Intro to Statistics Notes

Contents

Course Details

- **Dates:** Mon, Jun 1 – Mon Jul 13, 2020
- **Time:** 9:00am-12:00pm
- **Place:** Online! https://mcgill.zoom.us/j/98513054963
- **Course website:** MyCourses
- **Instructor email:** nandini.dendukuri@mcgill.ca
- **Assessment**:
  - 6 assignments: 60%
  - In-class quizzes: 10%
  - 1 group project: 30%

Project

- Goal is for you to learn the methods covered in this course by applying to a real problem you are familiar with
- You are required to work in groups of 3, identify a dataset, preferably related to your research and analyze it
- At the end of the course, submit a short report and make a 10-minute presentation
  - Use reporting guidelines from the Equator network as relevant
- The methods used will naturally limited to those covered during this introductory course. However, you can discuss what other methods would be needed to answer the research question satisfactorily.

Suggested References

- Many books available for free via McGill libraries, for example:

  - **Biostatistics with R: An Introduction to Statistics through Biological Data**, Babak Shahbaba, Springer, 2012
  - **A tiny handbook of R**, Mike Allerhand, Springer-Verlag, 2011

- Lecture material for this course is drawn from a variety of sources including:
  - Other courses:
    * **Tim Hanson's Course (Univ of South Carolina):** http://people.stat.sc.edu/hansont/stat205/stat205_spring2014.html
    * **Ingo Ruczinski's two-term course (Johns Hopkins)** http://www.biostat.jhsph.edu/~iruczins/teaching/140.615/140.615.index.html
    * **Lawrence Joseph's EPIB-607 Course (Dept of Epidemiology, Biostatistics and Occupational Health, McGill University)** http://www.medicine.mcgill.ca/epidemiology/Joseph/courses/EPIB-607/main.html
  - Text books (available via sites like amazon.com):
    * **Statistics for the life sciences**, Samuels, Wittmer and Schaffner, 2016
    * **Statistical data analysis for the life sciences**, Ekstrom and Sorensen, CRC press, 2010
    * **Data analysis for the life sciences**, Irizarry and Love, Leanpub, 2015
    * **Statistical ideas and methods**, Utts and Heckard, Thompson Brooks Cole, 2006

Chapter 1

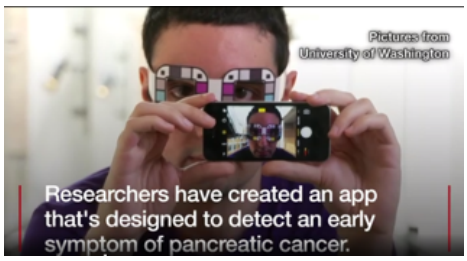
Lecture 1

1.1 Introduction

1.1.1 What is Statistics?

Statistics is a collection of procedures and principles for gathering data and analyzing information in order to help people make decisions when faced with uncertainty
- Utts & Heckard in 'Statistical Ideas & Methods'

1.1.2 A Motivating Example



News Clip Manuscript

- Pancreatic cancer has a very poor survival rate because it is often detected too late
- A new app promises to detect early symptoms of jaundice that may go unnoticed typically
- Should this "test" be adopted into routine practice?

1.1.3 What was the evidence behind this optimistic headline?

- In an initial study the app detected cases of "concern" correctly 89.7% of the time, and classified "negative" cases correctly 96.8% of the time
- The reference test was based on the total serum bilirubin level

1.1.4 What would a data detective ask?

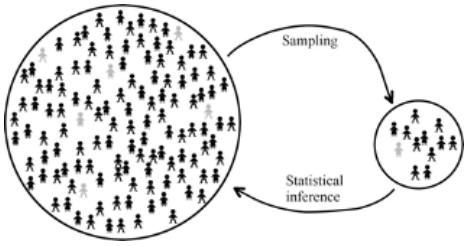
1. Are the statistical methods appropriate?
2. Is the study design appropriate?
3. Is there information external to the study that affects its interpretation?

1.1.5 Results reported in the study

	Borderline or Elevated Bilirubin	Normal Bilirubin
BiliScreen Positive	35 (89.7%)	1
BiliScreen Negative	4	30 (96.8%)
Total	39	31

- The statistics of interest when evaluating a diagnostic test are
  - Sensitivity = Probability(Positive result | Reference test positive) = 89.7%
  - Specificity = Probability(Negative result | Reference test negative) = 96.8%
- Do these data provide good estimates of BiliScreen accuracy?

1.1.6 Evaluating the quality of the statistical methods



- Is the study large enough?
- What is the uncertainty around the reported results?
- Were relevant statistics recorded?
- Do the statistics provided help make a decision about the next step?

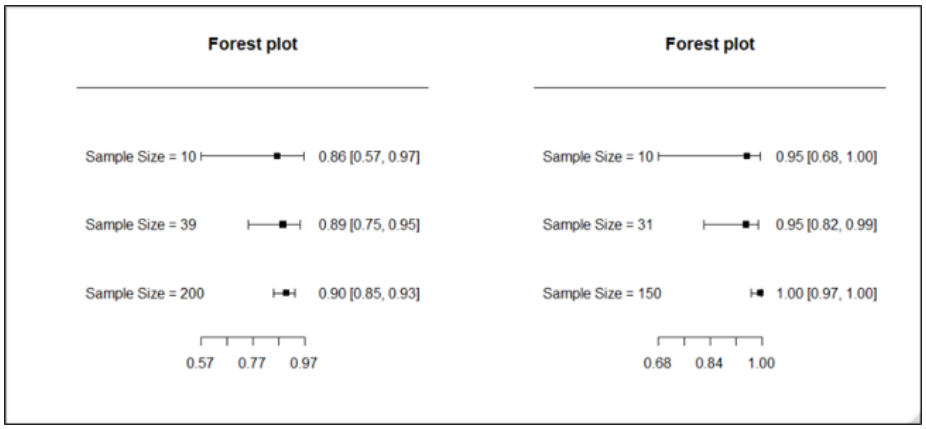
1.1.7 What if the sample size were smaller?

	Borderline or Elevated Bilirubin	Normal Bilirubin
BiliScreen Positive	9 (90%)	0
BiliScreen Negative	1	10 (100%)
Total	10	10

1.1.8 What if the sample size were larger?

	Borderline or Elevated Bilirubin	Normal Bilirubin
BiliScreen Positive	180 (90%)	0
BiliScreen Negative	20	150 (100%)
Total	200	150

1.1.9 Sample Size and Precision



1.1.10 Evaluating the quality of the statistical methods

- Notice that **the certainty we have in our conclusions depends on the sample size**. The extreme results were less convincing when the sample size was reduced.
- What sample size is needed to draw a definitive conclusion? That needs to be determined using appropriate statistical methods to obtain the desired precision. We will study this in Lectures 3 and 6

1.1.11 Evaluating the quality of the study design

- Are the subjects in the study representative?
- Is the reference standard relevant?
- Are the subjects in the study representative?
  - Healthy volunteers and patients from a medical centre were used
  - If the test accuracy is systematically better or worse in these patients than in patients on whom the test will be used, then the results are biased

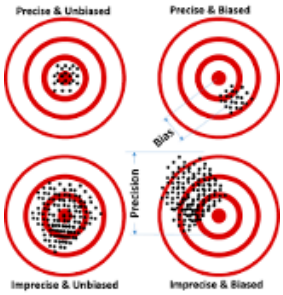
- Is the reference standard relevant?
  - Bilirubin level is a measure of jaundice, but not all cases of jaundice have pancreatic cancer
  - If the accuracy of the test with respect to bilirubin level is systematically different from the accuracy with respect to pancreatic cancer, then our results may be biased

1.1.12 The role of external (or prior) information

- Besides the sample size and study design, our conclusions may also be affected by information external to the observed results, for example from a previous study
- Statistical analyses should take into account the impact of this prior information. We will study how to do so in Lecture 6

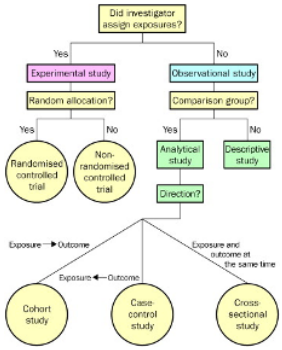
1.2 Reducing Bias in Research Studies

1.2.1 Bias vs. Precision



- Precision results in a random departure from the true value
- Bias is a systematic departure from the true value
- A large sample size can improve precision but not bias. Study design and analysis could reduce bias

1.2.2 Common study designs used in clinical research



- An analytical or experimental study can study the relation between an intervention and an outcome

- A descriptive study, with no control group, cannot

1.2.3 Randomized Controlled Trial

- Advantages:
  - unbiased distribution of confounders;
  - blinding more likely;
  - randomisation facilitates statistical analysis.
- Disadvantages
  - expensive: time and money;
  - study subjects not representative;
  - ethically problematic at times.

1.2.4 Reducing bias in research studies

- Different types of bias common in research studies have been enumerated

Type of bias	Description	Possibl
Selection bias	Sampling method results in sample not representative of the population	Rando
Measurement bias	Measurement method records outcome with systematic error	Statist
Detection bias	Measurement method differs between groups being compared	Blindi
Confounding	Risk factors distributed unequally in groups being compared	Rando

- Statistical methods are often used to reduce bias, either at the planning stage of a study or at the analysis stage
- In this lecture, we will look at random sampling and randomization. In Lecture 12 we will look at adjustment via regression

1.2.5 A second motivating example: Renal Denervation

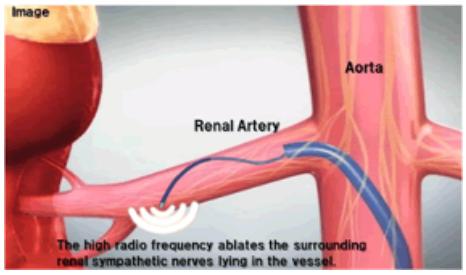


Image Source

- A surgical procedure called “renal denervation” was developed to help people with hypertension who do not respond to medication.

Baseline		3-month follow-up	
Number of patients	Blood pressure	Number of Patients	Change in blood pressure
153	176/98 [systolic/diastolic (mmHg) Mean]	135	-25/-11 [systolic/diastolic

	Baseline		3-mo	
	Number of patients	Blood pressure	Number of Patients	Char
Renal Denervation	45	176/98 [systolic/diastolic (mmHg) Mean]	39	-21/-
Control group*	5	173/98 [systolic/diastolic (mmHg) Mean]	3	+2/-

1.2.6 Example 4a: Results from a cohort study of renal denervation*

*Investigators Symplicity HTN-1. Catheter-based renal sympathetic denervation for resistant hypertension: durability of blood pressure reduction out to 24 months. Hypertension 2011;57(5):911-917.

- Can the large observed change be interpreted as being caused by renal denervation?
- This is an example of a before-after design that reports on change over a period of time, typically the change after an intervention.
- The primary drawback of this design is the lack of a control group.
- The observed change may simply be attributable to the participation in the study ('Hawthorne effect'). If so, then the same magnitude of change in the blood pressure would be observed in the control group. This would mean that the change was not due to renal denervation at all.
- Therefore this study cannot provide proof that renal denervation causes a decline in blood pressure.
- Another issue in the data presented here is that the variability around the mean change is not available. So we don't know if all patients experienced this benefit.

1.2.7 Example 4b: Results compared to a control group*

Catheter-based renal sympathetic denervation for resistant hypertension: a multicentre safety and proof-of-principle cohort study

Henry Krum, Markus Schlaich, Rob Whitbourn, Paul A Sobotka, Jerzy Sadowski, Krzysztof Bartus, Boguslaw Kapelak, Anthony Walton, Horst Sievert, Suku Thambar, William T Abraham, Murray Esler

Summary

Background Renal sympathetic hyperactivity is associated with hypertension and its progression, chronic kidney disease, and heart failure. We did a proof-of-principle trial of therapeutic renal sympathetic denervation in patients with resistant hypertension (ie, systolic blood pressure ≥160 mm Hg on three or more antihypertensive medications, including a diuretic) to assess safety and blood-pressure reduction effectiveness.

Lancet 2009; 373: 1275–81
Published Online
March 30, 2009
DOI:10.1016/S0140-6736(09)60566-3

	Baseline		
	Number of patients	Blood pressure	Number of Pat
Renal Denervation	49	178/96 [systolic/diastolic (mmHg) Mean]	
Control*	51	178/97 [systolic/diastolic (mmHg) Mean]	

*Patients excluded from renal denervation arm for anatomical reasons

*Catheter-based renal sympathetic denervation for resistant hypertension: a multicentre safety and proof-of-principle cohort study

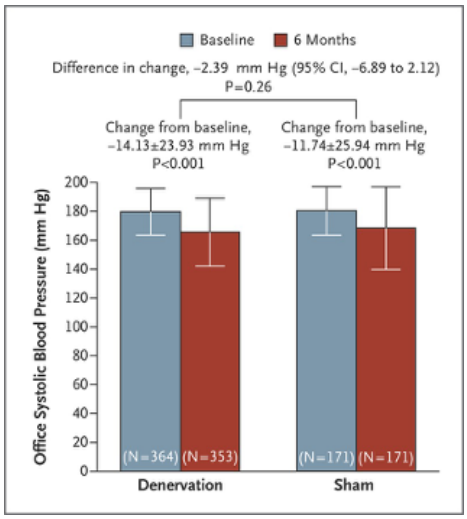
- The control group was of patients who were excluded for anatomical reasons.
- It is possible that, the control group may not have had the same risk of resistant hypertension as the treatment group, i.e. the 'anatomical reasons' were a confounding factor. This may explain why the control group had a worse mean change in blood pressure than the renal denervation group
- Therefore, once again, we don't have a conclusive result.
- Of course, the small size of the control group also does not help. Other concerns in this study include loss to follow-up. Only 18 patients completed the follow-up of 24 months.

1.2.8 Example 4c: Results from a randomized controlled trial (RCT) of renal denervation*

**Esler MD, Krum H, Sobotka PA, Schlaich MP, Schmieder RE, Bohm M. Renal sympathetic denervation in patients with treatment-resistant hypertension (The Symplicity HTN-2 Trial): a randomised controlled trial. Lancet 2010;376(9756):1903-1909

- The study concluded there was a statistically significant ($p<0.001$) difference between the intervention and control groups
- The randomization procedure gives us greater confidence in these results as patients had the same risk of a change in BP at the time of randomization
- However, the study was not perfect. Importantly, it was not blinded and the main outcome was office BP rather than ambulatory BP. Therefore, it is possible that the patients in the renal denervation arm reacted differently owing to the greater attention they received.
- Also, the follow-up of 6-months is very short and it is unknown whether the observed drop in BP is sustained in the long term.

1.2.9 Example 4d: Results from a second randomized controlled trial of renal denervation*



- "A significant change from baseline to 6 months in office systolic blood pressure was observed in both study groups.

The between-group difference (the primary efficacy end point) did not meet a test of superiority with a margin of 5 mm Hg.

The bars indicate standard deviations."

- The second RCT improved on the first one by using a sham procedure in the control group. This removed the concern about blinding.
- They found that there was no significant difference between the renal denervation and control groups.

*Bhatt et al. A controlled trial of renal denervation for resistant hypertension. N Engl J Med 2014;370:1393-401. DOI: 10.1056/NEJMoa1402670

1.2.10 Example 4: Renal Denervation as a treatment for resistant hypertension

- An early study suggested that renal denervation (which uses radiotherapy to destroy some nerves in arteries feeding the kidney) reduces blood pressure. In that experiment, patients who received surgery had an average improvement in systolic blood pressure of 33 mmHg more than did control patients who received no surgery.

- Later an experiment was conducted in which patients were randomly assigned to one of two groups. Patients in the treatment group received the renal denervation surgery. Patients in the control group received a sham operation in which a catheter was inserted, as in the real operation, but 20 minutes later the catheter was removed without radiotherapy being used. These patients had no way of knowing that their operation was a sham. The rates of improvement in the two groups of patients were nearly identical.(Samuels 10-11)

1.2.11 Lessons learnt from renal denervation example

- A control group is necessary to draw conclusions about the effect of a variable
- However, a randomized design is necessary to make a cause-effect conclusion
- A randomized, controlled trial is not automatically unbiased. Blinding is necessary

1.2.12 Health Technology Assessment of Renal Denervation

- The MUHC's Technology Assessment Unit evaluated Renal Denervation in 2013. The full report is available here
- We concluded:
  "… There is evidence, based mainly on observational data that this procedure results in a clinically significant reduction in blood pressure at 6 months. Weaker evidence suggests that the effect is sustained up to 2 years of follow-up. Some side-effects, none unmanageable or permanent, are reported.

It is recommended that this technology receive temporary (two-year) and conditional approval for use only in the context of a formal research study to be supported by the manufacturer as specified."

1.3 Random sampling and Randomization

1.3.1 Sample surveys

- A sample survey is a type of observational study
- In a **sample survey** a subgroup of a larger population is studied. Ideally, we wish to use methods to draw a representative sample to avoid bias
- **Surveys** are preferred because they are less expensive and time consuming than a census (or complete enumeration of a population)

1.3.2 Simple random sample

- A **simple random sample** is a sample of n items in which
  - every member of the population has an equal chance of being included,
  - members are chosen independently from each other
- The word random does not mean haphazard. Rather, it refers to a well-defined process whose outcomes are not fixed but are determined by a probability distribution

1.3.3 Sample surveys*

- Interestingly, if you use commonly accepted methods, a sample of size 1500 would be adequate to gauge the percentage of a population who have a certain trait or opinion to within $\pm 3\%$
- Further, this result does not depend on the size of the population. A sample size of 1500 is adequate whether the population size is 10 million or 4 billion, as long as a proper sampling technique has been used

1.3.4 Margin of error

- An obvious question is: how close is a sample estimate to the true value?
- The central limit theorem (which we will study in Lecture 3) we know that the margin of error around the sample mean is proportional to $\frac{\sigma}{\sqrt{n}}$, where σ is the standard deviation and n is the sample size

1.3.5 How to choose a simple random sample

- Create a sampling frame by listing all members of the population
- Find a method to randomly select from among these
  - e.g. a physical method, e.g. placing the names of members of the population in an opaque bowl and drawing the required number
  - e.g. a virtual method with a computer, e.g. using the sample() function in R
- The chosen members constitute the sample

1.3.6 Example: Drawing a random sample

- A respiratory researcher wants to estimate the amount of inflammation in the parenchyma of a mouse lung.
- She takes an image of a histological slide of the lungs of the mouse with staining of the inflammatory cells of interest.
- She divides the images in a grid of 100 rectangular areas, but excludes 10 areas because they include airways.
- She then counts the number of inflammatory cells in 40 areas randomly selected out of the remaining 90 areas
- What was the sampling frame in this study, and how did it differ from the population of interest?
- Explain why "using the wrong sampling frame" might lead to a biased estimate.
- Use R to propose to the researcher which rectangular areas she needs to study.

1.3.7 Practical concerns when random sampling

- For practical reasons, it may not be possible to obtain a simple random sample because it may not be possible to enumerate the entire population
  - e.g. how would we enumerate the population of people who need to be screened by Biliscreen?

- Then, it would be important to identify the population, and scrutinize the method of selection to ensure that the resulting sample satisfies the definition of a simple random sample
- Other sampling techniques such as **cluster sampling** or **stratified random sampling** may be easier to implement

1.3.8 Some typical biases that can arise during a survey

- **Selection bias**: Due to selecting non-representative sample
- **Non-response (or missingness bias)**: Arises when a representative sample was chosen but a subset could not or did not provide responses, e.g. a survey conducted during the evening would miss individuals who were working at that time
- **Response bias**: Occurs when participants respond differently from how they feel, e.g. response to sensitive questions such as smoking habits

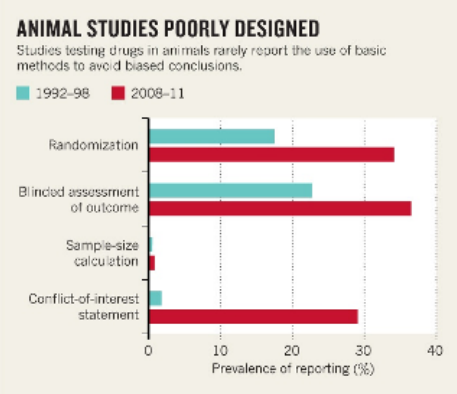
1.3.9 Randomization

- Random sampling can also be used in the context of an experiment, such that each subject has the same probability of receiving the different treatments under study
- Randomization ensures that any observed or unobserved confounding variables have a similar distribution in each treatment group

1.3.10 Simple randomization

- Like with random sampling, there are different techniques we can use to carry out randomization to a treatment group
- In **simple randomization**, subjects are assigned to groups based on a single sequence of random assignments
  - e.g. If there are two treatments, we can toss a coin to determine how to assign each patient recruited into the study (Heads – Treatment, Tails – Control)
  - Instead of a coin you can use a computer to generate the random sequence
- This method is suitable when the planned sample size is relatively large and the subjects to be sampled are relatively homogenous

1.3.11 Relevance of statistical methods to researchers in the life sciences



Nandini Dendukuri, McGill University

- Medical research is increasingly quantitative. Simultaneously, there is a move towards evidence-based medicine
- Statistical methods are necessary for designing and analyzing research studies that can answer relevant questions
- Knowledge of statistics is necessary for interpreting research publications

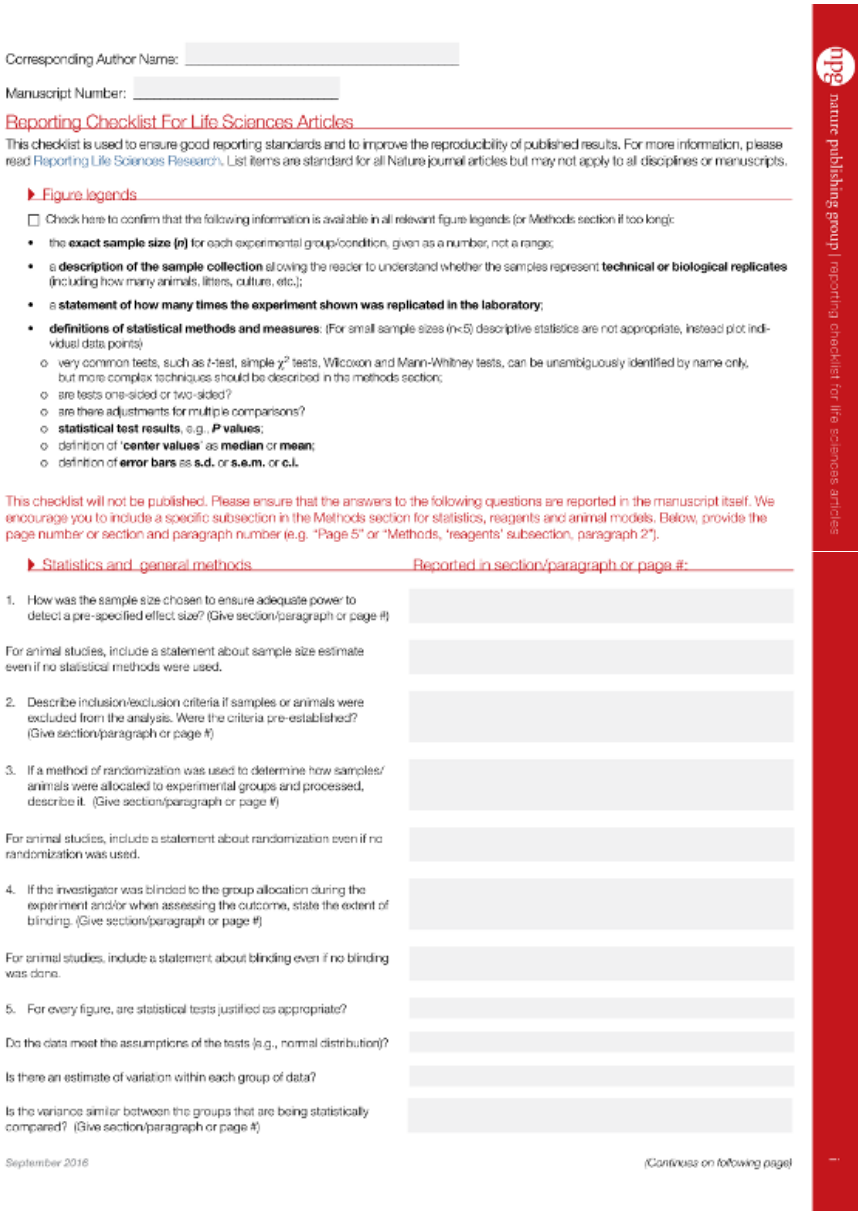
1.3.12 Organizations supporting transparent reporting of biomedical research & evidence-based decision making

Collaborative Approach to Meta Analysis and
·C·A·M·A·R·A·D·E·S·
Review of Animal Data from Experimental Studies

equator network

THE COCHRANE COLLABORATION®

1.3.13 Biomedical journals are insisting on appropriate statistical methods



Corresponding Author Name: _____

Manuscript Number: _____

Reporting Checklist For Life Sciences Articles

This checklist is used to ensure good reporting standards and to improve the reproducibility of published results. For more information, please read Reporting Life Sciences Research. List items are standard for all Nature journal articles but may not apply to all disciplines or manuscripts.

▸ Figure legends

☐ Check here to confirm that the following information is available in all relevant figure legends (or Methods section if too long):

- the **exact sample size (*n*)** for each experimental group/condition, given as a number, not a range;
- a **description of the sample collection** allowing the reader to understand whether the samples represent **technical or biological replicates** (including how many animals, litters, culture, etc.);
- a **statement of how many times the experiment shown was replicated in the laboratory**;
- **definitions of statistical methods and measures**: (For small sample sizes (n<5) descriptive statistics are not appropriate, instead plot individual data points)
  - very common tests, such as *t*-test, simple χ^2 tests, Wilcoxon and Mann-Whitney tests, can be unambiguously identified by name only, but more complex techniques should be described in the methods section;
  - are tests one-sided or two-sided?
  - are there adjustments for multiple comparisons?
  - **statistical test results**, e.g., ***P* values**;
  - definition of '**center values**' as **median** or **mean**;
  - definition of **error bars** as **s.d.** or **s.e.m.** or **c.i.**

This checklist will not be published. Please ensure that the answers to the following questions are reported in the manuscript itself. We encourage you to include a specific subsection in the Methods section for statistics, reagents and animal models. Below, provide the page number or section and paragraph number (e.g. "Page 5" or "Methods, 'reagents' subsection, paragraph 2").

▸ Statistics and general methods	Reported in section/paragraph or page #:
1. How was the sample size chosen to ensure adequate power to detect a pre-specified effect size? (Give section/paragraph or page #)	
For animal studies, include a statement about sample size estimate even if no statistical methods were used.	
2. Describe inclusion/exclusion criteria if samples or animals were excluded from the analysis. Were the criteria pre-established? (Give section/paragraph or page #)	
3. If a method of randomization was used to determine how samples/animals were allocated to experimental groups and processed, describe it. (Give section/paragraph or page #)	
For animal studies, include a statement about randomization even if no randomization was used.	
4. If the investigator was blinded to the group allocation during the experiment and/or when assessing the outcome, state the extent of blinding. (Give section/paragraph or page #)	
For animal studies, include a statement about blinding even if no blinding was done.	
5. For every figure, are statistical tests justified as appropriate?	
Do the data meet the assumptions of the tests (e.g., normal distribution)?	
Is there an estimate of variation within each group of data?	
Is the variance similar between the groups that are being statistically compared? (Give section/paragraph or page #)	

September 2016

(Continues on following page)

nature publishing group | reporting checklist for life sciences articles

1.3.14 FEV Example: Dataset

- The variables in the dataset include the following:

First few rows of FEV dataset					
id	age	fev	ht	sex	smoke
1	9	1.708	57.0	0	0
2	8	1.724	67.5	0	0
3	7	1.720	54.5	0	0
4	9	1.558	53.0	1	0
5	9	1.895	57.0	1	0
6	8	2.336	61.0	0	0

- fev (in liters)
- age (in years)
- height (in inches)
- gender (M/F)
- smoke (Y/N)

Chapter 2

Lecture 2: Types of Variables, Probability and Probability Distributions

2.1 Types of variables

- A variable is a characteristic of a person or a thing that can be assigned a number or a category
  - Age and sex are two variables that can be measured on a person
- Variables can be of different types

Qualitative	Quantitative
Nominal	Continuous
Ordinal	Discrete/Ordinal

- We need to distinguish between different types of variables because the statistical methods employed – whether descriptive or inferential - to study them depend on the type of variable we have studied

2.1.1 Some questions on types of variables

- Click on this link to answer a few questions on types of variables

2.1.2 Qualitative variables

- Qualitative variables are categorical and not measured on a numerical scale
  - Nominal variables do not have a particular ordering
    - Blood type of a person: A, B, AB, O
    - Sex of a fish: male, female
    - Research interest of students in EXMD 634: Cell Biology, Immunology, Cancer, …
- Qualitative variables are categorical and not measured on a numerical scale
  - Ordinal variables do have a particular ordering, but the gap between successive categories is not measureable and may not be equal
    - Likert-type scale:

Strongly agree	Agree	Neutral	Disagree	Strongly disagree

- Age in categories: + Infants, Toddlers, Gradeschoolers, Adolescents * Quantitative variables are measured on a numerical scale that allows us to measure the interval between any two observations + Continuous variables can take decimal values - Age of a patient - Cholesterol concentration in a blood specimen - Optical density of a solution + Discrete/Ordinal variables are reported as integers - Number of bacteria colonies in a petri dish - Number of cancerous lymph nodes detected in a patient - Length of a DNA segment in basepairs * The distinction between continuous and discrete variables is not a rigid one as measurements can be rounded off. e.g. age or birth weight can be reported as integers * In practice, if the number of unique integer values observed is small (say <10), then we would treat the quantitative variable as discrete/ordinal

2.2 Probability

- The conclusions of a statistical data analysis are often stated in terms of probability
- Probability models allow us to quantify how likely, or unlikely, an experimental result is, given certain modeling assumptions
- We will first look at probability and probability distributions for dichotomous and discrete variables, before proceeding to continuous variables

2.2.1 Definitions

- A probability is a numerical quantity that expresses the likelihood of an event

- The probability of an event E may be written as P(E)
- P(E) is always a number between 0 and 1, inclusive. May also be expressed as a percentage
- The higher the probability, the more certain we are that the event will occur
- We can speak meaningfully about a probability P(E) only in the context of a **chance operation** or a **chance experiment**—that is, an operation whose outcome is not pre-determined
- The chance operation must be defined in such a way that each time the chance operation is performed, the event E either occurs or does not occur. We refer to the event that E does not occur as E complement (E^c)
- The sample space enumerates all the possible events that a chance experiment gives rise to. The sum of their probabilities is 1

2.2.2 Example 1: Coin Tossing

- Consider the familiar chance operation of tossing a coin. The sample space is {Heads, Tails}. That means each time the coin is tossed, either it falls heads or Tails
- Define the event E:Heads. If the coin is fair (i.e. equally likely to fall heads or tails), then P(E) = 1/2 = 0.5
- If the coin is not fair (perhaps because it is slightly bent), then P(E) will be some value other than 0.5, e.g. P(E) = 0.6, suggesting it is more likely to see a head than a tail

2.2.3 Example 2: Coin Tossing again

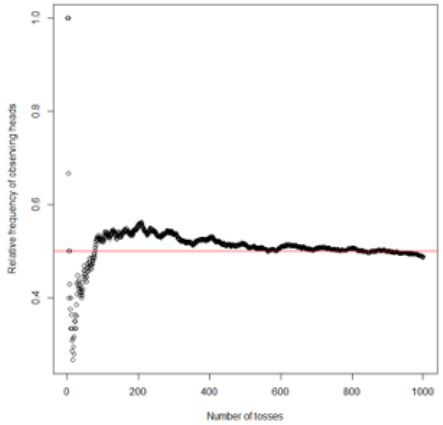
- Consider the event E: 3 heads in a row
- The chance operation that could give rise to this event is "Toss a coin 3 times"
- Notice that the sample space is now larger than when the operation was made of a single toss
- The sample space is now {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT} where H denotes Heads and T denotes Tails
- Assuming we have a fair coin, the probability of each of the 8 outcomes in the sample space is equally likely
- Therefore, P(E) = 1/8

2.2.4 Interpretation of probability

- How do we know P(Heads)=0.5 for a fair coin?
- For one, we know that there are two possible events resulting from a coin toss both of which are equally likely

- Another interpretation arises when a chance operation can be observed repeatedly. Then P(E) can be interpreted as the relative frequency of occurrence of E in an indefinitely long series of repetitions of the chance operation

2.2.5 Relative frequency interpretation of probability



Number of tosses	Outcome	Cumulative number of heads	Relative frequency of heads
1	H	1	1.000
2	H	2	1.000
3	T	2	0.670
4	T	2	0.500
5	T	2	0.400
10	H	4	0.400
200	H	111	0.555
500	H	257	0.514
750	H	378	0.504
1000	T	487	0.487

2.2.6 Subjective interpretation of probability

- It is not always possible to observe events repeatedly. In such cases, probability may be used to represent a subjective or personal degree of belief

e.g. There is an 80% chance it will rain tomorrow

e.g. It is believed that culture for M. tuberculosis in children has a <2% chance of being falsely positive

- There are very few restrictions place on personal probabilities besides that they must be coherent
  - e.g. If you say there is an 80% chance it will rain tomorrow, then you should also agree that there is a 20% chance it will not rain tomorrow
- Different individuals can have different personal probabilities, and may not necessarily agree,
  - e.g. members on a job interview committee may different views on the probability of a client being suitable

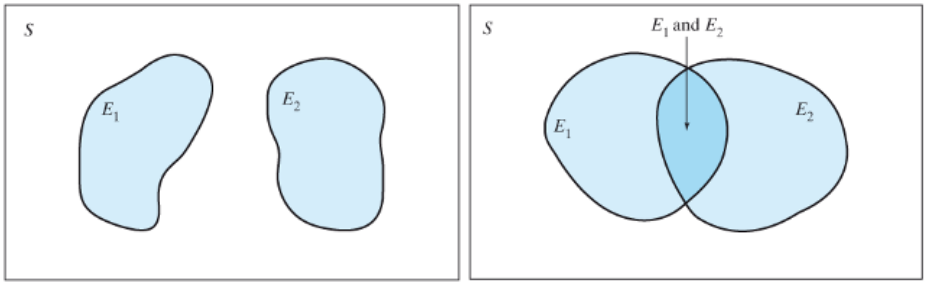
2.2.7 Compound events

- A compound event is defined by the joint occurrence of three simple events
  - e.g. Obtaining three heads on three successive tosses of a coin
  - e.g. Obtaining a true positive diagnostic test result for tuberculosis
- The different simple events may be independent or dependent
  - Each toss of a coin is independent of the previous toss
  - The probability of a positive result on a diagnostic test is dependent on whether the patient has the disease

2.2.8 Some questions on probability

- Click on this link to answer a few questions on probability

2.2.9 Combining probabilities: Addition rules



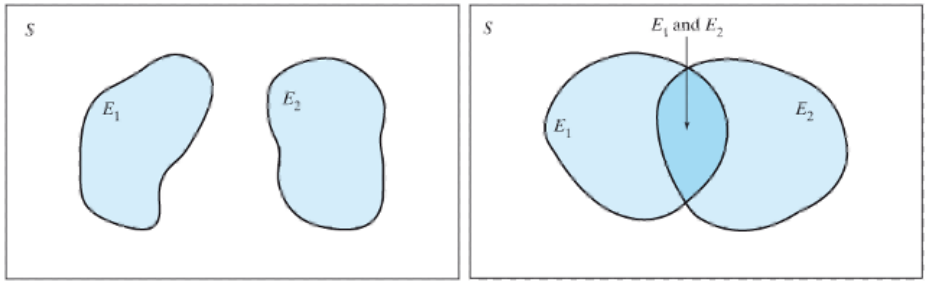
- When two events are independent (top panel) then

$$P\{E1 or E2\} = P\{E1\} + P\{E2\}$$

- When two events are dependent (bottom panel)

$$P\{E1 or E2\} = P\{E1\} + P\{E2\} - P\{E1 and E2\}$$

2.2.10 Combining probabilities: Multiplication rules



- When two events are independent (top panel) then

$$P\{E1 and E2\} = P\{E1\} \times P\{E2\}$$

- When two events are dependent (bottom panel)

$$P\{E1 and E2\} = P\{E1\} \times P\{E2|E1\} = P\{E2\} \times P\{E1|E2\}$$

2.2.11 Conditional probability

- P(E2|E1) is the conditional probability of E2 given E1, it is interpreted as the probability of observing E2 given that E1 has occurred

$$P(E2|E1) = \frac{P(E_1 and E_2)}{P(E_1)}$$

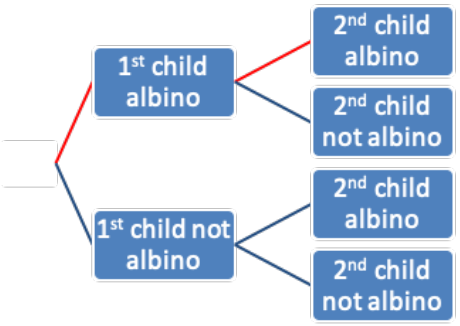
2.2.12 Probability Trees

- Often it is useful to depict a probability problem using a probability tree
- The following slides depict how we can enumerate the events in the sample space that arises from independent events, and how we can then calculate the corresponding probabilities of each event using probability rules

2.2.13 Example 3: Independent events

- If two carriers of the gene for albinism marry, each of their children has probability 1/4 of being albino.
- The chance that the second child is albino is the same (1/4) whether or not the first child is albino; similarly, the outcome for the third child is independent of the first two, and so on.
- We can use a probability tree to enumerate the sample space and corresponding probabilities

2.2.14 Albinism example



- Suppose two carriers of the gene for albinism marry and have two children. Then the probability that both of their children are albino is

$$P\{AA\} = 0.25 \times 0.25 = 0.0625$$

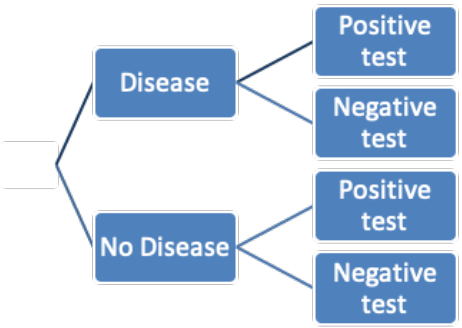
2.2.15 Albinism example: Sample space and probabilities

Number of Albino children	Probability
2	$0.25^2 = 0.0625$
1	$2 \times 0.25 \times 0.75 = 0.375$
0	$0.75^2 = 0.5625$
Total	1

2.2.16 Example 4: Medical testing

- The following is based on a scenario a statistician, David Eddy, (1982) posed to a 100 physicians* (see full text here):
- One of your patients has a lump in her breast. You are almost certain that it is benign, and believe there is only a 1% chance that it is malignant. Just to be sure you have the patient undergo a mammogram. Sadly for your patient the mammogram is positive.
- Suppose that the mammogram has the following characteristics
  - P{Testing positive | Person has disease} = Sensitivity = 80%
  - P{Testing negative | Person does not have disease} = Specificity = 90%
- What is the probability that a randomly chosen woman will test positive on a mammogram? What are the chances the lump is truly malignant?

2.2.17 Medical testing example



- The sample space is {D+T+, D+T-, D-T+, D-T-}
- The probability of a positive test

= P{true positive} + P{false positive}

= Pr{D+T+} + P{D-T+}

= $0.01 \times 0.8 + 0.99 \times 0.1$

= 0.107

- Probability that a person truly has the disease given they are positive

$= P\{D+ \mid T+\}$

$= P\{D+ \text{ and } T+\}/P\{T+\}$

$= (0.01 \times 0.8)/(0.01 \times 0.8 + 0.99 \times 0.1)$

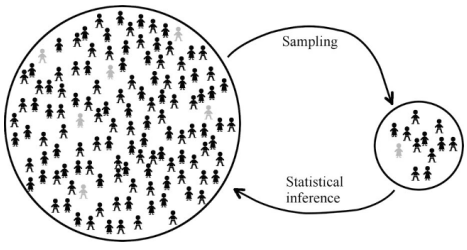
$= 0.075$

The above expression is formally referred to as Bayes Theorem

2.3 Probability Distributions

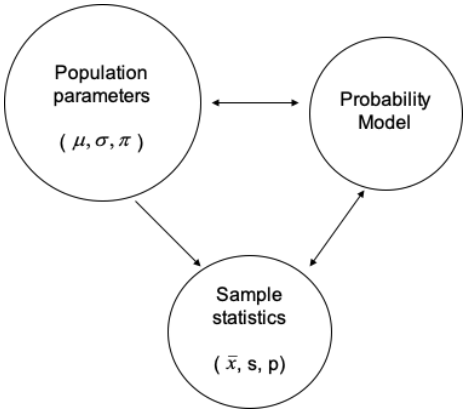
- A **probability distribution** is a mathematical function that provides the probabilities of occurrence of different possible values of a random variable
  - It follows the probability rules we studied earlier, e.g. the sum of the probabilities of all possible values of a random variable is 1
- A very large number(100s?) of probability distributions have been described – but we tend to use a much smaller number in common applications

2.3.1 Population and sample*



*From text by Ekstrom and Sorensen

2.3.2 Notation



2.3.3 Parameters, Statistics, Probability Distributions

- A **parameter** is a number that describes the **population**. A parameter is a fixed number; but in practice we do not know its value.
- A **statistic** is a number that describes a **sample**. The value of a statistic is known when we have taken a sample, but it can change from sample to sample. We often use a statistic to estimate an unknown parameter
- The **probability distributions** we will study in this lecture are examples of **probability models** that help us to make inference about the population based on observed statistics

2.3.4 Binomial Distribution

- Both the coin toss and the albinism examples were examples of random variables following a Binomial distribution. These variables are characterized by:
  - **Binary outcomes**: There are two possible outcomes for each trial (success and failure).
  - **Independent trials**: The outcomes of the trials are independent of each other.
  - **n is fixed**: The number of trials, n, is fixed in advance.
  - **Same value of π**: The probability of a success on a single trial is the same for all trials.

2.3.5 Binomial Distribution Function

Whereas both examples we looked at had n=2 trials, and were easy to illustrate with a probability tree, we can write a more general expression for the

probability of k successes in n independent trials as follows

$$Pr\{k|n,\pi\} = \frac{n!}{k!(n-k)!}\pi^k(1-\pi)^{n-k}$$

2.3.6 Albinism example for a couple with 5 children: Sample space and probabilities

Number of Albino children	Probability expression	Probability rounded value
0	$(1-\pi)^5$	0.24
1	$5\pi(1-\pi)^4$	0.40
2	$10\pi^2(1-\pi)^3$	0.26
3	$10\pi^3(1-\pi)^2$	0.09
4	$5\pi^4(1-\pi)$	0.01
5	π^5	0.00

These probabilities may be obtained from R with the following command:

dbinom(x=seq(0,5),size=5,prob=0.25)

2.3.7 Probability distributions in R

- One of the main advantages of R is that it has several functions related to statistical probability distributions that can be used to:
  - Obtain a random sample from a distribution
  - Calculate the density function
  - Calculate the cumulative probability
  - Obtain quantiles of the distribution

Distribution	Random Sample	Density function	Cumulative probability function	Quantiles
Binomial	rbinom	dbinom	pbinom	qbinom
Gamma	rgamma	dgamma	pgamma	qgamma
Poisson	rpois	dpois	ppois	qpois
Normal	rnorm	dnorm	pnorm	qnorm

2.3.8 Example: Binomial distribution in practice*

- The assumptions behind the use of the Binomial distribution may not always be satisfied in practice. For example:
- Let X represent the number of females in four children, among all couples in Canada with exactly four children

- The "Real World" data and the data predicted by a Binomial distribution model with n=4 and $\pi = 0.5$ are given on the following page, i.e. the predicted proportion was given by

$$Pr\{X = k\} = \frac{4!}{k!(4-k)!}\pi^k(1-\pi)^{4-k}$$

X	Predicted proportion obtained using dbinom in R	Observed proportion
0	0.0625	0.08
1	0.2500	0.26
2	0.3750	0.31
3	0.2500	0.27
4	0.0625	0.08

- Why do you think the observed values differ (slightly) from those predicted by a Binomial model?
- Which assumptions of the Binomial model may be violated here?

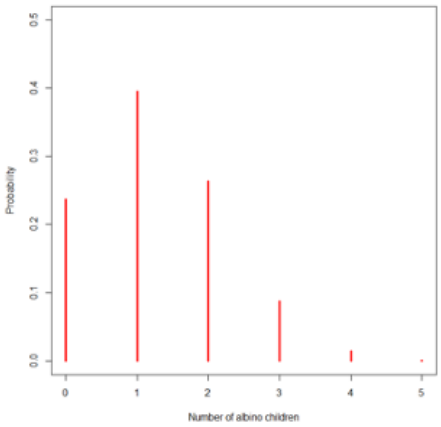
*From Lawrence Joseph's notes

2.3.9 Mean and variance of random variables

Let X be a discrete random variable taking values $\{x_1, x_2, ..., x_n\}$ with probabilities $\{p_1, p_2, ..., p_n\}$, respectively

- Population mean (or expectation) of a discrete random variable = $E(X) = \sum_{i=1}^{n} x_i p_i$
- Population variance of a discrete random variable = $\sum_{i=1}^{n}(x_i - E(X))^2 p_i$

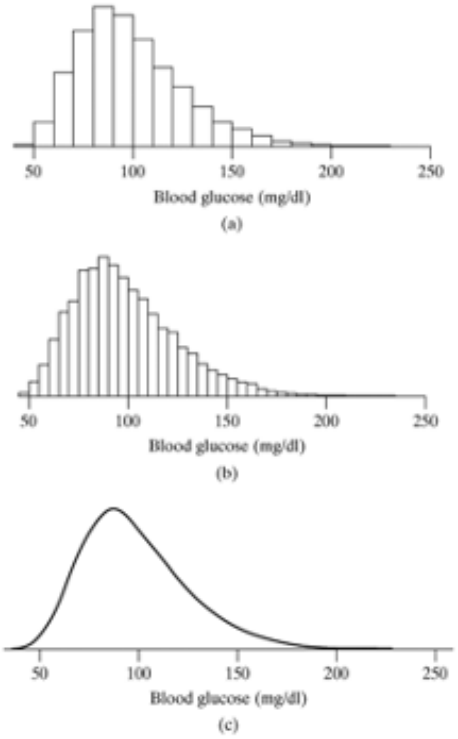
2.3.10 Mean and variance for a Binomial distribution



- Let X be a Binomial variable with n trials and probability of success π
- $E(X) = n\pi = 5 \times 0.25 = 1.25$
- $Var(X) = n\pi(1-\pi) = 5 \times 0.25 \times 0.75 = 0.9375$

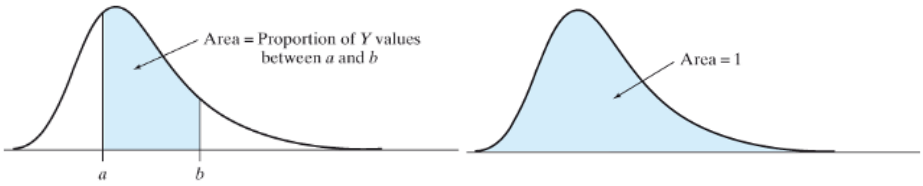
Therefore, standard deviation of X = 0.968

2.3.11 Probability of a continuous variable



- We can think of the relative frequency histogram of a continuous variable as an approximation of the underlying true population distribution from which the data came.
- A smooth curve representing a frequency distribution is called a probability density function
- On the x-axis we have different possible values of the variable (i.e. the sample space). On the y-axis we have the probability density corresponding to each value of the variable.

2.3.12 Probability density function for a continuous variable



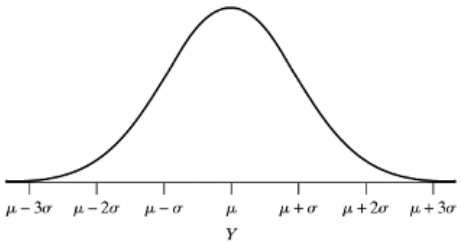
- If a variable is continuous, then we find probabilities by using the density curve for the variable.
- The probability that a continuous variable lies in a certain range equals the area under the density curve for the variable between two points
- This means the probability of a single value, say Pr{Y=a}=0. But the $Pr\{a - \delta < Y < a + \delta\}$, where δ is an infinitesimal quantity is non-zero and equal to the height of the density function at Y=a.
- The area under the entire curve is 1

2.3.13 Normal Distribution

- The most well know continuous distribution is the Normal (or Gaussian) distribution that is recognizable by its characteristic bell shape
- Probability density function of a continuous variable Y that follows a normal distribution with mean μ and variance σ^2

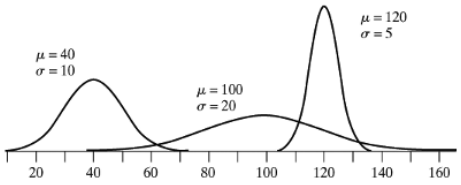
$$f(y|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right), -\infty < y < \infty$$

2.3.14 Normal probability density function

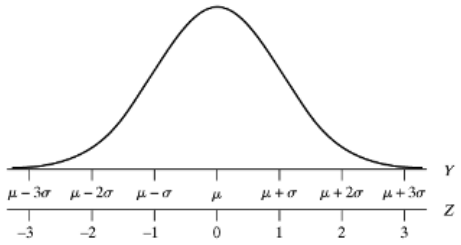


- The density function is symmetric about the mean μ (which also happens to be the median and mode of this distribution
- Though it is defined all the way from $-\infty$ to ∞, most of the probability lies in the range $\mu \pm 3\sigma$

2.3.15 Three normal curves with different means and standard deviations



2.3.16 Area under the normal curve



- $Pr\{a < Y < b\} = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$
- The values of the areas under the standard normal distribution (denoted $N(\mu = 0, \sigma^2 = 1)$) were typically published as tables in statistics books
- Today you can use a program like R to calculate this integral

2.3.17 Example: Distribution of serum cholesterol values

- The serum cholesterol levels of 12- to 14-year-olds follow a normal distribution with mean 155 mg/dl and standard deviation 27 mg/dl. What percentage of 12 to 14-year-olds have serum cholesterol values

a) 137 or less?

b) 186 or less?

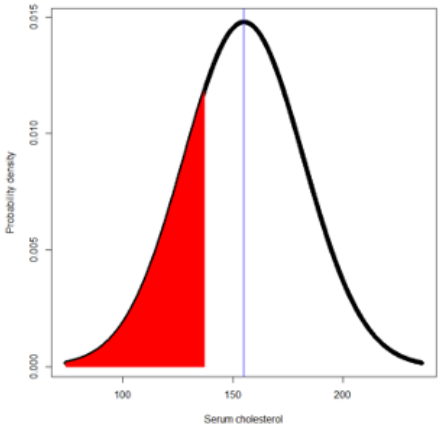
c) 164 or more?

d) 100 or more?

e) between 159 and 186?

f) between 100 and 132?

g) between 132 and 159?



Let Y denote the variable serum cholesterol.

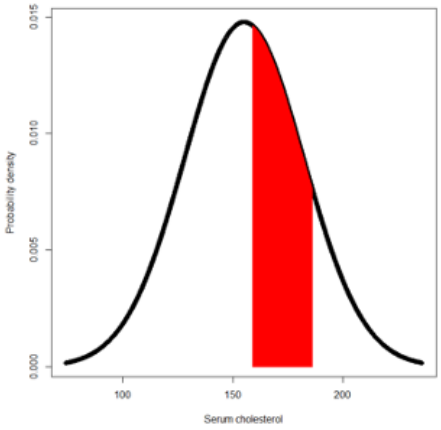
We know the distribution is symmetric about 155mg/dl and that most values of Y will lie between (74, 236)

a) P(Y 137) = pnorm(q=137,mean=155,sd=27) = 0.252

b) P(Y 164) = 1 – Pr(Y < 164) = 1 – pnorm(q=164,mean=155,sd=27) = 0.369

c) P(Y 186) = 0.875

d) P (Y 100) = 1 – 0.02 = 0.98



e) between 159 and 186?

P(159 Y 186) = P(Y 186) – Pr(Y 159) = pnorm(186,155,27) - pnorm(159,155,27)

= 0.875 – 0.559 = 0.316

Another way to answer this question is via the z-transformation of Y into a standard normal variable with mean=0 and standard deviation=1

P(159 Y 186) = P((159-155)/27 (Y-155)/27 (186-155)/27)

= Pr(0.148 Z 1.148) = Pr(Z 1.148) – Pr(Z 0.148)

= pnorm(1.148) - pnorm(0.148) = 0.875 – 0.559 = 0.316

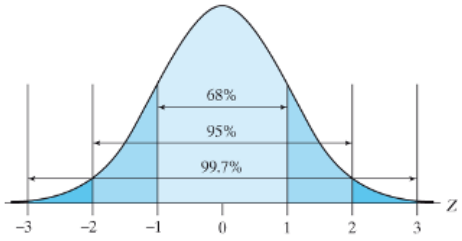
e) between 100 and 132?

P(100 Y 132) = 0.176

f) between 132 and 159?

P(132 Y 159) = 0.362

2.3.18 Area under the normal curve



- We can show that the probability of lying within 1 standard deviation of the mean is 0.68, within 2 standard deviations is 95% and within 3 standard deviations is 99.7%

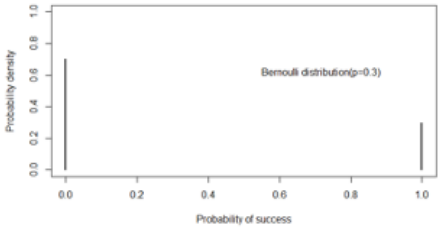
2.3.19 Mean and variance of the normal distribution

- The expressions for the mean and variance are similar to those for discrete variables, except that the summation sign is replaced by an integral sign
- $Expectation(Y) = \int_{-\infty}^{\infty} \frac{y}{\sigma\sqrt{2\pi}} exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) dy = \mu$
- $Variance(Y) = \int_{-\infty}^{\infty} \frac{(y-\mu)^2}{\sigma\sqrt{2\pi}} exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) dy = \sigma^2$

2.3.20 Examples of discrete distributions

- Bernoulli distribution
- Binomial distribution
- Poisson distribution
- Negative binomial distribution

2.3.21 Bernoulli distribution



- X is dichotomous
- Examples:
  - Single coin toss
  - Observation on an individual patient in a longitudinal study of survival following a treatment

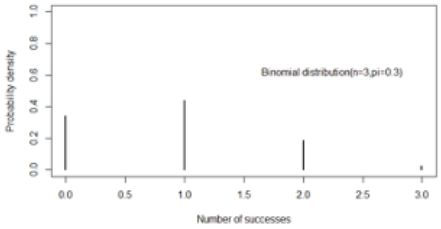
Probability density function

$$f(x|\pi) = \pi^x(1-\pi)^{1-x}, x = 0, 1$$

Mean $= \pi$

Variance $= \pi(1-\pi)$

2.3.22 Binomial distribution



- X is the sum of successes in n independent Bernoulli trials
- Examples:
  - three coin tosses
  - number of patients who will survive 1 year following a treatment

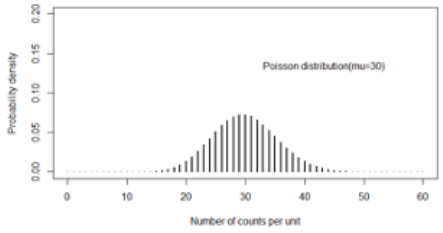
Probability density function

$$f(x|n,\pi) = \binom{n}{x}\pi^x(1-\pi)^{1-x}, x = 0, 1, ..., n$$

Mean = $n\pi$

Variance = $n\pi(1-\pi)$

2.3.23 Poisson distribution



- X takes discrete taking values 0, 1, to ∞ within a unit of time or space
- Examples
  - Number of downloads of an app in 1 minute
  - Number of cases of cancer reported in a square kilometre

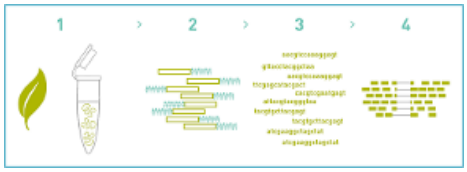
Probability density function

$$f(xevents|time = t, rate = \lambda) = \frac{exp^{-\mu}\mu^x}{x!}, x = 0, 1, ..., \infty$$

Mean = μ

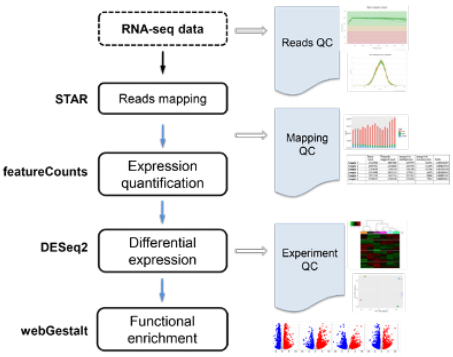
Variance = μ

2.3.24 Example: Transcriptomic Analyses*



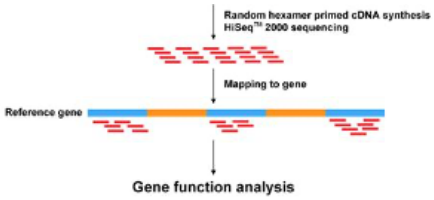
- RNA sequencing is a powerful and commonly used tool to analysis expression data
- The goal of most sequencing experiments is to identify differences in gene expression between biological conditions such as the influence of a disease-linked genetic mutation or drug treatment.

2.3.25 Underlying statistical principles of commonly used packages



2.3.26 How it works

- In a standard sequencing experiment (RNA-Seq), we map the sequencing reads to the reference genome and count how many reads fall within a given gene (or exon).

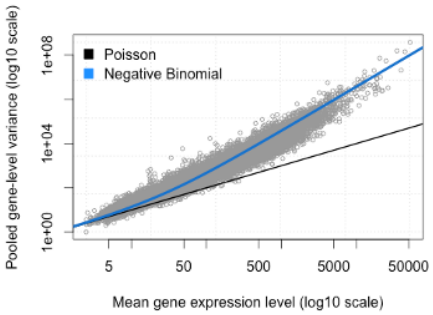


=> This means that the input for the statistical analysis are discrete non-negative integers ("counts") for each gene in each sample.

2.3.27 What would be a suitable probability distribution?

- The total number of reads for each sample tends to be in the millions, while the counts per gene vary considerably but tend to be in the tens, hundreds or thousands.
- The chance of a given read to be mapped to any specific gene is rather small.
- Discrete events that are sampled out of a large pool with low probability - sounds like a Poisson distribution would be suitable

Problem: The **variability of read counts** in sequencing experiments tends to be **larger than the Poisson distribution allows**.



- It is obvious that the variance of counts is generally greater than their mean, especially for genes expressed at a higher level. This phenomenon is called "**overdispersion**".
- The negative binomial distribution can model the greater variance

2.3.28 Poisson Distribution is limiting

- The Poisson distribution makes the restrictive assumption that the mean of the distribution is equal to its variance
- In terms of RNA-seq, Poisson distribution implies that for a certain gene, its expression profile follows a distribution with a mean expression equal to the variance in expression
- Empirical observations show that for highly expressed genes at least, this is not the case even in biological replicates
- Another degree of variation that removes this restriction

2.3.29 Negative Binomial Distribution

- The NB distribution is similar to a Poisson distribution but has an extra parameter ( ) called the "clumping" or "dispersion" parameter => **More variance**

$$\sigma^2 = \mu + \alpha\mu^2$$

- The NB distribution can be defined as a **Poisson-Gamma mixture distribution**
- This means that the NB distribution is a weighted mixture of Poisson distributions where the rate parameter (i.e. the expected counts) is itself associated with uncertainty following a Gamma distribution

2.3.30 Conceptual Justification

- When comparing samples of different conditions we usually have multiple independent replicates of each condition.
- Such replicates are called **"biological"** replicates because they come from independent animals, dishes, or cultures.
- Splitting a sample in two and running it through the sequencer twice would be a **"technical"** replicate.
- In general, there is more variance associated with biological replicates than technical replicates.
- As a result, the Poisson process in each biological replicate has a slightly different expected count parameter.

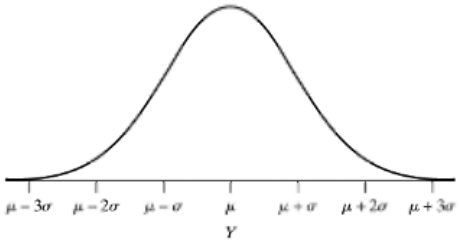
2.3.31 Additional Notes and Practical Implications

- In a standard sequencing experiments, we have to be content with few biological replicates per condition due to the high costs associated with sequencing experiments and the large amount of time that goes into library preparations.
- Modern RNA-Seq analysis tools such as DESeq2 and edgeR combine the gene-wise dispersion estimate with an estimate of the expected dispersion rate based on all genes.
- This **Bayesian "shrinkage"** of the variance has emerged as a powerful technique to mitigate the shortcomings of having few replicates.

2.3.32 Examples of continuous distributions

- Normal distribution
- Uniform distribution
- Student's t-distribution
- Gamma distribution
- Beta distribution

2.3.33 Normal distribution



- X is continuous and symmetrically distributed over a range that lies between $-\infty$ to ∞
- Example:
  - blood pressure
  - body mass index

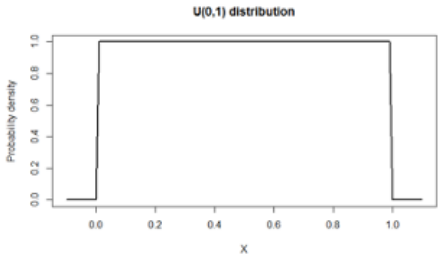
Probability density function

$f(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), -\infty < x < \infty$

Mean = μ

Variance = σ^2

2.3.34 Uniform distribution



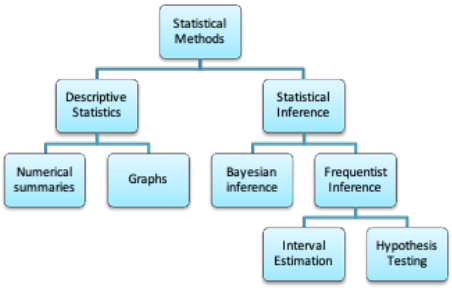
- X is continuous and equally likely to take values in the range (a,b)
- In the standard Uniform distribution, a=0, b=1
- Example:
  - X is a probability, such as disease prevalence or sensitivity of a test

Chapter 3

Lecture 3: Central Limit Theorem and Inference for Means

3.1 Mean and Standard Deviation

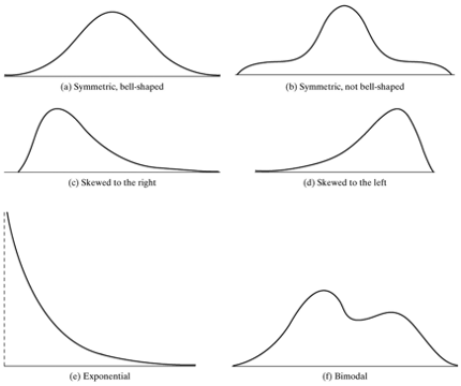
3.1.1 Descriptive statistics vs. Inferential Statistics



- Descriptive statistics help to describe the characteristics of the sample gathered

- Inferential statistics help to use these characteristics to draw conclusions about the target population

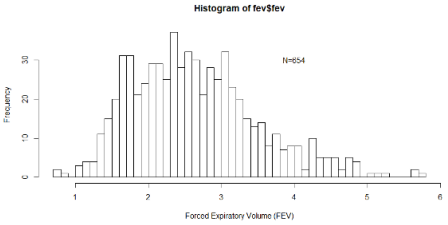
3.1.2 Some commonly encountered shapes of distributions of a variable



3.1.3 Descriptive statistics: Notation

- We use capital letters to denote a variable, and small letters to denote the values it takes. For example,
  - X = FEV (the variable),
  - x = 0.793 litres (an observed value)
- $\sum_{i=1}^{n} x_i$ means the sum of the observed values x on a sample of size n. x_i is the observed value for the i^{th} subject in the sample
- The next few slides list common measures of central tendency and spread

3.1.4 Histogram of FEV



3.1.5 Measures of central tendency

Measure	Definition
Mean or average	$(\frac{\sum_{i=1}^{n} x_i}{n})$
Median or 50% quantile	Quantile, below which 50% of values lie • When n is odd, it is the 0.5(n+1)th smallest value • When n is even, it is the average of the 0.5nth and the 0.5(n+2)th values
Mode	The most common value. May be difficult to determine from a frequency distribution

3.1.6 Summary of FEV variable

Measure	Definition
Mean FEV	$(\frac{\sum_{i=1}^{n} x_i}{n})$ = (1.708+1.724+ ... 3.211)/654 = **2.637 litres**
Median FEV	• Sort the values of FEV from 0.793 to 5.793 • The middle two values are the 327th and the 328th values, 2.545 and 2.550 • The median is their average, **2.548 litres**
Mode	The most common value is **3.082 litres**

For a symmetric distribution, the median=mean.

The values above suggest that the distribution of FEV may be slightly skewed to the right as the mean is higher than the mode

3.1.7 Robustness

- A statistic is said to be **robust** if the value of the statistic is relatively unaffected by changes in a small portion of the data, even if the changes

are dramatic ones. The median is a robust statistic, but the mean is not robust because it can be greatly shifted by changes in even one
- **Example:** In the FEV dataset, I replaced the last observation in the dataset of 3.211 by 6.211, an extreme value. This resulted in increasing the mean from 2.637 to 2.641 but the median remained at 2.548
- If the frequency distribution is skewed, both measures are pulled toward the longer tail, but the mean is usually pulled farther than the median

3.1.8 Mean vs. Median

- In some situations the mean makes very little sense. Suppose, for example, that the observations are survival times of cancer patients on a certain treatment protocol, and that most patients survive less than 1 year, while a few respond well and survive for 5 or even 10 years. In this case, the mean survival time might be greater than the survival time of most patients; the median would more nearly represent the experience of a "typical" patient. Note also that the mean survival time cannot be computed until the last patient has died; the median does not share this disadvantage. Situations in which the median can readily be computed, but the mean cannot, are not uncommon in bioassay, survival, and toxicity studies
- An advantage of the mean is that in some circumstances it is more efficient than the median. Efficiency is a technical notion in statistical theory; roughly speaking, a method is efficient if it takes full advantage of all the information in the data. Partly because of its efficiency, the mean has played a major role in classical methods in statistics

3.1.9 Quantiles

- Quantiles (also known as percentiles) help to demarcate different points of the distribution of a continuous variable
- The q% quantile is the number below which q% of observed values lie
- For example

  - The 10% quantile of FEV is the value below which 10% of FEV values lie = 1.612
    = $0.1n^{th}$ lowest value of FEV

3.1.10 Measures of spread

Measure	Definition
Inter-quartile range (IQR)	Middle 50% of the distribution. Difference between the 75% quantile (or 3rd quartile) and the 25% quantile (or 1st quartile) $IQR = Q_3 - Q_1$
Range	Minimum and maximum values
Variance	$\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
Standard deviation	$\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

3.1.11 Summary of FEV variable

Measure	Definition
Inter-quartile range (IQR)	$IQR = Q_3 - Q_1$ $= 3.1185 - 1.9810 = 1.1375$ litres
Range	Minimum = 0.791 litres Maximum = 5.793 litres
Variance	$\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = 0.752$
Standard deviation	$\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = 0.867$

3.1.12 Comparison of measures of spread

Range	Limited in the sense that it pertains only to the extremes. Can't tell where the majority of the sample is concentrated
Inter-quartile range	Advantage is that it is robust as it is based on quantiles. Particularly useful for skewed or multimodal distributions
Standard deviation (SD) and variance	Particularly suited for symmetric, unimodal distributions. Based on the properties of the normal distribution, which is the most commonly encountered distribution in practice, we can say that: • ~68% of the distribution will fall within $\bar{x}$± SD • ~95% of the distribution will fall within $\bar{x}$±2*SD • >99% of the distribution will fall within $\bar{x}$±3*SD
	However, like the mean, the standard deviation and variance are susceptible to extreme values and therefore not preferred for skewed distributions

3.1.13 Variance and Standard Deviation

- The standard deviation is more commonly reported than the variance because it is in the same units as the variable X and the mean of X
- Notice that we use the sum of the squared deviations. This is because the sum of the deviations themselves will always be 0. We need a way to get rid of the signs of the deviations. Alternatives to taking the squares include taking the absolute value. But squares are more popular because of their mathematical properties
- Why do we divide by n-1 rather than n? We do so because we are measuring the deviation from a quantity that is also defined using the sample, i.e. $\bar{x}$. It is as if we must penalize the sample size to correct for this. If we knew the true population mean (μ), then we would divide by n instead:

$$Populationvariance = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

3.1.14 Why n-1 rather than n?*

- Suppose the population has only 4 members {1,2,3,4}
  - The true mean is $\frac{1+2+3+4}{4} = 2.5$
  - The true variance is $\frac{(1-2.5)^2+(2-2.5)^2+(3-2.5)^2+(4-2.5)^2}{4} = 1.25$

- Now suppose we cannot view the whole population, but instead take a sample of size two. On the next slide, all possible samples are listed together with mean, the correct calculation for the sample variance dividing by n-1 and the incorrect calculation dividing by n. Each sample is equally likely to occur, assuming we are sampling with replacement from the population
- Notice that the incorrect expression for the sample variance results in an underestimate on the average across samples

Sample	Sample mean	Correct Sample variance	Underestimated Sample variance
(1,2)	1.5	0.50	0.250
(1,3)	2.0	2.00	1.000
(1,4)	2.5	4.50	2.250
(2,3)	2.5	0.50	0.250
(2,4)	3.0	2.00	1.000
(3,4)	3.5	0.50	0.250
(2,1)	1.5	0.50	0.250
(3,1)	2.0	2.00	1.000
(4,1)	2.5	4.50	2.250
(3,2)	2.5	0.50	0.250
(4,2)	3.0	2.00	1.000
(4,3)	3.5	0.50	0.250
(1,1)	1.0	0.00	0.000
(2,2)	2.0	0.00	0.000
(3,3)	3.0	0.00	0.000
(4,4)	4.0	0.00	0.000
Average across samples	2.5	1.25	0.625

*Lawrence Joseph's notes

3.2 Central Limit Theorem

3.2.1 Example 1: Serum cholesterol in children

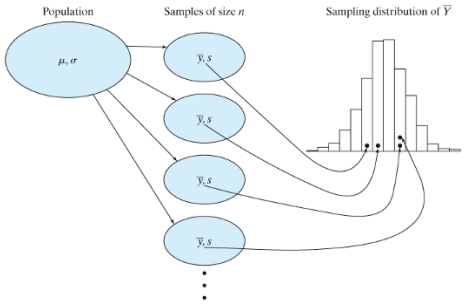
- Though we are more conscious of the relationship between cholesterol level and heart disease in adults, high levels of cholesterol are also a concern in children, particularly if they have risk factors like family history or obesity
- The American Academy of Pediatrics now recommends cholesterol testing in certain age groups
- To determine if a child is at risk of heart disease, we would need to compare the observed cholesterol level with the standard expected in a normal child. How large a sample size do we need to determine the normal level?
- The serum cholesterol levels (Y) of 12- to 14-year-olds follow a normal distribution with mean $=155$mg/dl and standard deviation $=27$ mg/dl

- You wish to estimate the true mean serum cholesterol in this population by using a sample of observations:
  - Should you prefer a sample of n=10, 30 or 100 observations?

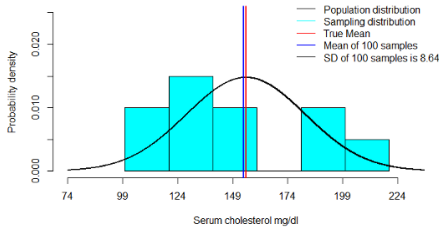
Sample size (n)	Sample mean ($\bar{y}$)	Sample standard deviation (s)
10	151.88	25.65
30	162.38	27.85
100	161.76	27.08

3.2.2 The sampling distribution of $\bar{Y}$

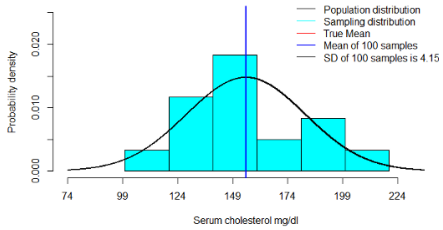
- The sample mean can be used, not only as a description of the data in the sample, but also as an estimate of the population mean .
- It is natural to ask, "How close to is $\bar{y}$?" We cannot answer this question for the mean $\bar{y}$ of a particular sample, but we can answer it if we think in terms of the random sampling model and regard the sample mean as a random variable $\bar{Y}$.
- The question then becomes: "How close to is $\bar{Y}$ likely to be?" and the answer is provided by the **sampling distribution of** $\bar{Y}$ - that is, the probability distribution that describes sampling variability in $\bar{Y}$
- In order to visualize the sampling distribution of $\bar{Y}$, imagine repeated samples of size n are drawn from a population with fixed mean μ and standard deviation . The variation of the $\bar{y}'s$ among the samples is specified by the sampling distribution of $\bar{Y}$



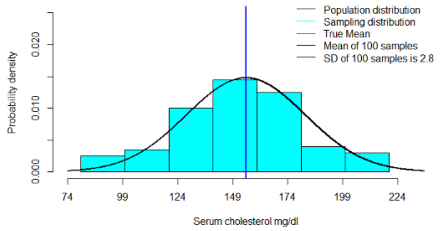
3.2.3 Example: Sampling distribution of $\bar{Y}$ when n=10



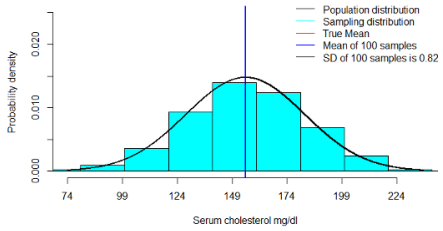
3.2.4 Example: Sampling distribution of $\bar{Y}$ when n=30



3.2.5 Example: Sampling distribution of $\bar{Y}$ when n=100



3.2.6 Example: Sampling distribution of $\bar{Y}$ when n=1000



3.2.7 Example 1

- We notice that the mean of the sampling distribution gets very close to μ even with smaller sample sizes. This only improves as n increases
- As n increases, there is a very clear decrease in the standard deviation of the means across a 100 samples
- Finally, we notice that the shape of the sampling distribution is increasingly like a normal distribution as n increases

3.2.8 The sampling distribution of $\bar{Y}$

- **Mean**: The mean of the sampling distribution of $\bar{Y}$ is equal to the population mean, i.e. $E(\bar{Y}) = \mu_{\bar{Y}} = \mu$
- **Standard deviation**: The standard deviation of the sampling distribution is equal to the population standard deviation divided by the square root of the sample size, i.e. $SD(\bar{Y}) = \sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$. Note that this implies the $Variance(\bar{Y}) = \sigma^2_{\bar{Y}} = \frac{\sigma^2}{n}$
- **Shape**
  - If the population distribution of Y is normal, then the sampling distribution is normal, regardless of the sample size n.
  - *Central Limit Theorem*: If n is large, then the sampling distribution is approximately normal, even if the population distribution of Y is not normal

3.2.9 Central Limit Theorem

- From the text by Moore and McCabe:

"The sampling distribution of $\bar{Y}$ is normal if the underlying population itself is normal.

What happens when the population distribution is not normal? It turns out that as the *sample size increases, the distribution of* $\bar{Y}$ *becomes closer to a normal distribution.* This is true no matter what the population distribution may be, as long as the population has a finite standard deviation . This famous fact of probability theory is called the *central limit theorem.* For large sample size n, we can regard $\bar{Y}$ as having the $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ distribution"

3.2.10 Example 1

- Applying the Central Limit Theorem, we can say that the sampling distribution of the mean serum cholesterol is:
  - $N\left(\mu_{\bar{Y}} = 155, \sigma_{\bar{Y}} = \frac{27}{\sqrt{10}} = 8.54\right)$ when n=10
  - $N\left(\mu_{\bar{Y}} = 155, \sigma_{\bar{Y}} = \frac{27}{\sqrt{30}} = 4.93\right)$ when n=30
  - $N\left(\mu_{\bar{Y}} = 155, \sigma_{\bar{Y}} = \frac{27}{\sqrt{100}} = 2.7\right)$ when n=100
  - $N\left(\mu_{\bar{Y}} = 155, \sigma_{\bar{Y}} = \frac{27}{\sqrt{1000}} = 0.85\right)$ when n=1000

Therefore, applying the rules pertaining to the normal distribution, we know that roughly 95% of the sampling distribution lies in the following ranges depending on the size of n:

Sample size	Range covering 95% of sample means
10	(155-2×8.54, 155+2×8.54) = (137.92, 172.08)
30	(145.14, 164.86)
100	(149.6, 160.4)
1000	(153.3, 156.7)

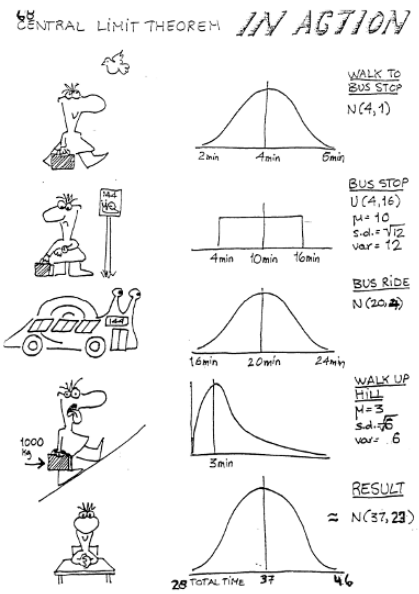
3.2.11 Theory related to the sums of random variables

- These two slides help to see how $\frac{\sigma}{\sqrt{n}}$ arises.
- Let X and Y be two arbitrary, independent random variables. Then from probability theory we know that:
  - E(X+Y) = E(X) + E(Y)
  - Var(X + Y) = Var(X) + Var(Y)
  - E(aX+bY) = aE(X) + bE(Y), where a and b are constants
  - $Var(aX + bY) = a^2Var(X) + b^2Var(Y)$
  - If $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ then $(X+Y) \sim N(\mu_X+\mu_Y, \sigma_X^2+\sigma_Y^2)$

3.2.12 Some examples related to the sums of independent random variables

1. If X ~ $N(\mu_X = 0, \sigma_X^2 = 1)$ and Y ~ $N(\mu_Y = 3, \sigma_Y^2 = 4)$, then X+Y ~ N(mean=3, variance=5)
2. If $X_1, X_2, ..., X_n$ ~ N(0,1), then $\sum_{i=1}^{n} X_i$ ~ N(0,n)
3. … and then, $\frac{1}{n}\sum_{i=1}^{n} X_i \sim N(0, \frac{1}{n})$
4. If $X_1, X_2, ..., X_n \sim N(\mu, \sigma^2)$, then $\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$
5. … and then, $\frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n})$

3.2.12.1 Excerpt from Lawrence Joseph's notes



3.2.13 Example 2: Central Limit Theorem in action

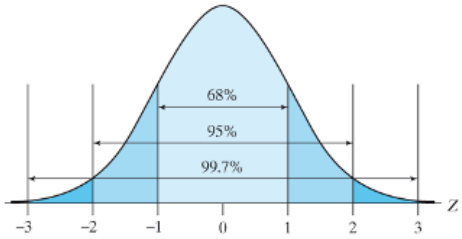
- What is the average time taken across the 50 students in the class?
- R code to replicate

x1 = rnorm(50,4,1) # walk to bus stop x2 = runif(50,4,16) # wait for bus x3 = rnorm(50,20,2) # bus ride x4 = rgamma(50,shape=3/2,scale=2) # trudge up hill

par(mfrow=c(2,3)) hist(x1);hist(x2);hist(x3);hist(x4) hist(x1+x2+x3+x4,xlab="Sum for 50 students",main="") hist((x1+x2+x3+x4)/4,xlab="Mean for 50 students",main=" ")

3.3 Confidence intervals for means

3.3.1 Confidence interval estimation for a single mean



- The construction of a confidence interval relies on the principal of the central limit theorem
- If, we can reasonably assume that the sample mean follows a normal distribution with mean µ and standard deviation $\frac{\sigma}{\sqrt{n}}$
- Then, across repeated samples, 95% of samples' means ($\bar{x}'s$) lie in the interval $\left(\mu - 2\frac{\sigma}{\sqrt{n}}, \mu + 2\frac{\sigma}{\sqrt{n}}\right)$
- This implies that 95% of the intervals $\left(\bar{x} - 2\frac{\sigma}{\sqrt{n}}, \bar{x} + 2\frac{\sigma}{\sqrt{n}}\right)$ will include μ. This interval is called the 95% confidence interval for µ
- More generally, $(1-\alpha)$% of the intervals $\left(\bar{x} - Z_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}}\right)$ will include μ.
- This interval is called the (1-)% equal-tailed confidence interval for µ, where $Z_{(1-\frac{\alpha}{2})}$ is the (1- /2) quantile of the standard normal distribution
- Equal-tailed refers to the fact that the probability of (1-) is divided equally in the two tails of the distribution
- Notice that the 95% or (1-)% in the definition refers to a percentage across repeated experiments
- We cannot say whether the 95% confidence interval estimated from the sample at hand is one of the ones that captured the true value of µ or not
- The population standard deviation ( ) is seldom known and must be substituted by the sample standard deviation (s)
- Does the assumption of 95% confidence still hold? It turns out that it does but we must replace the quantile $Z_{(1-\frac{\alpha}{2})}$ from the normal distribution by the $t_{(1-\frac{\alpha}{2})}$ quantile from the Student's t-distribution (or t-distribution for short)
- The resulting expression for the confidence interval is given by:

$$\left(\bar{x} - t_{(1-\frac{\alpha}{2}),n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{(1-\frac{\alpha}{2}),n-1}\frac{s}{\sqrt{n}}\right)$$

where $t_{(1-\frac{\alpha}{2}),n-1}$ is the (1- /2) quantile of the t-distribution with n-1 degrees of freedom

3.3.2 t-distribution

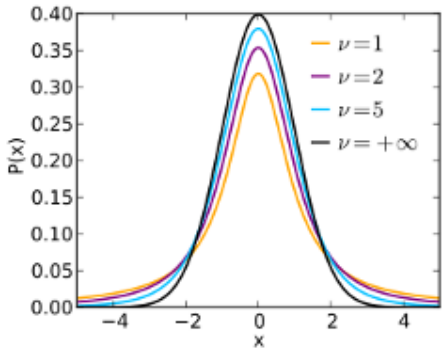


Image from Wikipedia

- The t-distribution was discovered by the British scientist W. S. Gossett who was employed by the Guiness Brewery.
  - He published his work in 1908 under the pseudonym Student
- The t-distribution is a bell-shaped, symmetrically distribution over the range -∞ to ∞. It resembles the normal distribution, but has a higher standard deviation.
- The exact shape of the distribution depends on a quantity called the degrees of freedom (in the illustration). The higher the value of the closer it is to a normal distribution

Probability density function centred at 0

$$f(x|v) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\Gamma\left(\frac{v}{2}\right)}\left(1 + \frac{x^2}{v}\right)^{-\frac{v+1}{2}}, -\infty < x < \infty$$

Mean=0

Variance=$\frac{v}{v-2}$

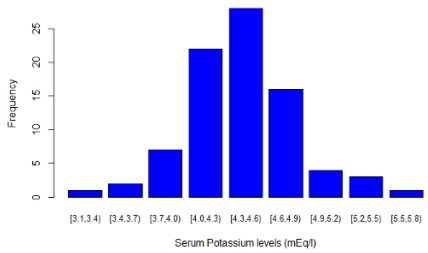
3.3.3 Example 1: Serum Potassium Concentration

Serum potassium (mEq/l)	Number of women
[3.1, 3.4)	1
[3.4, 3.7)	2
[3.7, 4.0)	7
[4.0, 4.3)	22
[4.3, 4.6)	28
[4.6, 4.9)	16
[4.9, 5.2)	4
[5.2, 5.5)	3
[5.5, 5.8)	1
Total	84

- As part of a study of natural variation in blood chemistry, serum potassium concentrations were measured in 84 healthy women.
- The mean concentration was 4.36 mEq/l, and the standard deviation was 0.42 mEq/l.
- The table presents a frequency distribution of the data
- Calculate the standard error of the mean
- Construct a histogram of the data and indicate the intervals mean $\pm$ SD and mean $\pm$ SE
- Construct a 95% confidence interval for the population mean. Interpret this confidence interval
- Would this interval be suitable to define "reference limits" for serum potassium in healthy women, i.e. the limits within which we would expect to find 95% of healthy people?
- Suppose a similar study is to be conducted the following year among 200 women. What would you predict would be

  - the SD of the new measurements?
  - the SE of the new measurements?

3.3.4 Example 1: Histogram of the data



3.3.5 Verifying assumptions behind the t-distribution confidence interval

- Does the central limit theorem hold?
- In other words, do at least one of the following conditions hold
  - the data follow an approximately normal distribution?
  - the sample size is large
- For the serum potassium example both conditions appear to hold

3.3.6 Example 1: Standard Error and 95% confidence interval

- The standard error of the mean (SE)
  $= \frac{SD}{\sqrt{n}} = \frac{0.42}{\sqrt{84}} = 0.05$ mEq/l, after rounding
- The 95% confidence interval
  $= \left(\bar{x} - t_{(1-\frac{\alpha}{2}),n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{(1-\frac{\alpha}{2}),n-1}\frac{s}{\sqrt{n}}\right)$
  $= (4.36 – t_{0.975,84-1}0.05, 4.36 + t_{0.975,84-1}0.05)$
  $= (4.36 - 1.98 \times 0.05,\ 4.36 + 1.98 \times 0.05)$
  $= (4.26,\ 4.46)$ mEq/l

3.3.7 Interpretation of the 95% confidence interval

- Assuming that the sample at hand is a random sample, there is a 95% probability that the procedure used to calculate the interval (4.26, 4.46) will capture the population mean serum potassium concentration
- It would **not** be correct to say: There is a 95% probability that the population mean serum concentration lies between 4.26 and 4.46 mEq/l

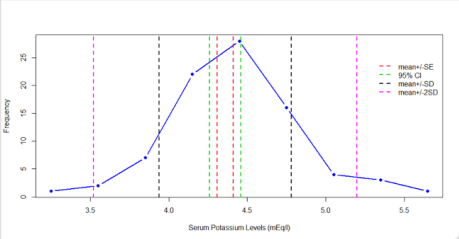
3.3.8 Confidence level

- The higher the confidence level, the wider the confidence interval would be

Confidence level	t-distribution quantile	Lower limit (mEq/l)	Upper limit (mEq/l)
90%	qt(0.950, 83)=1.66	4.28	4.44
95%	qt(0.975, 83)=1.98	4.26	4.46
99%	qt(0.995, 83)=2.64	4.23	4.49

- qt(prob,df) is the R function that returns the t-distribution quantile
  - Arguments provided are the cumulative probability and the degrees of freedom

3.3.9 Example 1: Distribution of the data (with intervals)



3.3.10 Interpreting the confidence interval

- *Would the 95% confidence interval be suitable to define "reference limits" for serum potassium in healthy women, i.e. the limits within which we would expect to find 95% of healthy people?*
- No. The 95% interval attempts to captures the uncertainty in the **mean** of the distribution.
- In the expression for the confidence interval, if we replaced the standard error by the standard deviation, we would get the desired reference limits

3.3.11 Standard error vs Standard deviation

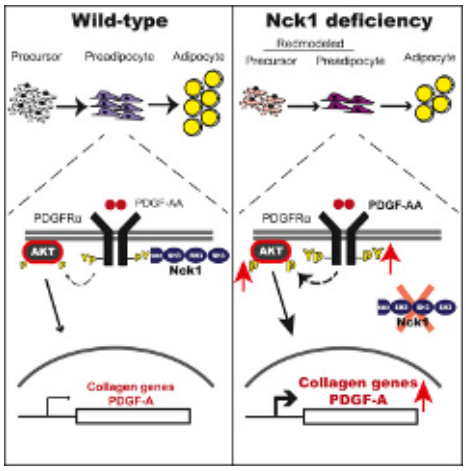
- *Suppose a similar study is to be conducted the following year among 200 women. What would you predict would be*

- *the SD of the new measurements?*
- *the SE of the new measurements?*

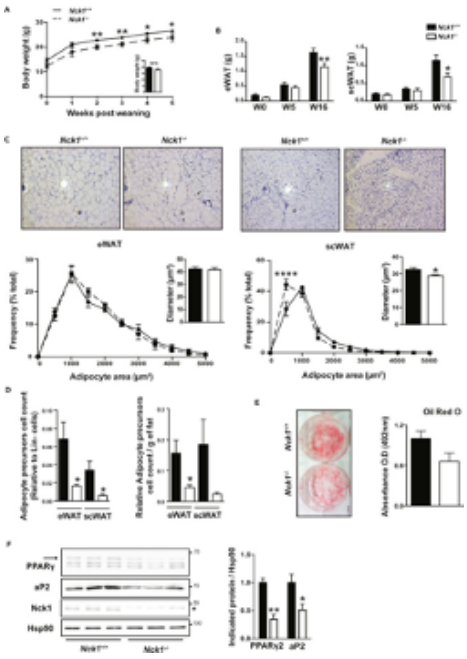
- Our best prediction for the SD would be the value in the smaller sample of 84, namely 0.42 mEq/l
- However, the SE of the new measurements would decrease from 0.05 to $\frac{0.42}{\sqrt{200}} = 0.03$ mEq/l

3.4 Confidence interval for the difference between two means

3.4.1 Example 2: Nck1 deficiency and adipogenesis



- Obesity results from an excessive expansion of white adipose tissue (WAT), which is still poorly understood from an etiologic-mechanistic perspective
- A study from the MUHC-RI reported on the role of the Nck1 adaptor protein during WAT expansion and in vitro adipogenesis
- Two outcomes of interest were body weight and adipose weight



- Nck1 wild type (Nck1+/+) and knock-out mice (Nck1-/-) were compared at baseline and at 16 weeks
- Two research questions of interest: Is there a difference in wild-type and knock-out mice in terms of
  - Body weight
  - Adipose weight
- What would be considered a meaningful change on these two outcomes?
- In order to apply the Central Limit Theorem we would ask:
  - Is it reasonable to assume that body weight and adipose weight follow an approximately normal distribution?
  - If not, is the sample size sufficiently large?
- The sample size is not large, so the approximate normality must hold to construct a t-distribution-based confidence interval

	Nck1 +/+	Nck1 -/-	Difference in means
Number of cases	16	9	
Body weight Mean (g)	38.2	35.7	2.5
Body weight SD (g)	5.4	5.6	
Adipose weight Mean (g)	1.6	1.1	0.5
Adipose weight SD (g)	0.5	0.4	

3.4.2 Confidence interval for the difference between means from two independent samples

The (1-)% confidence interval comparing two means from independent samples is given by

$$\bar{x}_1 - \bar{x}_2 - t_{(1-\alpha/2),df} s_{diff}, \bar{x}_1 - \bar{x}_2 + t_{(1-\alpha/2),df} s_{diff}$$

where

$\bar{x}_1 - \bar{x}_2$	denotes the difference in the two sample means
s_{diff}	denotes the standard deviation of this difference
$t_{(1-\alpha/2),df}$	denotes the (1- α/2) quantile of the t-distribution with df degrees of freedom

3.4.3 Variance of the difference in means

When the variance in the two groups can be assumed to be the same	$s_{diff} = s_p\sqrt{\frac{1}{n1}+\frac{1}{n2}}$ Where $s_p^2 = \frac{(n1-1)s_1^2+(n2-1)s_2^2}{n1+n2-2}$	Degrees of freedom of the t-distribution = n1+n2-2
When the variance in the two groups cannot be assumed to be the same	$s_{diff} = \sqrt{\frac{s_1^2}{n1}+\frac{s_2^2}{n2}}$	Degrees of freedom of the t-distribution = min(n1-1, n2-1) or Welch's approach

3.4.4 Calculating degrees of freedom of the t-distribution when variances are not equal

- The degrees of freedom can be set to min(n1-1, n2-1), which is a conservative value. This is a useful approach if you are doing the t-test by hand
- Alternatively, a computer program may use a more complex method called the Welch's method or Satterthwaite's method to calculate the degrees of freedom as follows:

$$\frac{(se_1^2 + se_2^2)^2}{\frac{se_1^4}{n1-1} + \frac{se_2^4}{n2-1}},$$

where $se_1 = se_1/\sqrt{n1}$ and $se_2 = se_2/\sqrt{n1}$

3.4.5 Example 2: Nck1 deficiency and adipogenesis

- Based on the sample estimates, and perhaps from information gathered previously, it may be reasonable to assume that the variance is the same in both groups being compared
- Since we are assuming that the variance is the same, it is reasonable to calculate a pooled variance that averages across both groups.

3.4.6 Calculating the pooled variance for body weight

- The pooled variance is given by

$$s_p^2 = \frac{(n1-1)s_1^2 + (n2-1)s_2^2}{n1+n2-2} = \frac{15*5.4*5.4 + 8*5.6*5.6}{16+9-2} = 29.9$$

- Therefore the pooled standard deviation is given by the square root of 29.9 or $s_p = 5.5$
- The value of $s_{diff} = s_p\sqrt{\frac{1}{n1}+\frac{1}{n2}} = 5.5\sqrt{\frac{1}{16}+\frac{1}{9}} = 2.3$

3.4.7 Confidence interval for difference in body weight

- The difference in mean body weight between Nck1+/+ and Nck1-/- mice is $\bar{y}_1 - \bar{y}_2 = 38.2 - 35.7 = 2.5$
- 95% confidence interval for the difference in means is

$\bar{y}_1 - \bar{y}_2 - t_{(1-\alpha/2),n1+n2-2}s_{diff}, \bar{y}_1 - \bar{y}_2 + t_{(1-\alpha/2),n1+n2-2}s_{diff}$

= (2.5 – 2.07 × 2.3, 2.5 + 2.07 × 2.3)

= (-2.3, 7.3)

3.4.8 Confidence intervals comparing the two groups

Comparison	Assumption	Degrees of freedom	95% Confidence Interval
Difference in body weight	Variance assumed equal	23	(-2.3, 7.3)
Difference in body weight	Variance assumed unequal	16.187	(-2.4, 7.4)
Difference in adipose weight	Variance assumed equal	23	(0.08, 0.91)
Difference in adipose weight	Variance assumed unequal	20.51	(0.1, 0.9)

- The assumption of unequal variance results in a lower value for the degrees of freedom and would typically be more conservative

3.4.9 Interpreting the confidence interval

- As in the case of a single mean, we have 95% confidence in the procedure used to construct the interval.
  - We cannot say if this interval based on our sample includes the true mean difference between Nck1 +/+ and Nck1 -/- mice
- Say we consider 5g to be a meaningful difference in body weight
  - This implies, though the confidence interval includes 0, the upper limit crosses 5g suggesting we cannot eliminate the possibility there is a meaningful difference. Ideally, the study should be repeated to obtain a more precise estimate
- Say we consider a 0.5g to be a clinically meaningful difference in adipose weight
  - The interval provides evidence for a statistically significant difference, but does eliminate the possibility that the difference may not be clinically meaningful difference as the lower limit lies below 0.5g

3.5 Sample size calculations

- Before collecting data for a research study, it is wise to consider in advance whether the estimates generated from the data will be sufficiently precise.
- It can be painful indeed to discover after a long and expensive study that the standard errors are so large that the primary questions addressed by the study cannot be answered.

3.5.1 An illustration

- https://www.youtube.com/watch?v=PbODigCZqL8

3.5.2 Sample size calculation

- The method one uses for the sample size calculation depends on the plan for the statistical inference
- Accordingly, depending on whether you intend to report a hypothesis test, or a confidence interval or a Bayesian analysis, your method for sample size calculation may change

3.5.3 Sample size calculation for reporting a confidence interval

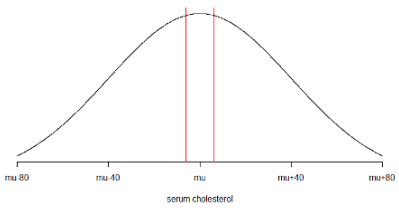
- This approach is relevant when we want to estimate a parameter within a certain precision, with a high level of confidence.
- For example, we might want to estimate
  - mean change in body weight in mice within ± 2.5g of the true value with 99% confidence
  - mean serum cholesterol in middle-aged men within ± 6mg/dL of its true value with 90% confidence

3.5.4 Example: Method for a single mean

- A medical researcher proposes to estimate the mean serum cholesterol level of a certain population of middle-aged men, based on a random sample of the population.
- He asks a statistician for advice. The ensuing discussion reveals that the researcher wants to estimate the population mean to within $= \pm 6$ mg/dl or less, with 95% confidence.

- Also, the researcher believes that the standard deviation of serum cholesterol in the population is probably about s=40 mg/dl.
- How large a sample does the researcher need to take?

3.5.5 The desired precision is much smaller than the standard deviation of the variable



3.5.6 Example: Method for a single mean

- The research question can be re-expressed as

"What is the sample size required to calculate a 95% confidence interval for the mean serum cholesterol which has half-width 6mg / dL?"

- Recall that the general expression for the (1-)% confidence interval is

$$\bar{x} - t_{(1-\alpha/2),n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{(1-\alpha/2),n-1}\frac{s}{\sqrt{n}}$$

- In other words, we need to find out how large n should be so that

$$t_{(1-\alpha/2),n-1}\frac{s}{\sqrt{n}} = \delta = 6$$

- To solve this expression for n, we need to know the values of $t_{(1-\alpha/2),n-1}$ and the value of s, the standard deviation
- Since $t_{(1-\alpha/2),n-1}$ itself depends on n, we cannot know its value without n! We therefore, replace it by the normal quantile $Z_{(1-\alpha/2)}$. In our example, $Z_{(1-\alpha/2)} = 1.96$
- The value of s could be a guess value or determined from the literature or an earlier pilot study. In our example, s=40
- Therefore, we wish to solve

$$Z_{(1-\alpha/2)}\frac{s}{\sqrt{n}} = 1.96\frac{40}{\sqrt{n}} = \delta = 6$$

- This implies $\sqrt{n} = Z_{(1-\alpha/2)}\frac{s}{\delta} = 1.96\frac{40}{6}$
- Or $n = (1.96\frac{40}{6})^2 \approx 171$

3.5.7 Alternative values of , s and

alpha	s	delta	n
0.05	40	6	171
0.01	40	6	240
0.05	30	6	96
0.01	30	6	135
0.05	40	12	43
0.01	40	12	60
0.05	30	12	24
0.01	30	12	24

- By varying the values of , s and we can see how they impact the sample size
- n increases if:
  decreases, s increases or decreases

- In practice, the sample size may be constrained by feasibility or cost. Using a table like this allows us to see how much precision we can 'buy' with the available sample size

3.5.8 Example: Sample size calculation for comparing two means

- Consider the study on body weight in Nck+/+ vs Nck-/- mice
- Lets say we wish to repeat the earlier study so that we can show more convincingly that there is a clinically meaningful difference
- Earlier in the lecture we found that the **pooled** standard deviation of the difference was $s_p = 5.5g$
- We desire to ensure that the observed mean change lies within $= \pm$ 2.5 g of the true mean change with 95% confidence.
- What is the sample size required in each group (assuming the sample size is equal in both groups)?

3.5.9 Example: Comparison of two means

- To calculate the sample size required to estimate a 95% CI with adequate precision we need to solve
  $Z_{(1-\alpha/2)}s_{diff} = Z_{(1-\alpha/2)}s_p\sqrt{\frac{1}{n}+\frac{1}{n}} = \delta$
  or $1.96 \times 5.5 \times \sqrt{\frac{1}{n}+\frac{1}{n}} = 2.5$
- This implies $\sqrt{n} = 1.96\frac{5.5\times\sqrt{2}}{2.5}$
- Or $n = 2(1.96\frac{5.5}{2.5})^2 \approx 37$ mice in each group

Chapter 4

Lecture 4: Inference for means continued

4.1 Hypothesis testing

4.1.1 Example 1: Nck1 and adipogenesis continued

- Whereas in the previous lecture we saw how to carry out statistical inference about the differences between Nck1 wild type and Nck1 knock out mice using confidence intervals, the manuscript relied on hypothesis testing

4.1.2 Hypothesis testing

- Hypothesis testing is an alternative approach to statistical inference that also relies on the Central Limit Theorem
- The research question takes the form of a decision making problem, e.g.
  - Does mobility improve 3-months after a stroke?
  - Is there a difference in the improvement in mobility between men and women after a stroke?
  - Does Treatment A improve life-expectancy compared to Treatment B?
- Each of these questions can be answered yes or no. Each response can be expressed as a specific statement
- We can view the response to the stroke mobility problem as a choice between the following two statements or hypotheses
  - H_0: There is no improvement in mobility 3 months after stroke

  - H_A: There is improvement in mobility 3 months after stroke

- In general, a decision-making problem can be framed in terms of a null hypothesis (H_0) and an alternative hypothesis (H_A). The H_A is the complement of the null hypothesis
- We typically focus on the null hypothesis, which is usually simpler than the alternative hypothesis, and decide whether or not to reject it.
- To this end, we examine the evidence that the observed data provide against the null hypothesis H_0
- If the evidence against H_0 is strong, **we reject** H_0
- If not, we state that the evidence provided by the data is not strong enough, and **we fail to reject** H_0.

4.1.3 Hypothesis testing for a single mean

- The hypothesis test may be set up with
  - a two-sided alternative
  - or a one-sided alternative
- resulting in 3 different possibilities mentioned in the following slides

4.1.4 Mobility after stroke: two-sided alternative hypothesis

Null Hypothesis
H_0: The **true mean** change in the STREAM score between 3-days and 3 months post stroke *

4.1.5 Mobility after stroke: one-sided alternative hypothesis I

Null Hypothesis
H_0: The **true mean** change in the STREAM score between 3-days and 3 months post stroke *

4.1.6 Mobility after stroke: one-sided alternative hypothesis II

Null Hypothesis
H_0: The **true mean** change in the STREAM score between 3-days and 3 months post stroke *

4.1.7 More generally, the hypothesis test for a single mean may be stated as follows

- The null and alternative hypotheses for a two-sided test may be stated as

$H_0 : = {}_0 vs H_A : \neq {}_0$

where µ denotes the true population mean ${}_0$ is a known constant

- The null and alternative hypotheses for a one-sided test can be stated as follows

$H_0 : \leq {}_0 vs H_A : > {}_0$

OR

$H_0 : \geq {}_0 vs H_A : < {}_0$

4.1.8 Example: Mobility after stroke

	Three days after stroke	Three months after stroke	Difference
Number of cases	235.00	235.00	235
Minimum	0.00	0.00	-22.22
Maximum	100.00	100.00	91.67
Mean	68.30	83.75	$\bar y = 15.45$
Standard deviation	30.12	22.74	s = 18.97

4.1.9 Defining the test statistic and the rejection region

- Recall that based on the Central Limit Theorem,
$\bar{Y} \sim N(\mu, \sigma^2/n)$ or $\bar{Y} \sim N(\mu, \sigma^2/235)$

- We can also express this as $\frac{\bar{Y}-\mu}{\frac{\sigma}{\sqrt{n}}}$ follow a standard normal distribution

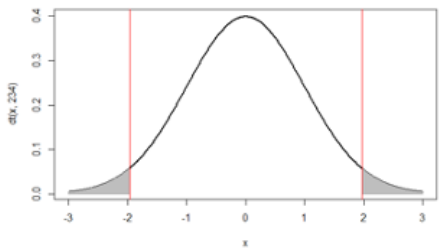
- We can use our knowledge of the sampling distribution of $\frac{\bar{Y}-\mu}{\frac{\sigma}{\sqrt{n}}}$ (the test statistic) to determine which values are likely under the null hypothesis

- We define a rejection region such that if test statistic falls in this region we reject the null hypothesis

4.1.10 Defining the t-test statistic

- As in the case of the construction of a confidence interval, we are faced, with the problem that we seldom know the true standard deviation.
- We can **estimate** the value of the unknown population standard deviation using the sample standard deviation $\hat{\sigma} = s = 18.97$
- The standardized test statistic is then $\frac{\bar{Y}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{15.45}{\frac{18.97}{\sqrt{235}}} = 12.49$
- This statistic is referred to as the **t-statistic** as it follows a t-distribution with n-1 degrees of freedom
- The corresponding hypothesis test is called the **t-test**.

4.1.11 Rejection region for the t-test



- Our goal is to select a rejection region such that it covers values that are unlikely under the null hypothesis
- The form of rejection region depends on the statement of the alternative hypothesis. * We first consider the two-sided alternative. Under this alternative hypothesis, the rejection region covers the extremes of the distribution on both sides
- The two areas each covering with 0.025 probability in the extremes are unlikely under the null hypothesis as illustrated by the diagram. They correspond to a **Type I error of 0.025+0.025 = 0.05**, which we will define shortly
- Under the t-distribution with degrees of freedom = n-1 = 234, these areas may be identified by the quantiles $Q_{0.025} = -1.97$ and $Q_{0.975} = 1.97$
  - Therefore, if the t-statistic is above 1.97 or less than -1.97 we reject the null hypothesis
- In our example, 12.49 is well above 1.97 so we **reject the null hypothesis**

4.1.12 Comparison to confidence interval

- The hypothesis testing approach resulted in a similar conclusion to the equal-tailed confidence interval derived earlier in that we concluded that

mobility improves 3 months after stroke

- In fact, the equal-tailed 95% confidence interval derived previously gives the range of possible values of the null hypothesis that cannot be rejected.
  - All values outside that interval will be rejected
  - That happens to include the value of 5 units which defines a clinically meaningful improvement

4.1.13 Determining the rejection region using R

We use the qt() function to obtain the quantiles of a t-distribution corresponding to the desired tail-area probability

```
qt(0.025,234)
[1] -1.970154
```

```
qt(0.975,234)
[1] 1.970154
```

The sample size is very large. Therefore, for all practical purposes the t-distribution with degrees of freedom n-1 = 234 is like a normal distribution

```
qnorm(0.025)
[1] -1.959964
```

```
qnorm(0.975)
[1] 1.959964
```

4.1.14 Type I and Type II errors

- With respect to our decision regarding the null hypothesis we can make two types of errors
  - Type I error ( ): We reject H_0 when it is true
  - Type II error ( ): We fail to reject H_0 when it is not true (i.e. when H_A is true)
- Clearly, we wish to minimize the chance of these errors. Typical values are =0.05 and =0.2

4.1.15 Hypothesis testing: A summary

1. Define null and alternative hypotheses
2. Define test statistic
3. Define rejection region with suitably selected Type I error
4. If test statistic lies in the rejection region then reject null hypothesis, otherwise conclude that you do not have enough evidence to reject the null hypothesis

4.1.16 Similarity between diagnostic testing and hypothesis testing

		True disease status		
		+	-	
Observed diagnostic test	+	A	B	A+B
	-	C	D	C+D
		A+C	B+D	

- Sensitivity = A / (A+C)
- Specificity = D / (B+D)
- A, B, C and D are numbers of individuals tested

		True population value		
		H_A	H_0	
Observed test statistic significant	+	A	B	A+B
	-	C	D	C+D
		A+C	B+D	

- 1-Type II error (Power) = A / (A+C)
- 1-Type I error = D / (B+D)
- A, B, C and D are values of the test statistic observed across repeated experiments

4.1.17 Defining the t-test statistic, for a one-sided test

- Consider the situation where we pose the null and alternative hypotheses as follows
  $H_0 : \leq {}_0 vs H_A : > {}_0$
- The test statistic is still evaluated at $= {}_0 = 0$ as before
- However, the rejection region is one-sided. In order to ensure that the rejection region has a 5% probability as in our previous example, we will define it as the region above $Q_{0.95} = 1.65$
- For our example, the t-statistic would remain unchanged at 12.49 and therefore would lie in the rejection region once again, leading to the same conclusion as before

4.1.18 Why did the test statistic not change for the one-sided hypothesis test?

- Notice that though our null hypothesis was $H_0 : \leq 0$, we calculated the test-statistic at µ=0
- This is because we know that rejection region under smaller values of µ below zero will be shifted to the left compared to $Q_{0.95} = 1.65$
- Therefore, if our test statistic results in rejecting µ=0, it will certainly result in rejecting values of µ less than 0

4.1.19 What is statistical significance?

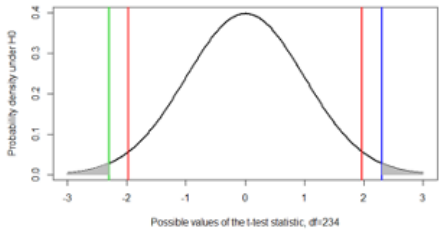
- As mentioned earlier, the rejection region is selected so that it is unlikely under the null hypothesis
- Therefore, when the test-statistic falls in the rejection region, we say it is statistically significant
- Traditionally, this region is selected to have 5% probability under the null hypothesis. However, 5% is arbitrary
- Note that in setting up the test statistic, only the null hypothesis came into play. The alternative hypothesis did not matter

4.1.20 What is a p-value?

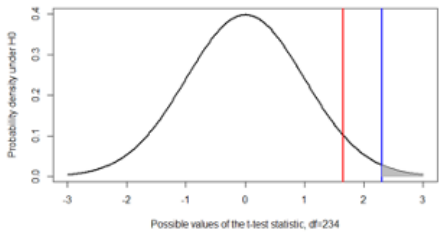
- The p-value is defined as the probability of being more extreme than the test statistic under the null hypothesis
  = **P(Test statistic is more extreme than its observed value | H0)**
- In our example, involving a **one-sided** test
  $p - value = P(T_{234} > 12.49 | H_0) = 1 - pt(12.49, 234) = 0$

- Clearly, when the test statistic is statistically significant, the p-value is less than 5% or more generally it is less than the Type I error
- This explains why the p-value is often compared to 5% to determine statistical significance

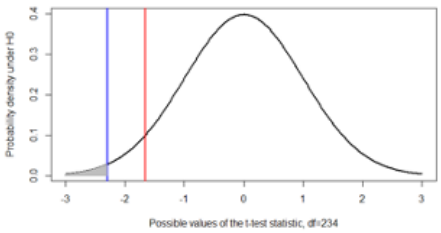
4.1.21 p-value illustrated



- Notice the difference between the p-value and the rejection region for a **two-sided** test
  - The red lines mark off the rejection region of =0.05 at ±1.97
  - The blue line is a hypothetical observed t-statistic=2.3
  - The shaded area marks off the p-value
  - The green line is at -2.3, was not observed. Yet, we use the area beyond it to obtain a two-sided p-value



- This figure illustrates the p-value for a one-sided test with $H_A : > {}_0$
  - Once again, the red line marks off the rejection region of =0.05 at 1.65
  - The blue line is the observed t-statistic=2.3 in this illustration
  - The shaded area marks off the p-value. Note that the p-value is half that of the one-sided test by definition



- Finally, this figure illustrates the p-value for a one-sided test with H_A : $<$ $_0$
  - This time, the red line marks off the rejection region of =0.05 at -1.65
  - The blue line is the observed t-statistic=-2.3 in this illustration
  - The shaded area marks off the p-value

4.1.22 Type I and Type II errors

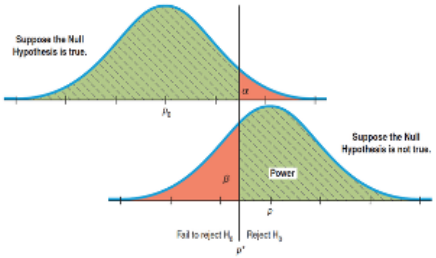


Figure 13.9 The power of a test is the probability that it rejects a false null hypothesis. The upper figure shows the null hypothesis model. We'd reject the null in a one-sided test if we observed a value in the red region to the right of the critical value, p^*. The lower figure shows the model if we assume that the true value is p. If the true value of p is greater than p_0, then we're more likely to observe a value that exceeds the critical value and make the correct decision to reject the null hypothesis. The power of the test is the green region on the right of the lower figure. Of course, even drawing samples whose observed proportions are distributed around p, we'll sometimes get a value in the red region on the left and make a Type II error of failing to reject the null.

- Recall
  - Type I error is the probability of rejecting the null hypothesis when it is true
  - Type II error is the probability of not rejecting the null hypothesis when the alternative is true

4.1.23 t-test

```
t.test(log10(a$IL)~a$Group)

        Welch Two Sample t-test

data:  log10(a$IL) by a$Group
t = -2.3309, df = 9.6793,
p-value = 0.04278

alternative hypothesis:
true difference in means is not
equal to 0

95 percent confidence interval:
 -2.52019044 -0.05115657


sample estimates:
mean in group 0 mean in group 1
       3.262180        4.547854
```

- The t.test function in R tells us that that we can reject the null hypothesis of no difference in the mean log10 interleukin levels in the two groups at the Type I error level of 0.05

4.1.24 A bit of history: Pearson vs. Fisher

- The hypothesis test and p-value were proposed by Karl Pearson and Ronald Fisher, respectively, who were contemporaries who strongly disagreed with each other
- It is ironic that today we use these two techniques together!
- As we will discuss in greater detail in later lectures, there has been a backlash against both these approaches and a move towards usage of confidence intervals or Bayesian methods

4.1.25 Inference for comparing two means

- The hypothesis test for comparing two means resembles the structure of the hypothesis test for a single mean
  - it can be two-sided or one-sided
  - the form of the test-statistics depends on the study design and assumptions, e.g.
    - Whether the study design involves paired or unpaired means
    - Assuming the variance in the two groups is equal or not
    - Assuming the variance is known or not

4.1.26 Stroke study: Question 2, two-sided alternative hypothesis

Null Hypothesis
H_0: The true mean change in the STREAM score between 3-days and 3 months post stroke **is the same for

4.1.27 Stroke study: Question 2, one-sided hypothesis I

Null Hypothesis
H_0: The true mean change in the STREAM score between 3-days and 3 months post stroke **is at most as gr

4.1.28 Stroke study: Question 2, one-sided hypothesis II

Null Hypothesis
H_0: The true mean change in the STREAM score between 3-days and 3 months post stroke **is at least as gr

4.1.29 Example: One-sided or two-sided test?

- We return to the second research question based on the stroke dataset. It is of interest to compare the change in mobility (from baseline to 3 months) between men and women
- One way to do this is to carry out a hypothesis test.
- We will begin with a two-sided hypothesis test:

$H_0 : \mu_1 = \mu_2 vs. H_A : \mu_1 \neq \mu_2$

where $_1$ is the true mean change in mobility in men and $_2$ is the true mean change in women

4.1.30 Difference in change in mobility between men and women

	Change in Men	Change in Women
Number of cases	144	91
Mean	$\bar y_1 = 17.09$	$\bar y_2 = 12.86$
Standard deviation	$s_1 = 19.25$	$s_2 = 18.31$

- Recall that we had assumed that the variance is the same in both groups being compared and calculate a pooled variance that averages across both groups of $s_{diff} = 2.53$

4.1.31 Comparing change in mobility between men and women

- The t-statistic is given by

$$\frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{s_{diff}} = \frac{\bar{Y}_1 - \bar{Y}_2}{s_{diff}} = \frac{17.09 - 12.86}{2.53} = 1.67$$

- Since we are working under the assumption that the variances are equal, the t-distribution used to define the rejection region has degrees of freedom n1+n2-2 (as we saw previously when defining a confidence interval for comparing two means)
- If we use a Type I error value of =0.05, we **would** reject the null hypothesis if it lies below -1.96 or above 1.96
- In our case, the t-statistic falls within this **region** so we say "we do not have enough evidence to reject the null hypothesis"
- The p-value is 0.09, which exceeds 0.05
- The p-value can be calculated as follows in R

```
2*(1-pt((17.09-12.86)/2.53,233))
[1] 0.09587913
```

4.1.32 What if our alternative hypothesis was one-sided instead?

- It is to be expected that men may experience a greater improvement in mobility than women. Therefore, we can restate our hypothesis test as:

$H_0 : {}_1 \leq {}_2 vs H_A : {}_1 > {}_2$

where µ1 is the true mean change in men and µ2 is the true mean change in women

- As in the case of hypothesis testing for a single mean, the test statistic remains the same
- However the rejection region is one-sided. Using the quantiles of the t-distribution with n1+n2-2=233 degrees of freedom, we can determine that the rejection region includes the region above $Q_{0.95} = 1.65$. Therefore, our test statistic of 1.67 lies in the rejection region
- In comparison with this rejection region, we would conclude that we have enough evidence to reject the null hypothesis that the mean change in mobility in men is less than or equal to that of women

4.1.33 What if our alternative hypothesis was one-sided in the other direction?

- Only for the purpose of illustrating how the rejection region is defined, let us restate our hypothesis test as:

$H_0 : {}_1 \geq {}_2 vs H_A : {}_1 < {}_2$

where µ1 is the true mean change in men and µ2 is the true mean change in women

- Once again, the test statistic remains the same
- However the rejection region is now the region below $Q_{0.95} = -1.65$. Therefore, our test statistic of 1.67 does not lie in the rejection region
- This would lead us to conclude we do not have enough evidence to reject the null hypothesis that the mean change in mobility men is less than or equal to that in women

4.1.34 What if our null hypothesis was one-sided instead?

- The p-value for this situation is
  P(Test statistic $> 1.67|$ H0)
- In R this can be calculated as (1-pt((17.09-12.86)/2.53,233)) = 0.04793956, which falls below the Type I error level of =0.05

4.1.35 Why did our conclusion change when we moved from a two-sided to a one-sided hypothesis?

- The two-sided test is a more stringent test, which makes it more difficult to reject the null hypothesis
- Under a two-sided alternative we have to consider the probability of being more extreme than the observed value on both sides of the null
- This would be relevant only if we thought that it were possible that the difference $\bar{Y}_1 - \bar{Y}_2$ could be either positive or negative
- If we have reason to believe that men are unlikely to have worse mobility than women, the one-sided test would make more sense in the context of our example

4.2 Hypothesis testing vs. confidence interval estimation

4.2.1 What is statistics?

Statistics is a collection of procedures and principles for gathering data and analyzing information in order to help people make decisions when faced with uncertainty
Utts & Heckard in 'Statistical Ideas & Methods'

4.2.2 Quantifying uncertainty vs. decision making

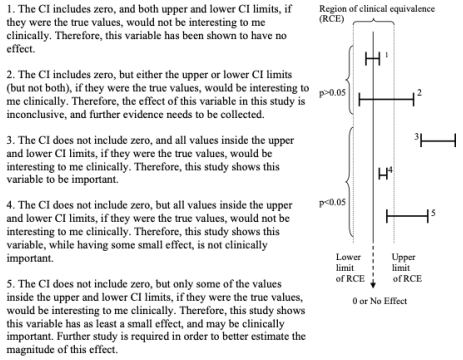
- The hypothesis testing framework is designed to support decision making, e.g.
  - Whether to take an umbrella to work
  - Whether the observed association between a predictor and an outcome is real
- Confidence interval estimation, on the other hand, conveys the uncertainty in our knowledge about a statistic, e.g.
  - There is a 60%-80% chance it will rain today
  - The difference in survival associated with treatment A vs. treatment B is 60%-80%

4.2.3 Interpreting Confidence Intervals vs. Hypothesis Tests*

- Suppose that you have just calculated a confidence interval for a certain parameter. There are five possible conclusions that can be drawn, depending on where the upper and lower confidence interval limits fall in relation to the upper and lower limits of the region of clinical equivalence.
- The region of clinical equivalence, sometimes called the region of indifference, is the region inside of which both treatments would be considered to be the same for all practical purposes.

*From Lawrence Joseph's notes

4.2.3.1 Interpreting confidence intervals: 5 possible conclusions



4.2.4 Notes on significance tests*

- We saw that there are two ways of reporting the results of a hypothesis test – either we can report **the decision** (reject vs. not reject which is the same thing as significant vs. not significant) or **the p-value**
- Reporting the p-value is more informative than merely reporting whether a test was "significant" or "not significant".
- The level of significance, α, is often set to 0.05, but it should be chosen according to the problem. There is nothing magical about $\alpha = 0.05$. There is no practical difference if p = 0.049 or p = 0.051.
- Even a very small p-value does not guarantee H_0 is false. Repeating the study is usually necessary for further proof, or to vary the conditions or population.
- Statistical significance (small p-value) is not the same as practical significance.
- The p-value is not everything. Must also examine your data carefully, data cleaning for outliers, etc. Remember – all tests carry assumptions that can be thrown off by outliers.
- Reporting a confidence interval for an effect is more informative than reporting a p-value.
- P-values are often misinterpreted. A p-value is not the probability of the null hypothesis.
- It is also not the probability that a result occurred by chance. . . .
- The p-value only tells you something about the probability of seeing your results given a particular hypothesis—it cannot tell you the probability that the results are true or whether they're due to random chance.

*From Lawrence Joseph's notes

4.3 Sample size calculations for studies of one or two means

4.3.1 Sample size for hypothesis tests

- This approach is relevant when we want to test a certain hypothesis
- For example, we might want to test
  - H_0: mean change in stroke mobility 10 points vs.
  - H_a: mean change in stroke mobility $>$ 10 points
  - H_0: mean serum cholesterol 200 vs.
  - H_a: mean serum cholesterol $>$ 200

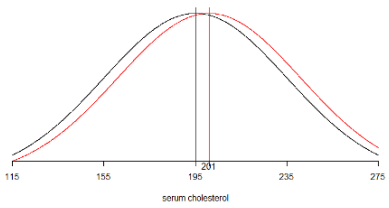
4.3.2 Example

- In the United States, appropriate levels of serum cholesterol in adults have been defined by the National Heart, Lung, and Blood Institute as follows:
  - **Good:** 200 mg/dL or lower
  - **Borderline:** 200 to 239 mg/dL
  - **High:** 240 mg/dL or higher
- Let's say the researcher in our earlier example posed the question differently.
- He or she wants to test the hypothesis that the mean cholesterol level in the population has fallen to 195 mg/dL such that it is now within the "Good" range

$$H_0 : \leq 195 vs. H_a : > 195$$

- How large a sample size is required to test this hypothesis such that
  - Type I error ( )= $P(RejectingH_0|H_0istrue) = 1\%$, and
  - Type II error ( )= $P(NotrejectingH_0|H_Aistrue) = 5\%$
- The researcher wishes to design the study such that the test is sufficiently sensitive to detect difference of 6 mg/dL or more (i.e. when μ=201 or more)

4.3.3 We are interested in detecting a shift in the mean of the distribution



4.3.4 Sample size required for a hypothesis test of a single mean

- Again, we rely on the quantiles of the normal distribution rather than the t-distribution
- The required sample size for a two-sided test is given by this expression:

$$n = \frac{s^2(Z_{1-\alpha/2} + Z_{1-\beta})^2}{(\mu_0 - \mu_A)^2}$$

- The required sample size for a one-sided test is given by this expression:

$$n = \frac{s^2(Z_{1-\alpha} + Z_{1-\beta})^2}{(\mu_0 - \mu_A)^2}$$

- From the expressions on the previous slide we can see that n increases as:
  - s increases
  - decreases or decreases
  - $\mu_0 - \mu_A$ decreases
- Once again, you may wish to calculate sample size under several different scenarios

4.3.5 Sample size required under different scenarios

α	β	s	$\mu_0 - \mu_A$	n
0.01	0.05	40	6	704
0.05	0.05	40	6	484
0.05	0.2	40	6	276
0.05	0.2	40	11	82

4.3.6 Example: Serum cholesterol

- The sample size required for a one-sided test is

$$n = \frac{s^2(Z_{1-0.01} + Z_{1-0.05})^2}{(\mu_0 - \mu_A)^2} = \frac{40^2(2.33 + 1.65)^2}{(195 - 201)^2} = 704$$

- **Impact of increasing to 0.05:**
  - If the type I error was increased to 0.05, we would replace $Z_{1-0.01} = 2.33$ by $Z_{1-0.05} = 1.65$.
  - $n = \frac{s^2(Z_{1-0.05}+Z_{1-0.05})^2}{(\mu_0-\mu_A)^2} = \frac{40^2(1.65+1.65)^2}{(195-201)^2} = 484$
- **Impact of increasing :**
  - If in addition to the above change, the type II error was increased to 0.2, as is commonly done in practice. Then, $Z_{1-0.2} = 1.65$ in the expression above would be replace by $Z_{1-0.2} = 0.84$
  - $n = \frac{s^2(Z_{1-0.01}+Z_{1-0.2})^2}{(\mu_0-\mu_A)^2} = \frac{40^2(1.65+0.84)^2}{(195-201)^2} = 276$
- **Impact of increasing $\mu_0 - \mu_A$**
  - If $_0$ were set to 190, then the difference between the two groups increases to 11
  - $n = \frac{s^2(Z_{1-0.01}+Z_{1-0.05})^2}{(\mu_0-\mu_A)^2} = \frac{40^2(1.65+0.84)^2}{(190-201)^2} = 82$

4.3.7 Summary: What do you need to calculate the sample size required?

Confidence interval	Hypothesis test
Confidence level 1-	Type I error
	Type II error
Guess value for standard deviation (s)	Guess value for standard deviation (s)
Desired precision (or half-width of interval) ( )	The minimum important difference to detect $(\mu_

4.3.8 Sample size calculation: Comparing two means

- Once again, we can define different methods depending on whether we plan to report confidence intervals or hypothesis tests

4.3.9 Example

- Consider the study on in-vivo efficacy of the single domain antibody P1.40 in Tg+ mice
- Lets say we wish to repeat the earlier randomized controlled trial.
- The authors reported that the mean change in cholesterol at 4 days after the intervention was 20 mg/dL and I guessed that the **pooled** standard deviation of the difference was s_p=9 mg/dL
- We desire to ensure that the observed mean change lies within $= \pm 5$ mg/dL of the true mean change with 95% confidence.
- What is the sample size required in each arm of the RCT (assuming the sample size is equal in both arms)?
- Alternatively, we may wish to carry out a one-sided hypothesis test of the difference between the two groups

$$H_0 : _{P1.40} - _{PBS} \leq 0 vs. H_a : _{P1.40} - _{PBS} > 0$$

- Recall, that the previous study reported that the mean change in cholesterol at 4 days after the intervention was 20 mg/dL and that the standard deviation was assumed to be s_p=9 mg/dL
- We desire to ensure that the test is sensitive enough to detect a difference greater than $_1 - _2 = 15$ mg/dL with Type II error = 20%. The Type I error is fixed at the traditional value of 5%.
- What is the sample size required in each arm of the RCT (assuming the sample size is equal in both arms)?

4.3.10 Sample size required to test $H_0 : \mu_1 = \mu_2$

- In the expressions below n = total sample size. If the sample size is the same in both groups, it is n/2 in each group
- The required sample size for a two-sided test is given by this expression:

$$n = \frac{4s_p^2(Z_{1-\alpha/2} + Z_{1-\beta})^2}{(\mu_1 - \mu_2)^2}$$