Preamble: This material is a — necessarily brief – introduction to some statistical concepts that are relevant in the interpretation of measurements (observations) made on an individual patient, and in the interpretation of the statistical material presented in research reports.

Learning Objectives

i. {Sections 1-4} To appreciate and be able to describe, in pictures and numbers, [observable] patterns of variation in a characteristic or measurement — from individual to individual, or from one measurement to another of the same individual – and the reasons for, and consequences, of this variability; to be familiar with the summary numbers used to describe these patterns of variation; and be able to identify which summaries are more appropriate in which circumstances.

ii. {Section 5} To appreciate the presence of the [necessarily unobservable] 'statistical noise' in a statistical estimate or summary obtained from a finite amount of data; to be able to quantify — probabilistically — the degree to which (say) a mean level, or a proportion, observed on a single individual, or in a sample of individuals might – just because of sampling variation – be an under-, or an over-estimate, of the level/proportion of interest; to understand and apply the concept of a Margin of Error, and the factors that affect it; to use this to construct a Confidence Interval.

iii. {Sections 6-7} To apply Confidence Intervals

iv. {Section 8} To understand the concepts of, and the proper interpretation of, P-value; test of hypothesis; statistically significant; statistical power.

v. {Section 9} To apply these concepts to published research based on data from aggregates of individuals.

Sources covering these topics

These concepts are covered more fully in formal course in statistics; you will also have encountered them if you have dealt with data in a research setting. Textbooks that cover them well include

AB Hill, *A short textbook of Medical Statistics, 1984* WA 950 H645s Life Sciences;

D Freedman et al., *Statistics, 1998* QA276 F683 Schulich Science & Engineering

J Ingelfinger et al.[1] *Biostatistics in Clinical Medicine, 1994* WA 950 B6165 Life Sciences

B Dawson-Saunders et al. *Basic & clinical biostatistics, 1994* WA 950 D272b Life Sciences

P Armitage et al. *Statistical methods in medical research, 2002* WA 950 A733s Life Sciences

B Rosner *Fundamentals of biostatistics, 2006* QH 323.5 R822f Life Sciences

M Pagano et al. *Principles of biostatistics, 2000* QH323.5 P34 2000 Schulich Sci. & Engineering

G van Belle et al.*Biostatistics: a methodology for the health sciences, 2004* <eBook>

Why is this material so lengthy? What to concentrate on...

Statistical concepts and an appreciation for variability are important for managing individual patients and for understanding published research.

Some of you will have taken college or undergraduate courses on statistics. Most such courses were unfortunately not very engaging or relevant at the time. In many of them, the tasks were to identify the formula to use, and the relevant inputs for it. Unless you typed in the raw data, you probably didn't get a good sense of variability; if you used a 'canned' routine, the calculations did not really illustrate the concepts. Moreover, most courses use the 'method-then-example' rather than the 'here's the case, now what do I do?' format that better prepares you for a problem-solving career. Moreover, you probably found the terminology – and even the logic – a bit strange.

It would take a much larger amount of time than we have in this course to motivate the need for statistical thinking in medicine, to fully explain the concepts, and to understand the basis for the methods. Many of us dislike having to rely on numerical lab results when we don't understand the scientific basis for them, and it's the same with statistical results. But we don't have the time – in one lecture and one small-group exercise – to learn all of what is behind the methods. However, as a gesture to those of you who would like to understand some of it, and can speed read, section 5 does go into more detail on the basis for standard errors and confidence intervals. Likewise, unless you have a particular liking for this kind of material, you can skip many of the technical statistical footnotes.

The key points are summarized – and when appropriate some pointers are given – at the end of each section, and again at the very end.

Unlike what you probably did in undergraduate statistics courses, you should approach this material in the same 'big picture - not so many detailed calculations' spirit that the small group exercises are meant to convey.

A word about the two clinical examples (angina and possible hypertension) in the earlier sections: clinicians generally do not use such formal explicit statistical calculations in the management of patients. Experienced clinicians do have a very good sense of variability, and of distributions, and so they do not go as far statistically as Ingelfinger et al. (from whom these examples are taken) would have us believe. However, at your stage, it is good to be explicit and to work through the statistical issues formally, even if the exercises appear a bit contrived. Where these issues will help you is when, in section 9, you focus on epidemiologic and medical studies involving the *collective* or *aggregated* experience of many subjects. In these applications, the variability and imprecision come mainly from the *inter*-individual differences, and even clinicians have a lot less experience with this type of 'research variability.' Fortunately, the central ideas of *statistic, standard error, margin of error, confidence interval, p-value, test of significance*, etc. remain the same as those used for the clinical examples.

[1] The Clinical examples, and accompanying text, in the following are adapted from this textbook.

1 Statistics and the Individual Patient

If the clinical course of some illness were always the same in the absence of treatment and if treatment always had the same effect, it would be easy to determine whether some new treatment was an improvement. We would need only to prescribe the treatment to see whether the outcome was changed. A similar approach is still possible when the course of disease is not precisely uniform, One has to observe a sufficient number of cases of the illness and record the frequency of each possible outcome in the absence of treatment, or in the presence of a standard treatment. Then after giving the new treatment to a sufficient number of additional cases, one can see whether and how the probabilities of various outcomes have changed. The following example — evaluating the response to treatment[2] in a patient with angina pectoris — is used to illustrate, and show the consequences of, the kinds of variability that may affect clinical observations. We also show how frequency distributions are useful in the study of clinical observations that may vary from patient to patient or from time to time.

1.1 Background and Clinical Problem

Background

Angina pectoris (substernal chest pain typically brought on by exercise and relieved by rest) is a common symptom of coronary vascular disease. Angina can cause substantial morbidity by limiting a patient's activity. For many years, nitroglycerin (NTG), administered sublingually, has been used to treat angina. Usually, NTG will relieve an attack in 1 to 3 min, and most patients find NTG helpful for most of their attacks. NTG acts for only 5 to 15 min; it is impossible to prescribe this drug frequently enough to have day-long prevention of angina. One therapeutic approach uses "long-acting" nitrate preparations, although some authorities question their effectiveness. Patients have also been treated with a β blocker in an effort to decrease the number of anginal attacks.

Clinical Problem 1. Does Long-Acting Nitrate Therapy Help?

Mr. Lewis is a 55-year-old man with angina. His attacks typically occur after he has climbed half a flight of stairs or walked a quarter of a mile. He has been having about six attacks each week.

His physician recently prescribed a nitrate preparation, isosorbide dinitrate (ISDN). ISDN has a much longer duration of action than NTG, which might give it substantial advantages if it is equally effective. **Mr. Lewis called his physician later to say that he had his usual angina halfway up his 14 stairs 1 h after taking ISDN, i.e. he climbed 6 steps without angina**. In addition, he experienced headache and palpitations (both being known side effects of ISDN). He wondered whether he should stop the ISDN as he has noted no change in his angina and the drug caused him bothersome side effects.

1.2 Gathering & Interpreting Evidence from Patient

To decide whether ISDN has value for Mr. Lewis, we need some idea of **how quickly his angina attacks occur without treatment**. (Note: The comparison proposed here is ISDN versus nothing, not ISDN versus NTG as discussed previously.) If Mr. Lewis or his physician kept records of his angina before treatment started, some available information might help answer the question. Mr. Lewis did keep a diary concerning his angina attacks. The diary reads as follows:

August 16: angina on 10th step.
August 18: angina on 3rd step.
August 19: angina on 6th step.
August 20: climbed all 14 stairs without angina.

The diary has some 50 entries for the most recent 2 months. We summarize the information in the **frequency distribution**[3] shown in the 'B' row of Table 1. We also converted the data on angina experience before ISDN to a **histogram**[4] in the left portion of Fig. 1. The values are the number of steps completed without angina. Thus if Mr. Lewis got angina on the first step, he completed 0 steps without angina. If he climbed the whole flight without angina, he completed 14 steps.

Table 1 Frequency distribution of number (No.) of steps before angina for Mr. Lewis before[B] and after[A] he started taking ISDN

No. steps →	0	1	2	3	4	5	6	7	7	9	10	11	12	13	14	total
B	0	0	2	3	5	4	9	6	4	4	3	1	1	0	8	50
A	1	0	0	1	0	2	2	3	3	2	2	1	0	1	2	20

Figure 1 has two frequency scales. The one gives the observed **frequency** or count of the number of times Mr. Lewis climbed the given number of steps

[2] This Ingelfinger e.g. was also used in earlier editions; other treatments may be available today.

[3] The numbers/frequencies of instances/occurrences of each possible value of the variable.

[4] Usually, histograms have the values (or intervals) of the variable on the horizontal, and frequencies on the vertical, axis. Here, we use a different orientation, putting the two distributions side by side.

without angina. The other [in parentheses] gives the **relative frequency**, the proportion of trials out of 50 (or out of 20). Usually, we are more interested in the proportion because we may be — as we are here — comparing two sets of data based on different total counts.

Relative frequencies can be used to estimate the probability of observing angina at each step. Thus for Mr. Lewis, we estimate the unknown probability that angina will occur just after the sixth step in the absence of treatment as 9/50 = 0.18, or 18 percent.

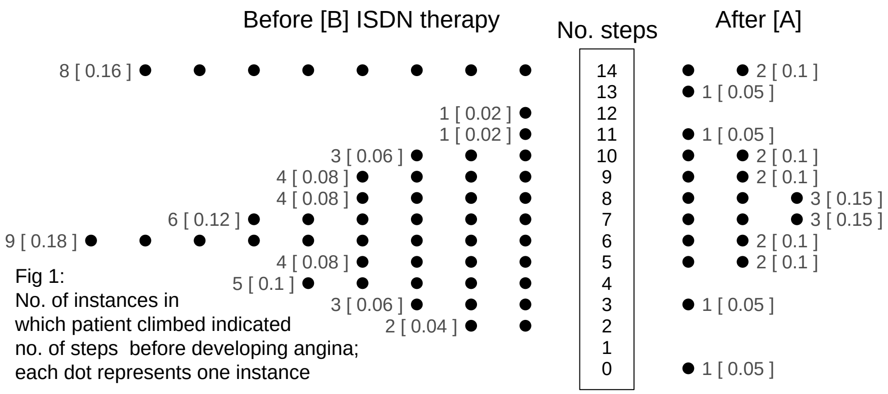


Fig 1:
No. of instances in which patient climbed indicated no. of steps before developing angina; each dot represents one instance

Interpreting One Observation

Looking at the data in the 'B' row of Table 1 or in the "Before" histogram, we see that Mr. Lewis's **new single observation of 6 steps without angina does not prove or disprove that the drug has some beneficial effect**. Already, though, we can make some estimate as to the possible response Mr. Lewis will get from ISDN. His first experience shows that ISDN does not completely prevent attacks. We do not know whether ISDN has changed the probability of angina after the sixth or any other step. To think about this, suppose for a moment that the probability of angina *at or before the seventh step* were very low, say, 1 in 100, with ISDN. Then we would have observed a rare event (probability, 1 percent) the very first time Mr. Lewis climbed his stairs after starting the new medication. Faced with this single observation, suppose that we must decide whether to believe that ISDN has reduced the probability of attacks at or before the seventh step to 1 percent. We have arbitrarily formulated two mutually exclusive possibilities:

Possibility 1. With ISDN, angina occurs at or before the seventh step only 1 percent of the time, and we have observed a rare event; or

Possibility 2. With ISDN, angina occurs on steps 1 to 7 just as often as without ISDN, and we have observed a commonplace event.

If forced to choose between these extreme possibilities on the basis of this one observation, most people would choose possibility 2. Thus even one measurement has produced a tentative conclusion about the extent of improvement from ISDN.[5]

Interpreting Several Observations

We **can learn more, of course, if we have more than one observation. Only very striking effects of treatment can be demonstrated with a single observation**. In Mr. Lewis's case, one observation would be insufficient to show a benefit even if ISDN were completely effective, because about 16 percent of the time he climbs to the top of his stairs without angina even while taking no medication. **In a given situation, the smaller the effect – gain or loss – of some treatment, the more observations will be needed to demonstrate that effect. The technical reason is that the variability of a mean — or median, or any other statistic — decreases with increasing sample size, and we measure our assurance in terms of the variability.**

Since Mr. Lewis had some unpleasant side effects from ISDN, it is a reasonable view that he should not take the drug unless he gets a "fairly large" benefit. What would be "fairly large" is hard to define in any precise way, but his physician believes that if the benefit is large enough to balance the side effects, it should be apparent after 20 or 25 observations. He instructs Mr. Lewis to continue the medication for 3 weeks and to keep a record of the point at which angina occurs each time he climbs the stairs.

1.3 Comparing Outcomes: summary statistics

Mr. Lewis returns with the data recorded in row 'A' of Table 1, which shows both the 50 observations before treatment was started and the 20 observations since then. *The general shapes of the histograms shown in Fig. 1 do not appear to differ a great deal.* Notice that the **scale of measurement for the raw frequencies differs in the two figures, out of 50 before, out of 20 after**, while **that for relative frequencies remains the same**. The observed fraction (**proportion**) of climbs without angina has gone down from 16 percent to 10 percent, a loss, and the **median** number of steps climbed before angina is now 8, whereas before it was 7, a slight gain.

[5] This is an example of two competing "hypotheses" (with ISDN, the probability of getting at least half way without angina is (1) 99% (2) 50%) to explain the observed data.

Median. The median is the *middle value* of a set of numbers when they are ordered according to size. If the number of values is odd, it is the middle number. If the number of values is even, it is the average of the two middle numbers.[6]

Examples The median of 4, 5, 5, 7, and 8 is 5. The median of 4, 5, 5, 7, 8, and 8 is 6.

Even without formal statistical analysis, it seems that Mr. Lewis has had no marked benefit from his ISDN and the continued presence of side effects suggests that it would be prudent to discontinue the medicine. (In a later section, we will use a formal way to test whether two frequency distributions differ.)

A second important approach to evaluating whether ISDN is beneficial for Mr. Lewis reviews how other patients respond to the drug. Two points are epecially important: the **proportion** of patients similar to Mr. Lewis who respond and the **degree** of improvement for those who do respond. If some patients are almost completely unresponsive, while responders tend to derive large benefits, this 2-week trial may be enough to conclude that the drug should be stopped. If almost all patients derive some benefit, but the average improvement is small, one might want to reconsider whether just 20 observations is enough to conclude that continued treatment is unwise.

1.4 Key Points

- Given the natural intra-patient variability in the (untreated) course of many diseases, conditions or risk indicators, one my need several observations of the patient to assess the effect of a treatment / intervention.
- Frequency distributions are helpful to appreciate the pattern of variability, and to assess effects of any change in management. Tables, graphs and summary statistics can be used to describe them.
- The setup and probabilities used to assess “Possibility 1. vs Possibility 2.” are a preview of the concept of a P-Value, to be discussed in section 8.
- Ingelfinger introduced the median, but didn’t say why it is sometimes preferred over the mean.[7] For small amounts of data, or data values that are already sorted by size, it is also easier to compute.

2 Biologic, Temporal, and Measurement Variation

2.1 Importance of Variation in Interpreting Outcomes

After the unsuccessful attempt to control Mr. Lewis’s angina with ISDN, he went without treatment. His clinical state was apparently unchanged for 5 months, at which time he told his physician that the angina had recently begun to appear more often and on a lower step than before. His physician must now consider three broad kinds of reasons for the change. First, Mr. Lewis may have suffered a biologic change (his coronary disease may have worsened). Second, Mr. Lewis’s angina may be temporarily worse for no apparent reason, just as exercise tolerance is higher on some days than on others. Finally, Mr. Lewis may have become a more (or less) accurate observer or reporter of his angina.

> **Generally, clinical observations are subject to three sources of change, which may be called biologic, temporal, and measurement variation.**

In evaluating the status of Mr. Lewis, his physician was at first concerned with whether ISDN caused a biologic variation in his anginal pattern. This evaluation was made difficult because angina has a great deal of temporal (day-to-day) variation and perhaps some measurement variation as well. In discussing the trial of ISDN reported by Danahy et al.,[8] we asked whether patients who responded well to ISDN were biologically different from the poor responders in some permanent way (interpatient variation), whether the observed variation in response to ISDN might reflect only the day-to-day variation of patient’s responsiveness (intrapatient variation), or whether some combination of these was at work.

When a series of observations is made on different individuals, the variation in responses is due to both **intersubject variation** (secondary to biologic, temporal, or measurement differences between the subjects) as well as **intrasubject variation** (also due to biologic, temporal, and measurement variation within a subject). To distinguish the contribution of each source to the overall variation, a series of separate observations on separate persons will not do. One has to study the same individuals more than one time to see whether the individual frequency distributions are similar to each other and, hence, to the frequency distribution for the population. For instance, obviously patients’ heights vary

[6] The median is *more resistant* to the influence of extreme observations than the mean is, and thus is a better indicator of the “middle” if the distribution is not symmetric.

[7] See footnote above.

[8] In section 4-2, omitted here. Section 4.4, also omitted here in the interests of time and space, uses data from a crossover study of several patients to asses whether Mr. Lewis is likely to respond to Propranolol, and if so the magnitude of his response.

widely, and the source is intersubject variation in height. Body temperature in patients at the outpatient clinic also varies. However, it is likely that most of that variation is due to intrasubject variation in body temperature, since unusual temperatures may be a symptom associated with going to the clinic.

2.2 Implications for Patient Care

A number of principles follow from recognizing that a clinical observation is subject to biologic, temporal, and measurement variation, and that each of these sources may be reflected in intra- as well as intersubject variability.

i. In conditions that have large temporal and/or measurement variation, therapeutic efficacy or other biologic changes may be difficult to detect even with large numbers of well-controlled observations.

ii. The "normal range," as determined by observing many individuals, is usually greater than that determined by observing one individual many times, unless there is little interperson variation. We often use the rather arbitrarily chosen range 2.5 to 97.5 percent, the central 95 percent of a sample of values obtained from normal subjects, as the normal range of a measurement. Therefore, it includes both inter-person and intra-person variability.

iii. Some patients seek their doctors' attention when their conditions seem to worsen. If the worsening simply represents temporal and not biologic variation in their illness, their illness is likely to improve irrespective of therapy. ("Most things, in fact, are better by morning."-Lewis Thomas.)[9]
The technical name in statistics for such changes is the "regression effect," meaning regression toward the mean. Thus a patient who feels spectacularly well today will probably not feel as well tomorrow.

iv. The physician who observes a patient numerous times, or orders numerous laboratory studies, may observe "abnormalities" that do not reflect a biologic variation but are due to temporal/measurement variation. These, too, are likely to be "better" or changed soon. This is why, when you are faced with a test result that does not seem to fit, it helps to repeat the test.

2.3 Key Points & Some Pointers

- Knowing the relative magnitudes of these various sources of variation is key. Points i-iv in section 2.2 highlight this importance.

3 Distributions

Frequency distributions, relative frequency distributions, and histograms are convenient (and equivalent) methods of summarizing collections of multiple observations. Typically, a frequency distribution is obtained by dividing observations into 10 to 20 classes such that each observation must fall into one and only one class. In a frequency distribution, the number of observations belonging to each class is recorded.

> **The relative frequency distribution assigns to each class an estimated probability (observed relative frequency) that an observation will be in that class. If conditions are constant, the larger the sample, the better the estimate. The estimated probability of each class is easily computed as the number of observations in that class divided by the total number of observations.**

Measures of Location

Many questions in medicine hinge on determining whether one probability distribution differs from another. Such a difference may be difficult to determine because the distributions themselves are unknown and must be estimated with some degree of error, also unknown. Estimated distributions can be compared in terms of many different properties. Perhaps the most important is the "centre" or location of the distribution, which may be defined as:

- The **mean**: the ordinary average of the observations, or
- The **median**: defined earlier, or
- The **mode**: the most popular (frequently occurring) value.

Measures of Spread

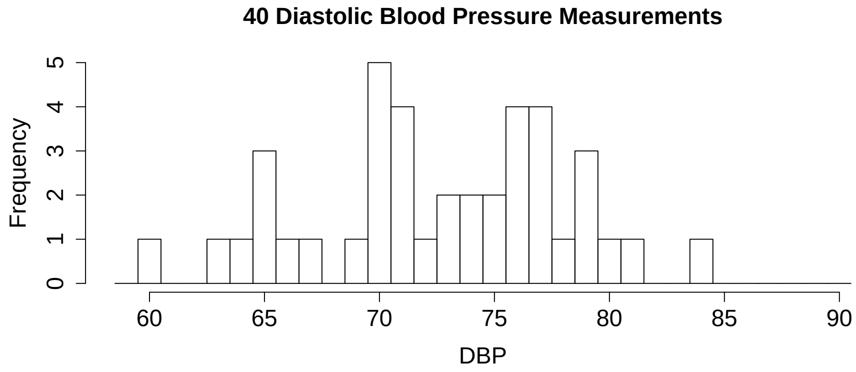
Another important property is the degree of "spread" or dispersion of observations about their centre. "Spread" may be defined in several ways, such as:

- The **range**: the difference between the largest and smallest observed values. This should not be confused with the "normal range" discussed earlier, although the ideas are similar. In the distribution shown below, the range is $84 - 60 = 24$ mmHg.
- The **interquartile range, "IQR"**: the range of values remaining when the largest 25 percent and smallest 25 percent are – temporarily – set aside.

[9] Or "if you go see you doctor about it, your cold will be better in a week. If you don't, it will be better in seven days".

Sometimes these *quartiles* are called Q_1 and Q_3, and sometimes Q_{25} and Q_{75}. In the distribution shown below, the 10th smallest value is 70 – as is the 11th; and the tenth largest is 77 – as is the 11th largest; setting these ten smallest and ten largest aside, we now have an IQR from $Q_1 = 70$, to $Q_3 = 77$ mmHg.

- The **standard deviation ("SD")**: a frequently used measure of spread, especially if the distribution is roughly bell-shaped. Technically, it is the square root of the average of the squared deviations from the mean. Thus, for more than a few values, it requires a scientific calculator, or the `STDEV` function in `Excel` to compute it. For most purposes, it can be estimated visually from a frequency distribution as the average absolute deviations of each value from the mean. Or, if the distribution is approximately Gaussian (bell-shaped), we can use the fact that about two-thirds of the observations lie between one standard deviation above and one standard deviation below the mean (or if we have enough observations, that about 95% of the observations lie between two standard deviations above and two standard deviations below the mean) to obtain a rough estimate.



Since about 1/6 of the observations for a Gaussian distribution fall beyond 1 standard deviation at each end, we might count in 1/6th of the observations from each end, get the distance between these two points, and divide by 2 as an estimate of the standard deviation. In the above example, with 40 blood pressure measurements, 1/6 is about 7. the seventh highest DBP is 78 and the seventh lowest is 66, so we estimate the standard deviation as (78-66)/2 = 6 (by calculator, the SD is 5.5).

- The **coefficient of variation ("CV")**: This is used when comparing the degree of measurement error or intra-person or inter-person variation between situations or persons with very different means or units. Examples include the intra-assay and inter-assay and true biologic short-term variation in the measurements of Prostate Specific Antigen (PSA), when the true (or average) value is 2.5 ng/mL vs. 10 ng/mL vs. 50 ng/mL; day to day variation in the calorie intake, or energy expenditure, of a 5 vs. a 25 year old, or the person-to-person variation in the amount of annual outdoor activity in Canada and Australia, or the person-to-person variation in the height (in inches) of adult females in the U.S. (SD measured in inches) vs Canada (SD measured in cm). The variability is more easily appreciated/compared if we *express the SD as a percentage of the mean* or *as a percentage of the known*[10] *value.*

$$CV = \frac{Standard\ Deviation}{Mean\ Value} \times 100\% \quad \text{or} \quad \frac{Standard\ Deviation}{Known\ Value} \times 100\% \ .$$

3.1 Key Points & Some Pointers

- Beware of the word 'average': during a past labour dispute – and before salary caps – the NHL owners told us the average NHL player's salary was one million dollars; the players' association, using the same data, told us it was half a million. You should be able to tell which group was using the *median* and which was using the *mean*. If need be, sketch the frequency distribution. The mean is further out in the longer tail.

- Don't fuss about the exact formula for the SD. The approx. method described above is good enough for the purposes here.

  The French term for SD is much more expressive: écart-type, *typical* deviation. Calling it the *standard* deviation doesn't enlighten us.

- Understand the CV. There are various versions, depending on the context. It is a useful measure, especially when dealing with the individual patient.

[10] If a lab assays specimens with a known concentration obtained from a Standards Bureau.

4 Biologic, temporal, and Measurement Variation - Example 2

The diagnosis and treatment of patients with high blood pressure is another clinical situation that forces us to consider the variability and distribution of blood pressure measurements in the individual patient. In this section we examine the blood pressure variation that might be observed in one office visit, and the variation from one visit to the next. In the next section, the concept of a confidence interval helps us when we try to **estimate a patient's average (mean) blood pressure**. Applying these ideas helps us to determine what we need to observe and why, and how to detect changes in blood pressure in response to antihypertensive therapy or other interventions.

4.1 Background and Clinical Problem

Clinical Problem 2. Moderately Elevated Blood Pressure at a Routine Physical

A company refers Mr. W.P., a 35-year-old computer programmer, to you for a pre-employment physical. He has a family history of stroke, he smokes one package of cigarettes a day, and his blood pressure is 130/95 mmHg.

Background

The following statements are excerpted from the recommendations of the Joint National Committee (JNC) on the evaluation and treatment of high blood pressure (Joint National Committee on Detection, Evaluation, and Treatment of High Blood Pressure, 1992) two years before the Ingelfinger text was written.[11]

> [The table] provides a new classification of adult blood pressure based on impact on risk.... All stages of hypertension are associated with increased risk of nonfatal and fatal CVD [cardiovascular disease] events and renal disease. The higher the blood pressure, the greater the risk.

Table: (1992) Classification of blood pressure for adults 18 years of age & older

Category	Systolic (mmHg)	Diastolic (mmHg)
Normal	< 130	< 85
High normal	130 – 139	85 – 89
Hypertension		
Stage 1 (mild)	140 – 159	90 – 99
Stage 2 (moderate)	160 – 179	100 – 109
Stage 3 (severe)	180 – 209	110 – 119
Stage 4 (very severe)	> 210	> 120

When systolic and diastolic pressure fall into different categories, the higher category should be selected to classify the individual's blood pressure status.

4.2 Variability of Blood Pressure in the Individual Patient

One's first impulse is to decide that Mr. W.P. has a diastolic blood pressure between 90 and 99 mmHg, placing him in the mild hypertension category. This view may turn out to be correct, but before settling on it, let us review the variability of blood pressure measurements.

Armitage and Rose (Clin Sci 1966, v.30, pp325-335) provide an instructive set of data showing how diastolic blood pressure varies in the individual. The Figure on the next page shows duplicate readings of casual diastolic blood pressure (10 subjects; 2 readings on 20 occasions). Even if we exclude the extreme right-hand measurement for subject 10, some subjects have ranges of measurements (largest minus smallest) of more than 30 mmHg. Therefore, Mr. W.P.'s **measurement of 95 could possibly be a high measurement for him, and perhaps he averages 15 mmHg lower, which would take him out of the hypertensive range. Or 95 might be a low measurement for him, and his average would be, say, 10 units higher, which would take him into the moderate category.**

The message from the measurements on these 10 patients – and from ambulatory BP monitoring (see e.g. in Ingelfinger) – is that Mr. W.P.'s diastolic blood pressure of 95 is ambiguous. Before we proceed, we need to understand the (im)precision of statistical estimates based on the mean of n values – including the above data where we have just $n = 1$. To keep the side-issues to a minimum, we use a simpler and more generic example to explain the **key statistical concept – a Confidence Interval** – before returning to the case of Mr. W.P.

[11] Last year, you saw the *current* recommendations; the *statistical* point remain unchanged.

4.3 Key Points & Some Pointers

- The pattern of variation in the data collected by Rose et al. is very instructive. Do you notice any digit preferences? How much do you think the pattern would be affected if we used an automated BP machine such as the one found in pharmacies?

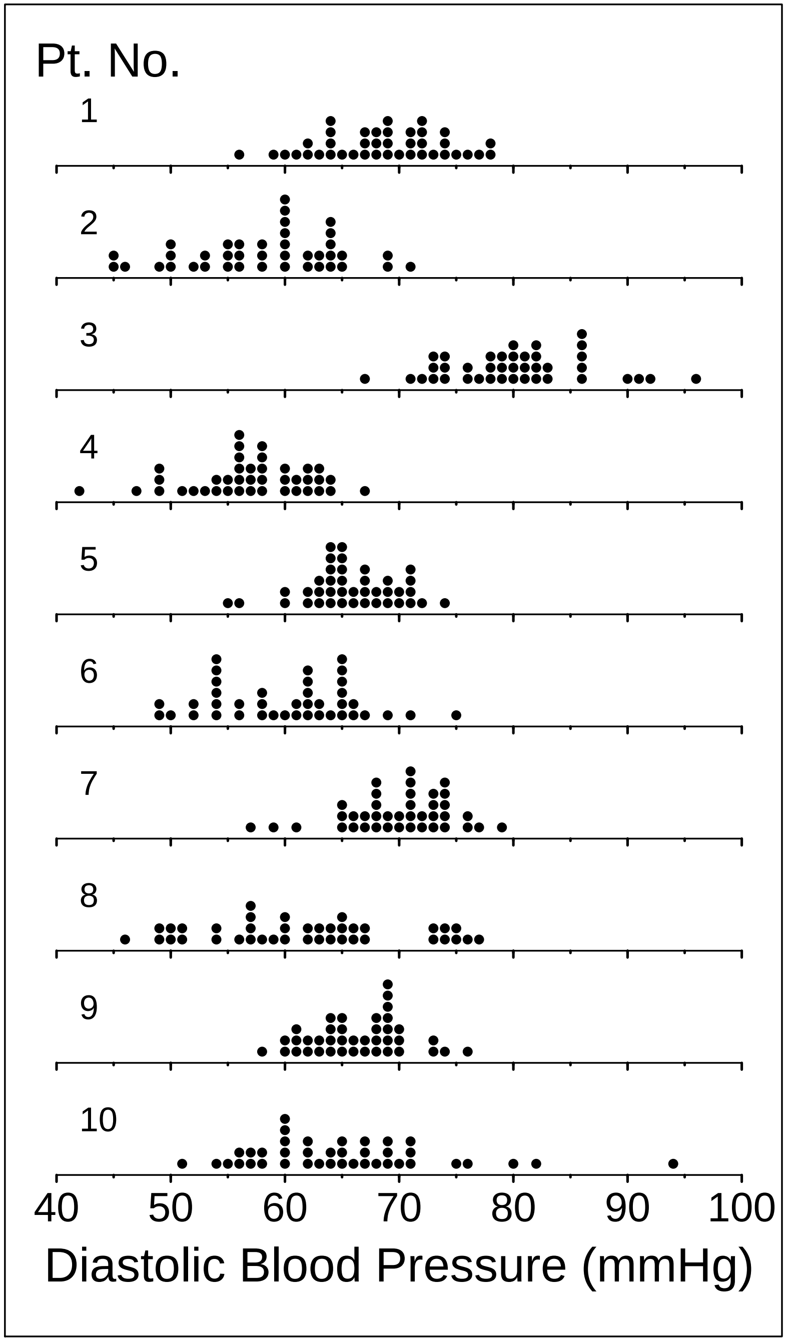
- The variation, especially the intra-patient variation, in BP is much more amenable to summarization using a mean and SD than the 'onset of angina' data in the earlier section. Part of this has to do with the 'ceiling effect' (no pun intended) with the angina data, since the patient did not continue climbing past 14 steps. Had he, we might have see a more 'unimodal' distribution.

- Trying to reliably classify the patient on the basis of one BP measurement is clearly impossible, just as it is to classify someone as an A or a B student on the basis of a single multiple choice exam in one course, or establish a taxi-driver's or waiter's income bracket on the basis of a single day's income.

- We need to appreciate how much we can reduce the noise by averaging several measurements, and the amount of 'noise' that remains. The next section (5) does so, in probably more detail than you will need, or have time to go through.

- For section 5, if you are short of time, focus on the pattern of variation in the diagrams on p. 11, and the SE formula in the second column of page 11, before going on to the **key sections 5.4 and 5.5**. If you find the 'chat' in the first column of page 12 a bit too much, just focus on the diagram on that page and the 'mean ± some multiple of SE' in the material below it.

- The distinction between Standard Error and Margin of Error is important, as is how the Margin of Error involves the degree of confidence.



5 Confidence Interval

QUANTIFYING THE (IM)PRECISION OF A STATISTICAL ESTIMATE

5.1 Statistic: sample ; Parameter: population

In *clinical* medicine, the focus is on the one ('single') patient, and on obtaining a good (i.e. precise, reproducible) estimate of the true – *but unknowable* – mean level of that patient's behaviour or symptom or sign, or level of some biological variable. We have seen the examples of the median or mean number of steps before a patient develops angina, or the proportion of times the patient can climb the entire stairs without angina, or a person's mean systolic or diastolic blood pressure. We could add another – mean cholesterol level or BP.[12]

Formal Terminology: Statisticians and statistical textbooks refer to the true mean level or proportion – or correlation or regression coefficient – as a ***parameter***, and often denote it by a Greek letter such as μ or π – or ρ or β. Mathematical statisticians typically denote a generic parameter value by the Greek letter theta (θ). It may help to remember this terminology by noting that both the words parameter, and 'population' (i.e. the 'universe') start with the letter p.[13] This is different to the lay use, where cholesterol itself would be a parameter, and blood pressure another. In statistics courses, the parameter refers to something more specific: it is some unknowable *property* of a distribution in some universe of (say cholesterol) values, such as the *mean* or the 95-th *percentile*, or *proportion* above some threshold.

Typically, the parameter value ("θ") is unknowable, because it is not practical to measure the level continuously or exhaustively, and thus have a perfectly precise estimate.[14] In *your own life*, you would probably not be willing to document daily activities and behaviours, such as alcohol consumption, commute time, time spent on the internet, how often you ate restaurant rather than home-cooked meals, or drove while talking on a cell phone, etc. in order for someone (even yourself) to obtain a precise picture of you. But you might be wiling to go through a few "24-hour recalls" (a *sample* of your experience) over the time-span of interest.

In statistical jargon, the summary value calculated from the values in a ***sample*** is called a ***statistic***, and is typically denoted by a Roman (Arabic) letter such as $\bar{y}$ (a mean) or p (a proportion) or r (a correlation) or b (a regression coefficient). It may help to remember this terminology by noting that both the words 'statistic,' and 'sample' start with the letter s. [appendix note 1.]

In *community medicine*, the focus is on a larger target – the *entire population* – and on obtaining a good (i.e. precise, reproducible) estimate of the true – *but, again for most variables, unknowable* – mean level of some biological variable, or activity, or behaviour, or the proportion (if an all-or-none, or otherwise binary, variable). Again, these mean levels or proportions are unknowable, because it is not practical for community medicine personnel to measure everybody, and thus have a precise estimate.[15] Thus, they depend on sample surveys. If McGill wished to estimate how much, on average, its students spend on accommodation, etc. it would probably have to do the same.

In the more precise sciences, one can control much more of the variability: some of the variation in measurements is unwanted and a nuisance, and thus is called measurement or experimental *error*; this unwanted measurement component is also present in clinical research, but is not always entirely separable from real – and often interesting in its own right – variation within and between individuals.

No matter the universe and quantity of interest (the mean for one individual, or for the population, or the value of a physical constant such as the speed of light, or a chemical determination), a summary number, such as a mean or a proportion, calculated from a small set (sample) of variable measurements or variable individuals will – despite the benefit of basing it on several observations, and of using scientific ways to decide which measurements or persons constitute the sample – not equal the (unknowable) value one would have obtained had one been able to make all of the possible measurements.

5.2 How far can an estimate from a *limited* amount of observation be from the "*true*" quantity of interest, e.g., how far can $\bar{y}$ be from μ? p from π?

Intuitively, if all of the possible measurements are highly variable about this true – *but unknowable* – mean value, one needs to average many independent values in order to arrive at a reproducible (precise) estimate of that true value; if they are highly concentrated about this true value, one needs to average fewer independent values in order to arrive at a reproducible (precise) estimate of that

[12] Monitoring cholesterol levels: measurement error or true change? Glasziou PP et al. Ann Intern Med. 2008 May 6;148(9):656-61. BP: Keenan K. et al. BMJ 2009

[13] If the focus is on one person, the 'population' analogy has less meaning; instead just think of the true mean value for the individual, or, if measuring the speed of light, think of '*c*' as the true value, or if quality-controlling assays, a true cholesterol concentration in the specimens supplied by the Standards Bureau.

[14] Except in a few instances such as continuous ambulatory monitoring of say BP or sugar levels, or activity, and even then we are limited to short spans of time.

[15] Some exceptions are data collected at the census, or in annual income tax returns, or motor vehicle registrations, or recorded in administrative databases such as RAMQ and MEDECHO, which document every medical claim for a physician visit, or hospital admission.

true value. Thus, e.g., if in order to estimate the average of entries per page in the telephone book, in order to multiply it by the total number of pages to obtain an estimate of the total number of entries in the book, I would imagine we would not have to sample (and count the entries in) very many pages, since the number of entries does not differ very much from page to page. But if we wanted to estimate precisely the average length of words in the New York Times, we might need several hundred. How many would you think we need in order to put Mr. W.P. fairly securely into one of the DBP categories in the JNC table above? Its a bit like GPS, and what level of technology you need to pinpoint the location of an object to within 5Km or 5m or 5cm of its true location.

The answer depends on how far 'off the target' an estimate one can live with, or how much effort and resources one is willing to spend to get an estimate closer to the true value. Unfortunately it is not guaranteed that a larger sample size will necessarily get you closer to the target, since by the luck of the draw it could turn out that an estimate based on $n = 4$ is closer to the target than another one based on $n = 8$. But the *probability* of being within a certain specified distance of the target is *higher* with a sample of $n = 8$ than one based on $n = 4$. So. its a matter of *probabilities*, not of *certainty.*

When we combine independent measurements, **the statistical (probability) laws governing how far a statistical estimate,** such as a sample mean $\bar{y}$ or a proportion p, **falls from the true value**, such as the true mean μ or proportion π, **are determined by surprisingly few factors**. The pattern of variation of *individual* measurements may be quite non-Gaussian. However, the distribution of the possible estimates can be remarkably close to a Gaussian distribution (bell-curve) – centered on the true value. Thus, we can make statements about the probability of any one estimate (the one we are going to produce) being within a certain distance of the truth.

You might well ask: Leger and Leger, and Gallup, and other 'measurers' can't know the true value[16] so how can they check whether their estimates are within the stated distance from the true value. How can they give those "95 of our estimates%[17] are within the stated distance from the truth" guarantees?

5.3 The 2 mathematical laws that quantify how much a sample mean or proportion can be an under- or an over-estimate - and a hypothetical example

Just as with eclipses to check Einstein's predictions, there are a few occasions (e.g., at elections) when we can directly check statistical predictions. These probabilistic guarantees derive from 2 statistical laws which apply equally to the mean/proportion in each of the possible samples of size n. Each sample would give a different "estimate" The *hypothetical* distribution of "all possible estimates" is called the **sampling distribution**. **The 2 laws are:**

- The possible estimates *would* fall around the true value in a pattern that *would* be close to a **Gaussian (bell-shaped) distribution**.
- The **standard deviation** of this sampling distribution *would* be $\sigma/\sqrt{n}$, where σ is the standard deviation of all of the individual units in the universe (context) of interest.

We now have **two standard *deviations*** – the original one that quantifies the **variation of individual values**, and now a new one (hypothetical) that quantifies the **variation of the statistic (estimate)** across all possible samples of size n. Even though theoretical statisticians are quite comfortable using the same term for both, applied scientists tend to use a different term – the **standard *error*** ("**SE**") to refer to the sampling error in the estimate.

In practice, since one usually does not know the value of σ, one cannot then calculate the value of $\sigma/\sqrt{n}$; so, instead they first *estimate* σ from the sample itself, and use this estimate, $\hat{\sigma}$, to instead calculate

$$SE(\text{estimate}) = \hat{\sigma}/\sqrt{n} = (SD \text{ of the } n \text{ individual sample values}) \div \sqrt{n}.$$

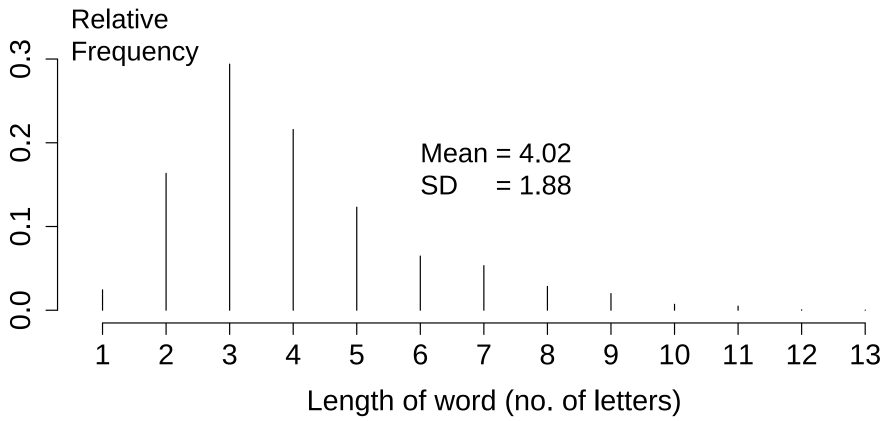
With a small sample size, say $n < 30$, the SD of the sample values may not be close to σ. To account for this additional uncertainty/'noise', we need to use a replacement for the Gaussian curve, nowadays called the t-distribution.[18] Having gotten out of this catch-22, we can now move ahead.

[16] After all, that is why they are estimating using a sample!

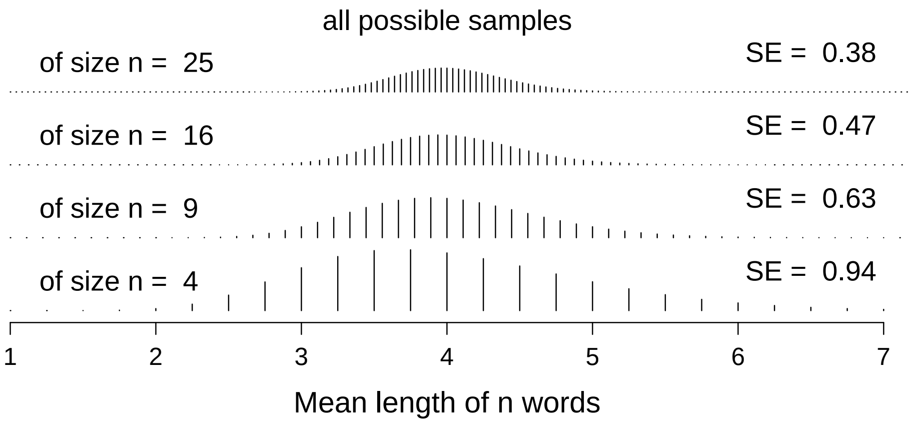
[17] They say "19 times out of 20."

[18] This replacement was worked out in a paper published 101 years ago, by a brewer/chemist who worked for the Guinness brewery in Dublin, and who was often working with samples as small as $n = 4$. His name was William Gosset, but he published under the nom-de-plume "Student." As you might imagine, there is a different distribution/curve for each n. The one for $n = 4$ is far wider than the Gaussian one that applies for the $n = \infty$ that allows σ to be perfectly estimated, whereas the one for $n = 30$ is only slightly wider than the Gaussian one. For example, for the 95% range, the limits are ±3.182 for $n = 4$, ±2.26 for $n = 10$, ±2.05 for $n = 30$, and ±1.96 for $n = \infty$ (the Gaussian range).

To see that n **doesn't necessarily have to be that large**, as long as the distribution of individual observations is not *too* skewed, consider the variation in the length of (number of letters in) individual words. In a famous text, several thousand words long, the mean, μ, is 4.02 letters; the SD, σ, is 1.88. The distribution, in the first figure, has a long right tail.



However, the distributions of the **means** of all the different possible samples of a given size, shown in the second figure, are much closer to Gaussian. When $n = 4$, the sampling distribution of all possible $\bar{y}$'s still has a slightly long right tail, but if one uses $n =$ **25, the sampling distribution is quite close to Gaussian**.



The two diagrams above show these these laws in action. The SD for the lengths of individual words is $\sigma = 1.88$. Note how skewed the distribution of the lengths of individual words is, and how close to Gaussian the distribution of possible sample means is.

An Aside...

Is this SD really an SD? Is there *this little* variation in head sizes?

Stephen Jay Gould's book "The Mismeasure of Ma" discusses a table from a 1978 article by Epstein. Gould read the original article and found that "a glance at Hooton's original table (The American Criminal, v. 1, Harvard U. Press, 1939) reveals that the *SE* column had been copied and re-labelled *SD*" Then, using this SD, and the n, to compute a much smaller-than-it-should-be SE, Epstein was able to "show" that the CI's for mean head circumference for people of varied vocational statuses did not overlap, and thus that there were "statistically significant" inter-group differences.

The astute reader would have noticed that the "SDs" in the table should not decrease with increasing n. Yes, SDs calculated from small n are *less stable* than those calculated from large ones, but the SD from a smaller n is as likely to be greater than the SD from a bigger n 1 as it is to be smaller. If SD's were smaller (some argue larger) in larger samples, then the SD of the diameters of red blood cells should be different for a large adult than a smaller adult!

The last column is in fact a column of S$\underline{\text{E}}$'s; if you back-multiply by each $\sqrt{n}$, you will find that the 7 SD's implied by the last column range from just 7.6 to 12.5, and with no obvious correlation with the n. Also, an SD of 10-12cm (CV≈2%) makes sense (think of hat-sizes!).

Mean and standard deviation of head circumference for people of varied vocational statuses*.

Vocational Status	N	Mean (in mm)	"S.D."
Professional	25	569.9	1.9
Semiprofessional	61	566.5	1.5
Clerical	107	566.2	1.1
Trades	194	565.7	0.8
Public service	25	564.1	2.5
Skilled trades	351	562.9	0.6
Personal services	262	562.7	0.7
Laborers	647	560.7	0.3

5.4 How do these laws help us?

Statistical laws help answer the question: how far a possible estimate might be from the true value? i.e., how far $\bar{y}$ might be from μ?, how far p might be from π?

Our 'mean length of words in a famous text' example in the appendix is a contrived one: why would we just use a sample of $n = 25$ or even $n = 100$ if we already *know* the parameter of interest, namely $\mu = 4.02$, in the full text? The point of showing it is to convince you that the formula can also be expected to work in situations where we use the mean or proportion in a sample to *estimate* a population mean μ or a population proportion π.

Since we know μ in this contrived example, we can use the Figure to verify that indeed (approximately) 68% of the possible estimates fall within 1 SE of μ, 80% fall within 1.28 SE's of μ, 95% within 1.96 SE's of μ, etc.

In practice, we are interested in the reverse: how far might the true value be from the estimate we actually observed in the sample ? i.e., how far μ might be from $\bar{y}$? how far π might be from p ?

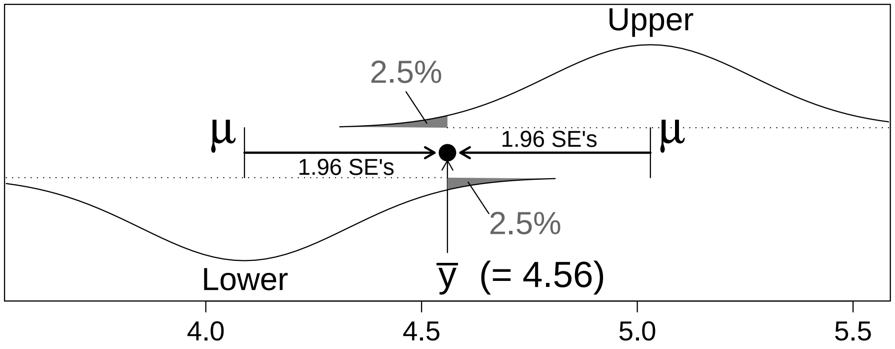
To estimate the average word length ("μ") in William Harvey's 1628 treatise *On The Motion Of The Heart And Blood In Animals*, JH took a random sample of $n = 100$ words from the treatise. The mean length of these 100 words was 4.56 letters, and the standard deviation of the 100 lengths was 2.40 letters. Thus, our "**point estimate**" of μ is 4.56, but this may be an **under-** or an **over-estimate** of μ. Can we work backwards and **bracket (put limits on)** μ?

Here is where the 'hypothetical' or 'what if' can help us. We will 'try out' various values of μ and see how 'far' or how 'extreme' – probabilistically speaking – our 4.56 is with these various trial values. We will keep ('*rule in*') those trial μ values against which the 4.56 is not extreme, and exclude ('*rule out*') those trial μ values against which it is.

We will say the sample mean is 'extreme' if the probability of a sample mean this far away, or further away, from μ is less than 2.5% in either direction. In a Normal (Gaussian) distribution, this corresponds to a value that is 1.96 standard deviations from the mean.[19] Conversely, 'not extreme' will thus denote any value that is less than 2 standard deviations from the mean. In our context, **since we are dealing with a statistic**, i.e., a value calculated from an aggregate of observations, we will use the more descriptive "*standard* ***error***" of the statistic, in keeping with our convention to reserve the term standard *deviation* for the variation of *individual* values, i.e. the lengths of individual words in our example.

Lets start with a trial μ of say 5.4. *If* the (unknown) μ were indeed 5.4, then our observed 4.56 would be an under-estimate. But is it plausible to have *this large* an under-estimate?

The 'laws' above tell us that the probability of obtaining an estimate as low as, or lower than the one we observed, *if indeed* μ were 5.4, can be calculated using a Normal distribution centered on 5.4, and with a SE of $2.40/\sqrt{100} = 2.40/10 = 0.24$. Under this scenario, the observed value of 4.56 is 0.84 letters below the 5.4 we are currently entertaining as the mean for the entire treatise. Since 1 SE is 0.24 letters, '0.84 letters below $\mu = 5.4$,' corresponds to an observation that is $0.84/0.24 = 3.5$ SE's below $\mu = 5.4$. This makes the 4.56 letters 'extreme' relative to this μ. Thus we need to move our trial value of μ downwards, in the direction of the 4.56, so that the 4.56 is no longer extreme relative to it.



In order to find the scenario in which the 4.56 *is just at the boundary between extreme and not*, we therefore need to have 4.56 be 1.96 SE's below μ. To do this, we solve

$$\mu - 4.56 = 1.96 \times SE,$$

to obtain

$$\mu = 4.56 + 1.96 \times SE = 4.56 + 1.96 \times 0.24 = 5.03$$

The value of 5.03 will thus serve as the '**upper limit**' for μ. It is written as $\mu_{upper} = 5.03$ or $\mu_U = 5.03$ for short.

Now, it is easy to see how to get the lower limit for μ, the value against which the observed 4.56 is just as extreme an over-estimate. We need to have 4.56 be 1.96 SE's above μ. To do so, we solve

$$4.56 - \mu = 1.96 \times SE,$$

[19] This 1.96 is often rounded to 2, but here JH prefers to leave it at 1.96, since a 2 might be mistaken for something else.

i.e.

$$\mu = 4.56 - 1.96 \times SE = 4.56 - 1.96 \times 0.24 = 4.09$$

Thus our **L**ower and **U**pper **limits** for μ are $\mu_L = 4.09$ and $\mu_U = 5.03$ letters, respectively. The interval, or range of parameter values, between these two limits is called a **Confidence Interval** for the parameter μ.

Properties of a Confidence Interval

As shown in the Fig. on p 12, these limits were constructed so that the lower 2.5% of the distribution centered at μ_U is excluded, along with the upper 2.5% of the distribution centered at μ_L. This allows us to say that the interval within these limits is a **"95% confidence"** interval. Confidence intervals (CI's) are often misunderstood, and so you need to appreciate exactly **how a CI should — and should not — be interpreted.** One helpful way to interpret it is to understand that 95% of *all* 95% confidence intervals – 'trap' or 'include' the parameter value one would obtain with an infinite sample size.[20] Thus, absent non-sampling biases, and selective disclosure/publication, sample survey companies, and scientists who publish results based on finite samples, would be justified in putting the claim

> "On average, for every 100 "95% CI's" we supply/publish, on average, 95 of them[21] include the true parameter value."

in their brochures. This wording emphasizes that the "95% confidence" arises from the daily applications of the statistical procedure.

The difference between this type of claim, and the ones made by surgeons ("the procedure is uneventful in 95% of patients like you"), or those who sell you a product ("its works for 95% of my customers") or give you a recipe for making a soufflé ("almost surefire") is that if you go for it, you will find out if you were the 1 in 20 or the 19 in 20. But, with a 95% CI, you don't usually get to find out if it was successful, i.e., if it did in fact 'trap' or 'include' the parameter value. The only exceptions are if someone subsequently measures the entire universe, or if your and all subsequent estimates made by others are combined (in what is known as a meta-analysis) so as to have a quite narrow confidence interval.

5.5 Confidence Intervals

Anatomy / Components of a Confidence Interval ("CI")

- *Isn't it always "your answer ± something"?*

In the exposition above, we took a somewhat tortuous route to arrive at what seems like a simple formula for the CI limits for a parameter θ:

$$\textit{Upper and Lower Limits} = \textit{estimate} \pm \textit{ some multiple of its SE}$$

The end result raises the obvious question: "why not simply *define* a CI by this formula?" The answer is that this simplistic formula does not always work. It works fine for a Statistics Canada survey when they observe that the proportion of 'positives' in a sample of $n = 900$ is $\hat{\pi} = 0.20$, and, using $\hat{\sigma} = 0.4$,[22] calculate the 95% CI as

$$0.20 \pm 1.96 \times 0.4/\sqrt{900} = 20\% \pm 3 \textit{ percentage points}$$

But what if, in a phase II study, an experimental treatment for advanced cancer showed no response in $n = 4$ consecutive patients, so that the estimated proportion of successes is $\hat{\pi} = 0/4 = 0.00$? Should the 95% CI for π be

$$0.0 \pm 1.96 \times 0.0/\sqrt{4} = 0\% \pm 0 \textit{ percentage points ?}$$

No! Even if this treatment would help an average of 1 patient in 3, i.e., if $\pi = 0.33$, it would not be that surprising to 'strike out' in the first 4: the probability of 0 successes would be $0.67^4 = 0.20$, or 20%, not *that* unlikely.[23] For extreme situations, we must determine the (asymmetric) limits by separate trial-and-error calculations for the upper, and lower limits, and using exact distributions (e.g. Binomial for proportions, Poisson for rates involving person-time denominators), rather than ill-fitting Gaussian-approximations. Thus, based on observing 0 successes out of 4, the 95% lower limit for π is (naturally) zero, while the 'exact' 95% upper limit is 60%.

[20] I hesitate to say the 'true' value, because non-sampling errors, such as biases in selection or participation of subjects, or in the measurement instruments. Unlike sampling variation, these biases can not be reduced by taking a larger sample size. Thus, if we used the question "Are you bilingual in Canada's two official languages" (rather than the more 'operational' definition "Can you keep up a conversation in both for at least 15 minutes?"), asking this question of *everyone,* e.g. at census time, would will not remove the over-statement that this type of question invites.

[21] Polling companies often use "19 times out of 20" instead of 95%.

[22] The SD of a sample of 0's and 1's, with 20% 1's and 80% 0's, is $\hat{\sigma} = \sqrt{0.8 \times 0.2} = 0.4$.

[23] See Hanley JA, Lippman-Hand A. If nothing goes wrong, is everything all right? Interpreting zero numerators. JAMA. 1983 Apr 1;249(13):1743-5.

What is the quantity after the ± called?

Answer: the **Margin of Error**.

What factors determine the magnitude of the Margin of Error?

The Margin of Error is a multiple of the Standard Error (SE), so the two determinants are

i. The multiple (i.e, *the number of SE's* in the table below), which in turn is determined by the the "degree of confidence" used. The multiples in the first row are for large enough sample sizes that the sampling distribution is closely approximated by a Normal (Gaussian, 'z') distribution; when the sample size is smaller, somewhat larger multiples are used:

Multiples of SE for different confidence levels:

Confidence →	50%	60%	70%	80%	90%	95%	99%	99.9%
Normal('z')	0.67	0.84	1.04	1.28	1.64	**1.96**	2.58	3.29
$t,\ n = 30$	0.68	0.85	1.06	1.31	1.70	**2.05**	2.76	3.66
$t,\ n = 15$	0.69	0.87	1.08	1.35	1.76	**2.14**	2.98	4.14
$t,\ n = 5$	0.74	0.94	1.19	1.53	2.13	**2.78**	4.60	8.61

ii. The SE, which in turn is proportional to σ (the variation of individual values) and inversely proportional to $\sqrt{n}$ (the square root of the sample size).

Thus, if one wishes **to halve the SE**, and thus halve the width of the CI, one would need to **quadruple (not double!) the sample size.** See the SE's for the sample mean of a sample of various numbers of words. The SD of *individual* word lengths was 1.88; the SD for *means* of all possible samples of size $n = 4$ was $1.88/\sqrt{4} = 1.88/2 = 0.94$; for means based on samples of size $n = 9$ it was $1.88/\sqrt{9} = 1.88/3 = 0.63$; etc.

You might be tempted to narrow the CI by taking a smaller multiplier (i.e., move to left in the table above). But, if you do, you also diminish the level of confidence. For example, if asked to give – out of your head, without checking with historical meteorological records – a confidence interval for the *mean* temperature in July in Montreal, you could give a quite narrow one (I might give "21.2C to 21.5 C" but I would not be very confident that this covers the true value. On the other hand, I could always give "15C to 35 C" and I could be virtually 100% confident that it would cover the true value. Without increasing the amount of information one puts into the estimate, you can simply trade more greater precision for less confidence, or vice versa.

"Error bars" in research articles

Reports routinely use error bars in graphs of their results. But in many of these, it is not explicitly stated what the error bars are. They could be...

- $\pm\mathbf{1}S\underline{E}$: thus – if the sampling distribution is Gaussian – it is a **67**% CI.
- $\pm\mathbf{1.96}S\underline{E}'s$: thus it is a **95**% CI.
- ± **some other number of** $SE's$, in which case it is a **??**% CI.
- $\pm\mathbf{1}S\underline{D}$, or $\pm\mathbf{1.96}S\underline{D}'s$: if so, it describes the variability of the *individual values* that went into the mean – rather than the statistical precision of the mean itself, a quantity that involves $\sqrt{n}$. Since the SD is $\sqrt{n}$ times larger than the SE, error bars are unlikley to be some ± number of SD's.

<u>Advice</u>: Always look in the legend, or methods section, to find out what the error bars refer to. If they are nor explained, but you have some sense of the SD, and know the n, you can often figure it out.

CI based on ±M.E. on log scale is asymmetric when back-converted.

To ensure a specified Margin of Error, how big should n be?

In our sampling of Harvey's treatise, suppose we wished to estimate the mean fairly precisely, with a margin of error in our 95% CI of say ±0.1 letters. To achieve this, we would need to have an n such that

$$1.96 \times SE = 1.96 \times 1.88/\sqrt{n} = 0.1.$$

We can solve this for n to obtain

$$n = \left\{ 1.96 \times \frac{\sigma}{0.1} \right\}^2.$$

To determine n, we also need an estimate of how large σ is . But we are not in the position we were in our first example, where we knew – from a full electronic source, and software that could measure all of the word lengths – that σ was exactly 1.88. After the fact, once we have our sample of n from the treatise, we will look at the variation *in the sample* to get a better estimate, $\hat{\sigma}$, of σ. But for now, we need to *project* what this is likely to be. We might want to base our projected $\hat{\sigma}$ on what we know – from all our past experience and intuition – about the lengths of words. We can also use the fact that σ was 1.88 in a real text, admittedly a simpler one – from many centuries earlier. So to be on the safe side, we might want to make a conservative projection, say

$\sigma = 2.5$. If we use this, we calculate that we will need a random sample of $n = (1.96 \times 2.5/0.1)^2 = 2400$ words.

Why so large a sample size in this example? The reasons are two-fold

- We asked for quite a narrow margin of error: If we are dealing with an average word length of say 4.5 letters, then the margin of error of ±0.1 letters in absolute terms represents just $(0.1/4.5) \times 100 = 2.2\%$ margin of error in relative terms.
- The word to word variation in length is substantial: the SD is approximately 2.5 letters. With respect to the average of 4.5 letters, 2.5 represents a coefficient of (inter-individual) variation (CV) of $(2.5/4.5) \times 100 = \underline{55}\%$!

What n would it take to determine the mean height of female students at McGill to within a margin of error of ±1% in a 95% CI?

Here we are dealing with maybe a SD of approx 7cm and a mean of say 165cm, i.e., a coefficient of (inter-individual) variation (CV) of $(7/165) \times 100 \approx \underline{4.25}\%$

Thus a sample of just $n = 100$ would give a (relative) SE of $4.25\% \div \sqrt{100} = 0.425\%$, and so the Margin of Error in a 95% CI would be $\pm 1.96 \times 0.425 = \pm 0.83\%$.

The reason we need fewer than 100 in this situation is the *narrow coefficient of (inter-individual) variation (CV)* of human heights to begin with.

Q: What is your estimate of the coefficient of (inter-individual) variation (CV) of human *weights* in the 18-25 age range?

In some circumstances, we need considerable precision.[24]

5.6 Key Points & Some Pointers

- Section 5.5 is the most important one.
- The precision of a sample mean is a function of the variation from observation to observation, and of the number of observations.
- Your statistics courses probably emphasized statistical 'tests' and P-values more than CI's. That's a pity: CI's are more useful, since they provide a measure of precision around the estimate. And, you have a better chance of correctly explaining them to your in-laws than you have of explaining a P-value. And over the years, more journals are starting to insist on CI's.
- The diagram on p12 makes it easy to start to describe the basis for a CI: start by worrying that your point estimate is an over-estimate; then worry that it is an under-estimate.
- For the exam, you should know that the SE of a sample mean or proportion is larger if the individual observations are highly variable, and that it involves the square root of the sample size, not the sample size itself.
- For the exam, you will not be expected to remember the various percentiles of the various t-distributions, but it might be good to remember the ±1.96 (or even just ± 2) for the 95% limits in a Gaussian ("Normal") distribution.

[24] *What sample size would be needed to determine the unemployment rate so that the margin of error in a 95% CI is ± 0.2 percentage points?*

± 0.2 *percentage points*, when converted to a *proportion*, is ±0.002. We are dealing with '0/1' data (person is employed/unemployed) where the mean is approx. 0.06, and so the interindividual variation in these 0's and 1's is approx. $\sigma = \sqrt{0.06 \times 0.94} = 0.24$. Thus the required sample size, if we were to take a simple random sample, is

$$n = (1.96 \times 0.24/0.002)^2 \approx 53,000 \text{ persons!}$$

StatsCan narrows the margins of error, especially those for changes in the unemployment rate, by using more sophisticated surveys that follow sampled persons over several months.

6 Back to Mr. W.P. [from Section 4]

TO WHICH CATEGORY DOES HIS AVERAGE DBP BELONG?

Recall that in his pre-employment physical, Mr. W.P.'s blood pressure was 130/95 mmHg. Suppose that in 5 new measurements, each one taken on a different occasion, the diastolic pressures were 99, 98, 101, 95, and 90. Thus their average is 96.6 (SD 4.3); this gives a SE of $4.3/\sqrt{5} = 1.9$ With $n = 5$, we need to go out 2.78 SE's in each direction to have a 95% CI. Thus

$$95\% \text{ CI for } \mu_{DBP} : 91.3 \text{ to } 101.9$$

These limits would put his mean rather firmly above 90-into the mild hypertensive range. Had the measurements been 89, 102, 97, 87 and 95 (mean 94, SD 6.1), the CI would have been a more equivocal 86.4 to 101.6. In such a case, a more extensive series would be needed to narrow the interval.[25]

7 Confidence Interval for Difference of 2 Means

Clinical Problem 3. Did Diuretic Therapy Lower the Blood Pressure?

A 50-year-old asymptomatic woman, Mrs. O.M., comes for a routine physical examination and you discover a blood pressure of 150 mmHg systolic and 105 mmHg diastolic. You start her on 50 mg of hydrochlorothiazide daily (a frequently used diuretic, antihypertensive drug), and 1 month later, her blood pressure is 140 mmHg systolic and 95 mmHg diastolic. She complains that she thinks the new medicine has made her slightly weak and she wants to stop taking it. Before you urge her to continue with the hydrochlorothiazide, you should be sure that it has lowered her blood pressure. Do the blood pressure measurements noted above convince you that the medicine has in fact lowered her blood pressure, or is there a reasonable chance that the observed difference in blood pressure might have occurred without therapy?

As we have seen, blood pressures are variable and we may well need measurements from more than one occasion to get a solid basis for decisions. In the case of Mrs. O.M., we have 2 diastolic measurements from each of $n_{pre} = 4$ pretreatment visits with average values: 102, 105, 110, and 103. On $n_{post} = 3$ recent visits since beginning the treatments, her averages have been: 95, 93, and 97. We want to use these two sets of measurements to assess the improvement.

CI for the difference between 2 means

The difference between the two sample means is

$$\bar{y}_{pre} - \bar{y}_{post} = \frac{102 + 105 + 110 + 103}{4} - \frac{95 + 93 + 97}{3} - = 105 - 95 = 10$$

Since there are now two sources of imprecision, the imprecision in the difference of two independently established means is greater than each one alone. Fortunately, the SE for the difference (or sum!) of two random quantities is *not* the sum (or difference) of the two SE's. Instead, ***SE's "add in quadrature",*** just like the rule for the length of the longest side of a right-anged triangle:

$$SE_{sum} \neq SE_1 + SE_2 \quad ; \quad SE_{diff.} \neq SE_1 - SE_2 \ .$$
$$(SE_{sum})^2 = (SE_1)^2 + (SE_2)^2 \quad ; \quad (SE_{diff.})^2 = (SE_1)^2 + (SE_2)^2 \ .$$
$$SE_{sum} = \sqrt{(SE_1)^2 + (SE_2)^2} \quad ; \quad SE_{diff.} = \sqrt{(SE_1)^2 + (SE_2)^2} \ .$$

Thus, the standard error of the difference of two sample means is

$$SE \text{ of } \{\bar{y}_1 - \bar{y}_2\} = \sqrt{(\text{SE of } \bar{y}_1)^2 + (\text{SE of } \bar{y}_1)^2} \ .$$

Thus, in the case of one μ, when the sample sizes are large enough, the 95% CI for $\mu_1 - \mu_2$ is of the same "*answer* ± *multiple of SE*" form, i.e.,

$$\{\bar{y}_1 - \bar{y}_2\} \pm 1.96 \times SE \text{ of } \{\bar{y}_1 - \bar{y}_2\} \ .$$

In our example, the sample sizes are just 4 and 3, and so we need to use ± somewhat more than 1.96 times the SE. The calculations are a bit more complicated, and so one one would normally use a statistical package, or Excel, to obtain the CI. However, a rough sketch is given here for those who don't like to use 'black-box' calculations.[26] The estimate of σ from these data is $\hat{\sigma} \approx 3$mmHg,[27] so the SE's for the two means are $3/\sqrt{4}$ and $3/\sqrt{3}$ respectively. Thus,

$$\text{SE for } \{\bar{y}_1 - \bar{y}_2\} \approx \sqrt{(3/\sqrt{4})^2 + (3/\sqrt{3})^2} \ \approx \ 2.3 \ .$$

[25] It may not always be possible – even with a quite large n – to have the interval fall unequivocally into a single 10-mmHg or 20-mmHg band – for example, if μ were truly 89.5, it would take a very large sample size to have a high probability that the CI did not overlap 90.

[26] In order to estimate σ, we take the square root of a weighted average of the squared SD's in each of the two samples, insert it into the last equation above to get the SE for the difference. We also use as a multiplier of the SE. These calculations apply when it can be assumed that the SD in 'universe' 1 is the same as it is in 'universe 2.' Matters become more complicated if one doesn't feel comfortable with this assumption. Fortunately, when the two n's are large, say 30 or more in total, the multiples from the Normal Table (1.96 for 95% confidence, etc) are accurate enough.

[27] If interested, see a statistics text under the index 'pooled estimate of variance'.

This is a bit more technical that can be covered in a single lecture, but, in order to estimate σ, we had to compute the 3 independent[28] variations around $\bar{y}_{pre} = 105$, and the 2 independent variations around $\bar{y}_{post} = 95$, so our estimate is based on $3 + 2 = 5$ "degrees of freedom"[29] Since this is a very small number, we can't take 1.96 as our multiple. Instead, the table for Student's 't' distribution, with 5 degrees of freedom, tells us that the multiple needs to be 2.57. **Thus, the 95% CI for $\{\mu_{pre} - \mu_{post}\}$** is approximately

$$10 \pm 2.57 \times \text{SE for } \{\mu_{pre} - \mu_{post}\} = 10 \pm 2.57 \times 2.3 = 10 \pm 5.9 = 4.1 \text{ to } 15.9\ .$$

We notice that 0 is not in the interval. This means that we are confident that Mrs. O.M.'s mean diastolic blood pressure is lower. The change is not readily accounted for by sampling variation. Because we are reasonably confident that the hydrochlorothiazide has reduced her pressure, we might urge her to continue it. Her weakness may be unrelated to the drug, and it may disappear. [The wording and conclusions in this last paragraph are Ingelfinger's, not JH's.]

The initial evaluation of a patient is often complicated by the fact that observations vary from minute to minute and day to day. The greater the potential for variation, the more the need for performing many observations over time to establish the patient's average condition.

7.1 Key Points & Some Pointers

- Its déjà vu all over again, once you get the generic pattern. All that changes is that the SE for a difference of two independent estimates has 2 components, one for each estimate.

- Don't try to remember the exact formula for the SE of a difference, since it will usually be computed by a statistical package. But do appreciate that if you subtract one independent estimate from another, the SE of the difference will be larger than the SE of each of the two estimates.

- It is better to calculate a single CI for the difference, rather than to compute 2 CI's and worry about their overlap. Two 95%'s don't translate into the single 95% CI you need: the 2 CIs can overlap slightly even though the difference is statistically significant.

- Don't fuss about the technicalities when the 2 sample sizes are small, and one has to use the t- distribution, and the concept of degrees of freedom. In the application in small-group exercise Q2 (ii), the relevant items have already been calculated for you so you can complete the hand-calculation. These technicalities were included in footnotes 26 and 27 above because those of you who have taken a statistics will remember it (even if not why) and will ask why we don't mention it.

[28] With 1 observation, one cannot assess variation. With 2, the two deviations from the mean add to zero, so effectively we have only 1 'independent' deviation; with 3, we have 2 – since all 3 must sum to zero; etc. Each independent assessment is one 'degree of freedom.'

[29] We have 5 'independent assessments' of the variation, 3 from the 4 'pre' observations , and 2 from the 3 'post' observations.

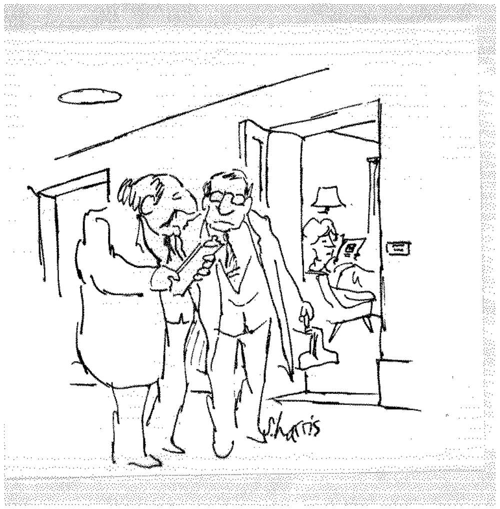
8 P-Values and Statistical 'Tests'

8.1 "P-Value"

Def$^{n.}$ A **probability concerning the observed data**, calculated under a **Null Hypothesis** assumption, i.e., assuming that the only factor operating is sampling or measurement variation.

Use To assess the evidence provided by the sample data in relation to a pre-specified claim or 'hypothesis' concerning some parameter(s) or data-generating process.

Basis As with a confidence interval, it makes use of the concept of a *distribution*.

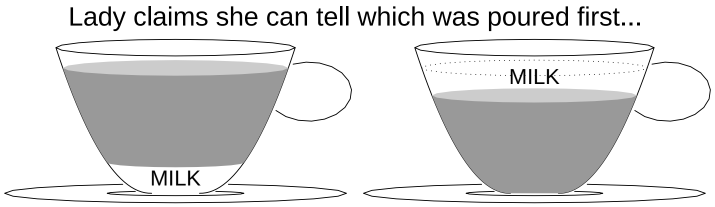


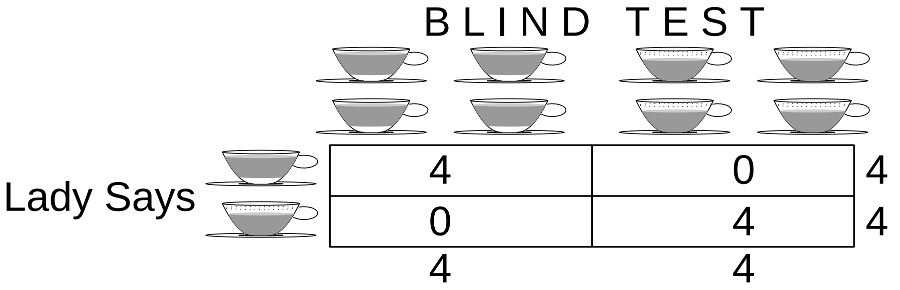
THE NULL HYPOTHESIS

"Find out who set up this experiment. It seems that half of the patients were given a placebo, and the other half were given a different placebo"

American Scientist 1982; 70:25.

Example 1 – **from *Design of Experiments*, by R.A. Fisher**





	4	0	4
	0	4	4
	4	4	

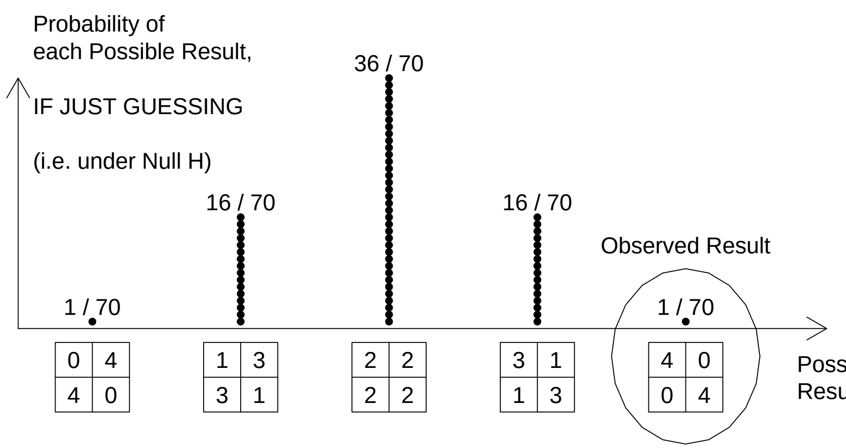
The "Null Hypothesis" (H_{null}) states that she can not tell them apart.[30] The "Alternative" Hypothesis (H_{alt}) is that she can (can you think of another?). We rank the possible test results according to the degree of evidence against the null hypothesis. **The "P-value" is the probability, calculated under the null hypothesis, of observing a result as extreme as, or more extreme than, the one that was obtained/observed.** In this case, the observed result is the most extreme, and so the P-value (cf next page) is

$$P_{value} = \text{Prob[correctly identifying all 4, IF merely guessing]} = 1/70 = 0.014.$$

The interpretation of such data is often couched in a rather simplistic way, as if these *data alone* should *decide*: i.e. if $P_{value} < 0.05$, we 'reject' H_{null}; if $P_{value} > 0.05$, we don't (or worse still, we 'accept' H_{null}). Try to avoid such simplistic 'conclusions'.

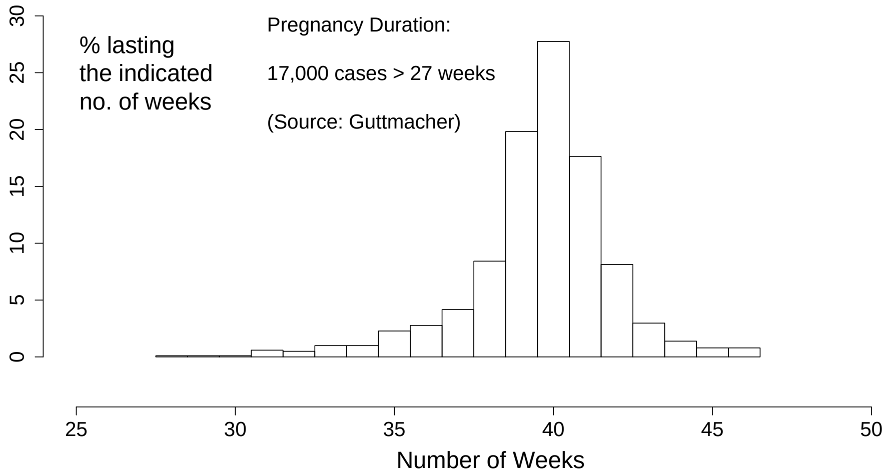
[30] Under this assumption, the blind test is equivalent to being asked to guess which 4 of the following 8 Gaelic words are the correctly spelled ones. You are told that 4 are correctly spelled and 4 are not.

1	2	3	4	5	6	7	8
madra	olscoil	cathiar	tanga	doras	cluicha	féar	bóthar



Example 2 – **[Preston-Jones vs. Preston-Jones, English House of Lords, 1949]** A divorce case in which the sole evidence of adultery was that a baby was born almost 50 weeks after the husband had gone abroad on military service. The appeal failed. To quote the court...

> The appeal judges agreed that the limit of credibility had to be drawn somewhere, but on medical evidence 349 (days) while improbable, was scientifically possible.



The P-value is calculated under the "Null" assumption that the husband was the father, and is thus the 'tail area' or probability correspond to an observation of '50 or more weeks' in the above distribution.

Effectively, one is asking: **What percent of the reference distribution does the observed value exceed?** The same system is used to report how extreme a lab value is – we are told where this value is located in the distribution of values from a healthy (reference) population.

In the reporting of statistical tests, it is common to define 'extreme' as *either* 'hyper' or 'hypo' and this to consider the 'alternative hypothesis' as 2-sided.

8.2 P-Value via the Normal (Gaussian) distribution.

The first example used a specialized mathematical distribution as the 'reference' (null) distribution while the second used an empirical population-based one. When judging the extremeness of a sample mean or proportion (or a difference between 2 sample means or proportions) calculated from an amount of information that is sufficient for the Central Limit Theorem to apply, one can use the Gaussian distribution to readily obtain the P-value. One simply calculate how many standard errors of the statistic, $SE_{statistic}$, the statistic is from where the null hypothesis states the true value should be. This "number of SE's" is in this situation referred to as a 'Z_{value}.'

$$Z_{value} = \frac{statistic - \text{its expected value under } H_{null}}{SE_{statistic}}.$$

The P-value can then be obtained by determining what percent of the values in a Normal distribution are as extreme or more extreme than this Z_{value}.

If the sample size is small enough that the value of the $SE_{statistic}$, is itself subject to some uncertainty, one would instead refer the "number of SE's" to a more appropriate reference distribution, such as Student's t- distribution.

8.3 The fallacy of the Transposed Conditional: the Prosecutor's Fallacy: What the P-value is NOT

The P-value is often mistaken for something very different.[31]

The P-value is a **probability concerning data**, *conditional on – i.e. given* – the Null Hypothesis being true.

Naive (and not so naive) end-users sometimes interpret the P-value as the probability that the Null Hypothesis is true, *conditional on – i.e. given – the data.*

Only very naive physicians mix up the complement of specificity (i.e. the probability of a 'positive' test result when in fact the patient does not have the disease in question) with the positive predictive value (i.e. the probability that a patient who has had a 'positive' test result does have the disease in question).

Statistical tests are often coded as 'positive' or 'negative' (or 'statistically significant' or not) according to whether the results are extreme or not with respect to a reference (null) distribution. Medical tests are also often coded as 'positive' or 'negative' according to whether the results are extreme or not with respect to a reference (healthy) distribution. But a test result is just one piece of data, and needs to be considered *along with all the rest of the evidence* before coming to a 'conclusion.' Likewise with statistical tests: the P-value as just one more piece of *evidence*, hardly enough to 'conclude' anything. The probability that the DNA from the blood of a randomly selected (innocent) person would match that from the blood on the crime-scene glove was $P=10^{-17}$. *Do not equate this* Prob[data | innocent] *with its transpose*: writing "data" as shorthand for "this or more extreme data", we need to be aware that

$$P_{value} = Prob[\ data \mid H_0] \neq Prob[\ H_0 \mid data].$$

The article " Are All Significant P Values Created Equal? The Analogy Between Diagnostic Tests and Clinical Research" by WS Browner & et al. in JAMA 1987;257:2459-2463 exploits the analogies between medical and statistical tests, and warns us not to transpose these two fundamentally different concepts.

The prosecutor's fallacy: Who's the DNA fingerprinting pointing at?

New Scientist, 29 Jan. 1994, 51-52. David Pringle

Pringle describes the successful appeal of a rape case where the primary evidence was DNA fingerprinting. In this case the statistician Peter Donnelly opened a new area of debate. He remarked that

> **forensic evidence answers the question "What is the probability that the defendant's DNA profile matches that of the crime sample, assuming that the defendant is innocent?"**
>
> **while the jury must try to answer the question "What is the probability that the defendant is innocent, assuming that the DNA profiles of the defendant and the crime sample match?"**

Donnelly suggested to the Lord Chief Justice and his fellow judges that they imagine themselves playing a game of poker with the Archbishop of Canterbury. If the Archbishop were to deal himself a royal flush on the first hand, one might suspect him of cheating. Assuming that he is an honest card player (and shuffled eleven times) the chance of this happening is about 1 in 70,000.

But if the judges were asked whether the Archbishop were honest, given that he had just dealt a royal flush, they would be likely to place the chance a bit higher than 1 in 70,000. The error in mixing up these two probabilities is called the "the prosecutor's fallacy," and it is suggested that newspapers regularly make this error.

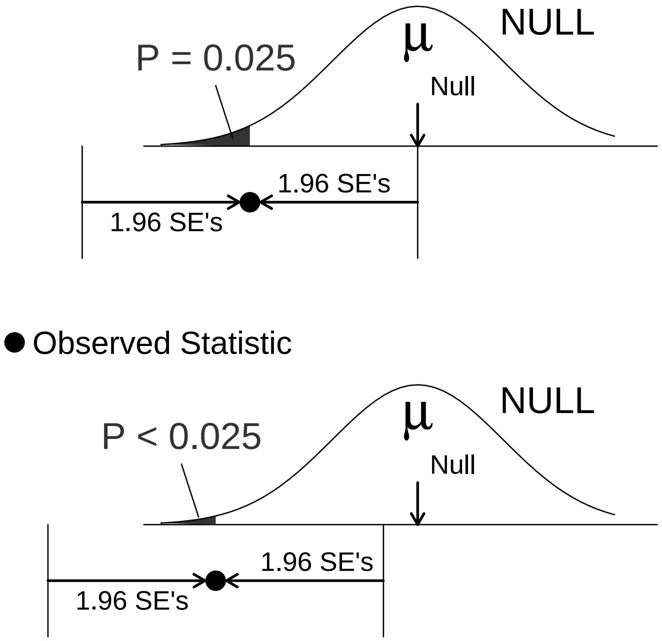
Apparently, Donnelly's testimony convinced the three judges that the case before them involved an example of this and they ordered a retrial.

from Vol 3.02 of Chance News.

[31] The larger text on this page is meant to convey the importance of this warning.

8.4 Relationship between P-value and CI

If you read the description of how a confidence Interval (CI) is formally derived, you will see that there is an intimate connection between P-values and CI's.



If, as in the upper e.g. in the graph, the upper limit of the 95% CI *just touches* the null value, then the 2- (1-) sided) P-value is 0.05 (0.025). If, as in the lower e.g., the upper limit *excludes* the null value, then the 2- (1-) sided) P-value is less than 0.05 (0.025). If (e.g not shown) the CI *includes* the null value, then the 2-sided P-value is greater than 0.05, and thus the observed statistic is "not statistically significantly different" from the hypothesized null value.

8.5 Key Points & Some Pointers

- P-values & 'significance tests' are widely misunderstood & misused. Very large or very small n's can influence what is 'statistically significant'. Use CI's instead. Pre study power calculations (the chance that results will be 'significant', as a function of the true underlying difference) can help, but *post-study*, i.e., *after the data have spoken*, a CI is much more relevant.

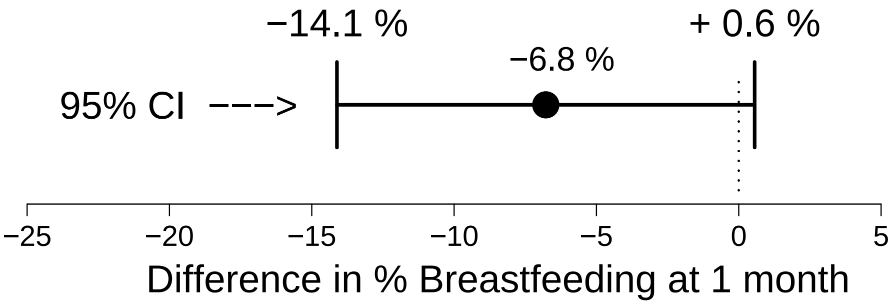
9 Statistical Inference: beyond the individual

"Statistical Inference" techniques (CI's, P-values, ...) are the same whether the focus is on the individual patient, as in the earlier e.g.'s, or on a larger universe as in the e.g.'s below. The only differences are what the parameters $(\mu, \pi, \ldots)$ stand for, and the fact that the main source of variability will probably be *inter*-individual variation. Because this variation can be considerable, n's tend to be larger, unless – as in the starch blocker e.g., – we can reduce it by careful lab-work and by matching on the large extraneous and unwanted sources of variation. In addition, if – as in the breast-feeding e.g., – the 'outcome' is measured on a (yes/no, all-or-none) scale, then the coefficient of inter-individual variation is larger than if a more refined quantitative scale is used.

Example 1 Do infant formula samples shorten the duration of breastfeeding? – [Bergevin Y, Dougherty C, Kramer MS. Lancet. 1983 May 21;1(8334):1148-51.]

Randomized Clinical Trial (RCT) which withheld free formula samples [given by baby-food companies to breast-feeding mothers leaving Montreal General Hospital with their newborn infants] from a random half of those studied.

	Mothers			
At 1 month	given sample	not given sample	Total	**Conclusion...**
Still Breast feeding	175 (*77%*)	182 (*84%*)	357 (80.4%)	P=0.07. So, ... the difference is "Not Statistically Significant" at 0.05 level
Not Breast feeding	52	35	87	
Total	227	217	444	



NO MATTER WHETHER THE P-VALUE IS "STATISTICALLY SIGNIFICANT" OR NOT, ALWAYS LOOK AT THE LOCATION AND WIDTH OF

THE CONFIDENCE INTERVAL. IT GIVES YOU A BETTER AND MORE COMPLETE INDICATION OF THE MAGNITUDE OF THE EFFECT AND OF THE PRECISION WITH WHICH IT WAS MEASURED.

THIS IS AN EXAMPLE OF AN **INCONCLUSIVE NEGATIVE** STUDY, SINCE IT HAS **INSUFFICIENT PRECISION** ("RESOLVING POWER") **TO DISTINGUISH** BETWEEN TWO IMPORTANT POSSIBILITIES – **NO HARM**, AND WHAT AUTHOROTIES WOULD CONSIDER A **SUBSTANTIAL HARM: A REDUCTION OF 10 PERCENTAGE POINTS** IN BREASTFEEDING RATES .

"**STATISTICALLY** SIGNIFICANT" AND "**CLINICALLY-**" (OR "**PUBLIC HEALTH-**") SIGNIFICANT ARE DIFFERENT CONCEPTS.

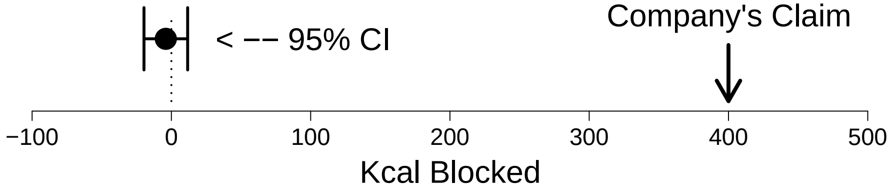
Example 2 Starch blockers – their effect on calorie absorption from a high-starch meal. Bo-Linn GW. et al New Eng J Med. 307(23):1413-6, 1982 Dec 2

Abstract: It has been known for more than 25 years that certain plant foods, such as kidney beans and wheat, contain a substance that inhibits the activity of salivary and pancreatic amylase. More recently, this antiamylase has been purified and marketed for use in weight control under the generic name "starch blockers." Although this approach to weight control is highly popular, it has never been shown whether starch-blocker tablets actually reduce the absorption of calories from starch.

Using a one-day calorie-balance technique and a high starch (100 g) meal (spaghetti, tomato sauce, and bread), we measured the excretion of fecal calories after $n = 5$ normal subjects in a cross-over trial had taken either placebo or starch-blocker tablets. If the starch-blocker tablets had prevented the digestion of starch, fecal calorie excretion should have increased by 400 kcal.

However, fecal calorie excretion was the same on the two test days (mean ± S.E.M., 80 ± 4 as compared with 78 ± 2).

We conclude that starch blocker tablets do not inhibit the digestion and absorption of starch calories in human beings.



EFFECT IS MINISCULE (AND ESTIMATE QUITE PRECISE)
AND VERY FAR FROM COMPANY'S CLAIM !!!

10 OVERALL SUMMARY

- The difference sources of variation have important implications in patient management.
- Descriptive statistics should be descriptive, and should suit the pattern of variation.
- Confidence intervals are preferable to P-values, since they are expressed in terms of the (comparative) parameter of interest; they allow us to judge the magnitude and its precision, and help us in 'ruling in / out' certain parameter values.
- A 'statistically significant' difference does not necessarily imply a clinically important difference.
- A 'not-statistically-significant' difference does not necessarily imply that we have ruled out a clinically important difference.
- Precise estimates allow us to distinguish between that which – if it were true – would be important and that which – if it were true – would not. Sample size is an important determinant of precision.
- A lab value that is in the upper 1% of the reference distribution (of values derived from people without any known diseases/conditions) does not mean that there is a 1% chance that the person in whom it was measured does not have some disease/condition; i.e., it doesn't mean than the a 99% chance that the person in whom it was measured does have some disease/condition.
- Likewise, a P-value is NOT the probability that the null hypothesis is true.
- The fact that

  $$Prob[\textit{the data} \mid \textit{Healthy}] \text{ is small}$$

  does not necessarily mean that

  $$Prob[\textit{Healthy} \mid \textit{the data}] \text{ is small}$$
- Ultimately, P-values, CIs and other evidence from a study need to be combined with other information bearing on the parameter or process.
- We should not treat any one study as the last word on the topic.
- We need to worry about distortions of a non-sampling kind that are not minimized by having a large 'n.' A larger sample size will not reduce systematic differences in a comparison.

11 Small-group Exercises

Q1: Serum Thyroxine (T_4) levels

In evaluating their method for determining serum thyroxine (T_4) Murphy and colleagues (1966) measured T_4 in more than 1000 patients. Table A shows the frequency distribution of serum T_4 in three series of patients, each series containing some patients judged to be hypothyroid, some euthyroid, and some hyperthyroid on clinical grounds.

Table A. Frequency distribution of T_4 in three patient series by type of thyroid disease (male and female patients combined)

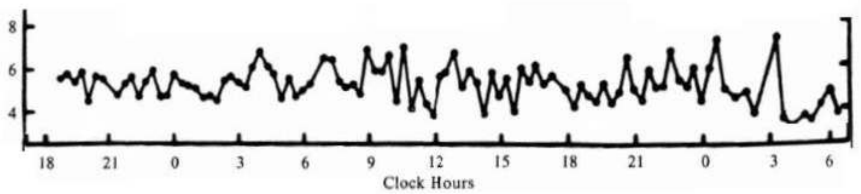
T_4	Series 1			Series 2			Series 3		
↓	Hypo	Eu	Hyper	Hypo	Eu	Hyper	Hypo	Eu	Hyper
> 16			2			6			1
15.1-16			3			2			0
14.1-15			0		1	1			1
13.1-14			1		1	2			2
12.1-13			2		0	5			1
11.1-12			2		3	2			0
10.1-11			1		6			3	0
9.1-10		10			29			9	1
8.1- 9		21			75			15	
7.1- 8		20			153			6	
6.1- 7		42			200			37	
5.1- 6		37			172			36	
4.1- 5		31			121			34	
3.1- 4	3	12		10	21			8	
2.1- 3	1	1		14	4		3		
1.1- 2	3	1		13	1		7		
0 - 1				11			2		
Mean T_4	2.6	6.3	13.8	2.1	6.5	14.4	1.7	6.5	13.4

T_4 = thyroxine; hypo = hypothyroid; eu = euthyroid; hyper = hyperthyroid.
From Murphy et al., *J Clin Endocrinol*, 26:247-256, 1966.

i. Construct probability histograms for hypo-, eu-, and hyperthyroid patients using the data from Series 2.

ii. What would you choose as the normal range? Why?

iii. Within the normal range there is considerable variation, both intraperson and interperson. The Figure below shows the variation in plasma T_4 in an individual subject observed on repeated occasions. How much of the normal range is explained by the intrasubject variability as seen in this patient? Answer by first visually estimating the standard deviation of (i) the values for the one individual subject and (ii) the values in Series 2, and then comparing them. (Note that the data from this figure represent a sampling over a 36-h period. Seasonal variation in serum T_4 level has also been reported.).



Baseline values of thyroxine (T_4) in male subject 205 during 36 h. [from Azukizawa et al. (1970) J Clin Endocrinol Metab, 43, 533-542.]

iv. You have a patient who is the sister of a woman with Graves disease and hyperthyroidism. You suspect that your patient may be more likely than usual to develop a similar illness. Her serum T_4 determined 1 year ago was 5 μg per 100 mL. On a recent office visit it was 10μg per l00mL. Does this change indicate that this patient is developing hyperthyroidism? Why or why not?

Q2: BP

The goal of treating patients with hypertension is to prevent morbidity and mortality associated with high blood pressure. The Canadian Hypertension Education Project (CHEP) 2008 Guidelines recommend that in general, blood pressure should be lowered to less than 140/90 mmHg and in those with diabetes or chronic kidney disease, to less than 130/80 mmHg. (See http://www.hypertension.ca/chep/resource-centre/publications/ if you wish to have more details on guidelines for the diagnosis and management of hypertension).

i. Mr. W.P. is started on treatment. He has the following diastolic blood pressures at his next 4 visits: 86, 92, 82, 84.

(a) The standard deviation of these 4 diastolic measurements is 4.3 mmHg. From these, one can compute the 95 percent confidence limits for his mean diastolic blood pressure using the multiple of 3.18 from the t distribution (The larger multiple – 3.18 rather than the more commonly used 1.96 – is to compensate for the fact that the 4.3 is estimated from just 4 observations). The 95 percent confidence interval is

$$\frac{86+92+82+84}{4} \pm 3.18 \times \frac{4.3}{\sqrt{4}}.$$

Complete the calculation and interpret the confidence interval.

(b) How many BP measurements would be needed to halve the width of the 95% confidence interval? [*Ignore the fact that the multiple would be closer to 1.96, and that the calculated standard deviation would not remain at 4.3 – it could go higher or lower*]

ii. You follow Mr. WP. and his diastolic blood pressure is consistently above 90 mmHg. His pulse on 3 visits is 80, 85, and 75. You prescribe propranolol (an antihypertensive agent which also slows the pulse). On the next 5 visits, his diastolic blood pressure is unchanged, but his pulse is 70, 65, 75, 60, and 65.

(a) In order to compute the 95 percent confidence limits for the *change* in Mr. WP.'s pulse, we use a weighed average of the SD of the $n = 3$ pulse measures pre propranolol (approx. 5) and the SD of the $n = 5$ post propranolol (approx. 5.7), to get an SD of ≈ 5.5. Since this SD estimate is based on only $2 + 4 = 6$ 'degrees of freedom' i.e. 6 'independent assessments of variation', we need to use a t- multiple of 2.47 (rather than 1.96) for the 95% CI. Using p.16 of the notes as a template, we can calculate the 95% CI for the *change* in Mr. WP.'s pulse as

$$\frac{70+65+75+60+65}{5} - \frac{80+85+75}{3} \pm 2.57 \times 5.5 \times \sqrt{\frac{1}{5}+\frac{1}{3}}.$$

Complete the calculation and interpret the confidence interval.

(b) Do you think the reason his blood pressure has not responded is that he has not taken the propranolol, or that the dose prescribed was not effective? Why? What if the difference had been 5 beats/min?

(c) Intuitively, inside the square root sign, why is the $\frac{1}{5}$ *added to* the $\frac{1}{3}$, thereby making the margin of error for the difference larger than the margin of error for each component of the difference?

iii. A patient with a sphygmomanometer at home reports to you that she measured her diastolic blood pressure once a day for the last 8 days, that the pressure varied, and the average was 85 mmHg. You check her diastolic blood pressure and observe a value of 95 mmHg. Ingelfinger asked you

Compute the *probability of observing such a large difference (10), given no true difference (H_0)*. Assume a gaussian distribution with $\sigma = 6$ mmHg. What do you suspect?

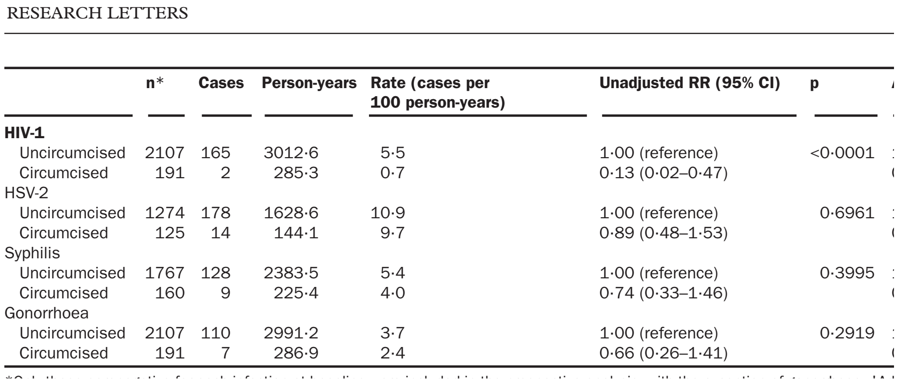
Here are the calculations, where Z stands for an observation from a Normal (Gaussian) distribution with mean 0 and SD 1:

$$Prob[\geq 10|H_0] = Prob\left[Z \geq \frac{95-85}{\sqrt{\frac{6^2}{8}+\frac{6^2}{1}}}\right] = Prob\left[Z \geq \frac{10}{6.4}\right] = Prob[Z \geq 1.6].$$

The probability, obtained by looking up what percentage of the Normal distribution is more than 1.6 SD's above the mean, is thus approximately 5%. What *do* you suspect?

Q3: Male circumcision and risk of HIV-1 and other sexually transmitted infections

The table below "Relative risk (RR) of HIV-1 and other STIs in circumcised and uncircumcised men" is from the article "Male circumcision and risk of HIV-1 and other sexually transmitted infections in India" by Reynolds SJ et al, *Lancet* 2004;363:1039-40 (*we will be using this article in more depth for the last small group session*). This was a prospective study of 2298 HIV-uninfected men attending sexually transmitted infection clinics in India. Only those seronegative for each infection at baseline were included in the prospective analysis, with the exception of gonorrhoea.



RESEARCH LETTERS

	n*	Cases	Person-years	Rate (cases per 100 person-years)	Unadjusted RR (95% CI)	p
HIV-1						
Uncircumcised	2107	165	3012·6	5·5	1·00 (reference)	<0·0001
Circumcised	191	2	285·3	0·7	0·13 (0·02–0·47)	
HSV-2						
Uncircumcised	1274	178	1628·6	10·9	1·00 (reference)	0·6961
Circumcised	125	14	144·1	9·7	0·89 (0·48–1·53)	
Syphilis						
Uncircumcised	1767	128	2383·5	5·4	1·00 (reference)	0·3995
Circumcised	160	9	225·4	4·0	0·74 (0·33–1·46)	
Gonorrhoea						
Uncircumcised	2107	110	2991·2	3·7	1·00 (reference)	0·2919
Circumcised	191	7	286·9	2·4	0·66 (0·26–1·41)	

i. How do you interpret the 95% CI for the RR for **HIV-1** in plain language?

ii. How much confidence do you have in this RR estimate of 0.13?

iii. Interpret the CI for the RR for Syphilis.

iv. What other factors besides male circumcision could explain the estimates?

v. What do the p-values (in the '**p**' column) add?