)
- Male five-number summary:
- Female five-number summary:
- To compare pulse rate after running with no running
  - Sort both columns by 'Ran' (and optionally by Pulse2 too) and obtain the separate five-number summaries
  - Use comparative box-plots





Outliers: what to do about them?

- Outlier = data point not consistent with the bulk of the data
  - Look for them via graphs, esp. box plots
  - Can have big influence on conclusions
  - Can cause complications in some statistical analyses
- Cannot discard without justification: not necessarily incorrect values
- Possible reasons for outliers and reasonable actions

- Mistake in measurement or data entry
- Individual in question belongs to a different group than bulk
- Outlier is legitimate data value - represents natural variability. Values may not be discarded — they provide important information about location and spread
- “Errors” should be approximately symmetric once outliers are excluded

Mean

- Another measure of **central tendency** (alternative to the median) is the **arithmetic mean**
- If n observations are denoted by x1, x2, x3,… xn, their sample mean or average is
  - x(overbar) = 1/n(x1+x2+x3+…xn)
- Note the “**overbar**” notation… **SAMPLE MEAN**

If *n* observations are denoted by $x_1, x_2, x_3, \ldots x_n$, their ***sample mean*** or ***average*** is

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + x_3 + \ldots + x_n)$$

or

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Example: Exercise data-rainfall

- Months 1-5: 31, 35, 29, 42, 33mm
- Months 6-11: 31, 35, 36, 30, 37, 35
- In excel: arithmetic mean =AVERAGE(cell range)
- BOTH samples have the same mean here (34mm)
- BUT **DIFFERENT SPREADS** or deviations from the mean: **STANDARD DEVIATION**

Variance and Standard Deviation

- The **Variance** is defined as the **mean squared-deviation.** This should mean dividing by the number in the sample. In fact we divide by **1 less** (otherwise we would tend to underestimate the true variability)
- The **sample variance** of n observations x1, x2, x3,… xn is:
- The **sample standard deviation**, s, is the **square root** of the **variance** (s^2)
  - Variance = s^2
- Has the same unit of measurement as the original observation. You should learn how to calculate x(overran), s using a scientific calculator

The **Variance** is defined as the mean squared-deviation. This should mean dividing by the number in the sample. In fact we divide by 1 less (otherwise we would tend to underestimate the true variability).

The ***sample variance*** of *n* observations $x_1, x_2, x_3, \ldots x_n$ is

$$s^2 = \frac{1}{n-1}[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2]$$

or

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

The ***sample standard deviation*** *s* is the square root of the variance s^2, and has the same unit of measurement as the original observations. You should learn how to calculate $\bar{x}, s$ using a scientific calculator.

Sample Standard Deviation, s

- For Months 1-5 calculated in lecture notes, s=5
- For Months 6-11, variance, s^2 = 8

Why is variance calculated using n-1 instead of n?

- Theoretical explanation: can show that if we:
  - Perform the experiment many times
  - Calculate sample variance each time using n instead of n-1
  - Average the variance estimates over the many experiments
  - We get approximately (n-1)/n times the true variance
- We say that the estimate is **biased by a factor of (n-1)/n**
- **Population** Variance vs **Sample** Variance

- *Population variance,* $\sigma^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$
- *Sample variance,* $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$

Robust Statistics and Linear Transformation (L2)

Robust Statistics

- Robust statistic: a statistic not much affected by outliers
- For position: **median** or mean?
- For spread: **IQR** or range or standard deviation?
  - Don’t change with outliers

Example: Travel time data

- Met travel times (mins)
- Are there any outliers?
- Calculate median, quartiles, IQR, mean, standard deviation for both travel time data sets: with and without outliers
- Which statistics are more robust? **Median, Q1, Q3, IQR**

Example: Travel time data

Met travel times (mins) (L5, p31)

Times A	29	35	30	34	30	36	29	**184**	31	34
Times B	29	35	30	34	30	36	29	**40**	31	34

- Are there any outliers?
- Calculate median, quartiles, IQR, mean, standard deviation for both travel time data sets: with and without outliers.

	median	Q1 , Q3	IQR	mean	Standard deviation
Times A	32.5	29.8, 35.2	5.4	47.2	48.1
Times B	32.5	29.8, 35.2	5.4	32.8	3.61

- Which statistics are more robust?