STAT3914 APPLIED STATISTICS LECTURE NOTES

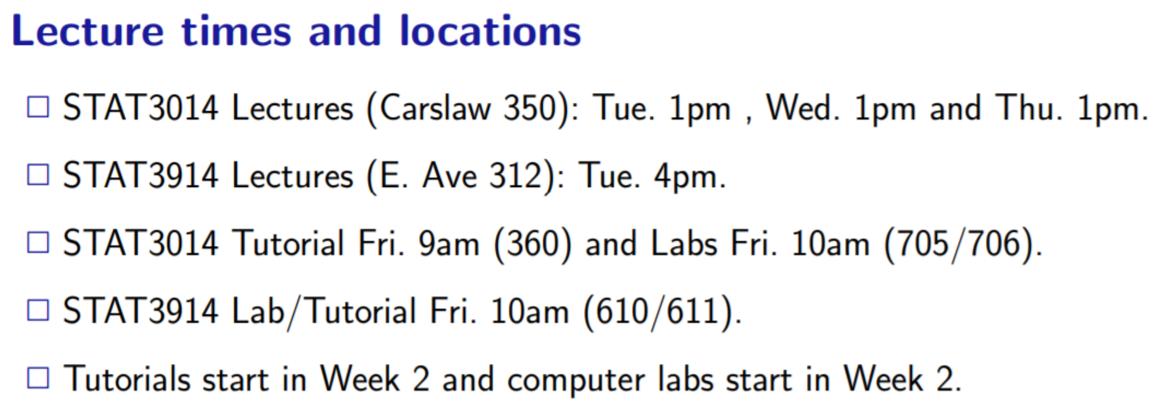
Contents

STAT3014 APPLIED STATISTICS Advanced

- This course Much more working with data
- Dr John Ormerod

Lecture 1.

Introduction



Lecture times and locations

□ STAT3014 Lectures (Carslaw 350): Tue. 1pm , Wed. 1pm and Thu. 1pm.

□ STAT3914 Lectures (E. Ave 312): Tue. 4pm.

□ STAT3014 Tutorial Fri. 9am (360) and Labs Fri. 10am (705/706).

□ STAT3914 Lab/Tutorial Fri. 10am (610/611).

□ Tutorials start in Week 2 and computer labs start in Week 2.

Content:

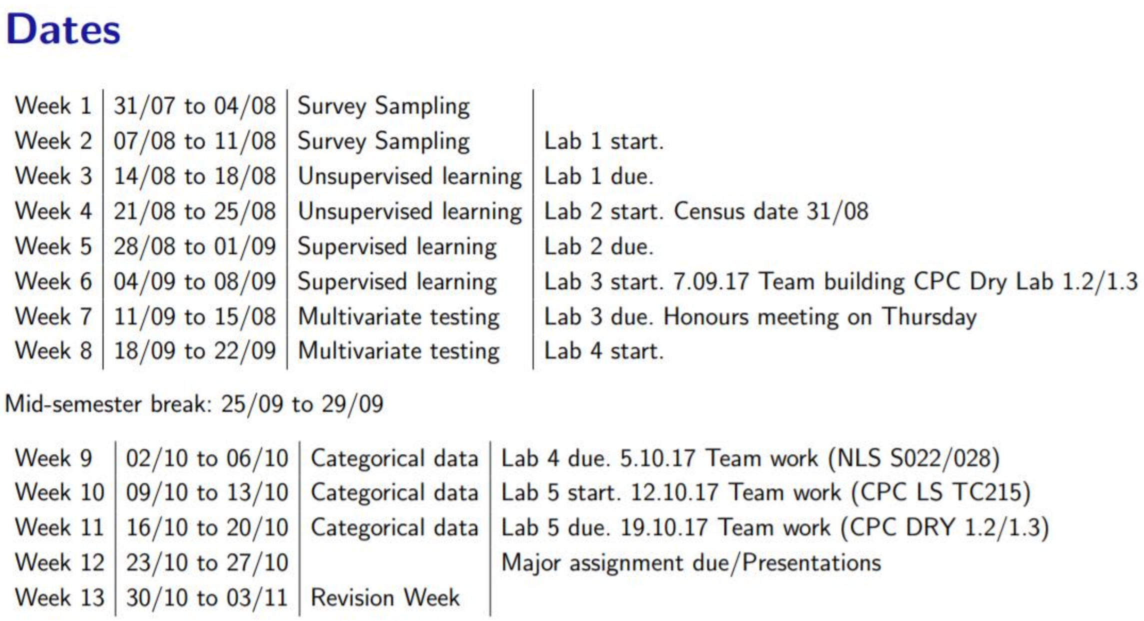
5 parts:

- Survey sampling (week 1-2)
- Unsupervised learning (week 3-4)
- Supervised learning (week 5-6)
- Multivariate testing (week 7-8)
- Categorical data analysis (week 9-11)

- Major presentations week 12
- Review week 13

Assessment:

- 5 labs on each of the 5 parts (20%)
  - 2 weeks for each lab; due on the Wednesday after the second Friday. Mini report written on each data set.
- Collaborative project with NUTM3001 (nutrition) students (weeks 6-12)
  - Major report (due week 12) (30%)
  - Presentations (to be given during week 12)
- STAT3914 has an additional assignment (5%) and quiz (5%)
- Final exam (40%); (40% in exam needed to pass course)



Dates

Week 1	31/07 to 04/08	Survey Sampling	
Week 2	07/08 to 11/08	Survey Sampling	Lab 1 start.
Week 3	14/08 to 18/08	Unsupervised learning	Lab 1 due.
Week 4	21/08 to 25/08	Unsupervised learning	Lab 2 start. Census date 31/08
Week 5	28/08 to 01/09	Supervised learning	Lab 2 due.
Week 6	04/09 to 08/09	Supervised learning	Lab 3 start. 7.09.17 Team building CPC Dry Lab 1.2/1.3
Week 7	11/09 to 15/08	Multivariate testing	Lab 3 due. Honours meeting on Thursday
Week 8	18/09 to 22/09	Multivariate testing	Lab 4 start.

Mid-semester break: 25/09 to 29/09

Week 9	02/10 to 06/10	Categorical data	Lab 4 due. 5.10.17 Team work (NLS S022/028)
Week 10	09/10 to 13/10	Categorical data	Lab 5 start. 12.10.17 Team work (CPC LS TC215)
Week 11	16/10 to 20/10	Categorical data	Lab 5 due. 19.10.17 Team work (CPC DRY 1.2/1.3)
Week 12	23/10 to 27/10		Major assignment due/Presentations
Week 13	30/10 to 03/11	Revision Week	

Why applied statistics?

- Data science

Data Science has recently captured the imagination of mainstream media for its impact on a wide range of academic disciplines and business endeavors. Data Science is an emerging interdisciplinary field whose aim is to solve real word problems in areas including (but not limited to) Information Technology, the Biological sciences, Medicine, Psychology, Neuroscience the Social Sciences and the humanities, Business and Finance.

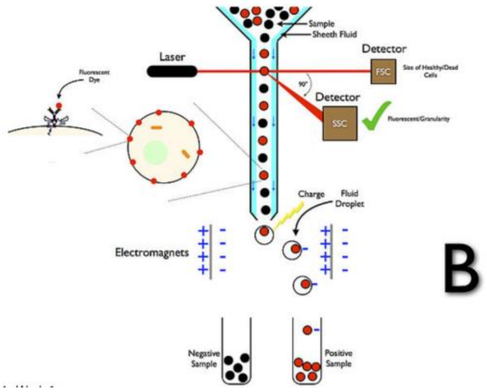
Data science success stories include Internet companies such as Google, eBay and Amazon.

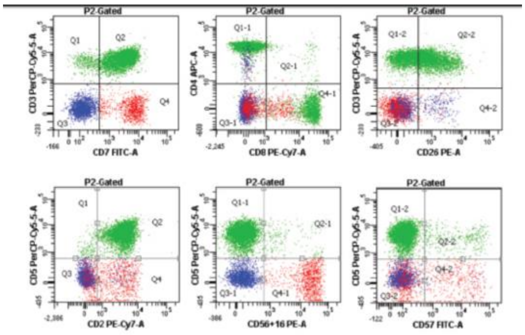
- Eg: politics; historical texts; medicine (IBM Watson)

Motivating data:

Flow cytometry:

- In biotechnology, flow cytometry is a laser-based, biotechnology for cell counting, cell sorting, biomarker detection and protein engineering.
- Works by suspending cells in a stream of fluid and passing them by an electronic detection apparatus.
- It allows simultaneous multivariate analysis of the physical and chemical characteristics of up to thousands of particles per second.
- Flow cytometry has been used in the diagnosis of health disorders, especially blood cancers.

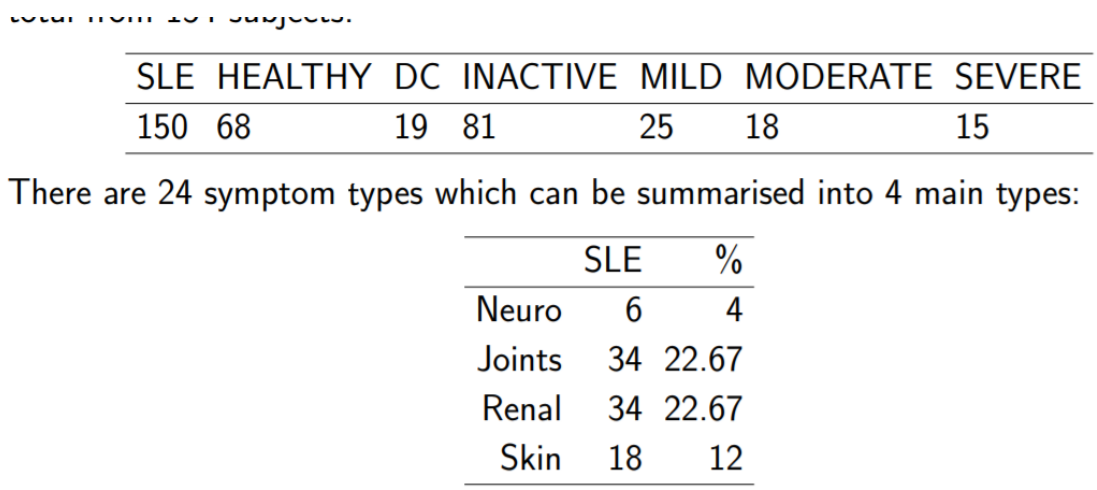




✷ Lupus (erythematosus) is the name given to a collection of autoimmune diseases in which the human immune system becomes hyperactive and attacks normal healthy tissues. ✷ Symptoms of these diseases can affect many different body systems, including joints, skin, kidneys, blood cells, heart, and lungs. ✷ Can even result in psychosis in rare cases

The data:

Loretta Lee (a PhD student over a period of 5 year) collected n = 236 samples in total from 154 subjects



total from 154 subjects.

SLE	HEALTHY	DC	INACTIVE	MILD	MODERATE	SEVERE
150	68	19	81	25	18	15

There are 24 symptom types which can be summarised into 4 main types:

	SLE	%
Neuro	6	4
Joints	34	22.67
Renal	34	22.67
Skin	18	12

Predictors:

- Treatment information (4 main types).
- Some results from blood tests taken from the hospitals the patients were taken from (14 variables).

- (Very) basic demographic variables (4 variables).
- Flow cytometry variables (77 variables).

Data cleaning:

- Measurement error
- Outliers and high leverage points.
- Missing values.
- Text in numeric fields.
- Covariate censoring.
- Highly non-normal covariates.

My former PhD student Dr. Chong You having found a job in a bioinformatics

research lab recently told me:

"Data in my UG degree was so clean compared to what I deal with now!"

Tidyverse lectures

Lecture 2. Wednesday, 2 August 2017

- Assume we have data
- It can answer our questions
- It is clean
- We interrogated the right aspects
- We communicate correctly

Data cleaning:

- Correupted column names
- Missing columns/rows
- Rows with missing values
- Rows with random values

Reading data

Base R functions are not sufficient for modern uses

- Readr functions superior in data import warning, column type handling, speed, scalability and consistitncy

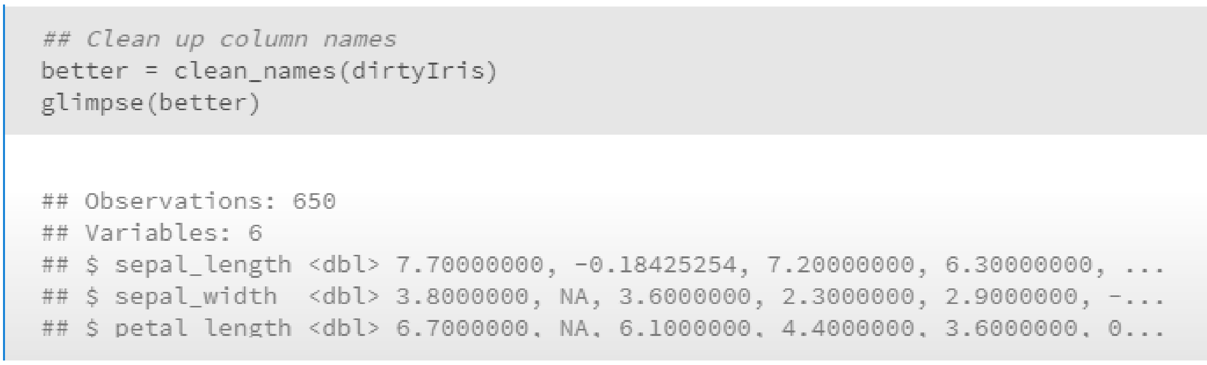
Cleaning data:

- Cleaning data allows us to do statistical modelling without extra processing
  - Good documentation
  - Each column is a variable

        - No bad characters
        - Inconsistent capitalisation or separators
    - Each row is an observation
        - No bad characters
        - No poorly esiregned row names
        - No repeated row names
- Clean data is well designed data.frame
    - Data cleaning using janitor

Janitor

- Clean up bad column names (clean_names)

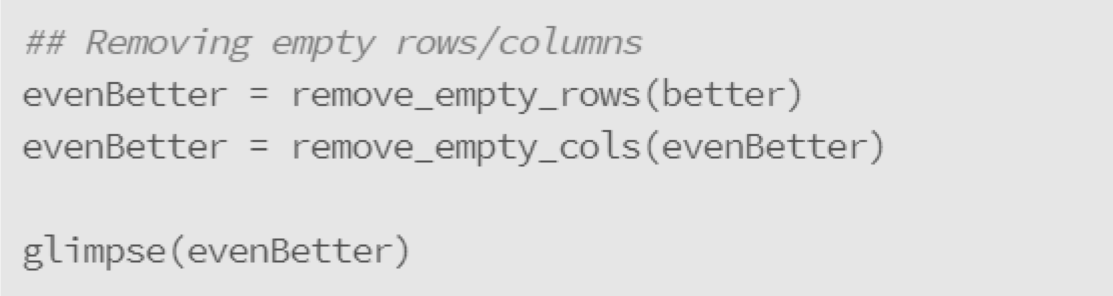


```
## Clean up column names
better = clean_names(dirtyIris)
glimpse(better)
```

```
## Observations: 650
## Variables: 6
## $ sepal_length <dbl> 7.70000000, -0.18425254, 7.20000000, 6.30000000, ...
## $ sepal_width  <dbl> 3.8000000, NA, 3.6000000, 2.3000000, 2.9000000, -...
## $ petal_length <dbl> 6.7000000, NA, 6.1000000, 4.4000000, 3.6000000, 0...
```

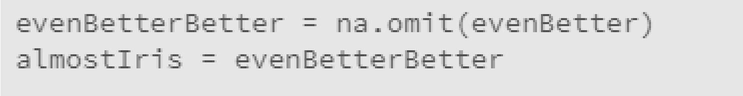
-

- Remove empty rows/cols

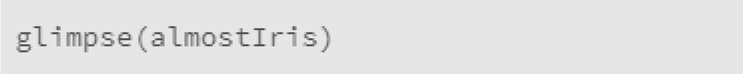


```
## Removing empty rows/columns
evenBetter = remove_empty_rows(better)
evenBetter = remove_empty_cols(evenBetter)

glimpse(evenBetter)
```

    - Genuine missing values should be mainted, in which case NA is added. (na.omit)



```
evenBetterBetter = na.omit(evenBetter)
almostIris = evenBetterBetter
```



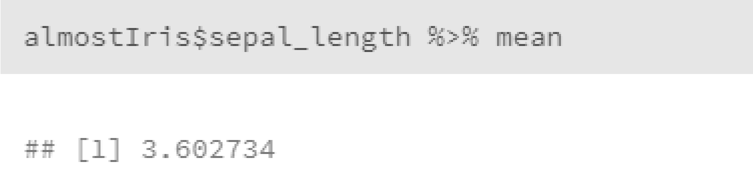
```
glimpse(almostIris)
```

-

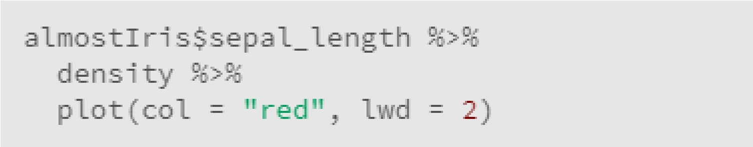
Cleaning Code

- ‘inside out’ code structure doesn’t read well

“x%>%f” means “f(x)” (called ‘x pipe f’)



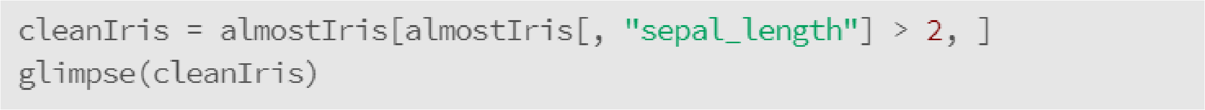
```
almostIris$sepal_length %>% mean

## [1] 3.602734
```



```
almostIris$sepal_length %>%
  density %>%
  plot(col = "red", lwd = 2)
```

Subsetting data in R

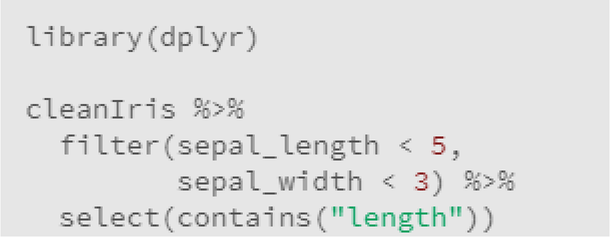
- If I want to remove obsercations of sepal_length <2



```
cleanIris = almostIris[almostIris[, "sepal_length"] > 2, ]
glimpse(cleanIris)
```

Using dplyr

- Rows and columns two separate different procedures
  - Select columns are operations on variables
  - Filter rows are operations on observations



```
library(dplyr)

cleanIris %>%
  filter(sepal_length < 5,
         sepal_width < 3) %>%
  select(contains("length"))
```

```
arrangeCleanIris = cleanIris %>%
  arrange(sepal_length, sepal_width, petal_length, petal_width)

## The true iris data
arrangeIris = iris %>%
  clean_names() %>%
  arrange(sepal_length, sepal_width, petal_length, petal_width)
```

- We sorted both the processed dirtyIris data and the arranged iris data

```
## The `Species` column is character or factor
all.equal(arrangeCleanIris, arrangeIris)
```

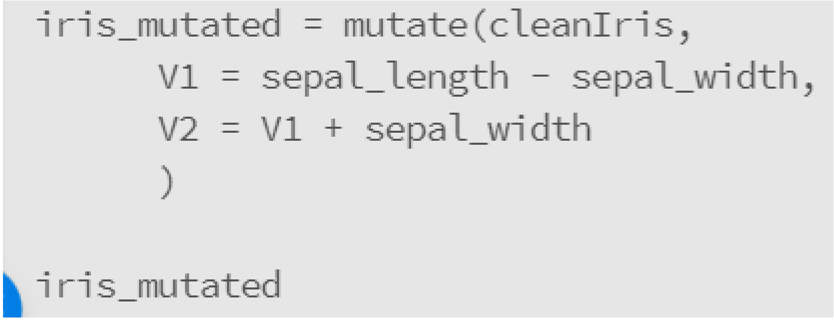
```
## [1] "Incompatible type for column `species`: x character, y factor"
```

```
arrangeIris = arrangeIris %>%
  mutate(species = as.character(species))

## Great!
all.equal(arrangeCleanIris, arrangeIris)
```

```
## [1] TRUE
```

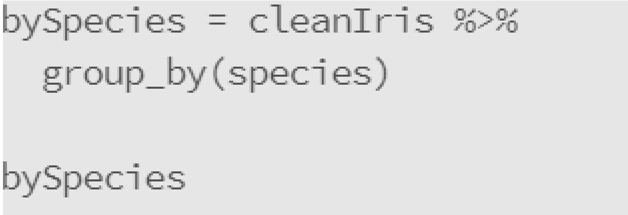
Modelling using dplyr

mutate to create new columns



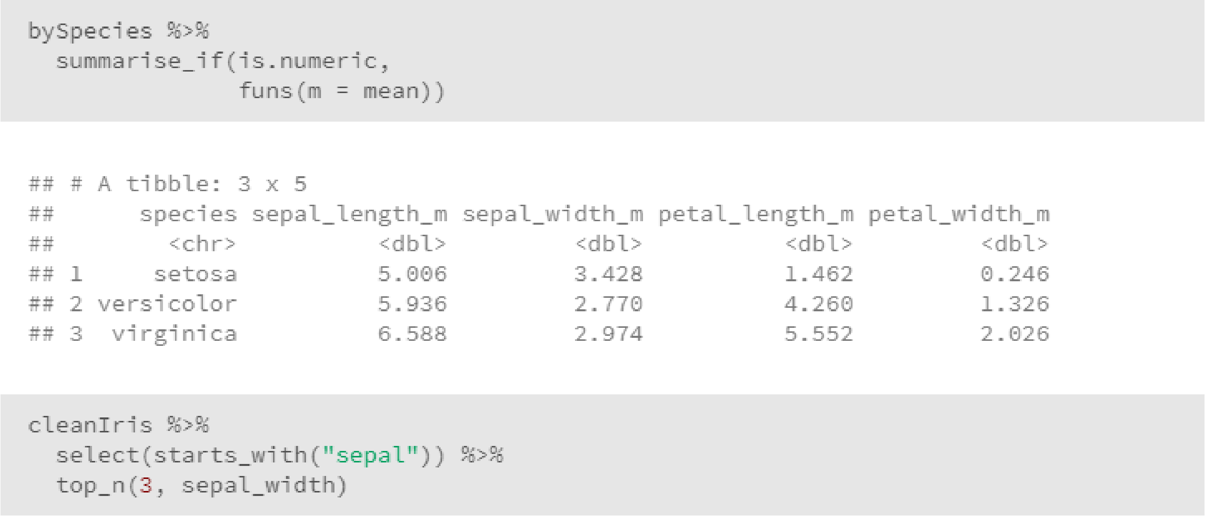
```
iris_mutated = mutate(cleanIris,
       V1 = sepal_length - sepal_width,
       V2 = V1 + sepal_width
       )

iris_mutated
```

- Group_by and summarise create summary statistics for grouped variables



```
bySpecies = cleanIris %>%
  group_by(species)

bySpecies
```

- `select` only if a column satisfy a certain condition

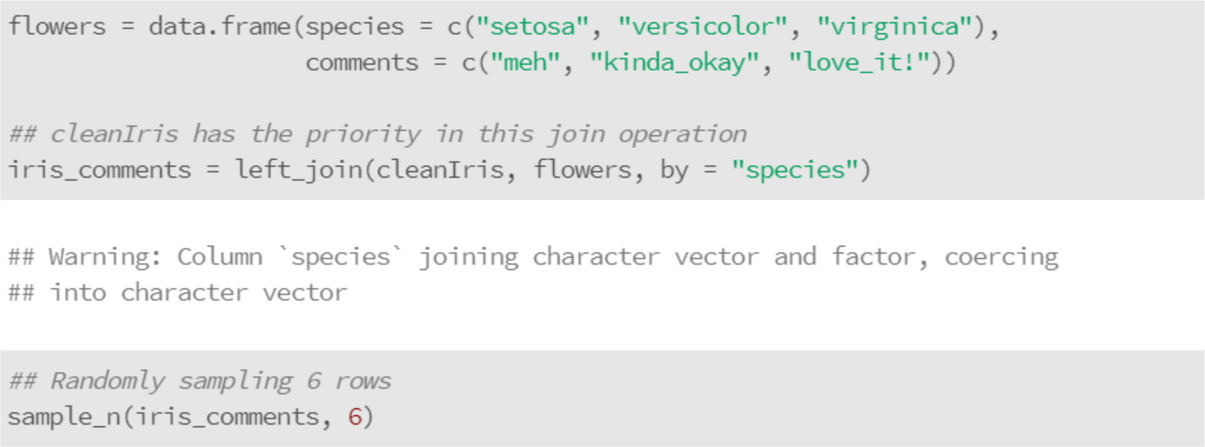


```
bySpecies %>%
  summarise_if(is.numeric,
               funs(m = mean))
```

```
## # A tibble: 3 x 5
##      species sepal_length_m sepal_width_m petal_length_m petal_width_m
##       <chr>           <dbl>         <dbl>          <dbl>         <dbl>
## 1     setosa          5.006         3.428          1.462         0.246
## 2 versicolor          5.936         2.770          4.260         1.326
## 3  virginica          6.588         2.974          5.552         2.026
```

```
cleanIris %>%
  select(starts_with("sepal")) %>%
  top_n(3, sepal_width)
```

- Left_join to merge data

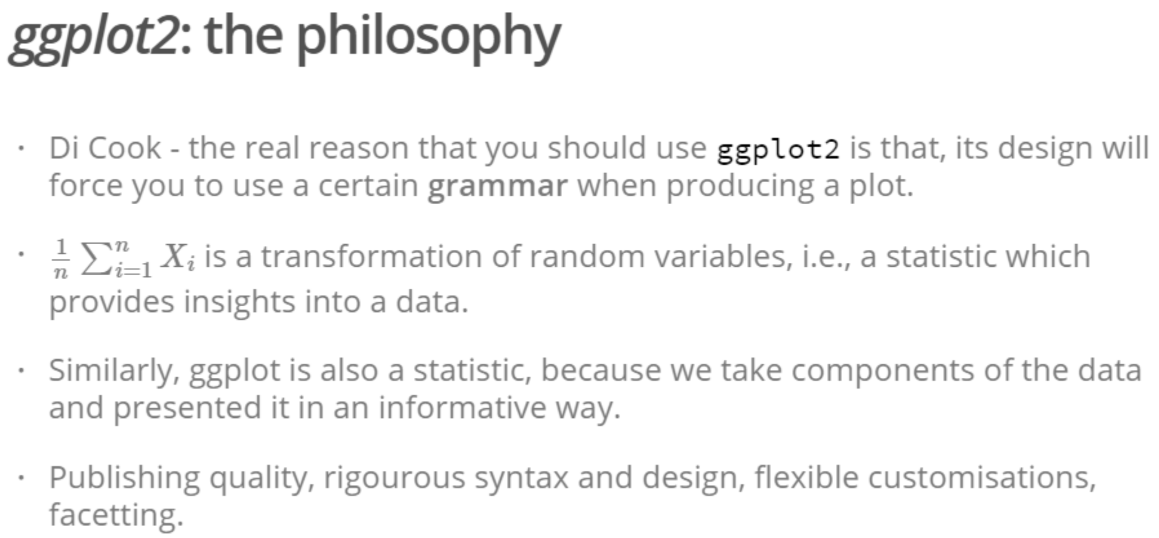


```
flowers = data.frame(species = c("setosa", "versicolor", "virginica"),
                     comments = c("meh", "kinda_okay", "love_it!"))

## cleanIris has the priority in this join operation
iris_comments = left_join(cleanIris, flowers, by = "species")
```

```
## Warning: Column `species` joining character vector and factor, coercing
## into character vector
```

```
## Randomly sampling 6 rows
sample_n(iris_comments, 6)
```

Ggplot2



ggplot2: the philosophy

- Di Cook - the real reason that you should use `ggplot2` is that, its design will force you to use a certain **grammar** when producing a plot.
- $\frac{1}{n}\sum_{i=1}^{n} X_i$ is a transformation of random variables, i.e., a statistic which provides insights into a data.
- Similarly, ggplot is also a statistic, because we take components of the data and presented it in an informative way.
- Publishing quality, rigourous syntax and design, flexible customisations, facetting.

Lecture 3.

Lecture 4. Tuesday, 8 August 2017

Survey Sampling

Topic:

✷ Only 5-6 lectures. ✷ Terminology and examples of uses of sampling. ✷ Advantages of sampling. ✷ Methods of sampling. ✷ Sources of bias. ✷ Census. ✷ Various sampling designs: advantages and designs.

Example 1:

An opinion poll on America's health concern was conducted by Gallup Organization between October 3-5, 1997, and the survey reported that 29% adults consider AIDS is the most urgent health problem of the US, with a margin of error of +/- 3%. The result was based on telephone interviews of 872 adults.

Where did the margin of error come from?

From the central limit theorem:

$$\hat{p} = \frac{s}{n} \sim N\left(p_0, \frac{p_0(1-p_0)}{n}\right)$$

The CI is

$$\hat{p} \pm 1.96\sqrt{\frac{p_0(1-p_0)}{n}}$$

Some issues part of the sample survey:

- Target population? (US pop)
- Sample pop? (adults reached by phone)
- How is the survey conducted? (phone)
- How is the sample selected? (randomly from phone list)
- How reliable are the estimates? Precision of the estimates?
- How much confidence do we have on the estimates?
- Were any bias introduced
- Why were n adults selected?
- Can an estimate on 1000 representatives represent 187 million adults?

From above:

If we have the standard error $\sqrt{\frac{p_0(1-p_0)}{n}} < \sqrt{\frac{1}{4n}}$

To get an error <1%; we need at least 1000 people

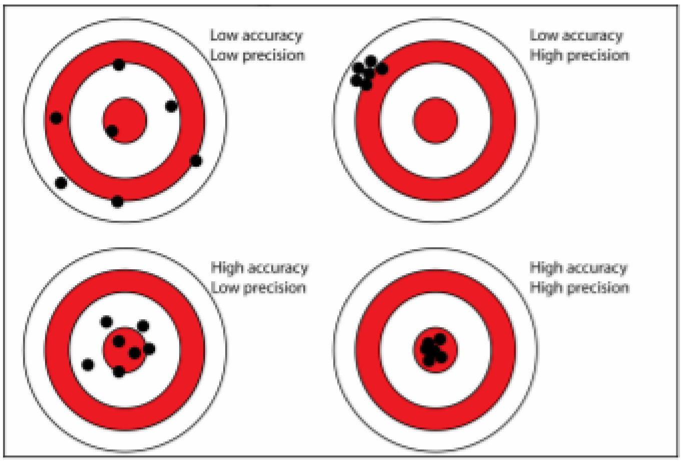
Revision:

- A sample is part of a population.
- A parameter is a numerical descriptor of a population.
- Usually a parameter cannot be determined exactly, but can only be estimated.
- A statistic can be computed from a sample, and used to estimate a parameter.
- A statistic is what the researcher know. A parameter is what the researcher wants to know.
- When estimating a parameter, one major issue is accuracy: how close is the estimate close to be?

Population decribed by a parameter; sample described by a statistic

We will also look at accuracy and precision (bias vs variance)

- Accuracy is bias
- Precision is variance



Example 2:

✷ Census data showed that in 1980, there were 227 million people in the US of whom 11.3% were 65 or older. ✷ In 2000, there were 281 million people, of whom 12.3% were 65 or older. ✷ Is the difference in the percentages statistically significant?

- Calling this statistically different is meaningless; as the census is the whole population, so there's no 'statistical significance'

Discussion:

- You can try to do a two sample test, but the result is close to meaningless. Why?
- We have census data, i.e., the whole population. There is no sampling variability to worry about. The collection of census data is subject to many different sources of error. We are no longer dealing with a chance model.
- The aging of the population is real and this may makes a difference to the health care system.
- If a test of significance is based on data from a whole population... watch out

We don't usually observe the entire population, due to:

- Time
- Resources
- Hard to observe
  - So we collect a sample instead

Sampling

The solution is for us to collect samples in the hope or expectation of making a general statement about the entire population

- Reduce the number of measurements

- Save time, money resources
- Can be essential in destructive testing (eg car crash testing)

Survey design:

- What survey design is appropriate
- How will it be conducted/implemented

Procedure:

- Sample size needed
- How does the design affect sample size

Appropriate survey design provides the best estimation with high reliability at the lowest cost with the available resources

$$Survey\ design + Survey\ Strategy = Survey\ sampling$$

Eg: opinion polls

Case study: literary digest

1936, poll for Roosevelt vs Landon

Background:

- 1936 Franklin Roosevelt was completing his term of office.

- The US was struggling with high unemployment (9 million), just at the end of the Great Depression (1929–1939).
- Took a list of 10 million subscribers.
- Received 20% response (2.4 million people reply)
- Digest was prestigious as they had called the winning election since 1916.
- They predicted victory for Landon.
- Results: Roosevelt won by 62% to 38%. Roosevelt won 46 of 48 states. What went wrong?

Whereas: a Gallup poll did a survey of 3000 from digest, and another of 500000 using a special method, and predicted Roosevelt's victory

What went wrong?

✷ How did Digest picked their sample ? ✷ Digest mailed questionnaires to 10 million people (sources like telephone books and club memberships). ✷ Sample from this list was not representative of the population.

Selection bias

- Affluent people
- Went selection procedure is biased, a larger sample compounds the bias; and **does not** help

Whether results from a sample can be extrapolated to the population depends on the design of the sampling process.

Bias in question wording: (measurement bias)

- Sensitive topics (eg abortion),
- How do you ask the questions?

✷ Should a woman have control over her own body, including her reproductive system? ✷ Should a doctor be allowed to murder unborn children who can't defend themselves?

- Plenty of literature on this
  - Order of questions
  - Emotive language
  - Who asks the questions
  - How they are asked

Measurement bias

Some sources can be:

- Recall bias
    - Forgetting the truth
- Sensitive questions
    - May not tell the truth
        - Use of illegal drugs
        - Paying taxes
- Misinterpreting questions
- Ambiguity of word choice
- Wording of questions and order
- Interview process

Example: race riots measurement bias

After major racial riots in 1968 in Detroit, a sample of African American were asked:“Do you personally feel that you can trust most white people, some white people, or none at all”

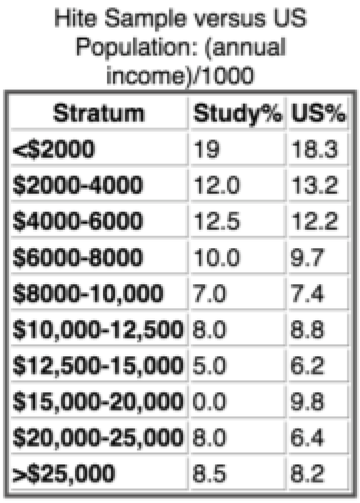
- White interview: 35%“most”
- African American interviewer: 7%“most”

Example: Hite’s Love Survey (non response bias)

✷ Shere Hite send out 100, 000 questionnaires to study how women feel about their relationship with men and reported the findings in a book called “Women and Love”. ✷ Kinsey (1953) 26% with extra-marital affair. ✷ Hite (1980) found 70%. ✷ Redbook (1991) survey found 31%

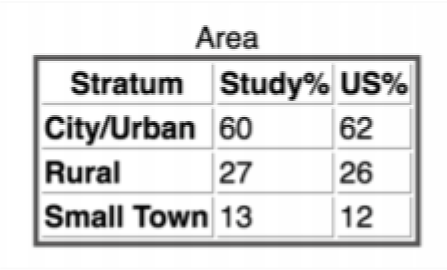
- Why is Hite’s different?

Looking at the sampling, it doesn’t look biased:



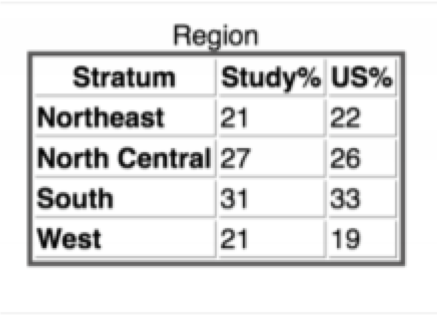
Hite Sample versus US Population: (annual income)/1000

Stratum	Study%	US%
<$2000	19	18.3
$2000-4000	12.0	13.2
$4000-6000	12.5	12.2
$6000-8000	10.0	9.7
$8000-10,000	7.0	7.4
$10,000-12,500	8.0	8.8
$12,500-15,000	5.0	6.2
$15,000-20,000	0.0	9.8
$20,000-25,000	8.0	6.4
>$25,000	8.5	8.2



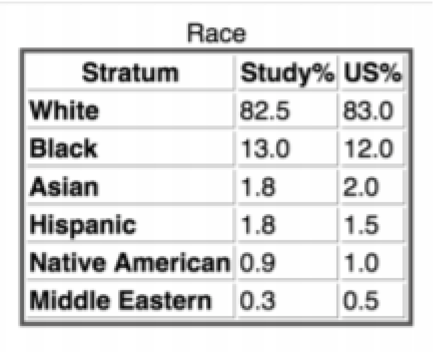
Area

Stratum	Study%	US%
City/Urban	60	62
Rural	27	26
Small Town	13	12



Region

Stratum	Study%	US%
Northeast	21	22
North Central	27	26
South	31	33
West	21	19



Race

Stratum	Study%	US%
White	82.5	83.0
Black	13.0	12.0
Asian	1.8	2.0
Hispanic	1.8	1.5
Native American	0.9	1.0
Middle Eastern	0.3	0.5

But:

In the 100,000 people asked, had a 95.5% non response rate

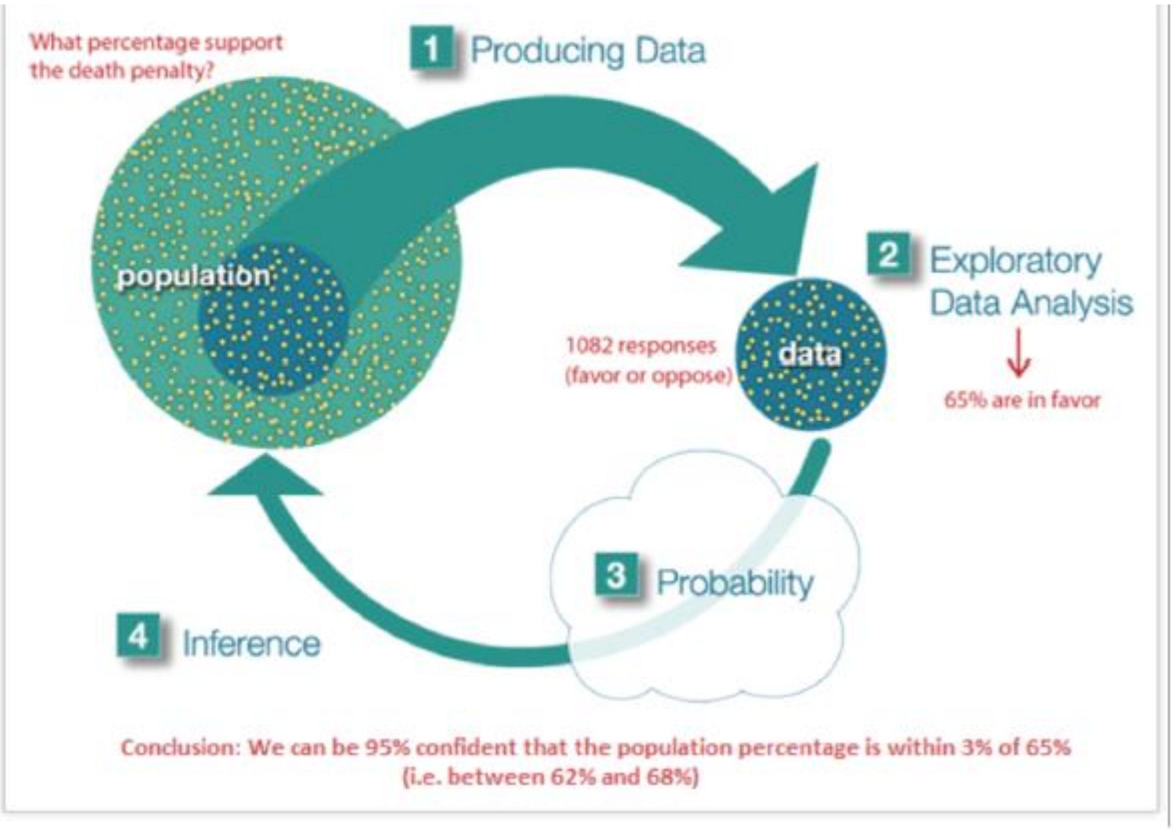
- Unhappy people are often more likely to respond
- Non response bias

Mode of survey administration:

- Mail
- Personal interview
- Telephone
- Combination of methods

Each can introduce its own biases

In an ideal world, we would like to see the characteristics of the samples matches that of the population. Ideally, we would like to have a list of everyone in the population. In practice, we often don't have this list. If the sample is not representative of the population, then the results may be inaccurate. If a survey is ambiguous, subjective, or biased, then the results may be inaccurate


What percentage support the death penalty?
1 Producing Data
population
2 Exploratory Data Analysis
1082 responses (favor or oppose)
data
65% are in favor
3 Probability
4 Inference
Conclusion: We can be 95% confident that the population percentage is within 3% of 65% (i.e. between 62% and 68%)

Sampling terms

Population: a population signifies the units that we are interested in studying. These units could be people, cases and pieces of data.

Unit: The element of the sample selected from the population. The population that you are interested consists of units, which can be people, cases or pieces of data. Consider that you want to examine the erect of health care facilities in a community on prenatal care. What is the unit of analysis: health facility or the individual woman?

Sample: A subset of the population selected for the study.

Sample size: number of units

Sampling frame

Sampling bias (selection bias)

Sampling techniques

- Ways to help select units
- Method for creating samples

Sampling frame

- A sampling framw is a clear and concise description of the population under study, from which the population units can be identified unambiguously and contacted, if desred for the purpose of study

For probability sampling, we should know the selection probability of an individual to be included in the sample. It is important that every individual is included in the selection process.
The sampling frame is very similar to the population you are studying, and may be exactly the same. When selecting units from the population to be included in your sample, it is sometimes desirable to get hold of a list of the population from which you select units. This list can be referred to as the sampling frame.
In additional to physical lists, this could also be used in procedures that can account for all the sampling units without the physical effort of actually listing them.

- Simple random sample refers to randomly selecting sample without replacement
  - How to select elements
  - Randomisation?

Example:

50000 households, 750 sample taken. The average number of TV sets is 1.86 with $\sigma = 0.8$. if possible, find 95% CI; if not explain

$$CI = \mu \pm 1.96\frac{\sigma}{\sqrt{n}} = (1.803{,}1.917)$$

Example 2:

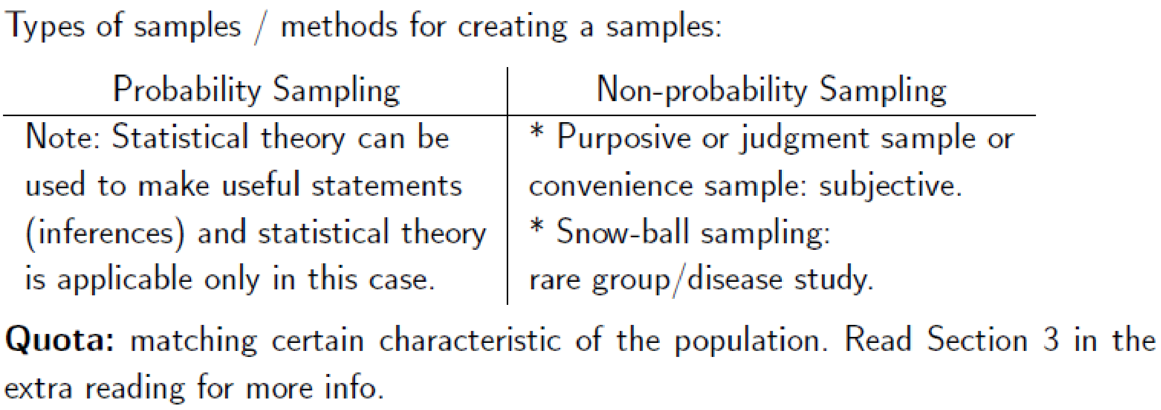
As part of the survey, all person aged 16 or over in the sample households are interviewed. This makes 1528 people, On the average, the sample people watched 5.2 hours of television the Sunday before this survey, and the SD was 4.5 hours. If possible find a 95% confidence interval for the average number of hours spent watching television on the Sunday by all person age 16 and over in the 50,000 households.

- This is a harder question, as not a simple random samble.
  - You can get everybody in a household or nobody, so the SE cannot be estimated (by previous year's methods)
- Household can be seen as a cluster and other methods could be used to calculate SE
- People in same house have similar watching patterns
- Usual problem with cluster samples is "chance error" and not bias

Lecture 5.

Sampling techniques

- Probability sampling
- Non probability sampling



Types of samples / methods for creating a samples:

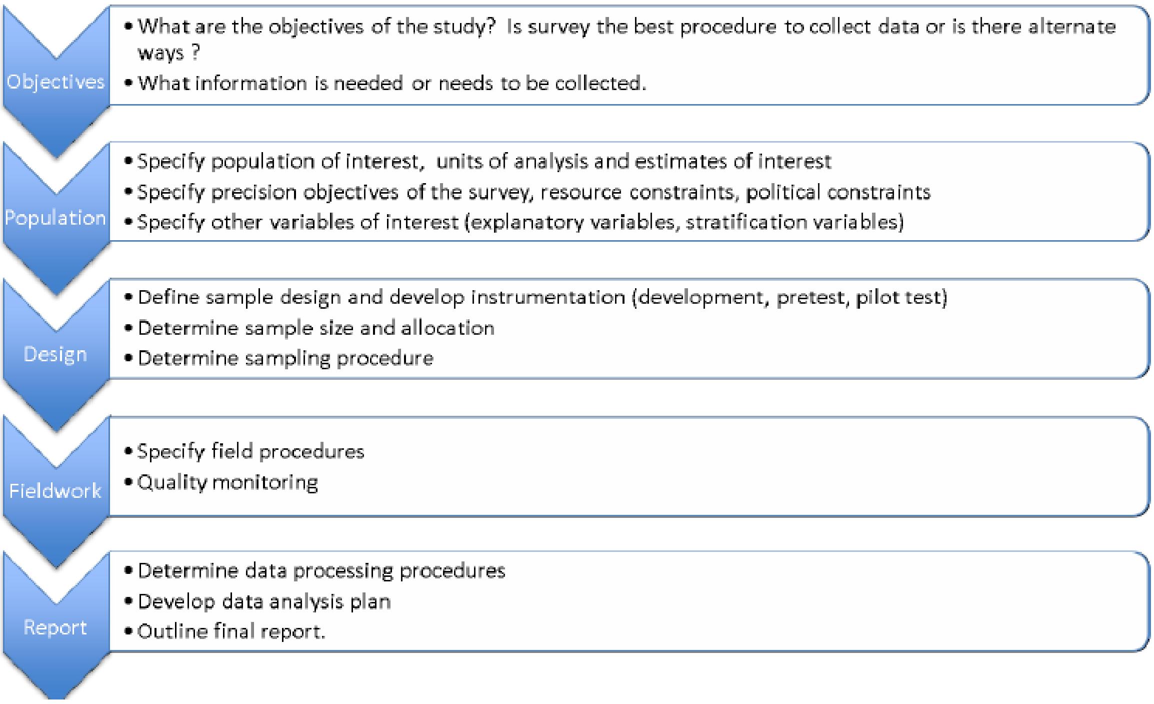
Probability Sampling	Non-probability Sampling
Note: Statistical theory can be used to make useful statements (inferences) and statistical theory is applicable only in this case.	* Purposive or judgment sample or convenience sample: subjective. * Snow-ball sampling: rare group/disease study.

Quota: matching certain characteristic of the population. Read Section 3 in the extra reading for more info.

Characteristics of good sampling

- Address study objectives
- Reliable/robust results
- Procedure understandable
- Realistic timeline
- Economical
- Accurate interpretation and representative results
- Acceptability

Survey design



Objectives
- What are the objectives of the study? Is survey the best procedure to collect data or is there alternate ways ?
- What information is needed or needs to be collected.

Population
- Specify population of interest, units of analysis and estimates of interest
- Specify precision objectives of the survey, resource constraints, political constraints
- Specify other variables of interest (explanatory variables, stratification variables)

Design
- Define sample design and develop instrumentation (development, pretest, pilot test)
- Determine sample size and allocation
- Determine sampling procedure

Fieldwork
- Specify field procedures
- Quality monitoring

Report
- Determine data processing procedures
- Develop data analysis plan
- Outline final report.

1. Specify population of interest.
2. Specify units of analysis and estimates of interest.
3. Specify precision objectives of the survey, resource constraints, and political constraints.
4. Specify other variables of interest (explanatory variables, stratification variables).
5. Review population characteristics (Distributional cost).
6. Develop instrumentation (development, pretest, pilot test).
7. Develop sample design.
8. Determine sample size and allocation.
9. Specify sample selection procedure.
10. Specify _eld procedures.
11. Determine data processing procedures.
12. Develop data analysis plan.
13. Outline _nal report.

Probability Samples:

Types

- Simple random sample (SRS)
    - ○
- Systematic
- Cluster
- Stratified random sample

(all multistage sampling if combined together)

Random sample

Every element of the population has equal probability of being chosen. If selected at random, said to be unbiased. However incorrect sampling may introduce bias.

Systematic sampling

Start your sampling by selecting an element from the list at random and then every kth; where k is known as the sampling interval or skip

Cluster sampling

- Selecting random clusters. You want the cluster to be heterogeneous as possible, so that each cluster is a good small scale representation of the population

Stratified sampling:

- Dividing members of the population into homogeneous, non-overlapping subgroups and then sample.
- Non population unit/elemnt can be excluded

Example: estimate size of UG class

- 100 rand samples from Sydney student and average
- Take a random 50 students, list the courses each student takes and average the list
- Take 50 instructors, list the courses each instructor takes and average the list?
- 5 from each school?
- Take rand sample of 5 schools from science and humanities; from each of these take 5 instructors and average the course sizes
  - Used by gallup
  - Multistage sampling
  - Definite procedure for selecting the sample and it involves the planned use of chance
  - Offers advantages over quota sampling and eliminates the worst feature of quota sampling which is selection bias on the part of the interviewer

Sample theory:

Objective: devise sample scheme which is

- Economical
- Easy to implement
- Unbiased estimators
- Minimise sample variations

Chance error and bias in theory (with SRS)

$$Observed\ Value = Expected\ Value + Chance\ Error$$
$$statistic = parameter + error$$

$$\hat{p} = p + \epsilon$$

In practice:

$$statistic = parameter + error + bias$$

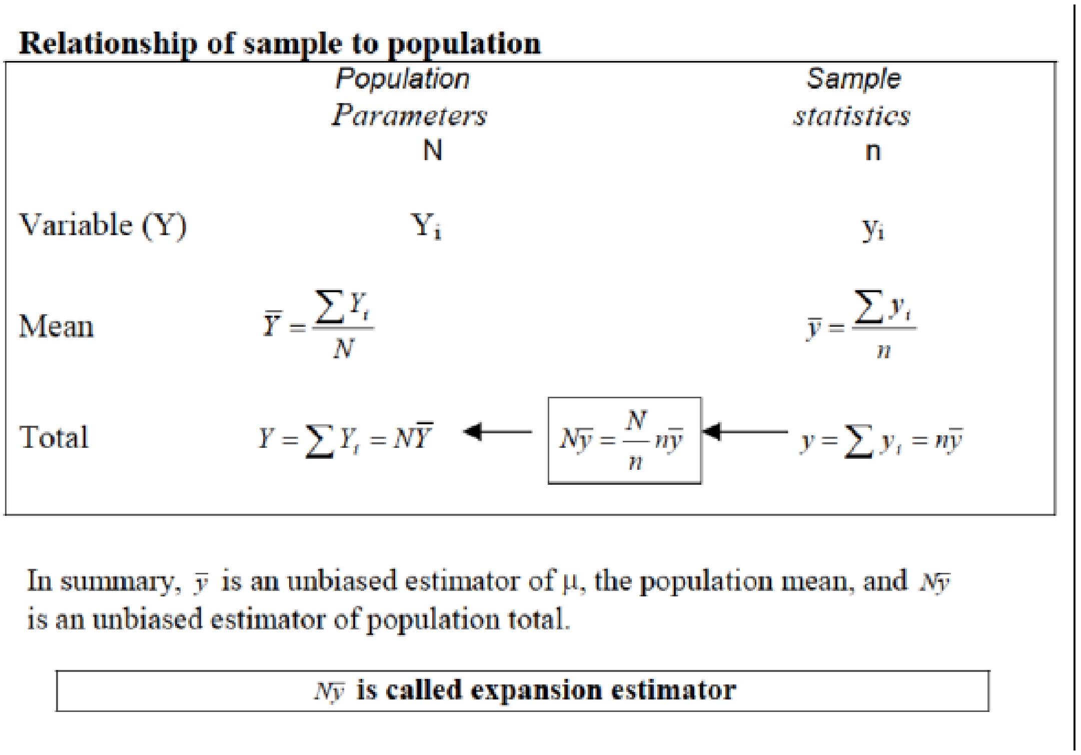
Eg:

- Selection bias
- Nonresponse bias
- Question bias
- Systematic error in estimate

Finite population:

- Contains finite number of items
- Compared to
  - Infinite population
  - Natural population (birth/death/immigration)

- Homogeneous vs heterogeneous population



Relationship of sample to population

	Population Parameters N		*Sample statistics* n
Variable (Y)	Y_i		y_i
Mean	$\bar{Y}=\frac{\sum Y_i}{N}$		$\bar{y}=\frac{\sum y_i}{n}$
Total	$Y=\sum Y_i = N\bar{Y}$ ←	$\boxed{N\bar{y}=\frac{N}{n}n\bar{y}}$ ←	$y=\sum y_i = n\bar{y}$

In summary, $\bar{y}$ is an unbiased estimator of μ, the population mean, and $N\bar{y}$ is an unbiased estimator of population total.

$N\bar{y}$ is called expansion estimator

With and without replacement

- Sample with replacement (SRSWR)
- Sample without replacement (SRSWOR)

How many ways can we select 2 elements from 4 things?

With replacement = 16

Without = 6

Element selection probability (inclusion probability)

From A, B, C, D; what is the probability of selective A (from 2 selections)

- With replacement?
- Without replacement?

Unbiased estimator

Unbiased estimator of $Var(p)$ from a SRS with replacement is

$$Var(p)=\frac{pq}{n-1}$$

Correction for sampling without replacement is

$$Var(p)=\frac{pq}{n-1}\left(1-\frac{n}{N}\right)$$

Lecture 6. Thursday, 10 August 2017

The Census:

Adjustment:

- 'capture recapture' method

$$Pop = Birth + Death + Immigration + Emigration$$

- Take cencus
- SRS and find how many were also in the sensus
- Estimate total number of people by $\frac{Number\ in\ census}{Fraction\ in\ sample\ found\ by\ census}$
-

Population

Finite number of elements, N; N is assumed known. This is a finite population

Post enumeration survey: (PES)

- 40000 households participate in PES, to estimate how many people were not included in the census

- The PES is designed to collect just enough information to determine whether someone was counted in the Census and to enable the ABS to produce a range of net undercount information.

- This included important demographic information (name, sex, age, date of birth, country of birth), the address where each person was on Census night, and any other addresses where each person may have been included on a Census form.

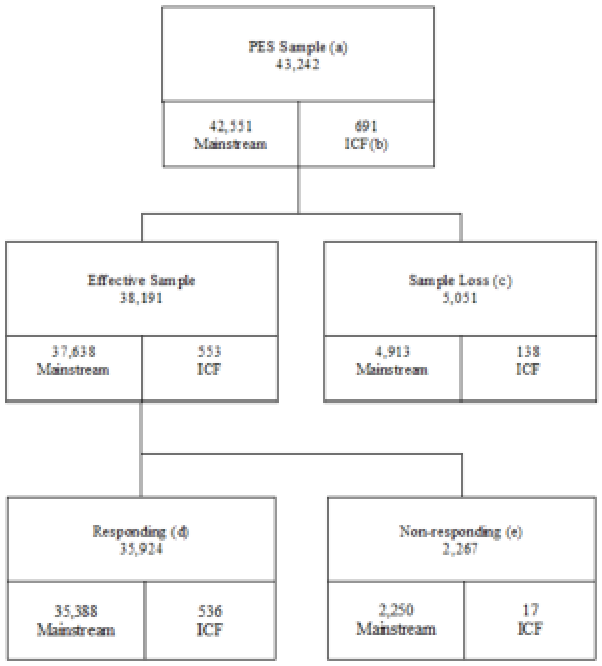
- While the PES asked whether each respondent thought they were included on a Census form, the ABS actually compared Census and PES data to determine whether someone was in fact counted or not.

Sampling:

- Specially trained PES interviewers collected data through face-to-face interviews which started around three weeks after Census night. Some telephone interviews were conducted by office staff where the respondent made contact with the officce and asked to complete the interview on the spot. All mainstream dwellings were enumerated using Computer Assisted Interviewing (CAI).

- Interviews were conducted with any responsible adult of the household who was asked to respond on behalf of all household members.
- This collection methodology differed to the way Census collected its information, where most forms were self-completed.
- A major advantage of interviewer-administered questionnaires is that people can be provided with assistance if they are uncertain about the meaning of questions, and help is also given to ensure no questions are left unanswered.
- To ensure a high response rate was achieved, the number of repeat visits made to non-contact dwellings was twice that of most other ABS household surveys

Figure 1. DWELLINGS, PES sample by main response type



The population:

$Y_1, \ldots, Y_N$ is the population. We adopt the Cochran notation where captials are characteristics of the population, small letters are used for corresponding characteristics of sample

Pop total:

$$Y = \sum_{i=1}^{N} Y_i$$

Mean:

$$\mu = Y\backslash bat = \frac{Y}{N} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

Variance:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 \,; S^2 = \frac{N}{N-1}\sigma^2$$

These are fixed population quantities. If we have to consider 2 numerical values Y, X, the ratio of totals is

$$R = \frac{\sum Y_i}{\sum X_i} = \frac{\bar{Y}}{\bar{X}}$$

Simple Random Sample:

Focus on numerical value of Y_i

Rand samp of size n taken without replacement, and $y_i, \dots, y_n$ are random variables that are stochastically dependent.

Sampling frame is a list of values of $Y_i; i = 1, .., N$

Estimatros

Natural estimator of μ

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$$

Hence, a natural estimator for $Y = N\mu$ is

$$\hat{Y} = N\bar{y}$$

Distributional properties of $\bar{y}$ are complicated by the dependence of the y_i's;

Sample variance:

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

Fundamental results

Sampling without replacement from a finite population:

$$E(\bar{y}) = \mu$$
$$Var(\bar{y}) = \frac{S^2}{n}\left(1 - \frac{n}{N}\right) = \frac{S^2}{n}(1-f) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

- Where f is the sampling fraction, and the finite population correction (fpc) is $1 - f$

$$E(s^2) = S^2$$

So s^2 is an unbiased estimator of S^2

- Remember the SE of an estimator $= \sqrt{Var}$

Comparrison:

- Derivations on Ed

$$Var(Y) = \sigma^2$$
$$Var(\bar{y}) = \frac{S^2}{n}\left(1 - \frac{1}{N}\right) = \frac{\sigma^2}{n} \;\; [for\ SRSWR]$$
$$Var(\bar{y}) = \frac{S^2}{n}\left(1 - \frac{n}{N}\right) \;\; [for\ SRSWOR]$$

Note that $1 - \frac{n}{N} = 1 - f$ is the fpc.

- This is approximately the multiplying factor to convert SRSWR→SRSWOR variance
- As the variance of SRSWOR is smaller, SRSWOR is always more efficient that SRSWR

Lecture 7. Tuesday, 15 August 2017

Sample Size Determination

“In the planning of a sample survey, a stage is always reached at which a decision must be made about the size of the sample. The decision is important. Too large a sample implies a waste of resources, and too small a sample diminishes the utility of the results.”

Purpose:

- One objective of smaple size determiniation is precision analysis
  - We like to measure with precision
  - Goal is to determine size of a sample that is large enough to have errors within desired limits

- We want a sample size to ensure that we can estimate a value, say p from the sample which corresponds to the population parameter P.
- ✷ We want the error in or estimate to be limited to a certain extent, that is, this error should be within a pre-specified **margin-of-error** d.

$$P = p \pm d$$

- More specifically we want a confidence interval of the form P = p ± d
- ✷ Another key element is to describe the confidence we have in our estimate. That is, we want some confidence limits say 95% to our error estimate of d. You will often see the term

$1 - \alpha$

Confidence interval:

If we have S^2 (population), with s^2 (sample), a 95% CI for μ is

$$y \pm 1.96 \frac{s}{\sqrt{n}} \sqrt{1-f}$$

$f = \frac{n}{N}$ for SRSWOR.

Corresponding CI for Y is N times greater

- Remember that capitals denotes population

$$N\left(y \pm 1.96 \frac{s}{\sqrt{n}} \sqrt{1-f}\right)$$

To calculate the sample size needed for sampling yet to be carried out, for specified precision, use past information, or information from a pilot survey.

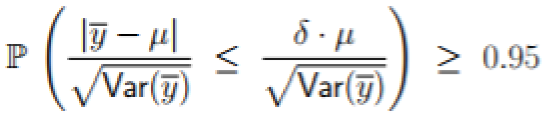
Eg: pilot survey

Suppose we want

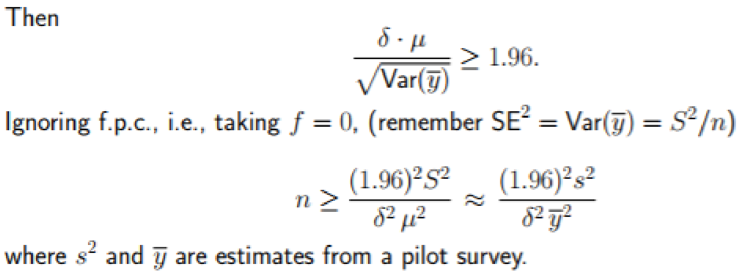
$$P\left(\frac{|\bar{y}-\mu|}{\mu} \le \delta\right) \ge 0.95$$

With 1% $\delta = 0.01$ of actual μ (relative error)

Then



$$\mathbb{P}\left(\frac{|\overline{y}-\mu|}{\sqrt{\mathsf{Var}(\overline{y})}} \le \frac{\delta \cdot \mu}{\sqrt{\mathsf{Var}(\overline{y})}}\right) \ge 0.95$$



Then

$$\frac{\delta \cdot \mu}{\sqrt{\mathsf{Var}(\overline{y})}} \ge 1.96.$$

Ignoring f.p.c., i.e., taking $f = 0$, (remember $\mathsf{SE}^2 = \mathsf{Var}(\overline{y}) = S^2/n$)

$$n \ge \frac{(1.96)^2 S^2}{\delta^2 \mu^2} \approx \frac{(1.96)^2 s^2}{\delta^2 \overline{y}^2}$$

where s^2 and $\overline{y}$ are estimates from a pilot survey.

Sample size calculations for p

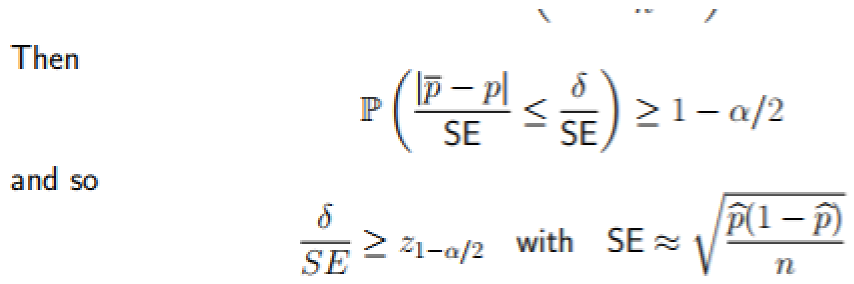
Suppose we want

$$P(|\bar{p} - p| \le \delta) \le 1 - \frac{\alpha}{2}$$

Using CLT:

$$\bar{p} \sim_{approx} N\left(p, \frac{p(1-p)}{n}\right)$$

Then



Then

$$\mathbb{P}\left(\frac{|\overline{p}-p|}{\text{SE}} \leq \frac{\delta}{\text{SE}}\right) \geq 1-\alpha/2$$

and so

$$\frac{\delta}{SE} \geq z_{1-\alpha/2} \quad \text{with} \quad \text{SE} \approx \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

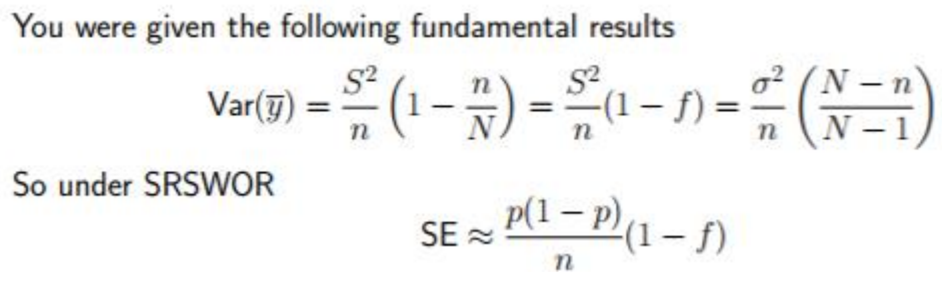
Giving sample size

$$n \geq z_{1-\frac{\alpha}{2}}^2 \frac{p(1-p)}{\delta^2}$$

  - ○ We want this to be at least the smallest integer greater than the RHS
- Note that, the sample size requirement is highest when p = 0.5. It is a common practice to take p = 0.5 when no information is available about p for a conservative estimation of sample size.
- You can justify this to yourself via calculus or in an ad-hoc way via computation.

Sample size calcs for p under SRSWOR

We know that



You were given the following fundamental results

$$\text{Var}(\overline{y}) = \frac{S^2}{n}\left(1-\frac{n}{N}\right) = \frac{S^2}{n}(1-f) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

So under SRSWOR

$$\text{SE} \approx \frac{p(1-p)}{n}(1-f)$$

So we get the sample size

$$n \geq z_{1-\frac{\alpha}{2}}^2 \frac{p(1-p)}{\delta^2}(1-f)$$

Example:

Example if we like to estimate p = 0.4 with 95% confidence with δ = 0.1 Then under SRSWR

$$n \geq 1.96^2 \times \frac{0.4 \times 0.6}{\delta^2} = 92.19$$

$$n = 93$$

Example 2:

Adjusting for SRSWOR:

- The adjustment is done looking at n' represents the number of sample under simple random sampling with replacement n represents the number of sample without replacement. What is the relationship between n and n'?
    - n' WR (from before)
    - n WOR (want to find now)

$$\frac{S^2}{n'} = \frac{S^2}{n}\left(1 - \frac{n}{N}\right)$$

Giving

$$n = \frac{n'}{1 + \frac{n'}{N}}$$

If $N = 200$; from before; we found $n' = 93$

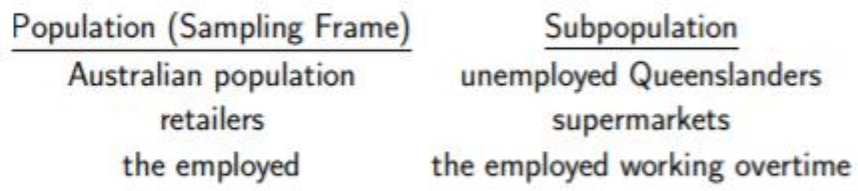
$$\therefore n = \frac{93}{1 + \frac{93}{200}} = 63.48$$
$$n = 64$$

SRS: Inference over Subpopulations

Population of size N individuals, SRS sample size n; measurements of which $Y_i i = 1, . . , N$ is considered.

Separate estimates wanted for one of a number of subclasses $\{C_1, C_2, \ldots\}$ are subsets of the population (sampling frame)

Eg:



Population (Sampling Frame)	Subpopulation
Australian population	unemployed Queenslanders
retailers	supermarkets
the employed	the employed working overtime

Notation and Estimators:

Denote number of items in each class C_j by N_j. (generally unkown)

Estimated by

$$\widehat{N}_J = N\frac{n_j}{n}$$

Where $\frac{n_j}{n}$ = preportion of total sample size in C_j

We find $\bar{Y}_{C_j}$ and Y_{C_j} as:

$$\bar{Y}_{C_j} = \sum_{i\in C_j}\frac{Y_i}{N_j}$$

$$Y_{C_j} = \sum_{i\in C_j} Y_i$$

We use

$$Y_i' = \begin{cases} Y_i; i \in C_j \\ 0; i \notin C_j \end{cases}$$

Correspondingly, in the sample

$$y_i' = \begin{cases} y_i; i \in C_j \\ 0; i \notin C_j \end{cases}$$

Then $\bar{y}'$ estimates $\bar{Y}'$; where $\bar{y}' = \sum_{i=1}^{n}\frac{y_i'}{n}$

So we haave:

$$N\bar{Y}' = Y_{C_j}$$

- So $N\bar{y}'$ estimates Y_{C_j}
- $\hat{\bar{Y}}_{C_j} = \frac{N\bar{y}'}{N_j}$ estimates $\bar{Y}_{C_j}$
    - These estiamteros are unbiased

We get that:

$$Var\left(\hat{\bar{Y}}_{C_j}\right) = \left(\frac{N}{N_j}\right)^2 Var(\bar{y}') = \left(\frac{N}{N_j}\right)^2 \frac{\left(S'_{C_j}\right)^2}{n}\left(1-\frac{n}{N}\right)$$

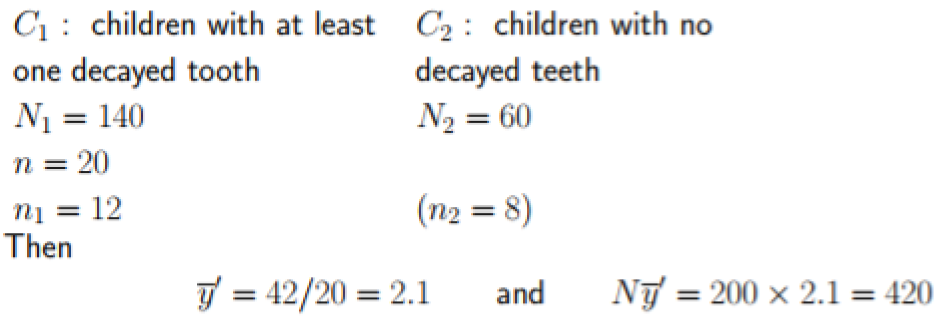
Where $S'^2_{C_j}$ is estimated by ${s'_{C_j}}^2$

$$(S'_{C_j})^2 = \sum_{i=1}^{N}\frac{(Y_i' - \overline{Y}')^2}{N-1}, \quad (s'_{C_j})^2 = \frac{\sum_{i=1}^{n}(y_i' - \overline{y}')^2}{n-1}.$$

Example:

There are 200 children in a village. One dentist takes a simple random sample of 20 and finds 12 children with at least one decayed tooth and a total of 42 decayed teeth. Another dentist quickly checks all 200 children and finds 60 with no decayed teeth

Estimate the total number of decayed teeth.

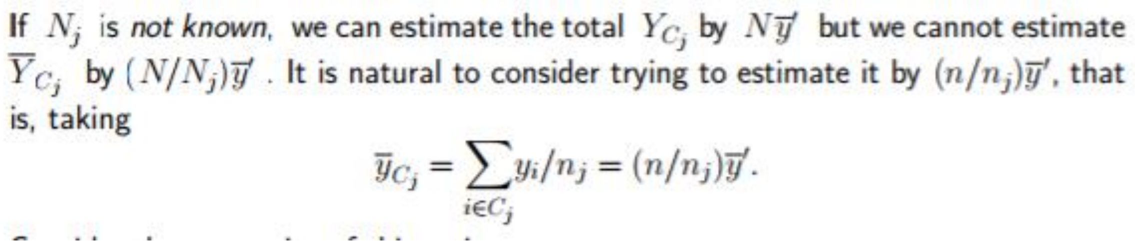


C_1 : children with at least one decayed tooth	C_2 : children with no decayed teeth
$N_1 = 140$	$N_2 = 60$
$n = 20$	
$n_1 = 12$	$(n_2 = 8)$

Then

$$\overline{y}' = 42/20 = 2.1 \qquad \text{and} \qquad N\overline{y}' = 200 \times 2.1 = 420$$

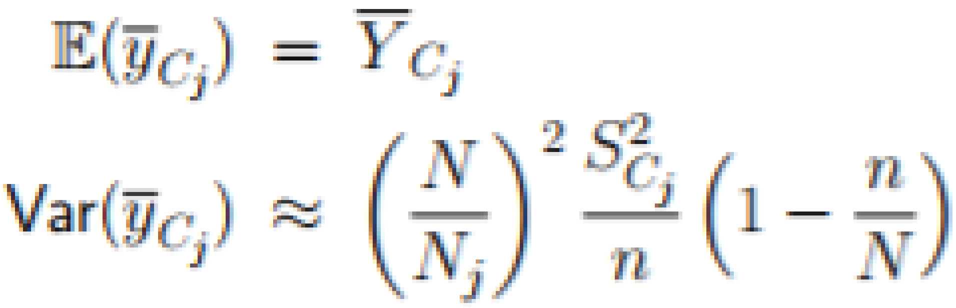
N_j not known



If N_j is *not known*, we can estimate the total Y_{C_j} by $N\overline{y}'$ but we cannot estimate $\overline{Y}_{C_j}$ by $(N/N_j)\overline{y}'$. It is natural to consider trying to estimate it by $(n/n_j)\overline{y}'$, that is, taking

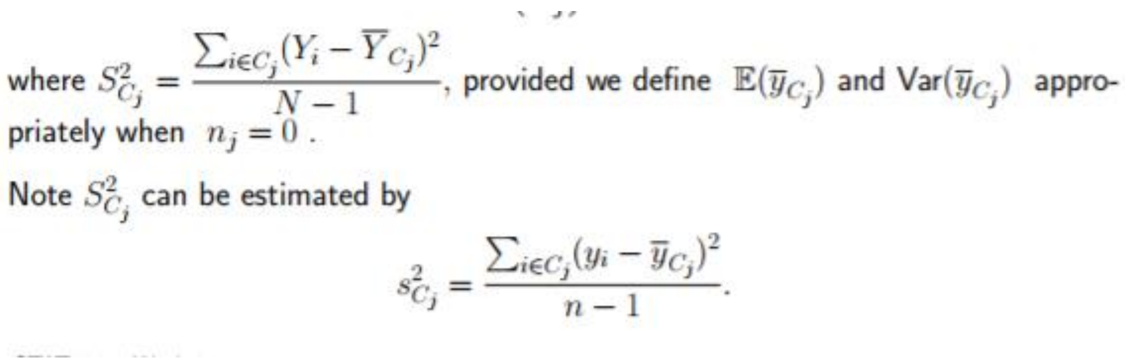
$$\overline{y}_{C_j} = \sum_{i \in C_j} y_i/n_j = (n/n_j)\overline{y}'.$$

Estimatros are then:



$$\mathbb{E}(\overline{y}_{C_j}) = \overline{Y}_{C_j}$$

$$\text{Var}(\overline{y}_{C_j}) \approx \left(\frac{N}{N_j}\right)^2 \frac{S^2_{C_j}}{n}\left(1 - \frac{n}{N}\right)$$



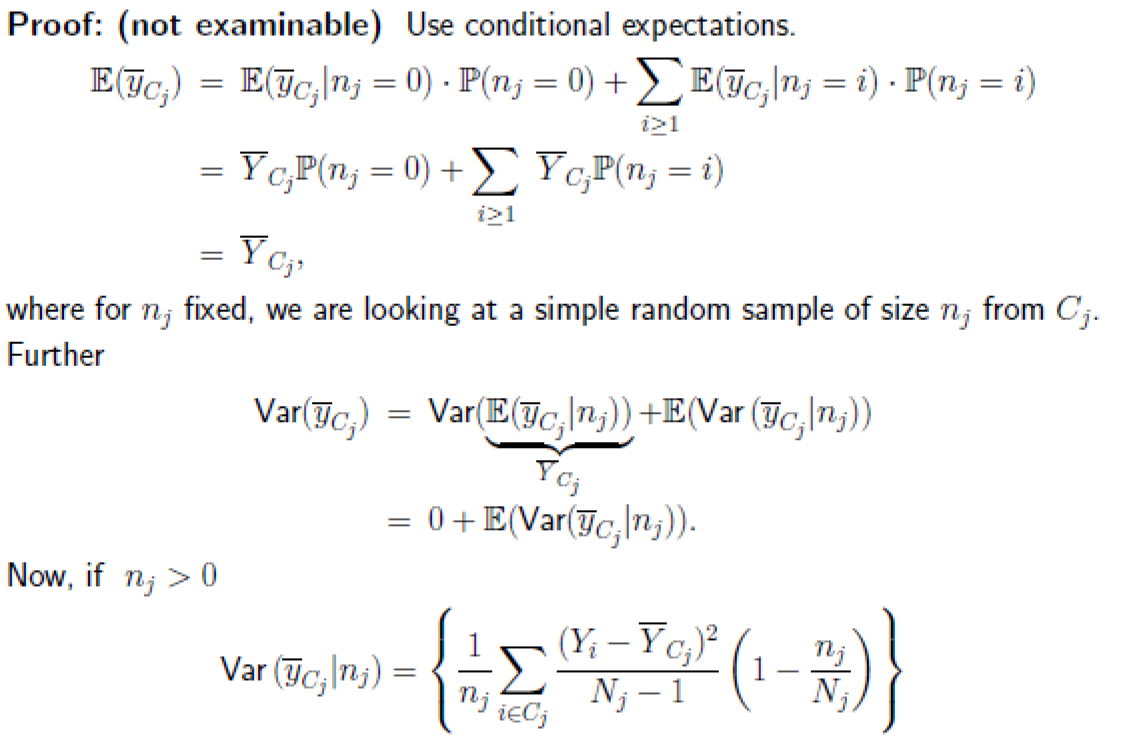
where $S^2_{C_j} = \dfrac{\sum_{i\in C_j}(Y_i - \overline{Y}_{C_j})^2}{N-1}$, provided we define $\mathbb{E}(\overline{y}_{C_j})$ and Var$(\overline{y}_{C_j})$ appropriately when $n_j = 0$.

Note $S^2_{C_j}$ can be estimated by

$$s^2_{C_j} = \frac{\sum_{i\in C_j}(y_i - \overline{y}_{C_j})^2}{n-1}.$$

Lecture 8.

Proof: (not examinable)



Proof: (not examinable) Use conditional expectations.

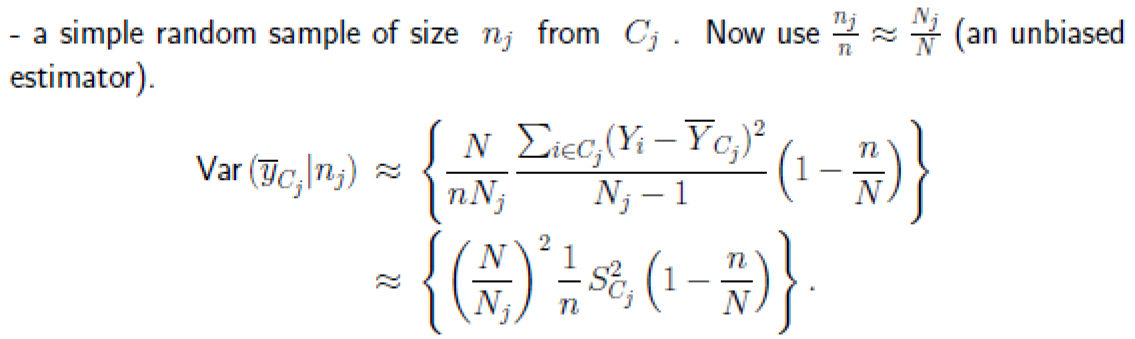
$$\begin{aligned}
\mathbb{E}(\overline{y}_{C_j}) &= \mathbb{E}(\overline{y}_{C_j}|n_j = 0)\cdot\mathbb{P}(n_j = 0) + \sum_{i\geq 1}\mathbb{E}(\overline{y}_{C_j}|n_j = i)\cdot\mathbb{P}(n_j = i)\\
&= \overline{Y}_{C_j}\mathbb{P}(n_j = 0) + \sum_{i\geq 1}\overline{Y}_{C_j}\mathbb{P}(n_j = i)\\
&= \overline{Y}_{C_j},
\end{aligned}$$

where for n_j fixed, we are looking at a simple random sample of size n_j from C_j. Further

$$\begin{aligned}
\text{Var}(\overline{y}_{C_j}) &= \text{Var}(\underbrace{\mathbb{E}(\overline{y}_{C_j}|n_j)}_{\overline{Y}_{C_j}}) + \mathbb{E}(\text{Var}\,(\overline{y}_{C_j}|n_j))\\
&= 0 + \mathbb{E}(\text{Var}(\overline{y}_{C_j}|n_j)).
\end{aligned}$$

Now, if $n_j > 0$

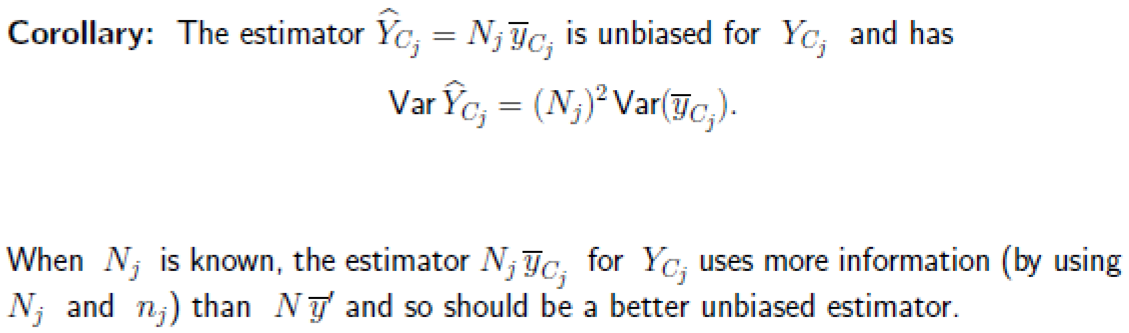
$$\text{Var}\,(\overline{y}_{C_j}|n_j) = \left\{\frac{1}{n_j}\sum_{i\in C_j}\frac{(Y_i - \overline{Y}_{C_j})^2}{N_j - 1}\left(1 - \frac{n_j}{N_j}\right)\right\}$$



- a simple random sample of size n_j from C_j . Now use $\frac{n_j}{n} \approx \frac{N_j}{N}$ (an unbiased estimator).

$$\begin{aligned}
\text{Var}\,(\overline{y}_{C_j}|n_j) &\approx \left\{\frac{N}{nN_j}\,\frac{\sum_{i\in C_j}(Y_i - \overline{Y}_{C_j})^2}{N_j - 1}\left(1 - \frac{n}{N}\right)\right\}\\
&\approx \left\{\left(\frac{N}{N_j}\right)^2\frac{1}{n}S^2_{C_j}\left(1 - \frac{n}{N}\right)\right\}.
\end{aligned}$$

Corollary

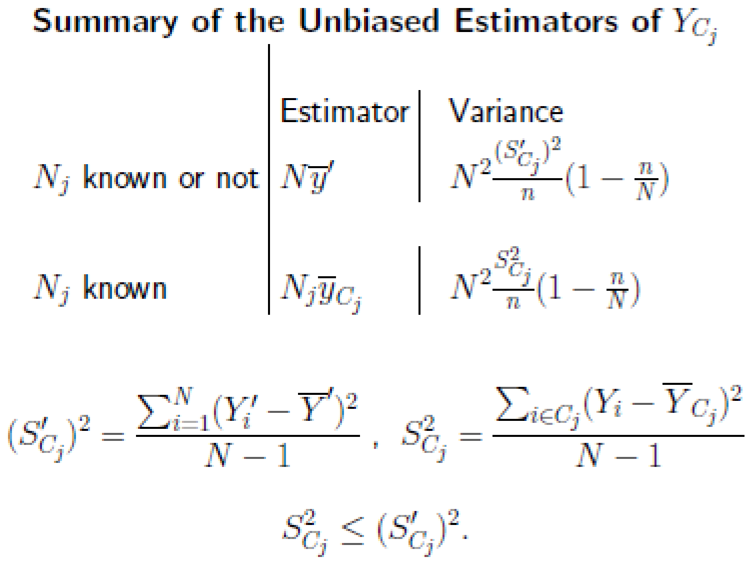


Corollary: The estimator $\widehat{Y}_{C_j} = N_j\,\overline{y}_{C_j}$ is unbiased for Y_{C_j} and has

$$\text{Var}\,\widehat{Y}_{C_j} = (N_j)^2\,\text{Var}(\overline{y}_{C_j}).$$

When N_j is known, the estimator $N_j\,\overline{y}_{C_j}$ for Y_{C_j} uses more information (by using N_j and n_j) than $N\,\overline{y}'$ and so should be a better unbiased estimator.

Summary of unbiased estimators



Summary of the Unbiased Estimators of Y_{C_j}

	Estimator	Variance
N_j known or not	$N\overline{y}'$	$N^2\frac{(S'_{C_j})^2}{n}(1-\frac{n}{N})$
N_j known	$N_j\overline{y}_{C_j}$	$N^2\frac{S^2_{C_j}}{n}(1-\frac{n}{N})$

$$(S'_{C_j})^2 = \frac{\sum_{i=1}^{N}(Y'_i - \overline{Y}')^2}{N-1}\,, \quad S^2_{C_j} = \frac{\sum_{i\in C_j}(Y_i - \overline{Y}_{C_j})^2}{N-1}$$

$$S^2_{C_j} \leq (S'_{C_j})^2.$$

Lecture 9.

Stratified sampling

In stratified sampling, the population is partitioned into groups, called strata, and sampling is performed separately within each stratum.

1. Stratum variables are mutually exclusive.
2. Aim to have homogenous population within-stratum and heterogenous between strata

Reasoning for stratification:

- Natural stratification may exist
- We can improce the precision of our estimator if the population can be divided into subpopulations where the variability within each subpopulation or strata is known. We can "redirect" observations from the less variable strata to the more variable ones (relative to a simple overall random sample) to reduce overall variability in our estimator.
- Administrative convenience, for example, if separate administrative units exist, such as states or local government areas, it may be easier to sample in each individually, and then put the information together.
- Cost. If the cost of taking observations varies across subpopulations (eg cost of surveys in remote rural areas vs the cities) then savings may be possible by adjusting the proportions of the subpopulations sampled to minimise overall costs.

Advantages

- Provides opportunity to study the stratum variations - estimation could be made for each stratum
- Disproportionate sample may be selected from each stratum
- The precision likely to increase as variance may be smaller than SRS with same sample size
- Field works can be organized using the strata (e.g., by geographical areas or regions)
- Reduce survey costs.

The principal objective of stratication is to reduce sampling errors.

Disadvantages:

- Sampling frame is needed for each stratum.
- Analysis method is complex such as correct variance estimation.
- Data analysis should take sampling weight into account for disproportionate sampling of strata.
- Sample size estimation is diffcult in practice.

Stratified random sampling

When sample is selected by SRS technique independently within each stratum, the design is called stratified random sampling.

Pop size N divided into subpopulations $N_1, \ldots, N_L$, the strata are mutually exclusive and exhaustive; N_h is known.

A SRS of size n_h is drawn from the h^{th} stratum; $h = 1, .., L$