Lecture 6
Study design & Review

- Study Design, Types of Studies, Variables
- Bias, Sample Size Issues
- Modules 1, 2 review

Study Design
Types of Studies
Variables

Objectives

- A primary *objective* of the statistician is to obtain information about a *target population*, using a *sample*.

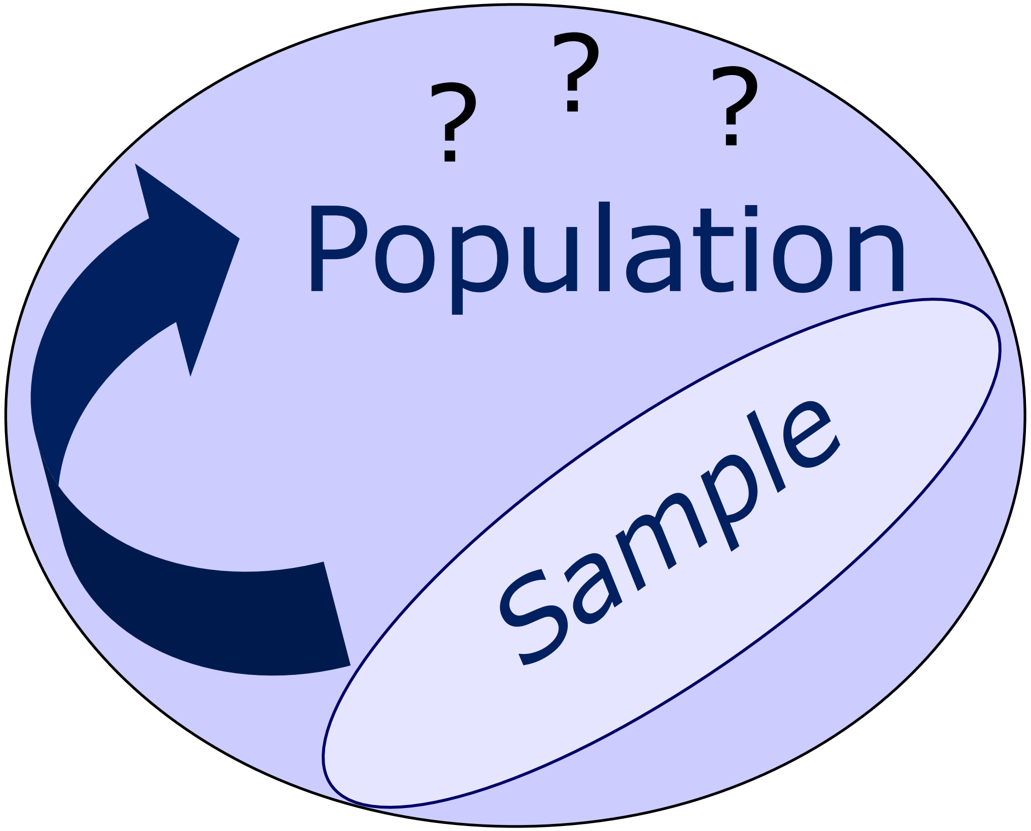
- The *target population* comprises all relevant subjects of interest.

- The *sample* is a manageable subset, selected to make the study feasible.

Answering Research Questions

Research Questions

We use a SAMPLE to answer questions about a target POPULATION



Study Design

- Formulate the question of interest
  (What? Why? Who?)
- Specify the target population
  (Who/What? Where? When?)
- Determine the measurements to be collected (the *variables*).
- Define the method of data collection
  (How? When? Where?)

Populations and Samples

- The *target population* should be well defined.
- The *sample* should be *representative* of the target population (not biased), and large enough to give accurate information about the population.
- Ideally, the *observations* should be *independent* of each other.

Selecting a sample

Only a *representative sample* should be used to make inferences about the target population. One way to ensure that a sample is representative of the target population is to obtain a *random sample*.

A *simple random sample*, of a given size n, is one in which *each set* of that size has the same chance of being selected from the target population.

A *random sample* is one where ***each member*** of the population has the ***same chance*** of being selected.

A Representative Sample

It is often difficult, or even impossible, to obtain a *simple random sample.*

However, researchers should ensure that the sample they obtain is a *representative sample*. That is, its characteristics should represent those of the target population without bias.

For example, if opinion on same sex marriage was sought from people living in Australia, our sample should have people living in Australia from different backgrounds, different age groups, different occupations etc.

Quiz 4

Indicate, for each of the following situations, whether the sample is random and/or representative.

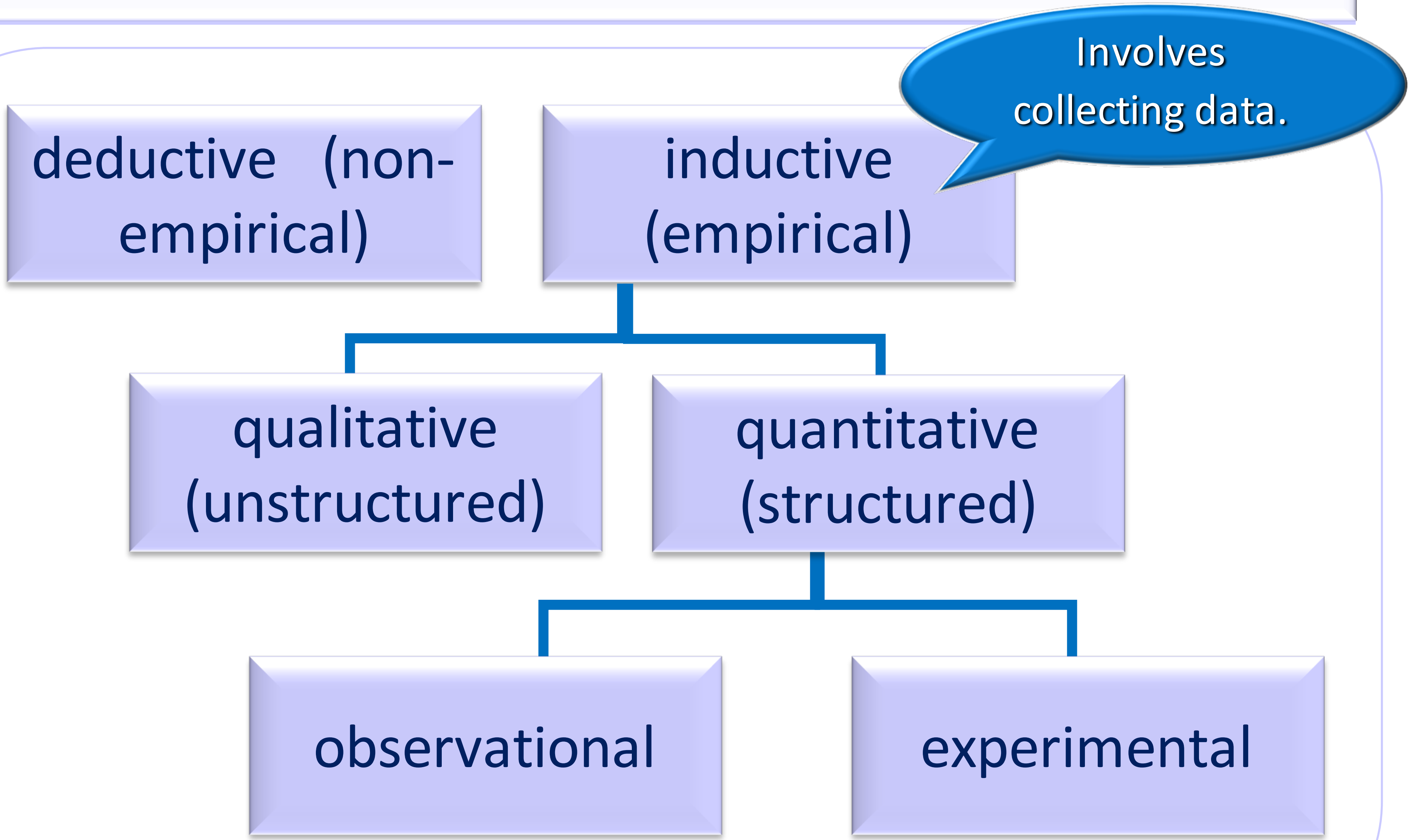
a. To investigate the association between alcohol consumption and sleeping patterns in adults, a medial centre selects 50 patients from among those claiming to be moderate drinkers. Each is asked to record information on their sleeping patterns on two separate nights – one night after drinking alcohol and a second night after drinking no alcohol.

b. A vet is considering offering a boarding service for pets while their owners are on holiday. The receptionist is asked to canvass opinions from pet owners who use the clinic. The receptionist decides to select as a sample all owners of pets vaccinated at the clinic in the following week. These owners are asked whether they would consider using the boarding service.

Solution to Quiz 4b

Random	Yes/No
Representative	Yes/No

Types of Studies



Involves collecting data.

deductive (non-empirical)

inductive (empirical)

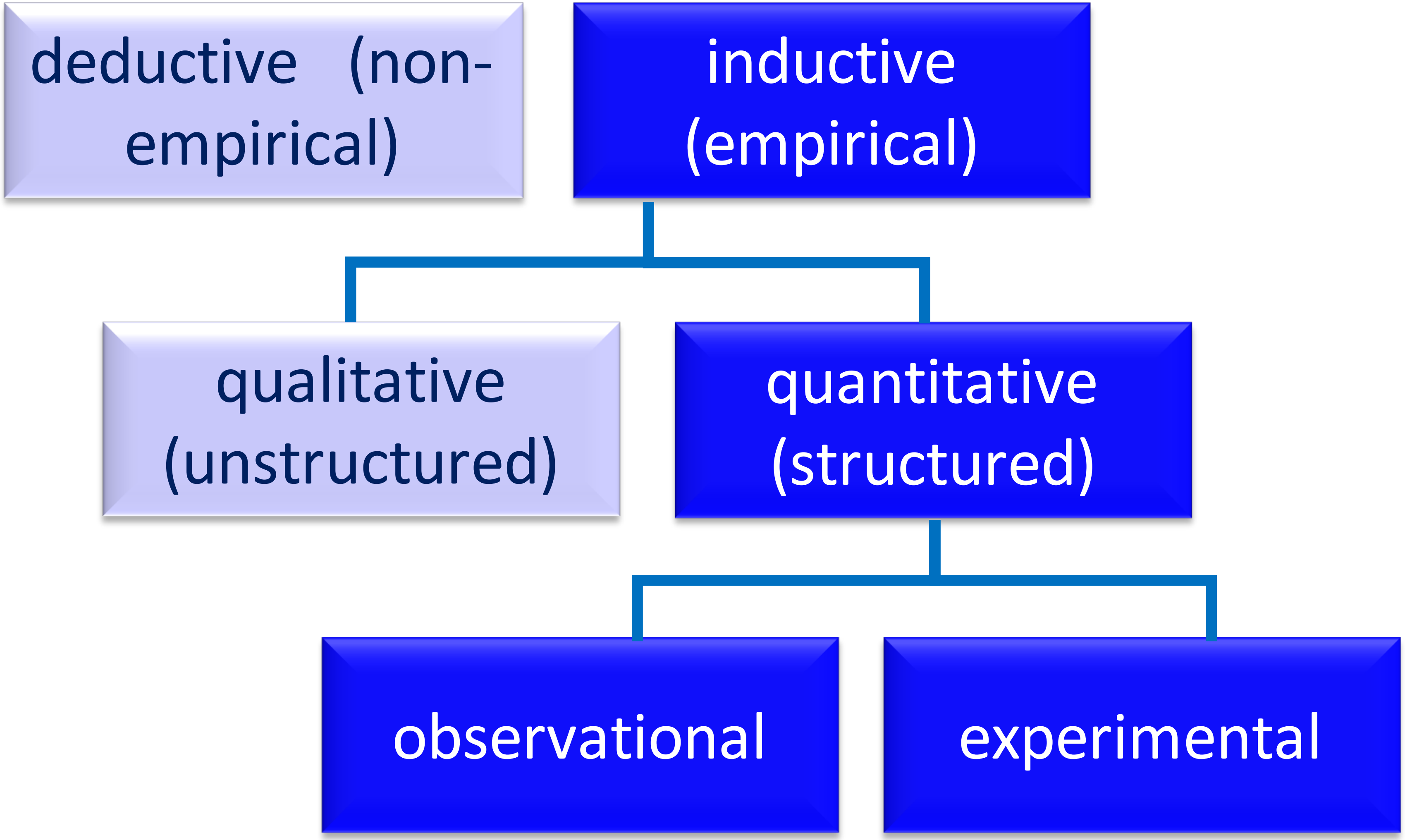
qualitative (unstructured)

quantitative (structured)

observational

experimental

Types of Studies



deductive (non-empirical)

inductive (empirical)

qualitative (unstructured)

quantitative (structured)

observational

experimental

Types of Studies

An **observational** study is one in which there is no intervention by the investigator nor is there any treatment imposed.

An **experimental** study is one in which the investigator has some control over the determinant.

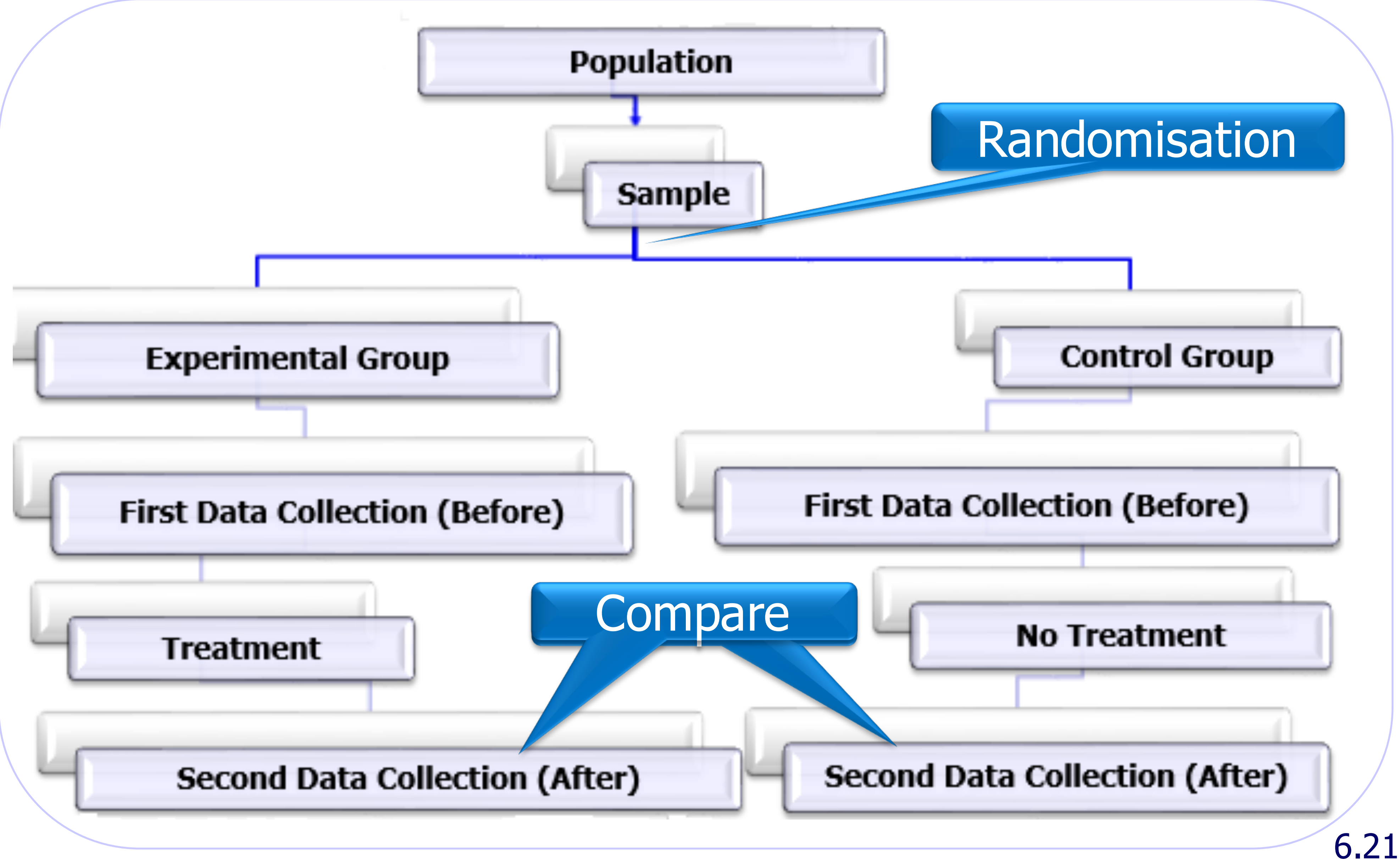
Examples of Observational Studies

- A German study (Stang et al., 2001) investigated the association between the use of mobile phones and a rare form of eye cancer called uveal melanoma. 118 subjects with the eye cancer were compared to 475 subjects who did not have the eye cancer. Subjects with eye cancer had significantly higher mobile phone usage.
- A researcher takes blood samples from students to measure blood alcohol levels during Monday morning lectures in Week 1 of semester.

Examples of Experimental Studies

- The **Physicians' Health Study** (1982) was a clinical trial conducted to determine whether low-dose aspirin reduced the risk of cardiovascular disease. 22,071 male physicians between the ages of 40 and 84 were randomly assigned into one of two groups taking either aspirin or a placebo. In follow up, heart attack rates were compared in the two groups.
- A researcher randomly assigns law students into two groups. Members of one group are all given an alcoholic drink. Each student is asked to argue on a topic and the quality of their arguments are rated. Ratings are compared for the two groups.

General Design of an Experimental Study


Population
Sample
Randomisation
Experimental Group
Control Group
First Data Collection (Before)
First Data Collection (Before)
Treatment
Compare
No Treatment
Second Data Collection (After)
Second Data Collection (After)

Quiz 5

For the following examples, indicate the type of study:

a. To investigate whether the antidepressant, Zyban, was useful for smoking cessation, 429 smokers who wished to quit were randomly assigned to one of two groups: Zyban or Control. After 12 months, the study observed whether each subject had successfully abstained from smoking or relapsed.

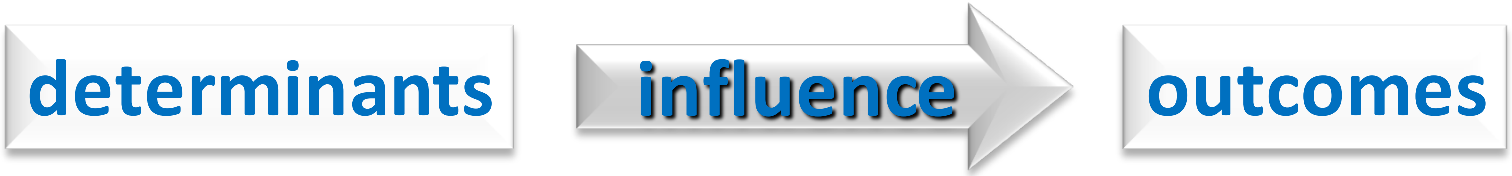
b. To investigate a link between exposure to lead and tooth decay, a study of 24901 American children showed that the greater a child's exposure to lead, the more missing or decayed teeth.

Variables

- Measurements are taken on subjects in a study according to the *variables* of interest. The measurements will vary from one subject to another.
- In any study, variables take on specific **roles** and these **roles** may be classified as:

outcomes (responses) or

determinants (may influence responses)



determinants

influence

outcomes

Examples of Variables

For example, in Quiz 4:

Part a: ***outcomes*** may be: number of hours sleep the previous night/difference between number of hours sleep with and without alcohol the previous night….

Part b: ***outcome*** may be: whether client uses a pet boarding service

For each of these studies, any other measurements recorded on the subjects such as age, sex, etc may be possible ***determinants***.

For part a. number of alcoholic drinks may be a possible ***determinant***. For part b. how many pets/how often client goes on holiday may be possible ***determinants***.

Bias

Sample size

Bias

Bias may be defined as ***any systematic error*** (ie. not occurring randomly) which results in an incorrect estimate of a parameter or an incorrect association between variables in a study. Studies can be affected by various types of bias including:

- selection bias
- measurement bias
- response bias
- confounding

Selection Bias

Selection bias refers to any systematic differences occurring ***in the way that subjects are selected*** for a study.

For example, suppose we wish to estimate the proportion of 18 to 25 year olds in Australia who have private health insurance.

Selecting a sample from a student database could produce a biased result, since the proportion of students with private health insurance may differ from the proportion of other young adults with private health insurance.

Measurement Bias

Measurement bias refers to ***systematic differences in the measurement of variables***. For example:

- in a comparison of influenza rates among people with and without chronic illnesses, the responses for people with chronic illnesses may be more accurate as their past illnesses may be better documented and/or recalled.
- people collecting information from subjects may do so more carefully in the morning than in the afternoon.

Response Bias

Response bias can occur when the ***response rate to a survey is too low***.

It is well known that those who respond to a survey often have different characteristics than those who don't respond.

Ideally, the response rate should be at least 75% to ensure that a study is not significantly affected by response bias.

Confounding

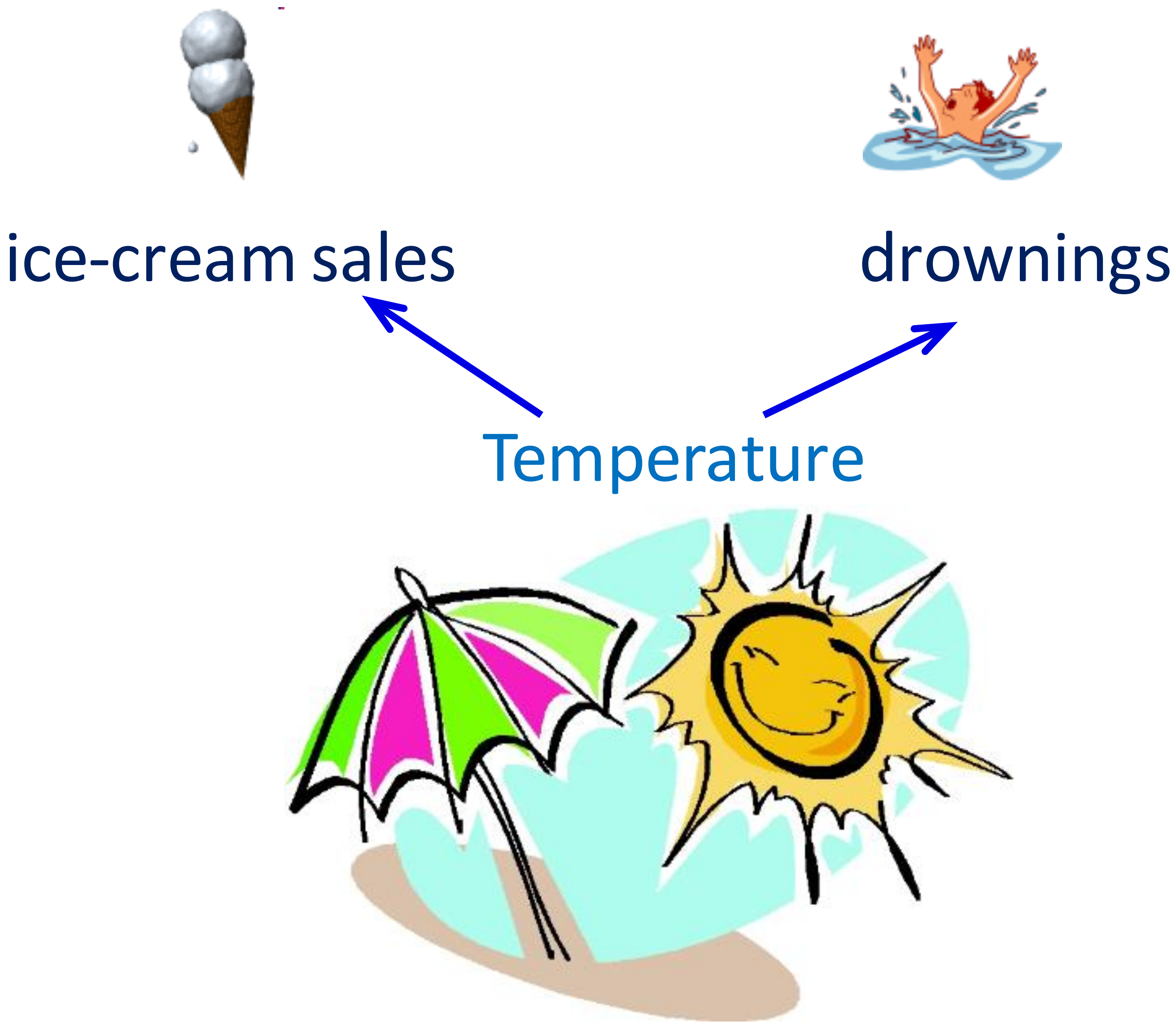
A confounder is a variable that ***distorts*** (increases or decreases) the apparent effect of one variable (determinant) on another (outcome).

For example:

It has been suggested that watching more than four hours of TV per day is associated with an increased risk of heart disease.

However, it is likely that those who watch a lot of TV do ***not exercise much and it is the lack of exercise which leads to the increased risk of heart disease***.

A Confounder



ice-cream sales

drownings

Temperature

Quiz 6

a. The Centre for Addiction and Substance Abuse sent a mail out questionnaire to 5000 randomly selected teenagers and elicited 1297 responses. 130 reported that parents are never present at parties they attend and that marijuana is available at the parties they attend. *Is it reasonable to conclude that approximately 10% of teenagers attend parties where parents are not present and where marijuana is available?*

b. In 2013 a digital magazine company, CatalogSpree, conducted an online survey and claimed, as a result, that consumers use tablets to shop for 66% of all Christmas presents. The survey was conducted using the CatalogSpree app which can be installed on tablets. *Can we conclude that two thirds of Christmas presents in 2013 were bought online?*

✓ Solution to Quiz 6

Sample Size

A sample needs to be *sufficiently large* to give an *accurate* representation of the target population.

The *accuracy* of a sample for determining a population characteristic depends on two factors:

- ○ the ***sample size*** (n) used for the study
- ○ the ***variability*** (spread) of the measurements

The sample size needed depends on the kind of data which are of interest.

Sample Size Requirements

Sample size for determining a **proportion**:

- Most opinion polls are based on surveys of **at least 500 persons**. We need this number to ensure a reasonable degree of accuracy.

Sample size for determining a **mean**:

- Smaller samples are often sufficient for estimating characteristics of populations of ***numerical*** (measured) data.

Sample Size

The sample needs to be large enough:

However, a small sample may be sufficient when the population is homogenous (ie. the population does not vary much).

Quiz 7

A study was undertaken to investigate the link between red wine consumption and heart disease. Research had indicated that people living in France, particularly in the Burgundy area, experience a lower rate of heart disease than people living in Australia.

Researchers wish to estimate the average number of glasses of red wine consumed per week by people living in France, which has a population of more than 60 million, and people in Australia, which has a population of more than 20 million. Researchers have decided to sample 400 subjects for the study.

*Assuming the variability in the number of glasses of red wine consumed per week is the same in France and Australia, how many subjects should be sampled from each country such that **each sample provides the same degree of accuracy**?*

Solution to Quiz 7

Number sampled from each country?

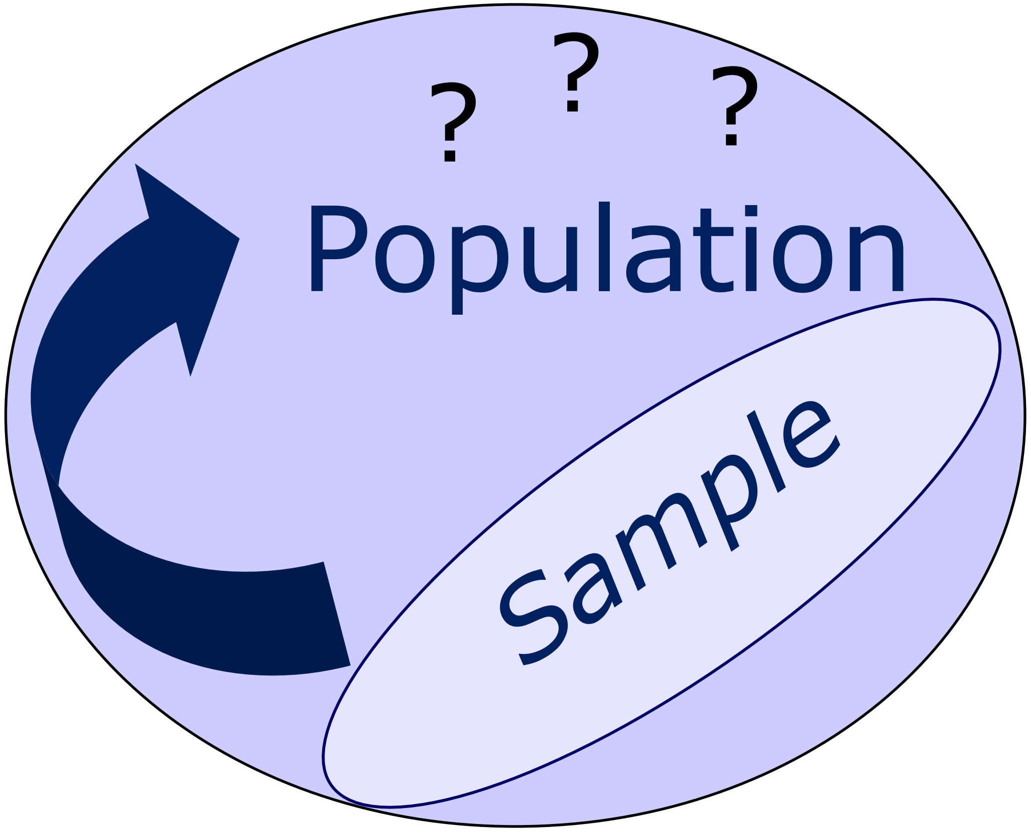
from Australia:	
from France:	
total:	400

Modules 1, 2 review

Answering Research Questions

Research Questions

We use a SAMPLE to answer questions about a target POPULATION



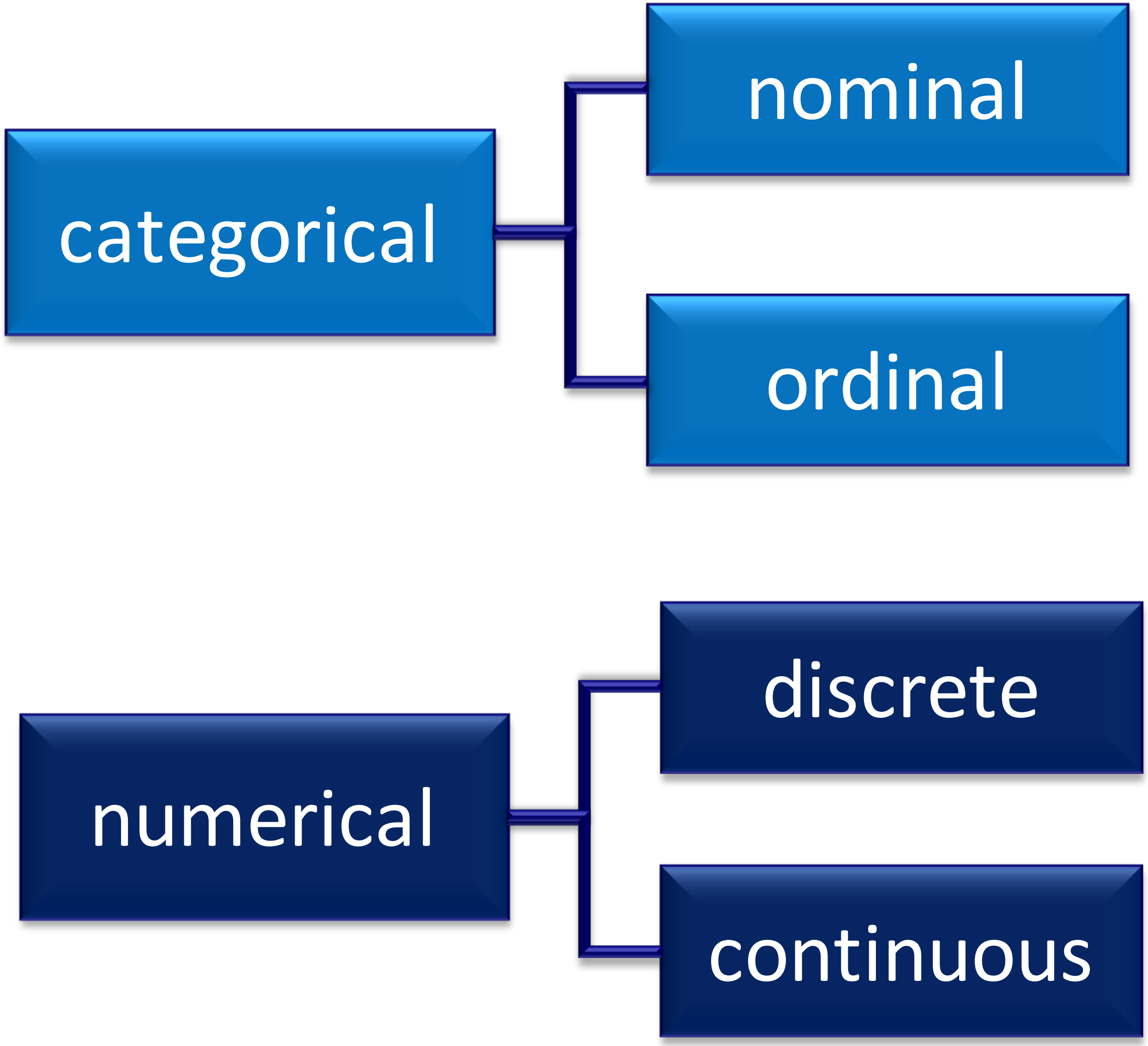
Notation

Sample Statistics		estimate →	Population Parameters
Mean	$\bar{y}$	→	μ
Median	$\tilde{y}$	→	$\tilde{\mu}$
Std.Dev	s	→	σ
Variance	s^2	→	σ^2
se(mean)	$s/\sqrt{n}$	→	$\sigma/\sqrt{n}$
Proportion	p	→	π
se(proportion)	$\sqrt{\frac{p(1-p)}{n}}$	→	$\sqrt{\frac{\pi(1-\pi)}{n}}$
Least Squares Regression Line	$\hat{y} = a + bx$	→	$\mathrm{y} = \alpha + \beta x$
Slope	b	→	β

Sampling

- Only a *representative sample* should be used to make inferences about the target population. One way to ensure that a sample is representative of the target population is to obtain a *random sample*.
- A *simple random sample*, of a given size n, is one in which *each set* of that size has the same chance of being selected from the target population.
- A *random sample* is one where ***each member*** of the population has the ***same chance*** of being selected.

Data Classification



categorical

nominal

ordinal

numerical

discrete

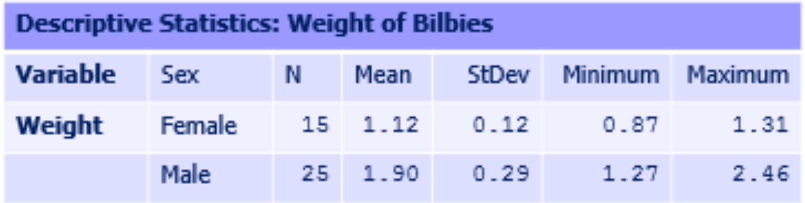
continuous

Summarising Data Graphically

Variable	categorical	numerical	
categorical	clustered bar charts	comparative box plots	bar charts or pie charts
numerical	comparative box plots	scatter plots	histograms
	bar charts or pie charts	histograms	

Summarising Data Numerically

Numerical Variables:



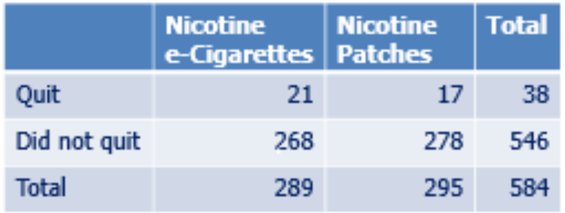
Descriptive Statistics: Weight of Bilbies

Variable	Sex	N	Mean	StDev	Minimum	Maximum
Weight	Female	15	1.12	0.12	0.87	1.31
	Male	25	1.90	0.29	1.27	2.46

Measures of Centre: mean/median

Measures of Spread: range/interquartile range/ standard deviation

Categorical Variables:

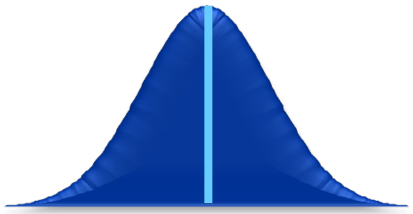


	Nicotine e-Cigarettes	Nicotine Patches	Total
Quit	21	17	38
Did not quit	268	278	546
Total	289	295	584

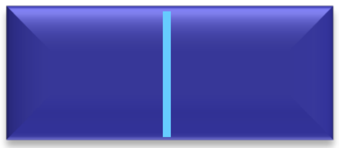
Tables showing counts/proportions/percentages

Population Distributions

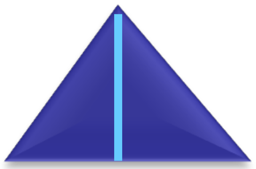
Symmetric



Normal



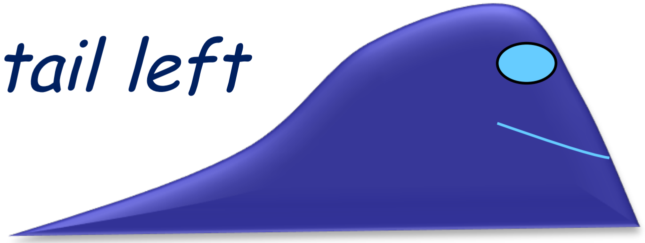
Uniform



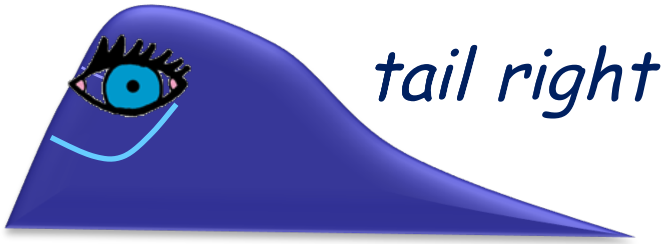
Triangular

Skewed

tail left



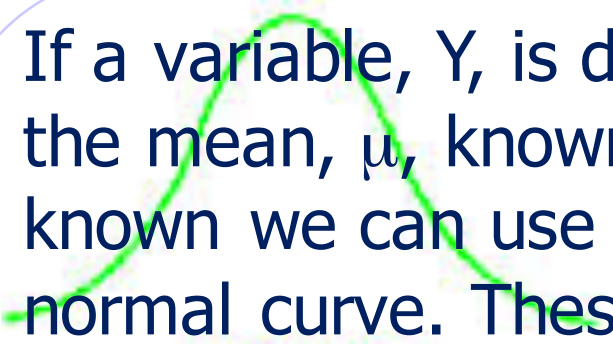
Skewed Left



tail right

Skewed Right

The normal (z) distribution

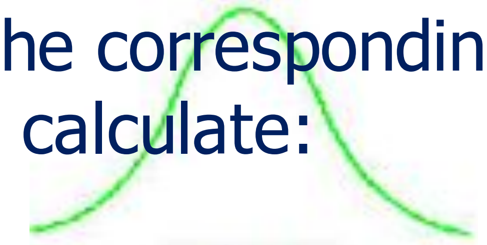


If a variable, Y, is drawn from a normal distribution with the mean, μ, known and the standard deviation, σ, known we can use the normal table to find areas under a normal curve. These areas represent probabilities.

- To find the probability corresponding to a value y, from a normal distribution, calculate its z-score: $z = \frac{y-\mu}{\sigma}$, and look up the corresponding area in the normal (z) table.

- To find a percentile (a y value corresponding to an area under a normal curve), find the corresponding z-score from the normal table and calculate:

$$y = \mu + z \times \sigma$$



The Central Limit Theorem

The Central Limit Theorem for sample means: in repeated sampling, sample means will follow an approximately normal distribution if the sample size, n, is 'large'. This approximation improves as the sample size increases.

- The standard deviation of sample means (standard error) = $\sigma/\sqrt{n}$
- The z-score for finding a probability for a sample mean, z = $\frac{\bar{y}-\mu}{\sigma/\sqrt{n}}$

The Central Limit Theorem also applies to sample proportions: sample proportions will follow an approximately normal distribution in repeated sampling if the sample size, n is 'large' ie. if both nπ and n(1–π) are ≥ 5.

- The standard deviation of sample proportions (standard error) = $\sqrt{\frac{\pi(1-\pi)}{n}}$
- The z-score for finding a probability for a sample proportion, z = $\frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$

Confidence Intervals

Confidence Intervals use sample statistics for estimating population parameters with a given level of confidence.

- ○ 95% Confidence Interval for μ (σ known) = $\bar{y} \pm z_{crit} \times {}^{\sigma}/_{\sqrt{n}}$
- ○ 95% Confidence Interval for μ (σ unknown) = $\bar{y} \pm t_{crit} \times {}^{s}/_{\sqrt{n}}$ *(df = n - 1)*

For 95% intervals: z_{crit} = 1.96, t_{crit} depends on the sample size

- ○ 95% Confidence Interval for π = $p \pm z_{crit} \times \sqrt{\frac{p(1-p)}{n}}$
- ○ 95% Confidence Interval for μ_d = $\bar{y}_d \pm t_{crit} \times {}^{s_d}/_{\sqrt{n_d}}$ *(df = n_d - 1)*
- ○ 95% Confidence Interval for μ_1 - μ_2 = $(\bar{y}_1 - \bar{y}_2) \pm t_{crit} \times s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

(df = n_1 + n_2 - 2)

- ○ 95% Confidence Interval for β = $\mathrm{b} \pm t_{crit} \times se(b)$ *(df = n - 2)*

A Study

A medical trial involved recording the blood pressure and heart rates of a sample of adults who were aged between 18 and 45 years. One of the reasons for the trial was to study how measurements of blood pressure taken in doctors' clinics were related to measurements taken by a device worn by the patient at home, so that future observations could be taken from home measurement. The following information was recorded for the sample of patients at the start of the trial (baseline measurements).

Some of the variables recorded are listed below.

Variable	**Description**
Sex	Sex (Male, Female)
Smoke	Number of Cigarettes smoked per day (None, 1-5, 6-10, 11-20)
BMI	Body Mass Index (100 * weight (kg) / height (m)^2)
Exercise	Amount of usual exercise (None, Light, Moderate)
Age	Age (years)
SBP_Clinic	baseline Systolic Blood Pressure measured in clinic (mm Hg)
DBP_Clinic	baseline Diastolic Blood Pressure measured in clinic (mm Hg)
HR_Clinic	baseline Heart Rate measured in clinic (beats per minute)
SBP_Home	baseline Systolic Blood Pressure measured at home (mm Hg)
DBP_Home	baseline Diastolic Blood Pressure measured at home (mm Hg)
HR_Home	baseline Heart Rate measured at home (beats per minute)

Source OzDasl and Schork, M. A., and Remington, R. D. (2000). Statistics with Applications to the Biological and Health Sciences 3rd Edition. Prentice-Hall (adapted)

Question 1

Question 1 (7 marks)

a. If possible, list one variable which is:

Numerical

Ordinal

Nominal

b. Would it be possible for the variable 'Age' to be recorded as an ordinal variable? How would the observations be recorded in this case?

Question 1

c. If all the information had been recorded from patients of one doctor, describe an appropriate target population for this trial

d. What would be an appropriate graph to use to investigate whether there is an association between the amount of usual exercise (none/light/medium) and the number of cigarettes smoked per day (none/1-5/6-10/11-20) among adults?

e. What would be an appropriate graph to use to investigate whether there is a difference between average systolic blood pressure for males and females?

Information for Questions 2 and 3

Information for Question 2 and Question 3

The following table summarises the data from a sample of 200 patients in this trial for the variables Smoke and Sex. Use the data in this table in answering Questions 2 and 3.

```
Tabulated statistics: Sex, Smoke
Rows: Sex    Columns: Smoke
           None    1-5    6-10    11-20    All
Female       47      5       0        2     54
Male        111     10       9       16    146
All         158     15       9       18    200
Cell Contents:        Count
```

Question 2

a. What proportion of females in the sample do not smoke?

b. Calculate a 95% confidence interval to estimate the true proportion of females in the target population who do not smoke.

Question 3

Recall from lectures that adult IQ scores are known to be normally distributed, with a mean of 100 and a standard deviation of 15.

a. What is the probability that a random sample of 25 adults would have an average IQ of between 105 and 110?

Question 3

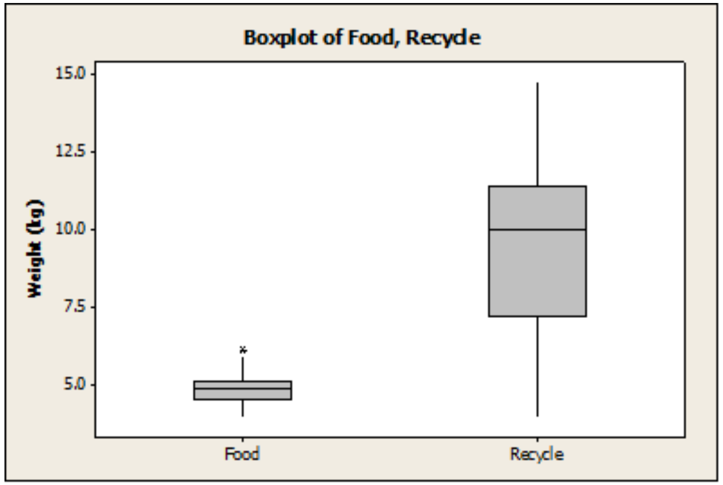
b. A research group wants to recruit adults of moderately high intelligence (defined as IQ greater than 135) for a study. They can select their sample from 1000 adults who have agreed to participate. How many adults would they expect to meet their criteria?

Question 4

a. Recent research has suggested that the true proportion of households in a particular area which have soil with a high level of lead contamination is 0.20. If we take a random sample of 50 households, what is the probability that between 8 and 15 of them will have soil with a high level of lead contamination.

Question 4

b. The boxplot below displays the weights of food waste and recycling materials in a garbage collection for a sample of households. Indicate whether each of the statements below is likely to be true or false.



Statement	True or False?
The weight of food waste is typically higher than the weight of recycling materials	
The weight of food waste is less variable than the weight of recycling materials	
The weight of recycling materials is typically higher than the weight of food waste	
One household has an unusually low weight of food waste	

Homework problems

(Only for new topics)

Homework Question 1

For the two following studies identify:

i. when the study was conducted

ii. the target population

iii. the variables recorded and indicate the type of each variable

a. A study was undertaken to investigate the link between TV violence and aggressive behaviour (Johnson et al., *Science, 2002*). Researchers randomly sampled 707 teenagers from New York State. Of the 88 who watched less than one hour of TV per day, 5 were reported to have later committed an aggressive act whilst of the 619 who watched at least one hour of TV per day, 154 were reported to have committed an aggressive act.

Homework Question 1 continued

b. An experiment was conducted to investigate the link between mobile phone usage and drivers’ reaction times (Strayer et al., *Psych Science, 2001*). 64 university students were randomly assigned to one of two groups, a treatment group and a control group, both of which used a machine to simulate driving situations. The treatment group carried out a conversation on a mobile phone throughout the simulation. Participants were instructed to press a brake button as soon as a red light flashed. The mean response time was compared for the two groups.

Solution to Homework Question 1

a. Teenagers

   i.

   ii.

   iii.

b. Drivers

   i.

   ii.

   iii.

Lecture 6 Summary

- *Studies* are needed to resolve questions of interest.
- Studies are *deductive/inductive;*
  - inductive studies are *qualitative/quantitative;*
  - quantitative studies are *observational/experimental;*
    - observational studies are those where there is no intervention by the investigator, nor is any treatment imposed
    - experimental studies are those in which the investigator has some control over the determinant

Lecture 6 Summary

- *Statistics* involves determining *population* characteristics, using data from *samples*.
- Samples should be *unbiased* (they should represent the target population).
- A *random sample* is one where *each member* of the population has the same chance of being selected.
- The *accuracy* of an estimate depends on the *sample size* and the *population variability*, but not on the population size.

Textbook References

Further information on the topics discussed in this lecture can be found in:

Modern Statistics: An Introduction
by Don McNeil and Jenny Middledorp
(ISBN 9781486007011).

- Chapter 1: Pages 2 to 25 (bits)