STATISTICAL CONCEPTS FOR CLINICIANS

James A. Hanley

Department of Epidemiology, Biostatistics and Occupational Health
McGill University

Lecture 24, Unit 8b - Epidemiology
Basis of Medicine
McGill University, November 6, 2008

OUTLINE

An introduction to some statistical concepts that are relevant in the interpretation of

- measurements (observations) made on an individual patient
- the statistical material presented in research reports.

LEARNING OBJECTIVES: TO...

- Appreciate and describe patterns of intra- and inter-individual variability in measurements; and understand the reasons for, and consequences, of this variability;
- Appreciate the '(im)precision of a mean level (or a proportion) measured on an individual, or group of individuals; understand and apply the concept of a Margin of Error used in "Confidence Intervals."
- Apply Confidence Intervals.
- Understand the concepts of, and proper interpret:- (statistical) "P-value"; test of hypothesis; "Statistically significant"; "statistical power."
- Apply these concepts to published research based on data from aggregates of individuals.

STATISTICS AND THE INDIVIDUAL PATIENT

- If clinical course of some illness were always the same in absence of treatment and if treatment always had same effect, would be easy to determine whether a new treatment was an improvement.
- The following example is used to **illustrate**, and show the **consequences** of, the **kinds of variability** that may affect clinical observations.

  Frequency distributions are useful in study of clinical observations that vary from patient to patient / time to time.

Background: Angina pectoris

- Substernal chest pain typically brought on by exercise and relieved by rest.
- Common symptom of coronary vascular disease.
- Causes substantial morbidity by limiting patient's activity.
- Nitroglycerin (NTG), administered sublingually, used to treat it.
- Impossible to prescribe NTG frequently enough for day-long prevention of angina.
- How about "long-acting" nitrate ?

Does Long-Acting Nitrate Therapy Help Mr Lewis?

- 55-year-old man with angina.
- Attacks typically occur after has climbed 1/2 flight of stairs or walked 1/4 mile. About 6 attacks each week.
- His MD recently prescribed isosorbide dinitrate (ISDN).
- Mr Lewis called his MD later that day:-
  - Usual angina halfway up his 14 stairs 1 h after taking ISDN.
  - Experienced headache & palpitations (known side effects).
- Should he stop the ISDN?

Gathering & Interpreting Evidence from Patient

- How quickly do angina attacks occur $\bar{s}$ Tx ?
- Records of his angina (50 entries for 2 mo. before Tx)
- Summarize information in **frequency distribution**
- No. steps before angina before & after started taking ISDN

No. Steps	0	1	2	3	4	5	6	7	7	9	10	11	12	13	14	total
Before	0	0	2	3	5	4	9	6	4	4	3	1	1	0	8	50
After	1	0	0	1	0	2	2	3	3	2	2	1	0	1	2	20

Alternative Display: Histogram

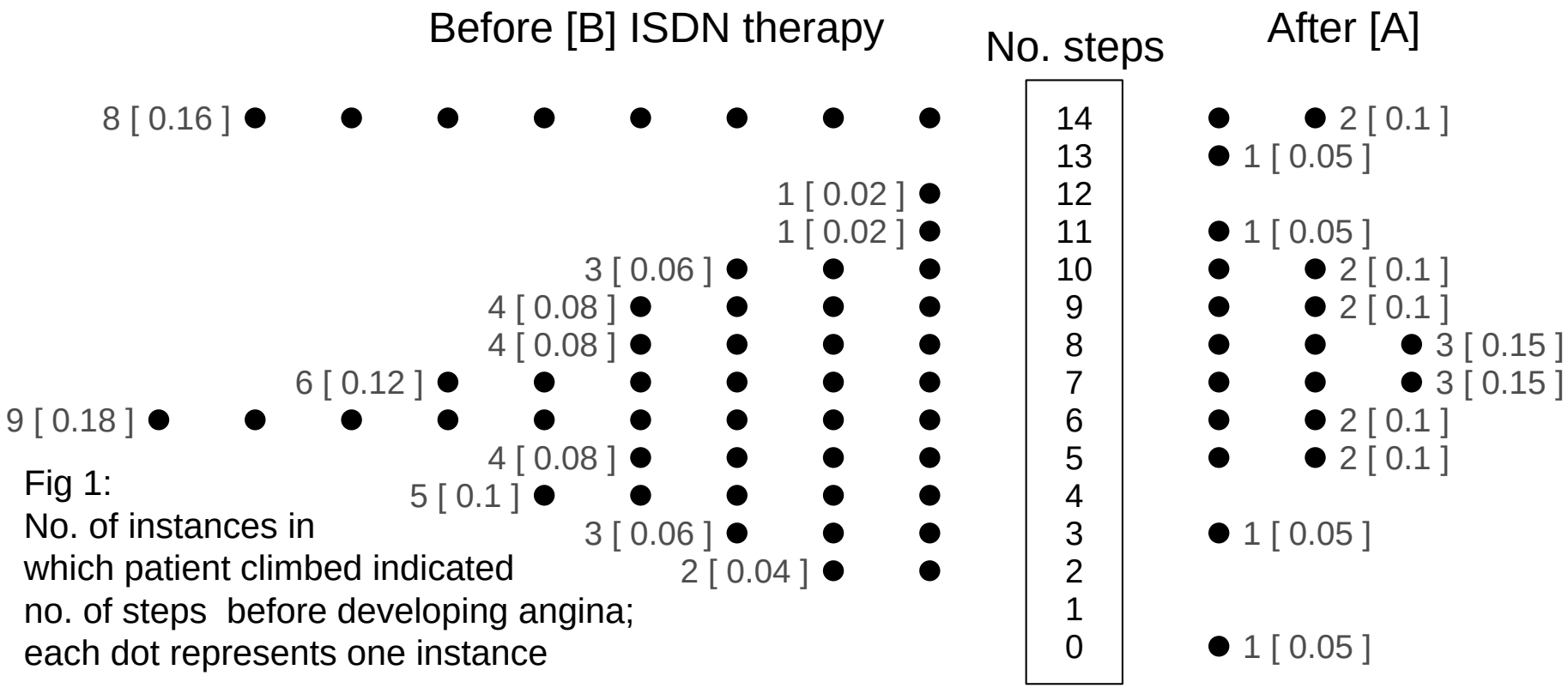


Fig 1:
No. of instances in which patient climbed indicated no. of steps before developing angina; each dot represents one instance

- **Observed frequency or count** of the no. of times Mr. Lewis climbed the given number of steps without angina.
- **[Relative frequency]** : proportion of trials out of 50 (20).
- Proportion: can compare 2 datasets with different n's

Interpreting One / Several Observation(s)

- Mr. Lewis' new single obs^n of 6 steps without angina does not prove/disprove that drug has some beneficial effect.
- Can learn more if have $>$ 1 obs^n. Only very striking effects of treatment can be demonstrated with 1 obs^n.
- 1 obs^n insufficient to show benefit even if ISDN were completely effective: $\approx$ 16% of time he climbs to top of stairs without angina $\bar{s}$ medication.
- The smaller the effect – gain or loss – of some Tx, the more obs^{ns} will be needed to demonstrate that effect.
- In light of side effects, require "fairly large" benefit of ISDN. "Fairly large" hard to define precisely, but MD believes that a benefit large enough to balance side effects should be apparent after 20-25 observations. Instructs Mr. L to **continue ISDN for 3 weeks** & to record the point at which angina occurs each time he climbs the stairs.

Comparing Outcomes:

3 weeks after starting to use ISDN ...

No. Steps	0	1	2	3	4	5	6	7	7	9	10	11	12	13	14	total
Before	0	0	2	3	5	4	9	6	4	4	3	1	1	0	8	50
After	1	0	0	1	0	2	2	3	3	2	2	1	0	1	2	20

Comparing Outcomes:

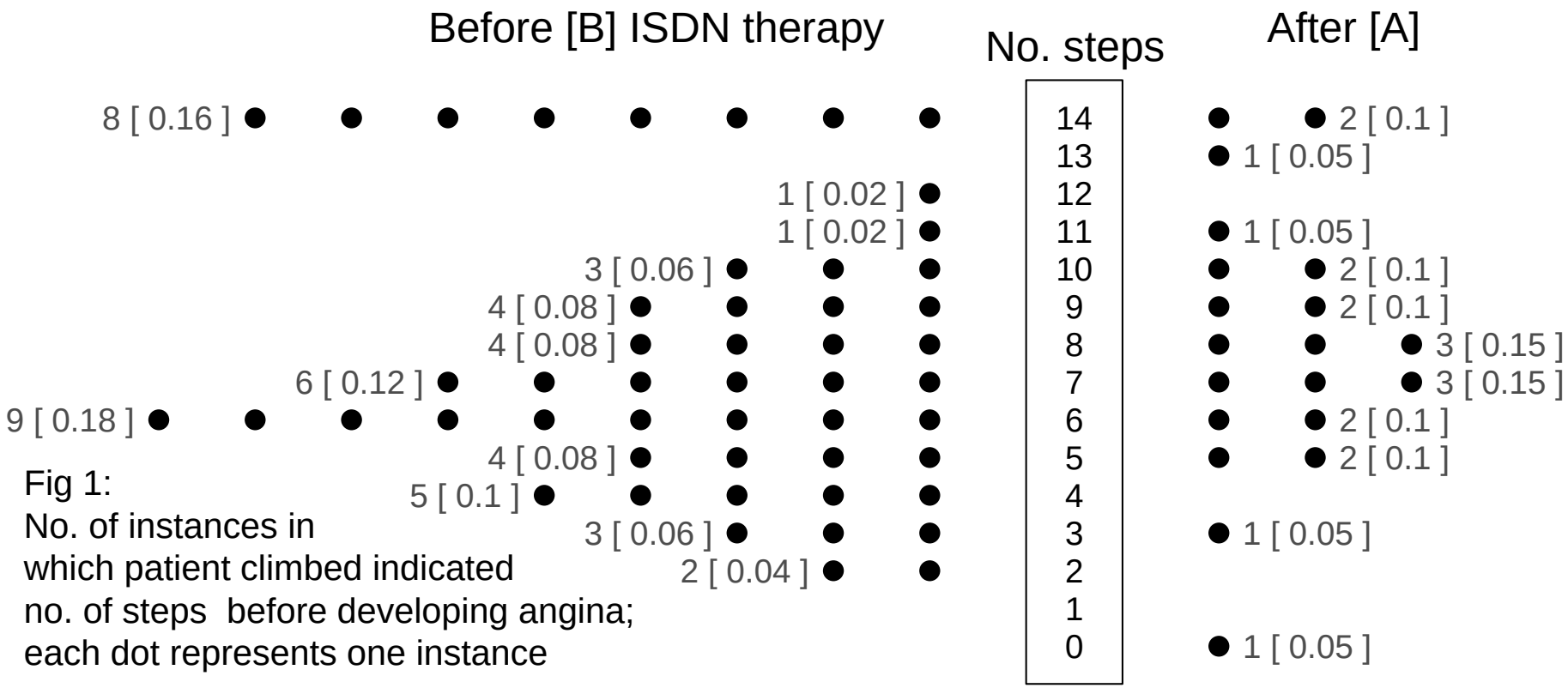


Fig 1:
No. of instances in which patient climbed indicated no. of steps before developing angina; each dot represents one instance

- Shapes of histograms do not appear to differ a great deal.
- The observed proportion of climbs without angina has gone down from 16 percent to 10 percent, a loss.
- **median** number of steps climbed before angina is now 8, whereas before it was 7, a slight gain.

Summary statistic: Median

Definition

Middle value of a set of numbers when ordered by size

– If number of values is odd → middle number

– If number of values is even → average of 2 middle numbers.

Examples:

Numbers	Median
4, 5, 5, 7, 8	5
4, 5, 5, 7, 8, 8	6

Advantages (over mean):

- *More resistant* to influence of extreme obs^{ns}
- Better indicator of "middle" if distribution not symmetric.

Will Long-Acting Nitrate Therapy Help Mr Lewis?

(**Inference is to the future**; "best guide to future is the past")

- Even without formal statistical analysis, it seems that Mr. Lewis has had no marked benefit from ISDN
- Continued presence of side effects $\longrightarrow$ discontinue Tx?
- How do **other** patients respond to ISDN?
  - **proportion** of patients similar to Mr. Lewis who respond
  - **degree** of improvement for those who do respond
    - If some patients almost completely unresponsive, while responders tend to derive large benefits, 2-week trial may be enough to conclude drug should be stopped.
    - If almost all patients derive some benefit, but average improvement is small, 20 observations may not be enough to conclude that continued treatment is unwise.

Key Points so far

- Natural intra-patient variability in (untreated) course of many diseases / conditions / risk indicators → need several observations of patient to assess Tx effect.
- Frequency distributions help us to
  - appreciate pattern of variability
  - assess effects of any change in management.
- Frequency distributions described using tables, graphs and summary statistics.

Key Points so far

- Natural intra-patient variability in (untreated) course of many diseases / conditions / risk indicators $\rightarrow$ need several observations of patient to assess Tx effect.
- Frequency distributions help us to
  - appreciate pattern of variability
  - assess effects of any change in management.
- Frequency distributions described using tables, graphs and summary statistics.

Key Points so far

- Natural intra-patient variability in (untreated) course of many diseases / conditions / risk indicators $\rightarrow$ need several observations of patient to assess Tx effect.
- Frequency distributions help us to
  - appreciate pattern of variability
  - assess effects of any change in management.
- Frequency distributions described using tables, graphs and summary statistics.

Key Points so far

- Natural intra-patient variability in (untreated) course of many diseases / conditions / risk indicators $\rightarrow$ need several observations of patient to assess Tx effect.
- Frequency distributions help us to
  - appreciate pattern of variability
  - assess effects of any change in management.
- Frequency distributions described using tables, graphs and summary statistics.

Biologic, Temporal, and Measurement Variation

cf. §2 for update on Mr. Lewis's angina, and discussion of ...

- **intersubject variation** (secondary to biologic, temporal, or measurement differences between the subjects).
- **intrasubject variation** (also due to biologic, temporal, and measurement variation within a subject).

To distinguish contribution of each source to overall variation, a series of separate observations on separate persons not sufficient. One has to study same individuals more than once.

Wide variation in..

- Heights of patients $\longrightarrow$ mostly **inter**subject var^n.
- Body temp. of outpatients $\longrightarrow$ mostly **intra**subject var^n.
- Serum T_4 levels $\longrightarrow$ Q1, small group assignment.

Biologic, Temporal, and Measurement Variation

cf. §2 for update on Mr. Lewis's angina, and discussion of ...

- **intersubject variation** (secondary to biologic, temporal, or measurement differences between the subjects).
- **intrasubject variation** (also due to biologic, temporal, and measurement variation within a subject).

To distinguish contribution of each source to overall variation, a series of separate observations on separate persons not sufficient. One has to study same individuals more than once.

Wide variation in..

- Heights of patients $\longrightarrow$ mostly **inter**subject var^n.
- Body temp. of outpatients $\longrightarrow$ mostly **intra**subject var^n.
- Serum T_4 levels $\longrightarrow$ Q1, small group assignment.

Variability: Implications for Patient Care

1. Large temporal / measurement var^{n} $\rightarrow$ Tx efficacy / biologic changes difficult to detect even with large n of well-controlled obs^{ns}.
2. "Normal range," as determined by observing many individuals, usually $>$ range in 1 individual observed often, unless little interperson variation.
3. Often (arbitrarily) use central 95% of sample of values obtained from normal subjects, as **normal range** of measurement. Includes both inter- & intra-person variability.
4. Some pts. seek MD attention when conditions seem to worsen. If worsening simply represents temporal and not biologic variation in illness, illness likely to improve irrespective of therapy.
   ... "Most things, in fact, are better by morning." - Lewis Thomas
   ... "If see MD, cold will be better in a week. If don't, better in 7 days."
5. Technical name: "regression effect" ; "regression toward the mean"
6. MD who observes a pt. many times, or orders many lab studies, may observe "abnormalities" that do not reflect a biologic variation but are due to temporal/measurement variation.
7. These, too, are likely to be "better" or changed soon. When faced with a test result that does not seem to fit, it helps to repeat the test.

Variability: Implications for Patient Care

1. Large temporal / measurement var^n $\rightarrow$ Tx efficacy / biologic changes difficult to detect even with large n of well-controlled obs^{ns}.
2. "Normal range," as determined by observing many individuals, usually $>$ range in 1 individual observed often, unless little interperson variation.
3. Often (arbitrarily) use central 95% of sample of values obtained from normal subjects, as **normal range** of measurement. Includes both inter- & intra-person variability.
4. Some pts. seek MD attention when conditions seem to worsen. If worsening simply represents temporal and not biologic variation in illness, illness likely to improve irrespective of therapy.
   ... "Most things, in fact, are better by morning." - Lewis Thomas
   ... "If see MD, cold will be better in a week. If don't, better in 7 days."
5. Technical name: "regression effect" ; "regression toward the mean"
6. MD who observes a pt. many times, or orders many lab studies, may observe "abnormalities" that do not reflect a biologic variation but are due to temporal/measurement variation.
7. These, too, are likely to be "better" or changed soon. When faced with a test result that does not seem to fit, it helps to repeat the test.

Variability: Implications for Patient Care

1. Large temporal / measurement var^n $\rightarrow$ Tx efficacy / biologic changes difficult to detect even with large n of well-controlled obs^{ns}.
2. "Normal range," as determined by observing many individuals, usually $>$ range in 1 individual observed often, unless little interperson variation.
3. Often (arbitrarily) use central 95% of sample of values obtained from normal subjects, as **normal range** of measurement. Includes both inter- & intra-person variability.
4. Some pts. seek MD attention when conditions seem to worsen. If worsening simply represents temporal and not biologic variation in illness, illness likely to improve irrespective of therapy.
   ... "Most things, in fact, are better by morning." - Lewis Thomas
   ... "If see MD, cold will be better in a week. If don't, better in 7 days."
5. Technical name: "regression effect" ; "regression toward the mean"
6. MD who observes a pt. many times, or orders many lab studies, may observe "abnormalities" that do not reflect a biologic variation but are due to temporal/measurement variation.
7. These, too, are likely to be "better" or changed soon. When faced with a test result that does not seem to fit, it helps to repeat the test.

Variability: Implications for Patient Care

1. Large temporal / measurement var^n $\rightarrow$ Tx efficacy / biologic changes difficult to detect even with large n of well-controlled obs^{ns}.
2. "Normal range," as determined by observing many individuals, usually $>$ range in 1 individual observed often, unless little interperson variation.
3. Often (arbitrarily) use central 95% of sample of values obtained from normal subjects, as **normal range** of measurement. Includes both inter- & intra-person variability.
4. Some pts. seek MD attention when conditions seem to worsen. If worsening simply represents temporal and not biologic variation in illness, illness likely to improve irrespective of therapy.
   ... "Most things, in fact, are better by morning." - Lewis Thomas
   ... "If see MD, cold will be better in a week. If don't, better in 7 days."
5. Technical name: "regression effect" ; "regression toward the mean"
6. MD who observes a pt. many times, or orders many lab studies, may observe "abnormalities" that do not reflect a biologic variation but are due to temporal/measurement variation.
7. These, too, are likely to be "better" or changed soon. When faced with a test result that does not seem to fit, it helps to repeat the test.

Variability: Implications for Patient Care

1. Large temporal / measurement var^n $\rightarrow$ Tx efficacy / biologic changes difficult to detect even with large n of well-controlled obs^{ns}.
2. "Normal range," as determined by observing many individuals, usually $>$ range in 1 individual observed often, unless little interperson variation.
3. Often (arbitrarily) use central 95% of sample of values obtained from normal subjects, as **normal range** of measurement. Includes both inter- & intra-person variability.
4. Some pts. seek MD attention when conditions seem to worsen. If worsening simply represents temporal and not biologic variation in illness, illness likely to improve irrespective of therapy.
   ... "Most things, in fact, are better by morning." - Lewis Thomas
   ... "If see MD, cold will be better in a week. If don't, better in 7 days."
5. Technical name: "regression effect" ; "regression toward the mean"
6. MD who observes a pt. many times, or orders many lab studies, may observe "abnormalities" that do not reflect a biologic variation but are due to temporal/measurement variation.
7. These, too, are likely to be "better" or changed soon. When faced with a test result that does not seem to fit, it helps to repeat the test.

Variability: Implications for Patient Care

1. Large temporal / measurement var^n $\rightarrow$ Tx efficacy / biologic changes difficult to detect even with large n of well-controlled obs^{ns}.
2. "Normal range," as determined by observing many individuals, usually $>$ range in 1 individual observed often, unless little interperson variation.
3. Often (arbitrarily) use central 95% of sample of values obtained from normal subjects, as **normal range** of measurement. Includes both inter- & intra-person variability.
4. Some pts. seek MD attention when conditions seem to worsen. If worsening simply represents temporal and not biologic variation in illness, illness likely to improve irrespective of therapy.
   ... "Most things, in fact, are better by morning." - Lewis Thomas
   ... "If see MD, cold will be better in a week. If don't, better in 7 days."
5. Technical name: "regression effect" ; "regression toward the mean"
6. MD who observes a pt. many times, or orders many lab studies, may observe "abnormalities" that do not reflect a biologic variation but are due to temporal/measurement variation.
7. These, too, are likely to be "better" or changed soon. When faced with a test result that does not seem to fit, it helps to repeat the test.

Variability: Implications for Patient Care

1. Large temporal / measurement var^n $\rightarrow$ Tx efficacy / biologic changes difficult to detect even with large n of well-controlled obs^{ns}.
2. "Normal range," as determined by observing many individuals, usually $>$ range in 1 individual observed often, unless little interperson variation.
3. Often (arbitrarily) use central 95% of sample of values obtained from normal subjects, as **normal range** of measurement. Includes both inter- & intra-person variability.
4. Some pts. seek MD attention when conditions seem to worsen. If worsening simply represents temporal and not biologic variation in illness, illness likely to improve irrespective of therapy.
   ... "Most things, in fact, are better by morning." - Lewis Thomas
   ... "If see MD, cold will be better in a week. If don't, better in 7 days."
5. Technical name: "regression effect" ; "regression toward the mean"
6. MD who observes a pt. many times, or orders many lab studies, may observe "abnormalities" that do not reflect a biologic variation but are due to temporal/measurement variation.
7. These, too, are likely to be "better" or changed soon. When faced with a test result that does not seem to fit, it helps to repeat the test.

Variability: Implications for Patient Care

1. Large temporal / measurement var^n $\rightarrow$ Tx efficacy / biologic changes difficult to detect even with large n of well-controlled obs^{ns}.
2. "Normal range," as determined by observing many individuals, usually $>$ range in 1 individual observed often, unless little interperson variation.
3. Often (arbitrarily) use central 95% of sample of values obtained from normal subjects, as **normal range** of measurement. Includes both inter- & intra-person variability.
4. Some pts. seek MD attention when conditions seem to worsen. If worsening simply represents temporal and not biologic variation in illness, illness likely to improve irrespective of therapy.
   ... "Most things, in fact, are better by morning." - Lewis Thomas
   ... "If see MD, cold will be better in a week. If don't, better in 7 days."
5. Technical name: "regression effect" ; "regression toward the mean"
6. MD who observes a pt. many times, or orders many lab studies, may observe "abnormalities" that do not reflect a biologic variation but are due to temporal/measurement variation.
7. These, too, are likely to be "better" or changed soon. When faced with a test result that does not seem to fit, it helps to repeat the test.

Distributions: Measures of Central Tendency

- Frequency distr^{ns}, relative frequency distr^{ns}, & histograms help summarize collections of multiple observations.
- Frequency distribution:
  - divide obs^{ns} into 10 - 20 classes.
  - Record no. observations in each class.
- Distributions can be compared w.r.t. different features; important: "**centre**" or **location**:
  - The **mean**: the ordinary average of the observations, or
  - The **median**: defined earlier, or
  - The **mode**: the most popular (frequently occurring) value.

Distributions: Measures of Spread [cf. notes]

Degree of dispersion of obs^{ns} about their centre, defined by:

- **Range**: difference b/w largest & smallest observed values.
- **Interquartile range, "IQR"**: the range of values remaining when largest 25% and smallest 25% are set aside.
  These *quartiles* called Q_1 & Q_3, or Q_{25} & Q_{75}.
- **Standard deviation ("SD")**: frequently used, especially if distribution is roughly bell-shaped.
  ... Technically, $\sqrt{\text{average of squared deviations from the mean}}$.
  ... Français: écart-type, *typical* deviation.
- **Coefficient of variation ("CV")**: to compare degree of measurement error / intra-person or inter-person variation b/w situations / persons with v. different means / units.
  ... Unitless. See Notes for examples.

Distributions: Measures of Spread [cf. notes]

Degree of dispersion of obs^{ns} about their centre, defined by:

- **Range**: difference b/w largest & smallest observed values.
- **Interquartile range, "IQR"**: the range of values remaining when largest 25% and smallest 25% are set aside.
  These *quartiles* called Q_1 & Q_3, or Q_{25} & Q_{75}.
- **Standard deviation ("SD")**: frequently used, especially if distribution is roughly bell-shaped.
  ... Technically, $\sqrt{\text{average of squared deviations from the mean}}$.
  ... Français: écart-type, *typical* deviation.
- **Coefficient of variation ("CV")**: to compare degree of measurement error / intra-person or inter-person variation b/w situations / persons with v. different means / units.
  ... Unitless. See Notes for examples.

Distributions: Measures of Spread [cf. notes]

Degree of dispersion of obs^{ns} about their centre, defined by:

- **Range**: difference b/w largest & smallest observed values.
- **Interquartile range, "IQR"**: the range of values remaining when largest 25% and smallest 25% are set aside.
  These *quartiles* called Q_1 & Q_3, or Q_{25} & Q_{75}.
- **Standard deviation ("SD")**: frequently used, especially if distribution is roughly bell-shaped.
  ... Technically, $\sqrt{\text{average of squared deviations from the mean}}$.
  ... Français: écart-type, *typical* deviation.
- **Coefficient of variation ("CV")**: to compare degree of measurement error / intra-person or inter-person variation b/w situations / persons with v. different means / units.
  ... Unitless. See Notes for examples.

Distributions: Measures of Spread [cf. notes]

Degree of dispersion of obs^{ns} about their centre, defined by:

- **Range**: difference b/w largest & smallest observed values.
- **Interquartile range, "IQR"**: the range of values remaining when largest 25% and smallest 25% are set aside.
  These *quartiles* called Q_1 & Q_3, or Q_{25} & Q_{75}.
- **Standard deviation ("SD")**: frequently used, especially if distribution is roughly bell-shaped.
  ... Technically, $\sqrt{\text{average of squared deviations from the mean}}$.
  ... Français: écart-type, *typical* deviation.
- **Coefficient of variation ("CV")**: to compare degree of measurement error / intra-person or inter-person variation b/w situations / persons with v. different means / units.
  ... Unitless. See Notes for examples.

Distributions: Measures of Spread [cf. notes]

Degree of dispersion of obs^{ns} about their centre, defined by:

- **Range**: difference b/w largest & smallest observed values.
- **Interquartile range, "IQR"**: the range of values remaining when largest 25% and smallest 25% are set aside.
  These *quartiles* called Q_1 & Q_3, or Q_{25} & Q_{75}.
- **Standard deviation ("SD")**: frequently used, especially if distribution is roughly bell-shaped.
  ... Technically, $\sqrt{\text{average of squared deviations from the mean}}$.
  ... Français: écart-type, *typical* deviation.
- **Coefficient of variation ("CV")**: to compare degree of measurement error / intra-person or inter-person variation b/w situations / persons with v. different means / units.
  ... Unitless. See Notes for examples.

Biologic, temporal, and Measurement Variation - 2

Clin. Problem. Moderately ↑ BP at Routine Physical

A company refers Mr. W.P., a 35-year-old computer programmer, to you for pre-employment physical. Has a family history of stroke, is a 1 pack a day smoker, and his blood pressure is 130/95 mmHg.

- Goal of hypertension Tx: prevent morb. & mort. associated $\bar{c}$ high BP.
- Canadian Hypertension Education Project (CHEP) 2008 Guidelines: In general, BP should be lowered to less than 140/90 mmHg & in those with diabetes / chronic kidney disease, to less than 130/80 mmHg.

(`http://www.hypertension.ca/chep/resource-centre/publications/`)

Variability of Blood Pressure in Individual Patient

- First impulse: Mr. W.P. has a diastolic blood pressure between 90 and 99 mmHg, placing him in mild hypertension category (1992 table)
- This view may turn out to be correct, but before settling on it, we should review the variability of BP measurements.
- Armitage and Rose data ... next slide

Variability of Blood Pressure in Individual Patient

- First impulse: Mr. W.P. has a diastolic blood pressure between 90 and 99 mmHg, placing him in mild hypertension category (1992 table)
- This view may turn out to be correct, but before settling on it, we should review the variability of BP measurements.
- Armitage and Rose data ... next slide

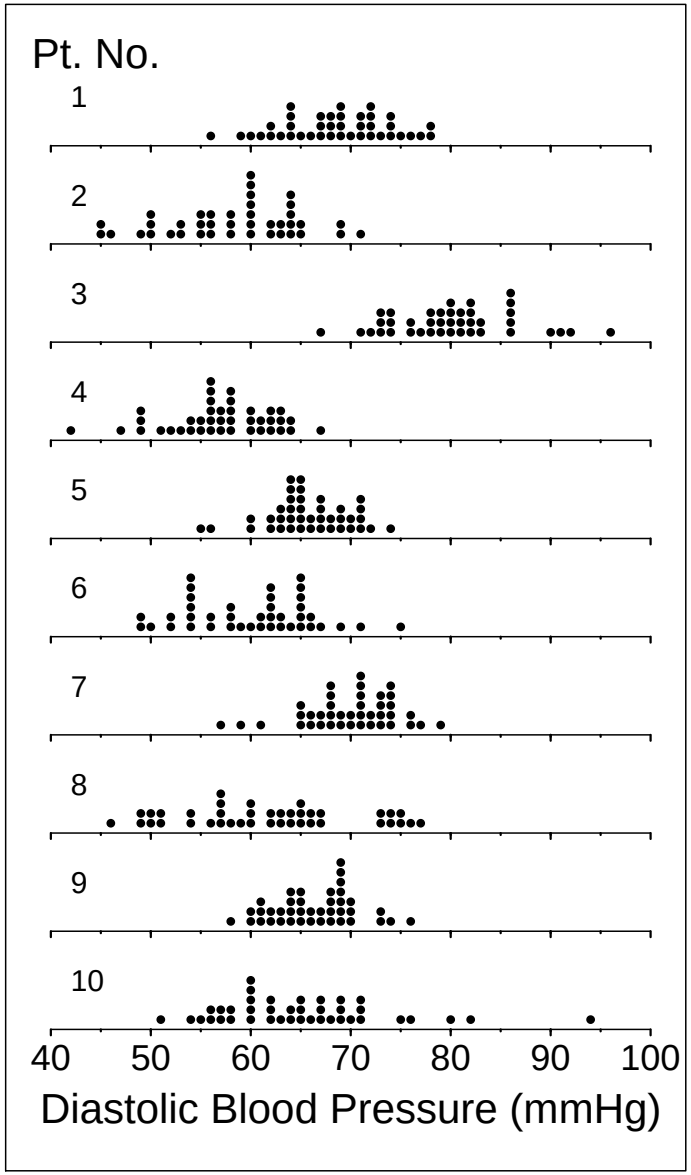
Variability of Blood Pressure in Individual Patient

- First impulse: Mr. W.P. has a diastolic blood pressure between 90 and 99 mmHg, placing him in mild hypertension category (1992 table)
- This view may turn out to be correct, but before settling on it, we should review the variability of BP measurements.
- Armitage and Rose data ... next slide

Variability of Blood Pressure in Individual Patient

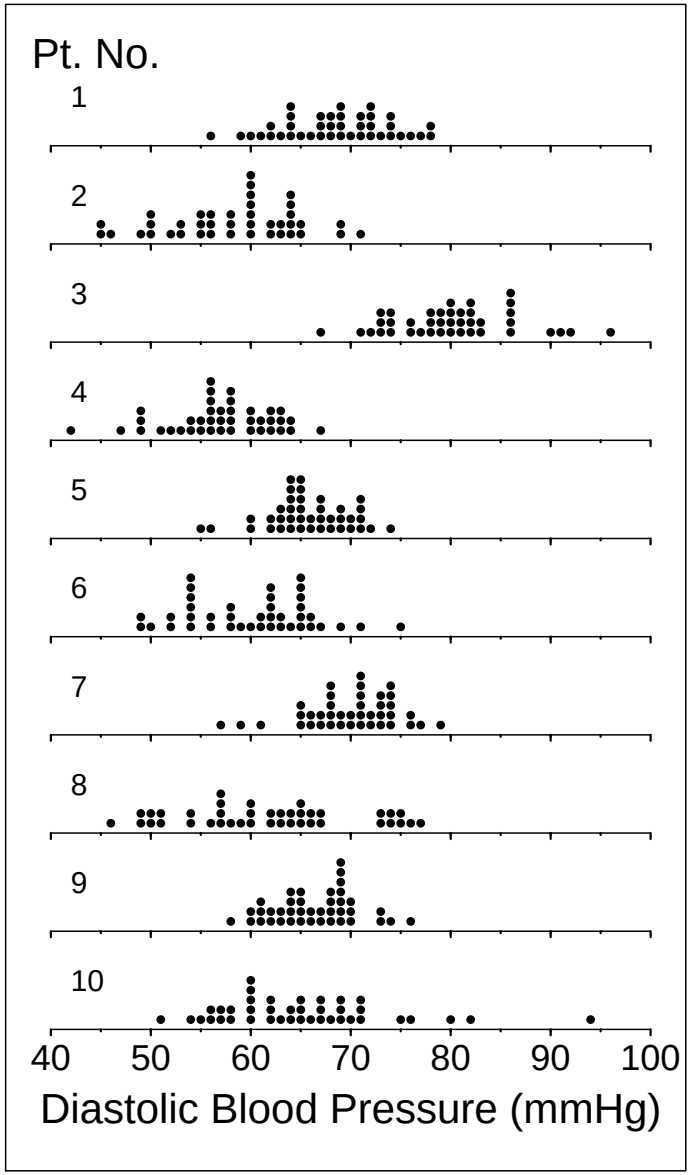
- First impulse: Mr. W.P. has a diastolic blood pressure between 90 and 99 mmHg, placing him in mild hypertension category (1992 table)
- This view may turn out to be correct, but before settling on it, we should review the variability of BP measurements.
- Armitage and Rose data ... next slide

DBP readings:10 ss, 2 readings on 20 occasions



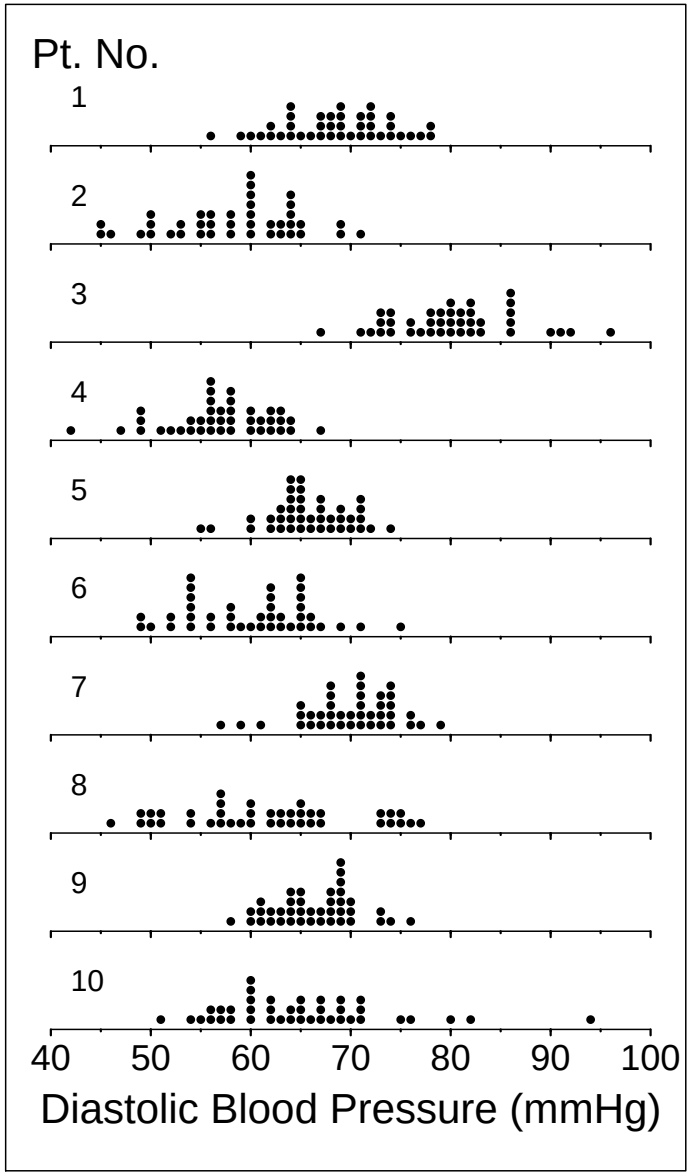
- Some subjects have ranges of measurements (largest minus smallest) of more than 30 mmHg.
- Mr. W.P.'s measurement of 95 could possibly be a high measurement for him, and perhaps he averages 15 mmHg lower, which would take him out of the hypertensive range.
- Or 95 might be a low measurement for him, and his average would be, say, 10 units higher, which would take him into the moderate category.
- This 'possibly an **over**-estimate / possibly an **under**-estimate' thinking is central to concepts of margin of error and confidence interval (later).

DBP readings:10 ss, 2 readings on 20 occasions



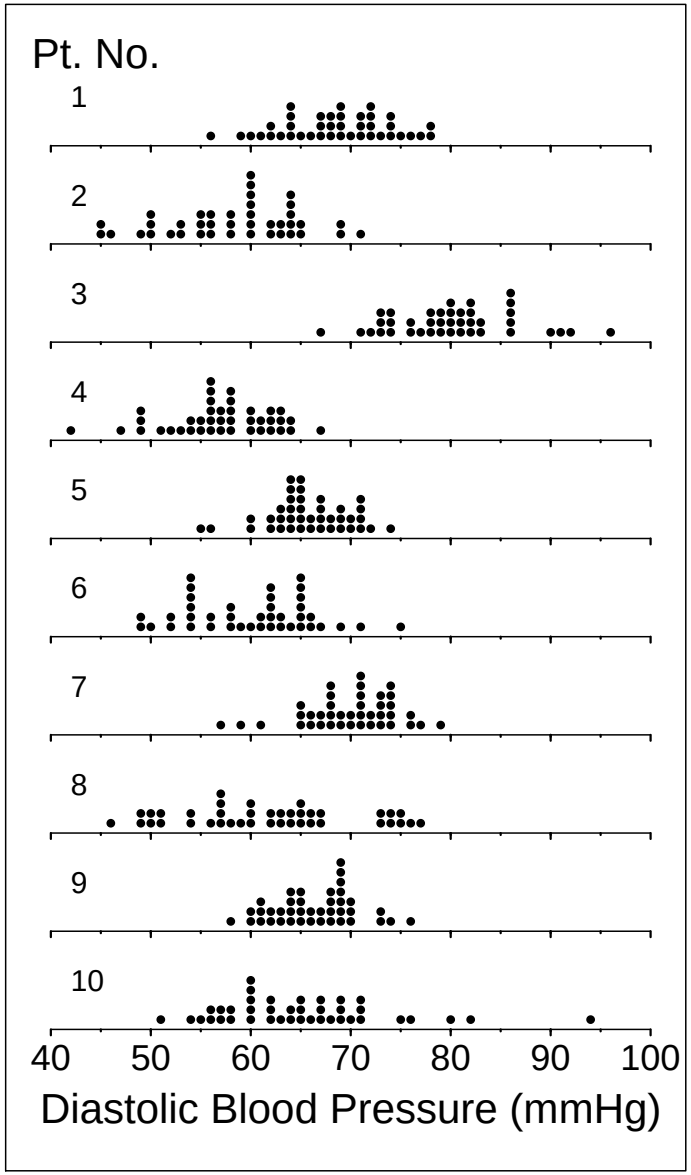
- Some subjects have ranges of measurements (largest minus smallest) of more than 30 mmHg.
- Mr. W.P.'s measurement of 95 could possibly be a high measurement for him, and perhaps he averages 15 mmHg lower, which would take him out of the hypertensive range.
- Or 95 might be a low measurement for him, and his average would be, say, 10 units higher, which would take him into the moderate category.
- This 'possibly an **over**-estimate / possibly an **under**-estimate' thinking is central to concepts of margin of error and confidence interval (later).

DBP readings:10 ss, 2 readings on 20 occasions



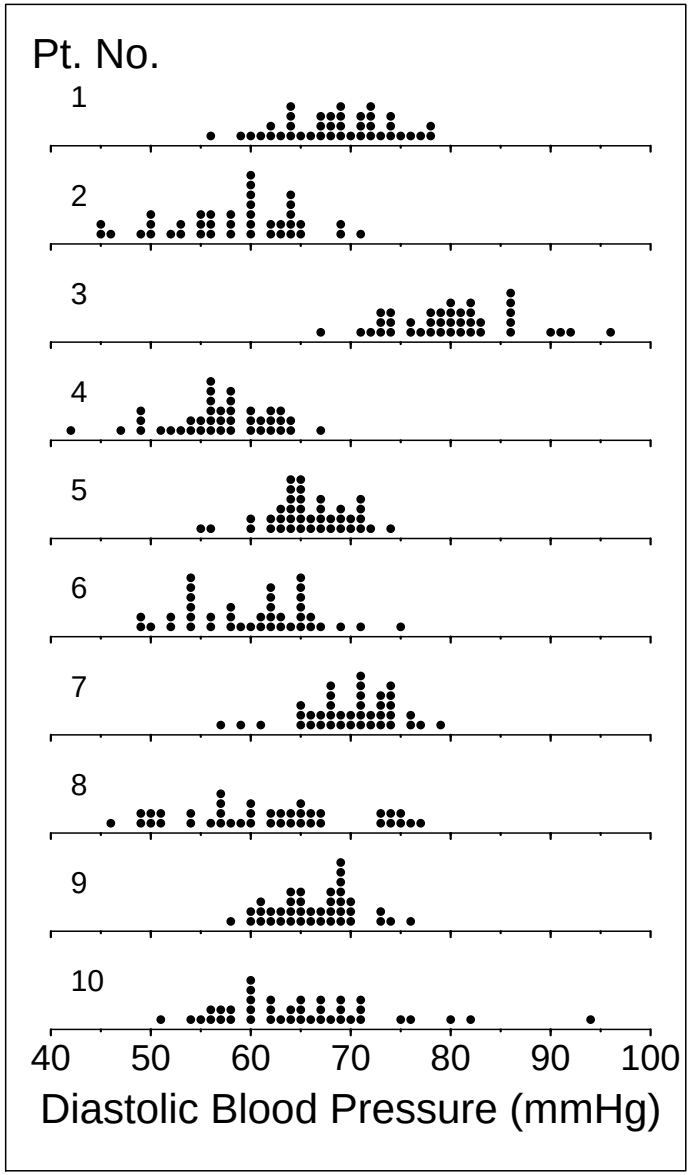
- Some subjects have ranges of measurements (largest minus smallest) of more than 30 mmHg.
- Mr. W.P.'s measurement of 95 could possibly be a high measurement for him, and perhaps he averages 15 mmHg lower, which would take him out of the hypertensive range.
- Or 95 might be a low measurement for him, and his average would be, say, 10 units higher, which would take him into the moderate category.
- This 'possibly an **over**-estimate / possibly an **under**-estimate' thinking is central to concepts of margin of error and confidence interval (later).

DBP readings:10 ss, 2 readings on 20 occasions



- Some subjects have ranges of measurements (largest minus smallest) of more than 30 mmHg.
- 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.
- This 'possibly an **over**-estimate / possibly an **under**-estimate' thinking is central to concepts of margin of error and confidence interval (later).

DBP readings:10 ss, 2 readings on 20 occasions



- Some subjects have ranges of measurements (largest minus smallest) of more than 30 mmHg.
- 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.
- This 'possibly an **over**-estimate / possibly an **under**-estimate' thinking is central to concepts of margin of error and confidence interval (later).

Messages from measurements on these 10 pts.

- Mr. W.P.'s diastolic blood pressure of 95 is ambiguous.
- Trying to reliably classify him on basis of $n = 1$ BP measurement is like trying to
  - classify someone as an A or a B student on basis of 1 multiple choice exam in 1 course
  - establish a taxi-driver's or waiter's income bracket on basis of 1 day's income.
- How much can we $\downarrow$ 'statistical noise' by averaging several measurements?
- Need to understand (im)precision of statistical estimates based on the mean of n values – in above data, $n = 1$!
- To mimimize side-issues, will use a simpler more generic e.g. to explain **key statistical concept – a Confidence Interval** – before returning to the case of Mr. W.P.

Imprecision of a sample mean or proportion

Key concepts / terminology :

- ***parameter*** The **true** mean level or proportion (often denoted by a Greek letter – μ or π or θ). **Value is unknowable**: not practical to measure level continuously or exhaustively; cannot obtain perfectly precise estimate.
- ***statistic***: Summary value calculated from values in a ***sample*** (Roman/Arabic letter – $\bar{y}$ (mean) or p (proportion).
- **Summary number** calculated from a small set (**sample**) of variable measurements or variable individuals **will not equal the (unknowable) value** one would have obtained had one been able to make all possible measurements
- In statistical shortand ...

$$\hat{\theta} \neq \theta \; ; \qquad \hat{\theta} = \theta + \textit{sampling var}^n \; ; \qquad \theta = \hat{\theta} + \textit{sampling var}^n .$$

Imprecision of a sample mean or proportion

In order to have a reproducible (precise) estimate of true – *but unknowable* – mean value, one needs to average ...

- many independent values if all of the possible measurements are highly variable about this true value;
- fewer independent values if they are highly concentrated about this true value.

A larger sample size does not guarantee that you will be closer to the target, since by luck of the draw an estimate based on $n = 4$ could be closer to it than another one based on $n = 8$.

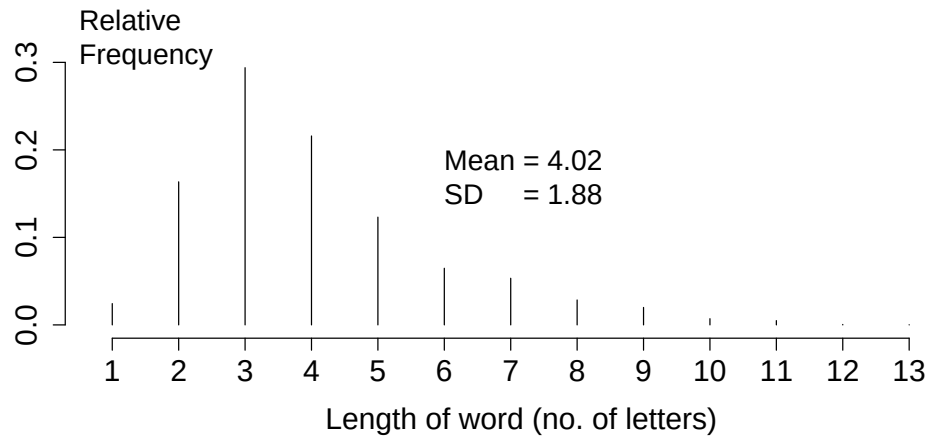
- **But** the ***probability*** of being within a certain specified distance of the target is *higher* with a sample of $n = 8$ than with one based on $n = 4$.
- It's a matter of probabilities, not of certainty.

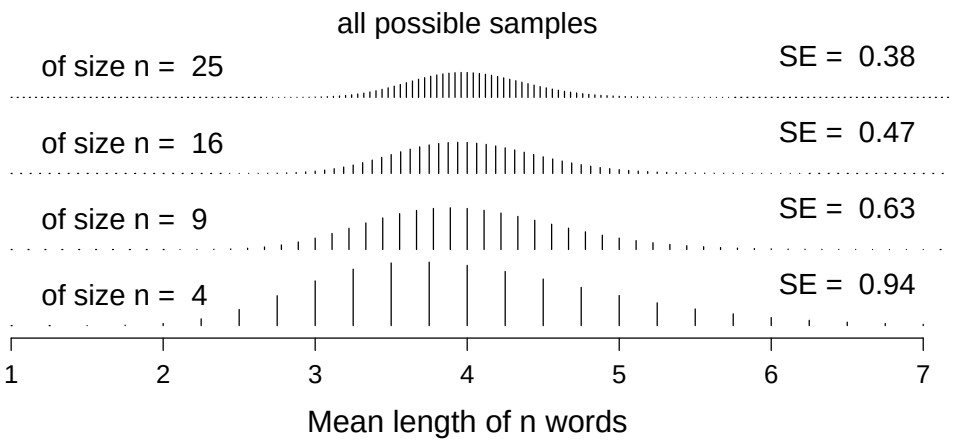
The good news...

The statistical (probability) laws governing the degree and frequency of under- or over-estimation (due to 'sampling' variation) are determined by surprisingly few factors.

- The pattern of variation of *individual* measurements may be quite non-Gaussian. However, the distribution of the **possible sample means** can be remarkably close to a Gaussian distribution (bell-curve) – centered on true value.
- The spread of this distribution is a function only of ...
  (1) the SD of individual measurements in the 'universe'
  (2) the sample size (n).

These laws in action: no. letters in individual words





- Distr^n of individual word lengths has long right tail.
- BUT, distr^{ns} of **means** of all *different possible samples of a given size*, much closer to Gaussian.
- When $n = 4$, sampling distribution still has a slightly long right tail
- If use $n = 25$, sampling distr^n close to Gaussian
- The spread (SD) of possible sample means is

$1.88/\sqrt{4} = 0.94$,
$1.88/\sqrt{9} = 0.63$,
$1.88/\sqrt{16} = 0.47$,
$1.88/\sqrt{25} = 0.38$, i.e.

SD(all possible $\bar{y}$'s) = $\frac{SD}{\sqrt{n}}$.

How do these probability laws help us?

- They answer Q: how far could a possible $\bar{y}$ be from μ?
- Our e.g. is a contrived one: why would we just use a sample if we already *know* $\mu = 4.02$ in full text?
- BUT, since they work in *known* situations where we can check their performance, they can also be expected to work in *unknown* situations where we don't know the truth.
- In practice, once we have observed our sample mean, $\bar{y}$, we are interested in the reverse Q:

How far might μ *be from* $\bar{y}$?

A ('toy') example where these laws can help

- μ = average word length in William Harvey's 1628 treatise *On The Motion Of The Heart And Blood In Animals.*
- μ is **UNKNOWN**; it would take a lot of work to determine it.
- In a random sample of $n = 100$ words from treatise,

$$\overline{length} = \bar{y} = 4.56 \textit{ letters}; \quad SD(100 \textit{ lengths}) = 2.40 \textit{ letters}.$$

- Our "**point estimate**" of μ is 4.56
- But this may be an **under-** or an **over-estimate.**
- Can we *work backwards* & 'bracket' (i.e., put limits on) μ?

A ('toy') example where these laws can help

- μ = average word length in William Harvey's 1628 treatise
  On The Motion Of The Heart And Blood In Animals.
- μ is **UNKNOWN**; it would take a lot of work to determine it.
- In a random sample of $n = 100$ words from treatise,

$$\overline{length} = \bar{y} = 4.56 \textit{ letters}; \quad SD(100 \textit{ lengths}) = 2.40 \textit{ letters}.$$

- Our "**point estimate**" of μ is 4.56
- But this may be an **under-** or an **over-estimate.**
- Can we *work backwards* & 'bracket' (i.e., put limits on) μ?

A ('toy') example where these laws can help

- μ = average word length in William Harvey's 1628 treatise *On The Motion Of The Heart And Blood In Animals.*
- μ is **UNKNOWN**; it would take a lot of work to determine it.
- In a random sample of $n = 100$ words from treatise,

$$\overline{length} = \bar{y} = 4.56 \text{ letters}; \quad SD(100 \text{ lengths}) = 2.40 \text{ letters}.$$

- Our "**point estimate**" of μ is 4.56
- But this may be an **under-** or an **over-estimate**.
- Can we *work backwards* & 'bracket' (i.e., put limits on) μ?

A ('toy') example where these laws can help

- μ = average word length in William Harvey's 1628 treatise *On The Motion Of The Heart And Blood In Animals.*
- μ is **UNKNOWN**; it would take a lot of work to determine it.
- In a random sample of $n = 100$ words from treatise,

$$\overline{length} = \bar{y} = 4.56 \text{ } letters; \quad SD(100 \text{ } lengths) = 2.40 \text{ } letters.$$

- Our "**point estimate**" of μ is 4.56
- But this may be an **under-** or an **over-estimate**.
- Can we *work backwards* & 'bracket' (i.e., put limits on) μ?

A ('toy') example where these laws can help

- μ = average word length in William Harvey's 1628 treatise *On The Motion Of The Heart And Blood In Animals.*
- μ is **UNKNOWN**; it would take a lot of work to determine it.
- In a random sample of $n = 100$ words from treatise,

$$\overline{length} = \bar{y} = 4.56 \text{ } letters; \quad SD(100 \text{ } lengths) = 2.40 \text{ } letters.$$

- Our "**point estimate**" of μ is 4.56
- But this may be an **under-** or an **over-estimate.**
- Can we *work backwards* & 'bracket' (i.e., put limits on) μ?

A ('toy') example where these laws can help

- μ = average word length in William Harvey's 1628 treatise *On The Motion Of The Heart And Blood In Animals.*
- μ is **UNKNOWN**; it would take a lot of work to determine it.
- In a random sample of $n = 100$ words from treatise,

$$\overline{length} = \bar{y} = 4.56 \text{ } letters; \quad SD(100 \text{ } lengths) = 2.40 \text{ } letters.$$

- Our "**point estimate**" of μ is 4.56
- But this may be an **under-** or an **over-estimate**.
- Can we *work backwards* & 'bracket' (i.e., put limits on) μ?

A ('toy') example where these laws can help

- μ = average word length in William Harvey's 1628 treatise *On The Motion Of The Heart And Blood In Animals.*
- μ is **UNKNOWN**; it would take a lot of work to determine it.
- In a random sample of $n = 100$ words from treatise,

$$\overline{length} = \bar{y} = 4.56 \text{ } letters; \quad SD(100 \text{ } lengths) = 2.40 \text{ } letters.$$

- Our "**point estimate**" of μ is 4.56
- But this may be an **under-** or an **over-estimate**.
- Can we *work backwards* & 'bracket' (i.e., put limits on) μ?

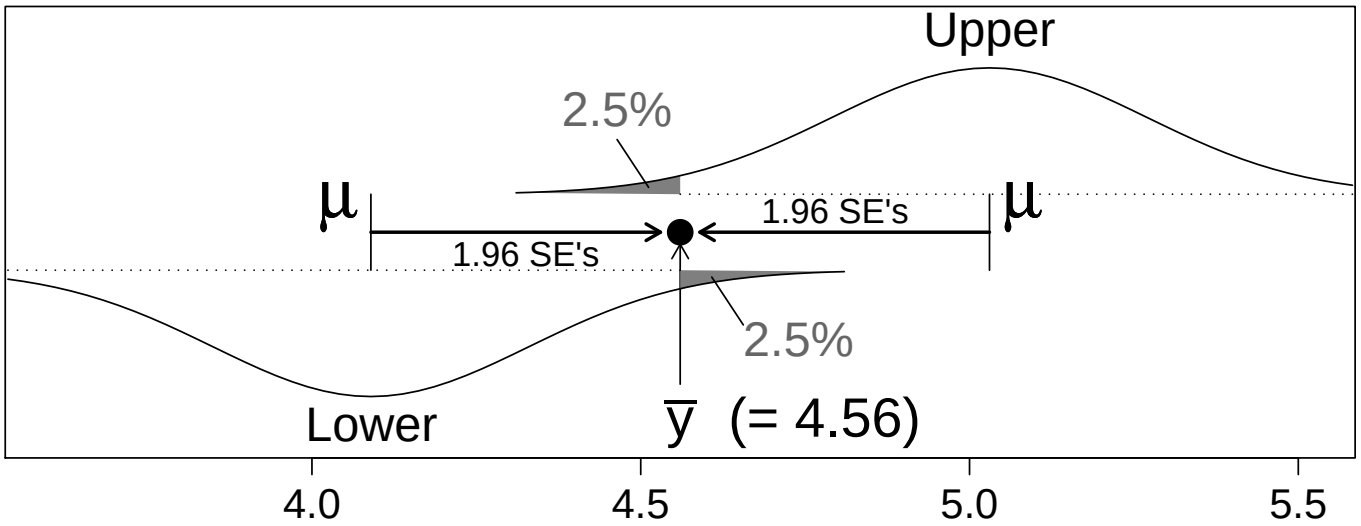
Reasoning behind a Confidence Interval

- Use 'hypothetical' or 'what if' logic.
- '**Try out**' various values of μ and calculate how 'far away' or how 'extreme' – probabilistically speaking – our observed 4.56 is in relation to these various trial values.
- Keep ('*rule in*') those trial μ values against which 4.56 is not extreme, and exclude ('*rule out*') those trial μ values against which it is.
- Our def^n: sample mean is '*extreme*' if 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~~ errors* from μ.
- Conversely, '*not extreme*' : any value that is less than 2 standard ~~deviations~~ errors from μ.
- * We use the term "*standard **error***" of the **statistic** – in keeping with our convention to **reserve the term** standard **deviation** for the variation of **individual** values.

Example of a 'What If?'

- Trial value: $\mu = 5.4$
- *If* μ were indeed 5.4, then our observed 4.56 would be an under-estimate. Is *this large* an under-estimate possible?
- 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, with a SE of $2.40/\sqrt{100} = 2.40/10 = 0.24$.
- Under this scenario, observed mean of 4.56 is 0.84 letters below the 5.4 we are currently entertaining as the μ for entire treatise. Since 1 *SE* = 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 observed 4.56 letters 'extreme' relative to this trial $\mu = 5.4$.
- Move trial value of μ towards 4.56, so 4.56 is not so extreme relative to μ.

In pictures rather than words – upper limit for μ



To find the μ at which 4.56 *is just at boundary b/w extreme and not*, need to have 4.56 be 1.96 *SE*'s below μ, i.e.,

$$\mu - 4.56 = 1.96 \times SE, \text{ or}$$

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

5.03 is '**upper limit**' for μ. **Shorthand:** $\mu_{upper} = 5.03$ or $\mu_U = 5.03$.

Lower and Upper limits for μ

- Similarly, lower limit for μ is value against which observed 4.56 is just as extreme an over-estimate.
- 4.56 is 1.96 *SE*'s above μ.

$$4.56 - \mu = 1.96 \times SE, \quad i.e.,$$

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

- **L**ower & **U**pper **limits**: $\mu_L = 4.09$ and $\mu_U = 5.03$ letters.
- Interval, or *range of parameter values*, b/w these two limits called a **Confidence Interval** for the parameter μ.

Properties of Confidence Intervals

- Limits constructed so that lower 2.5% of distr$^{n\cdot}$ centered at μ_U and upper 2.5% of distr$^{n\cdot}$ centered at μ_L are excluded.
- Interval b/w them is a **"95% confidence"** interval ("CI").
- CI's often misunderstood: need to appreciate exactly **how a CI should — and should not — be interpreted.**
- **Correct**: 95% of *all* 95% CI's 'trap' or 'include' the parameter value one would obtain with an infinite *n*.
- Thus, absent non-sampling biases, & selective disclosure/publication, sample survey companies, and scientists who publish results based on finite samples, might claim

  "Of every 100 "95% CI's" we supply/publish, on average, 95 of them include the true parameter value."

- "95% confidence" refers to *applications* of stat$^{l\cdot}$ *procedure*.

Anatomy / Components of a CI - 1

Q: Isn't a CI simply "your answer $\pm$ something"?

A: This simplistic formula does not always work.

- Statistics Canada Survey:
  - $n = 900$ *Canadians surveyed*
  - 20% "yes"
  - 95%*CI for %yes among Canadians* : $20\% \pm 3$ %*points* ✓
- Phase II study of experimental Tx:
  - $n = 4$ *patients*
  - 0% (0/4) "successes"
  - 95%*CI for %success in future pts.* :
    - $0\% \pm 0\%$ ×
    - 0% *to* 60% ('exact' 95% CI) ✓

Anatomy / Components of a CI - 2

The quantity after the $\pm$ is called the *Margin of Error*

Q: What determines the magnitude of the Margin of Error?

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

1. The *multiple* (*number of SE's* in table in next slide), which in turn is determined by "*degree of confidence*" used.
2. The SE, which in turn is proportional to σ (the SD of individual values) and inversely proportional to $\sqrt{n}$.

   Thus, if wish **to halve the SE**, and thus **halve width of CI**, need to **quadruple (not double!) the sample size.**

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

- You might be tempted to narrow the CI by taking smaller multiplier
- But, if you do, you also diminish the level of confidence.
- Without increasing amount of information that goes into the estimate, can only trade greater precision for less confidence, or vice versa.

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

- You might be tempted to narrow the CI by taking smaller multiplier
- But, if you do, you also diminish the level of confidence.
- Without increasing amount of information that goes into the estimate, can only trade greater precision for less confidence, or vice versa.

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

- You might be tempted to narrow the CI by taking smaller multiplier
- But, if you do, you also diminish the level of confidence.
- Without increasing amount of information that goes into the estimate, can only trade greater precision for less confidence, or vice versa.

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

- You might be tempted to narrow the CI by taking smaller multiplier
- But, if you do, you also diminish the level of confidence.
- Without increasing amount of information that goes into the estimate, can only trade greater precision for less confidence, or vice versa.

"Error bars" in research articles

Reports routinely use error bars in graphs of their results. Many of these do not explicitly state what error bars are. Could be...

- $\pm$**1***SE*:- if sampling distribution Gaussian – it is a **68**% CI.
- $\pm$**1.96***SE*$'s$:- thus it is a **95**% CI.
- $\pm$ **some ?? # of** *SE*$'s$, in which case it is a **??**% CI.
- $\pm$**1***SD*, or $\pm$**1.96***SD*$'s$:- if so, it describes variability of *individual values* that went into mean – rather than statistical precision of mean itself; the latter involves $\sqrt{n}$. Since SD is $\sqrt{n}$ times *larger* than SE, error bars are unlikely to be some $\pm$ number of SD's.

Advice: Always look in legend, or methods section, to find out what the error bars refer to. If they are not explained, but you have some sense of the SD, and know *n*, can often figure it out.

Why some CI's not symmetric about point estimate

"Relative risk (RR) of HIV-1 and other STIs in circumcised and uncircumcised men" from 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 study this article in more depth for last small group session*).

	n*	Cases	Person-years	Rate (cases per 100 person-years)	Unadjusted RR (95% CI)
HIV-1					
Uncircumcised	2107	165	3012·6	5·5	1·00 (reference)
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)
Circumcised	125	14	144·1	9·7	0·89 (0·48–1·53)

- CI of 0.02 to 0.47 is based on an **exact** (not a Gaussian-based) **CI** since numerators are so small that ratio is unstable.
- Gaussian-based CI's for ratios usually calculated on **log scale**, and become **asymmetric when back-converted**.
- (Incidentally) **Width of a CI for a rate ratio is a function of the numerators (#'s of cases)** in the two rates , not the denominators. Small numbers of cases (e.g. 2 HIV-1 among the circumcised) make rate ratio unstable.

The *n* required for a desired Margin of Error

In our sampling of Harvey's treatise, suppose we wished to estimate the mean with ME in a 95% CI of ± 0.1 letters.

- To achieve this, would need an n such that

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

- Can solve this for n to obtain $n = \{1.96 \times \sigma/0.1\}^2$.
- If use $\sigma = 2.5$ for planning, need random sample of

$$n = (1.96 \times 2.5/0.1)^2 = 2400 \textit{ words}.$$

- Why so large a sample size in this example?
  ... We stipulated a narrow ME: If average word length is approx 4.5 letters, ME of ± 0.1 letters represents just $(0.1/4.5) \times 100 = 2.2\%$ relative margin of error.
  ... 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 $(2.5/4.5) \times 100 = \underline{55}\%$!

To **halve** the ME ($\bar{s}$ changing % confidence) must **quadruple** n.

To which category does Mr. W.P.'s DBP belong?

- At pre-employment physical, BP was 130/95 mmHg.
- Suppose that in $n = 5$ new measurements, each taken on different occasion, the DBP's were 99, 98, 101, 95, and 90. Thus their mean is 96.6 (SD 4.3);Thus SE = $4.3/\sqrt{5} = 1.9$.
- With $n = 5$, 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 measurements been 89, 102, 97, 87 and 95 (mean 94, SD 6.1), CI would have been 86.4 to 101.6.

To which category does Mr. W.P.'s DBP belong?

- At pre-employment physical, BP was 130/95 mmHg.
- Suppose that in $n = 5$ new measurements, each taken on different occasion, the DBP's were 99, 98, 101, 95, and 90. Thus their mean is 96.6 (SD 4.3);Thus SE = $4.3/\sqrt{5} = 1.9$.
- With $n = 5$, 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 measurements been 89, 102, 97, 87 and 95 (mean 94, SD 6.1), CI would have been 86.4 to 101.6.

To which category does Mr. W.P.'s DBP belong?

- At pre-employment physical, BP was 130/95 mmHg.
- Suppose that in $n = 5$ new measurements, each taken on different occasion, the DBP's were 99, 98, 101, 95, and 90. Thus their mean is 96.6 (SD 4.3);Thus SE = $4.3/\sqrt{5} = 1.9$.
- With $n = 5$, 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 measurements been 89, 102, 97, 87 and 95 (mean 94, SD 6.1), CI would have been 86.4 to 101.6.

To which category does Mr. W.P.'s DBP belong?

- At pre-employment physical, BP was 130/95 mmHg.
- Suppose that in $n = 5$ new measurements, each taken on different occasion, the DBP's were 99, 98, 101, 95, and 90. Thus their mean is 96.6 (SD 4.3);Thus SE = $4.3/\sqrt{5} = 1.9$.
- With $n = 5$, 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 measurements been 89, 102, 97, 87 and 95 (mean 94, SD 6.1), CI would have been 86.4 to 101.6.

To which category does Mr. W.P.'s DBP belong?

- At pre-employment physical, BP was 130/95 mmHg.
- Suppose that in $n = 5$ new measurements, each taken on different occasion, the DBP's were 99, 98, 101, 95, and 90. Thus their mean is 96.6 (SD 4.3);Thus SE = $4.3/\sqrt{5} = 1.9$.
- With $n = 5$, 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 measurements been 89, 102, 97, 87 and 95 (mean 94, SD 6.1), CI would have been 86.4 to 101.6.

To which category does Mr. W.P.'s DBP belong?

- At pre-employment physical, BP was 130/95 mmHg.
- Suppose that in $n = 5$ new measurements, each taken on different occasion, the DBP's were 99, 98, 101, 95, and 90. Thus their mean is 96.6 (SD 4.3);Thus SE = $4.3/\sqrt{5} = 1.9$.
- With $n = 5$, 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 measurements been 89, 102, 97, 87 and 95 (mean 94, SD 6.1), CI would have been 86.4 to 101.6.

Did Diuretic Tx Lower Mrs. O.M.'s Blood Pressure?

- 50-y-o asymptomatic F; at routine physical, BP is 150/105.
- Started on a diuretic; 1 mo. later, BP is 140/95.
- Complains Tx has made her slightly weak; wants to stop taking it. Before urging her to continue Tx, has it $\downarrow$ BP ?
- Is there reasonable chance that observed difference in BP might have occurred $\bar{s}$ Tx?
- May need measurements from $>$ 1 occasion to have solid basis for decision.
- Have 2 diastolic measurements from each of $n_{pre} = 4$ pre-Tx visits with average values: 102, 105, 110, and 103.
- On $n_{post} = 3$ recent visits since beginning Tx, her averages have been: 95, 93, and 97.
- Use these 2 sets of measurements to assess improvement.

CI for the difference between 2 means

- Difference between the two sample means:

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

- Now 2 sources of imprecision, which "add in quadrature" :

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

- 95% CI for $\mu_1 - \mu_2$ is of the same "*answer* $\pm$ *multiple of SE*" form, i.e.,

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

- Estimated σ from data is $\hat{\sigma} \approx 3$, so SE's for 2 $\bar{y}$'s are $3/\sqrt{4}$ and $3/\sqrt{3}$. Thus,

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

Since *n*'s small, can't use 1.96 as multiple. Table for Student's '*t*' distribution, 5 df, tells us multiple should be 2.57. **Thus, 95% CI for** $\{\mu_{pre} - \mu_{post}\}$ is approx.

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

- **0 is not in the CI**. "We are confident that Mrs. O.M.'s mean DBP is lower. The ↓ of 10 mmHg is not readily accounted for by sampling variation. Because we are reasonably confident that Tx has ↓ her BP, we might urge her to continue it. Her weakness may be unrelated to the Tx, and it may disappear." [Ingelfinger]

CI for the difference between 2 means

- Difference between the two sample means:

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

- Now 2 sources of imprecision, which "add in quadrature" :

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

- 95% CI for $\mu_1 - \mu_2$ is of the same "*answer* $\pm$ *multiple of SE*" form, i.e.,

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

- Estimated σ from data is $\hat{\sigma} \approx 3$, so SE's for 2 $\bar{y}$'s are $3/\sqrt{4}$ and $3/\sqrt{3}$. Thus,

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

Since *n*'s small, can't use 1.96 as multiple. Table for Student's '*t*' distribution, 5 df, tells us multiple should be 2.57. **Thus, 95% CI for** $\{\mu_{pre} - \mu_{post}\}$ is approx.

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

- **0 is not in the CI**. "We are confident that Mrs. O.M.'s mean DBP is lower. The ↓ of 10 mmHg is not readily accounted for by sampling variation. Because we are reasonably confident that Tx has ↓ her BP, we might urge her to continue it. Her weakness may be unrelated to the Tx, and it may disappear." [Ingelfinger]

CI for the difference between 2 means

- Difference between the two sample means:

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

- Now 2 sources of imprecision, which "add in quadrature" :

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

- 95% CI for $\mu_1 - \mu_2$ is of the same "*answer* $\pm$ *multiple of SE*" form, i.e.,

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

- Estimated σ from data is $\hat{\sigma} \approx 3$, so SE's for 2 $\bar{y}$'s are $3/\sqrt{4}$ and $3/\sqrt{3}$. Thus,

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

Since *n*'s small, can't use 1.96 as multiple. Table for Student's '*t*' distribution, 5 df, tells us multiple should be 2.57. **Thus, 95% CI for** $\{\mu_{pre} - \mu_{post}\}$ is approx.

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

- **0 is not in the CI**. "We are confident that Mrs. O.M.'s mean DBP is lower. The ↓ of 10 mmHg is not readily accounted for by sampling variation. Because we are reasonably confident that Tx has ↓ her BP, we might urge her to continue it. Her weakness may be unrelated to the Tx, and it may disappear." [Ingelfinger]

CI for the difference between 2 means

- Difference between the two sample means:

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

- Now 2 sources of imprecision, which "add in quadrature" :

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

- 95% CI for $\mu_1 - \mu_2$ is of the same "*answer* $\pm$ *multiple of SE*" form, i.e.,

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

- Estimated σ from data is $\hat{\sigma} \approx 3$, so SE's for 2 $\bar{y}$'s are $3/\sqrt{4}$ and $3/\sqrt{3}$. Thus,

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

Since *n*'s small, can't use 1.96 as multiple. Table for Student's '*t*' distribution, 5 df, tells us multiple should be 2.57. **Thus, 95% CI for** $\{\mu_{pre} - \mu_{post}\}$ is approx.

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

- **0 is not in the CI**. "We are confident that Mrs. O.M.'s mean DBP is lower. The ↓ of 10 mmHg is not readily accounted for by sampling variation. Because we are reasonably confident that Tx has ↓ her BP, we might urge her to continue it. Her weakness may be unrelated to the Tx, and it may disappear." [Ingelfinger]

CI for the difference between 2 means

- Difference between the two sample means:

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

- Now 2 sources of imprecision, which "add in quadrature" :

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

- 95% CI for $\mu_1 - \mu_2$ is of the same "*answer* $\pm$ *multiple of SE*" form, i.e.,

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

- Estimated σ from data is $\hat{\sigma} \approx 3$, so SE's for 2 $\bar{y}$'s are $3/\sqrt{4}$ and $3/\sqrt{3}$. Thus,

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

  Since *n*'s small, can't use 1.96 as multiple. Table for Student's '*t*' distribution, 5 df, tells us multiple should be 2.57. **Thus, 95% CI for** $\{\mu_{pre} - \mu_{post}\}$ is approx.

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

- **0 is not in the CI**. "We are confident that Mrs. O.M.'s mean DBP is lower. The $\downarrow$ of 10 mmHg is not readily accounted for by sampling variation. Because we are reasonably confident that Tx has $\downarrow$ her BP, we might urge her to continue it. Her weakness may be unrelated to the Tx, and it may disappear." [Ingelfinger]

Key Points & Some Pointers

- Its déjà vu (point estimate $\pm$ ME) all over again, i.e. the generic CI structure. All that changes is that SE for a difference of two independent estimates has 2 components, one for each estimate.
- Don't fuss about formula for SE of a difference, since it will usually be computed by a statistical package.
- **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 technicalities when 2 n's small, and one has to use t- distribution, & concept of degrees of freedom.

  This part was included because those of you who have taken a statistics course will remember it (even if not why) & will ask why it wasn't mentioned.

P-Values and Statistical 'Tests'

"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*.

P-Values and Statistical 'Tests'

"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*.

P-Values and Statistical 'Tests'

"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*.

P-Values and Statistical 'Tests'

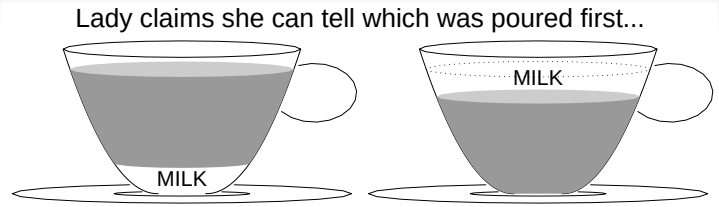
"P-Value"

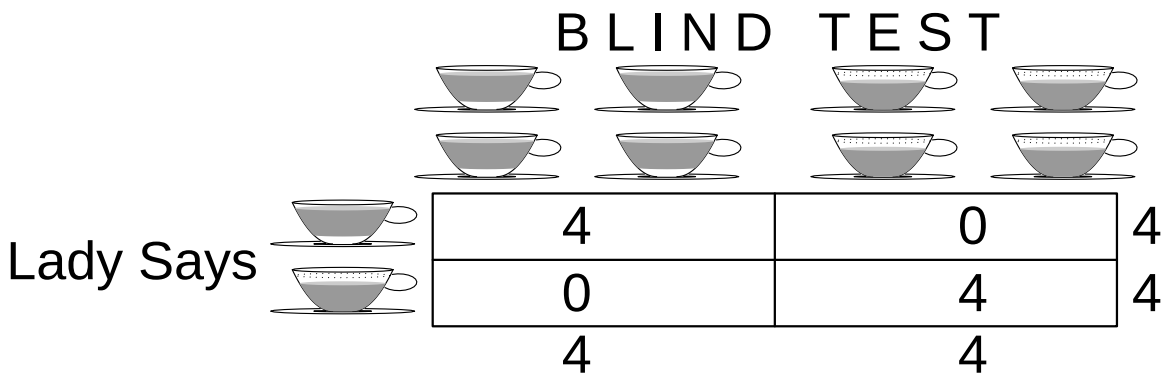
Def$^{n.}$ A **probability concerning the observed data**, calculated under a **Null Hypothesis** assumption, i.e., assuming that the only factor operating is sampling or measurement variation.

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*.

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



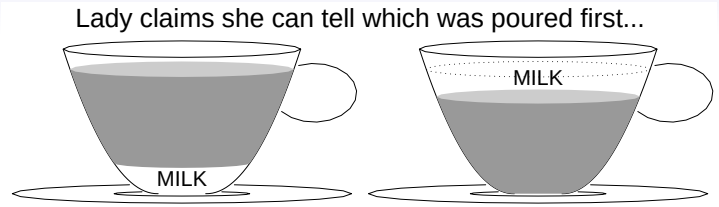


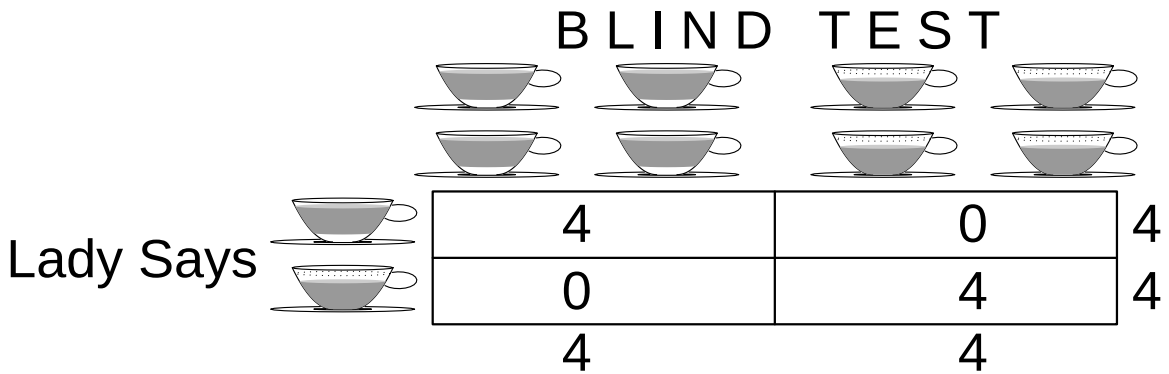
- "Null Hypothesis" (H_{null}): she can not tell them apart.
- "Alternative" Hypothesis (H_{alt}): she can (can you think of another "H" ?).

Blind test is equivalent to being asked to say **which 4** of the following 8 Gaelic words are the **correctly spelled** ones. You are told that **4 are correctly spelled & 4 are not**.

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

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



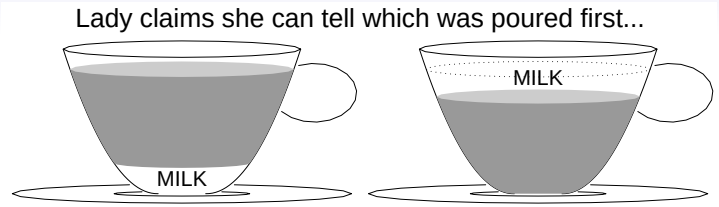


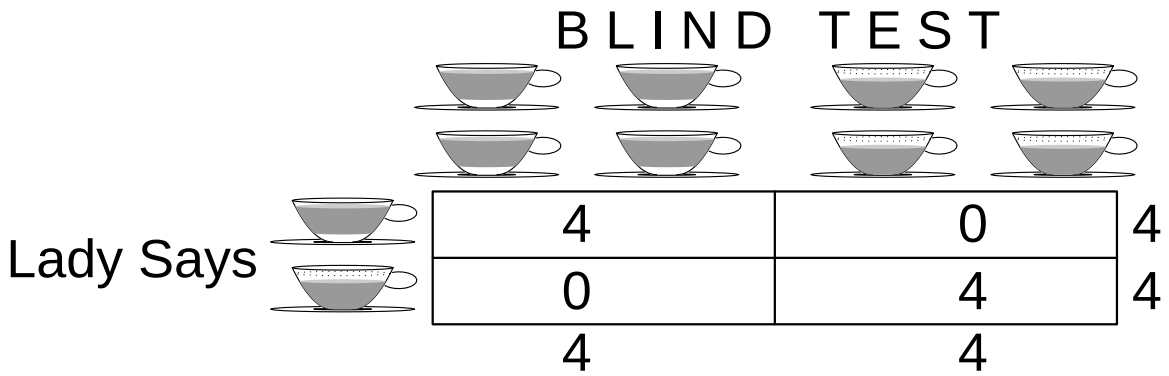
- "Null Hypothesis" (H_{null}): she can not tell them apart.
- "Alternative" Hypothesis (H_{alt}): she can (can you think of another "H" ?).

Blind test is equivalent to being asked to say **which 4** of the following 8 Gaelic words are the **correctly spelled** ones. You are told that **4 are correctly spelled & 4 are not**.

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

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



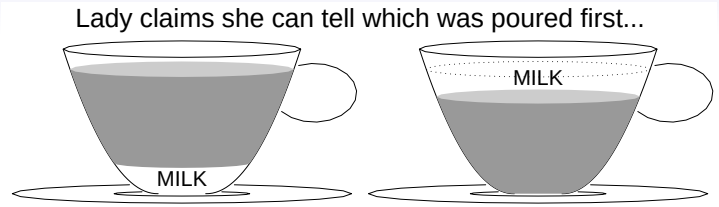


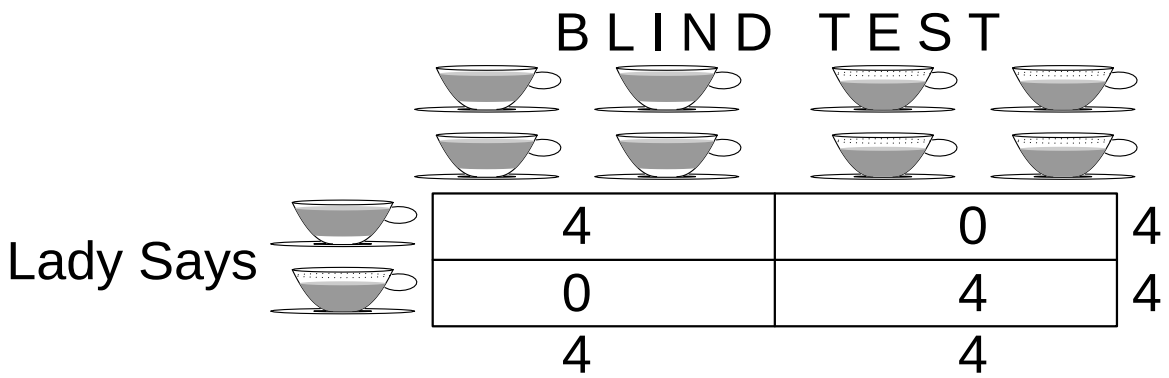
- "Null Hypothesis" (H_{null}): she can not tell them apart.
- "Alternative" Hypothesis (H_{alt}): she can (can you think of another "H" ?).

Blind test is equivalent to being asked to say **which 4** of the following 8 Gaelic words are the **correctly spelled** ones. You are told that **4 are correctly spelled & 4 are not**.

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

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





- "Null Hypothesis" (H_{null}): she can not tell them apart.
- "Alternative" Hypothesis (H_{alt}): she can (can you think of another "H" ?).

Blind test is equivalent to being asked to say **which 4** of the following 8 Gaelic words are the **correctly spelled** ones. You are told that **4 are correctly spelled & 4 are not**.

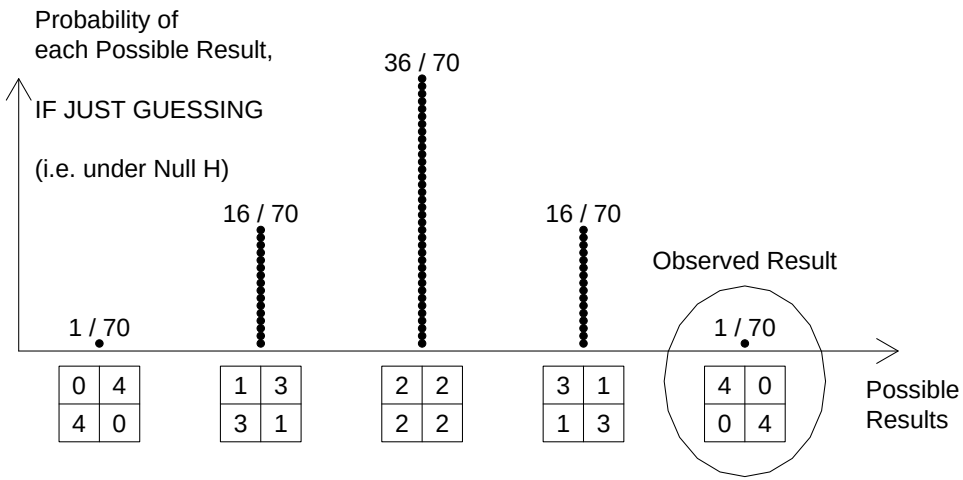
1	2	3	4	5	6	7	8
madra	olscoil	cathiar	tanga	doras	cluicha	féar	bóthar

These words of the following 8 Gaelic words are the **correctly spelled** ones.

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

The evidence provided by the test

- Rank possible test results by degree of evidence against H_{null}.
- "P-value" is the probability, calculated under null hypothesis, of observing a result as extreme as, or more extreme than, the one that was obtained/observed.



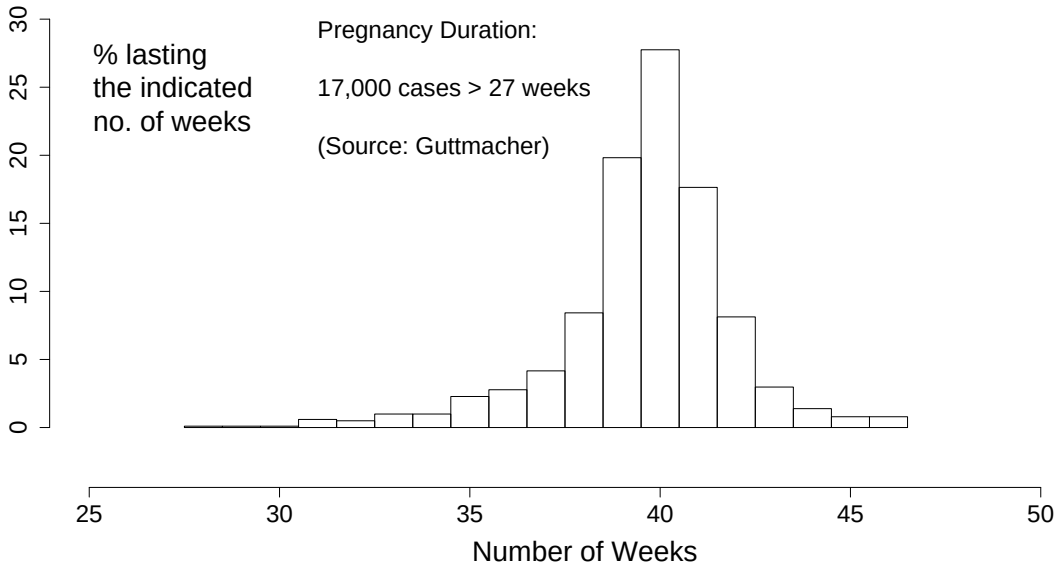
In this e.g., observed result is the most extreme, so

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

- Interpretation of such data often rather simplistic, 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, we ~~'accept' H_{null}~~). Avoid such simplistic 'conclusions'.

e.g. 2: Preston-Jones vs. Preston-Jones, English House of Lords, 1949

Divorce case: sole evidence of adultery was that a baby was born almost 50 weeks after husband had gone abroad on military service. Appeal failed. To quote 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.*"



• P-value, calculated under "Null" assumption that husband was father, = 'tail area' or probability corresponding to an observation of '50 or more weeks' in above dist^{rn}.

• Effectively asking: **What % of reference distribution does observed value exceed?** Same system used to report how extreme a lab value is – are told where value is located in distribution of values from healthy (reference) population.

P-Value via the Normal (Gaussian) distribution.

- 1st e.g. used specialized mathematical dist$^{rn.}$ as 'reference' (null) dist$^{rn.}$
- 2nd used an empirical population-based one.
- When judging extremeness of a sample mean or proportion (or difference b/w 2 sample means or proportions) calculated from an amount of information that is sufficient for the Central Limit Theorem to apply, one can use Gaussian distribution to readily obtain the P-value.
- Calculate how many standard errors of the statistic, $SE_{statistic}$, the statistic is from where null hypothesis states true value should be. This "number of SE's" is in this situation referred to as a 'Z_{value}.'

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

  P-value can then be obtained by determining what % of values in a Normal distribution are as extreme or more extreme than this Z_{value}.
- If n is small enough that value of $SE_{statistic}$, 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.

What the P-value is NOT

- P-value often mistaken for something very different.
- 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 Null Hypothesis is true**, *conditional on – i.e. given – the data.*
- Very few MDs mix up complement of specificity (i.e. probability of a 'positive' test result when in fact patient does not have disease in question) with positive predictive value (i.e. probability that a patient who has had a 'positive' test result does have disease in question).
- Statistical tests often coded '+ve' or '+ve' ('statistically significant' or not) according to whether results are extreme or not with respect to a reference (null) $\text{dist}^{rn\cdot}$. Medical tests also often coded as '+ve' or '-ve' according to whether results are extreme or not with respect to a ref. (healthy) $\text{dist}^{rn\cdot}$. But a test result is just one piece of data, and needs to be considered *along with rest of evidence* before coming to a 'conclusion.' **Likewise with statistical 'tests': the P-value is 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 blood on 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 prosecutor's fallacy

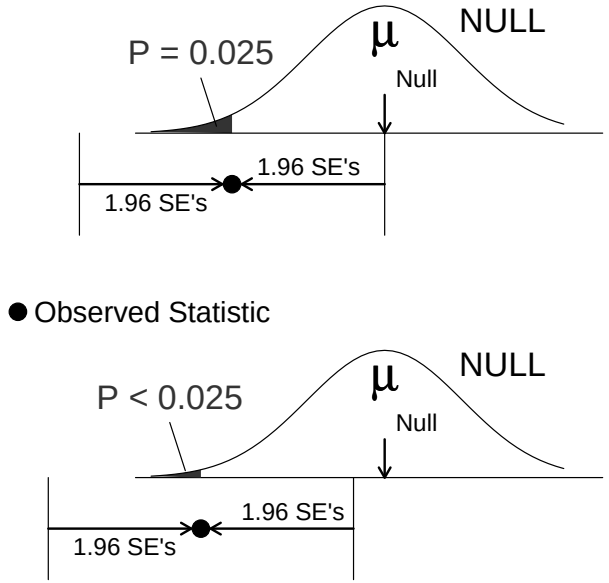
Who's the DNA fingerprinting pointing at? New Scientist, 1994.01.29, 51-52.

- David Pringle describes successful appeal of a rape case where primary evidence was DNA fingerprinting.
- Statistician Peter Donnelly opened new area of debate, remarking 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?"***

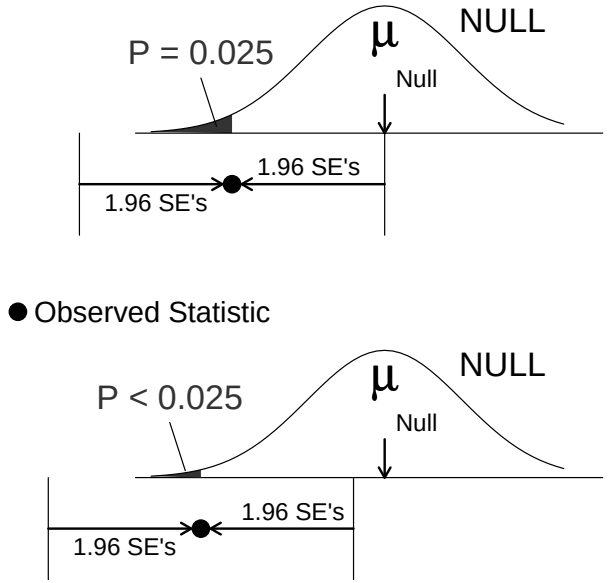
- The error in mixing up these two probabilities is called "**the prosecutor's fallacy**," and it is suggested that newspapers regularly make this error.
- Donnelly's testimony convinced the judges that the case before them involved an example of this and they ordered a retrial.

(Intimate) Relationship between P-value and CI



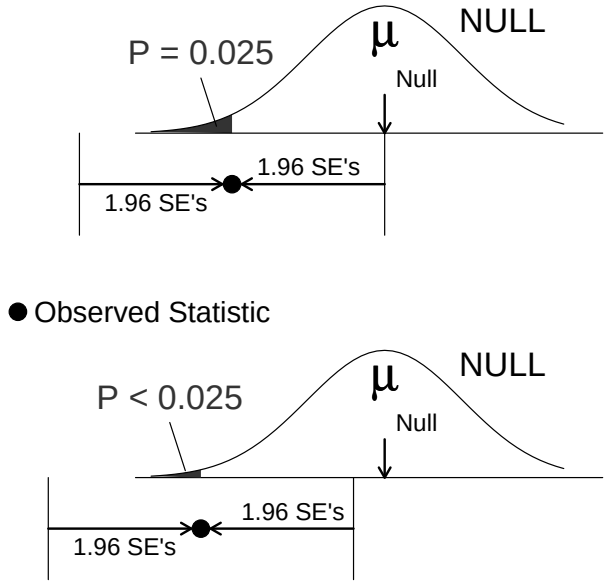
- If, as in the upper e.g. in graph, upper limit of 95% CI *just touches* null value, then the 2- (1-) sided) P-value is 0.05 (0.025).
- If, as in lower e.g., upper limit *excludes* null value, then the 2- (1-) sided) P-value is less than 0.05 (0.025).
- If (e.g not shown) CI *includes* null value, then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

(Intimate) Relationship between P-value and CI



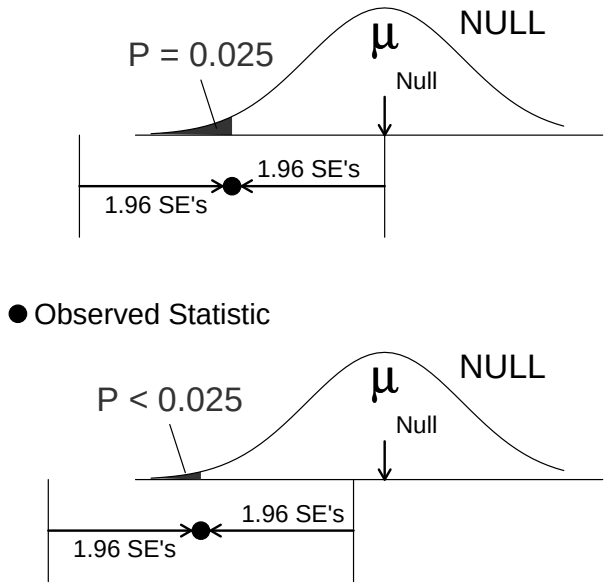
- If, as in the upper e.g. in graph, upper limit of 95% CI *just touches* null value, then the 2- (1-) sided) P-value is 0.05 (0.025).
- If, as in lower e.g., upper limit *excludes* null value, then the 2- (1-) sided) P-value is less than 0.05 (0.025).
- If (e.g not shown) CI *includes* null value, then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

(Intimate) Relationship between P-value and CI



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

(Intimate) Relationship between P-value and CI



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

Don't be overly-impressed by P-values

- P-values and 'significance tests' widely misunderstood and misused.
- Very large or very small *n*'s can influence what is / is not 'statistically significant.'
- Use CI's instead.
- *Pre study* power calculations (the chance that results will be 'statistically significant', as a function of the true underlying difference) of some help.
- *post-study* (i.e., *after the data have 'spoken'*), a CI is much more relevant, as it focuses on magnitude & precision, not on a probability calculated under H_{null}.

Statistical Inference: beyond the individual

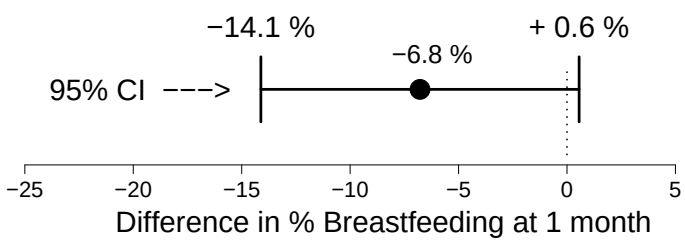
- "Statistical Inference" techniques (CI's, P-values, ...) **same** whether the focus is on individual patient, as in earlier e.g.'s, or on a larger universe as in e.g.'s below.
- Differences: what parameters ($\mu, \pi, \ldots$) stand for, and fact that main source of variability may be *inter*-individual.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

Do infant formula samples $\downarrow$ $dur^{n\cdot}$ of breastfeeding?

[Bergevin Y, Dougherty C, Kramer MS. Lancet. 1983 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.

At 1 month	Mothers given sample	Mothers 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	

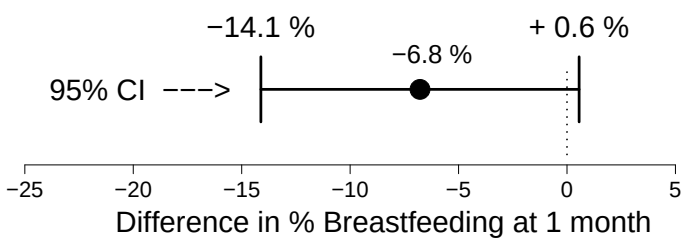


Do infant formula samples $\downarrow$ $dur^{n\cdot}$ of breastfeeding?

[Bergevin Y, Dougherty C, Kramer MS. Lancet. 1983 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.

At 1 month	Mothers given sample	Mothers 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	

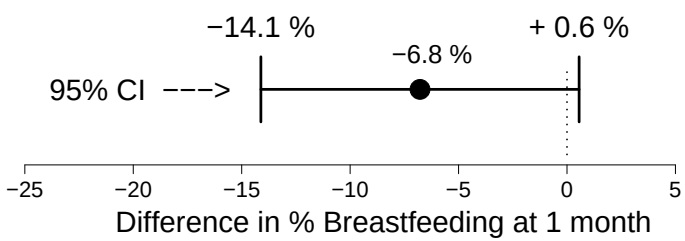


Do infant formula samples $\downarrow$ $dur^{n\cdot}$ of breastfeeding?

[Bergevin Y, Dougherty C, Kramer MS. Lancet. 1983 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.

At 1 month	Mothers given sample	Mothers 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	



Messages

- 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.
- (Msg.from 1st au. :) Plan to have **enough statistical power**. His study had only 50% power to detect a difference of 10 percentage points)

Messages

- 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.
- (Msg.from 1st au. :) Plan to have **enough statistical power**. His study had only 50% power to detect a difference of 10 percentage points)

Messages

- 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.
- (Msg.from 1st au. :) Plan to have **enough statistical power**. His study had only 50% power to detect a difference of 10 percentage points)

Messages

- 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.
- (Msg.from 1st au. :) Plan to have **enough statistical power**. His study had only 50% power to detect a difference of 10 percentage points)

Messages

- 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.
- (Msg.from 1st au. :) Plan to have **enough statistical power**. His study had only 50% power to detect a difference of 10 percentage points)

Do starch blockers really block calorie absorption?

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

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

Do starch blockers really block calorie absorption?

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

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

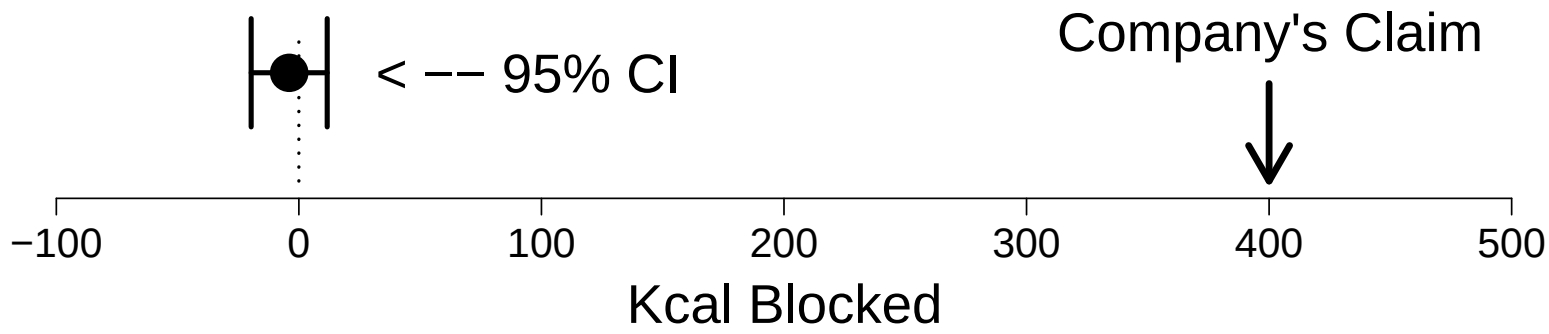
Do starch blockers really block calorie absorption?

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

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

Do starch blockers really block calorie absorption?

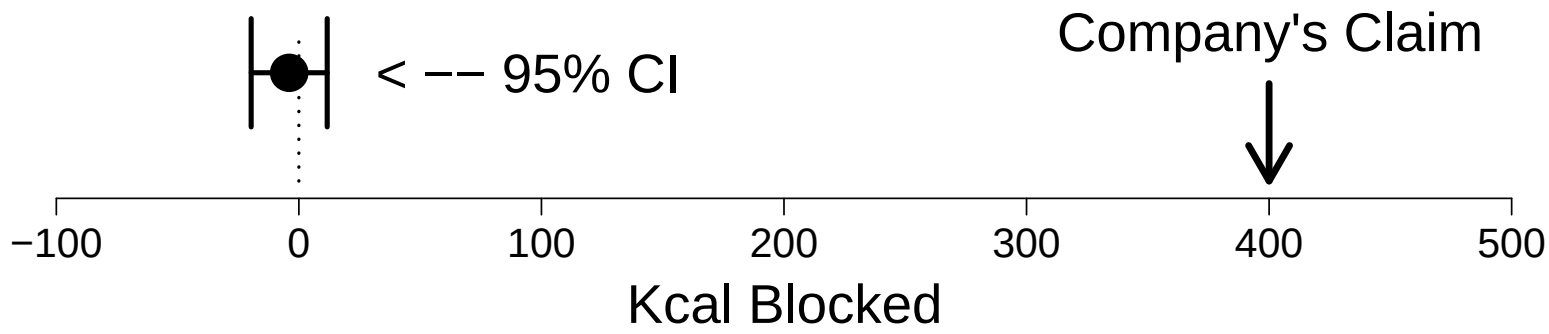
- However, fecal calorie excretion was same on the 2 test days (mean $\pm$ S.E.M., 80 $\pm$ 4 as compared with 78 $\pm$ 2).



- 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 !!!
- A '**DEFINITIVELY NEGATIVE**' STUDY.

Do starch blockers really block calorie absorption?

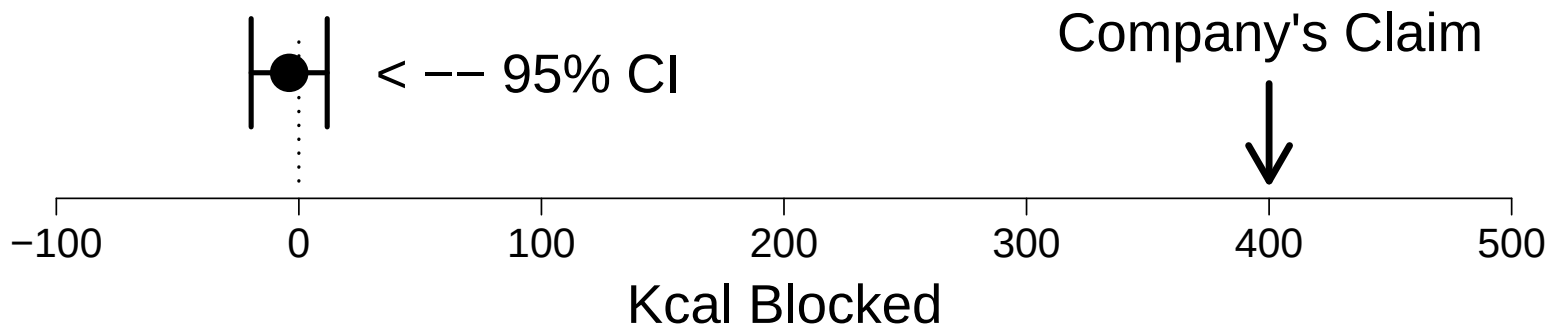
- However, fecal calorie excretion was same on the 2 test days (mean $\pm$ S.E.M., 80 $\pm$ 4 as compared with 78 $\pm$ 2).



- 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 !!!
- A '**DEFINITIVELY NEGATIVE**' STUDY.

Do starch blockers really block calorie absorption?

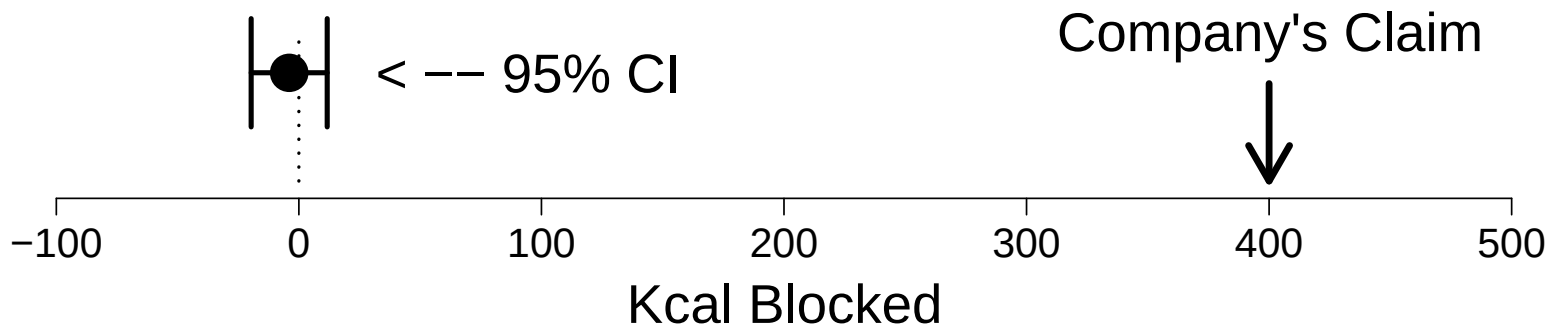
- However, fecal calorie excretion was same on the 2 test days (mean $\pm$ S.E.M., 80 $\pm$ 4 as compared with 78 $\pm$ 2).



- 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 !!!
- A '**DEFINITIVELY NEGATIVE**' STUDY.

Do starch blockers really block calorie absorption?

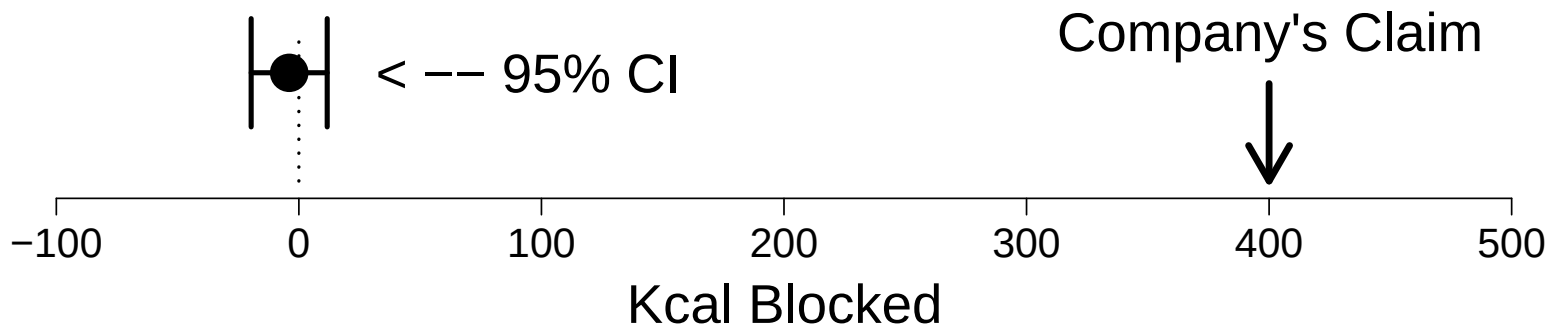
- However, fecal calorie excretion was same on the 2 test days (mean $\pm$ S.E.M., 80 $\pm$ 4 as compared with 78 $\pm$ 2).



- 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 !!!
- A '**DEFINITIVELY NEGATIVE**' STUDY.

Do starch blockers really block calorie absorption?

- However, fecal calorie excretion was same on the 2 test days (mean $\pm$ S.E.M., 80 $\pm$ 4 as compared with 78 $\pm$ 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 !!!
- A '**DEFINITIVELY NEGATIVE**' STUDY.

SUMMARY - 1

- The difference sources of variability have important implications in patient management.
- Descriptive statistics should be descriptive, and should suit the pattern of variation.
- Confidence intervals preferable to P-values, since they are expressed in terms of (comparative) parameter of interest; they allow us to judge 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.

SUMMARY - 1

- The difference sources of variability have important implications in patient management.
- Descriptive statistics should be descriptive, and should suit the pattern of variation.
- Confidence intervals preferable to P-values, since they are expressed in terms of (comparative) parameter of interest; they allow us to judge 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.

SUMMARY - 1

- The difference sources of variability have important implications in patient management.
- Descriptive statistics should be descriptive, and should suit the pattern of variation.
- Confidence intervals preferable to P-values, since they are expressed in terms of (comparative) parameter of interest; they allow us to judge 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.

SUMMARY - 1

- The difference sources of variability have important implications in patient management.
- Descriptive statistics should be descriptive, and should suit the pattern of variation.
- Confidence intervals preferable to P-values, since they are expressed in terms of (comparative) parameter of interest; they allow us to judge 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.

SUMMARY - 1

- The difference sources of variability have important implications in patient management.
- Descriptive statistics should be descriptive, and should suit the pattern of variation.
- Confidence intervals preferable to P-values, since they are expressed in terms of (comparative) parameter of interest; they allow us to judge 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.

SUMMARY - 1

- The difference sources of variability have important implications in patient management.
- Descriptive statistics should be descriptive, and should suit the pattern of variation.
- Confidence intervals preferable to P-values, since they are expressed in terms of (comparative) parameter of interest; they allow us to judge 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.

SUMMARY - 2

- Precise estimates distinguish b/w that which – if it were true – would be important and that which – if it were true – would not. '*n*' an important determinant of precision.
- A lab value in upper 1% of reference dist$^{rn.}$ (of values derived from people without known diseases/conditions) does not mean that there is a 1% chance that person in whom it was measured is healthy; 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, P-value $\neq$ probability that null hypothesis is true.
- The fact that

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

does not necessarily mean that

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

SUMMARY - 2

- Precise estimates distinguish b/w that which – if it were true – would be important and that which – if it were true – would not. '*n*' an important determinant of precision.
- A lab value in upper 1% of reference dist$^{rn.}$ (of values derived from people without known diseases/conditions) does not mean that there is a 1% chance that person in whom it was measured is healthy; 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, P-value $\neq$ probability that null hypothesis is true.
- The fact that

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

does not necessarily mean that

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

SUMMARY - 2

- Precise estimates distinguish b/w that which – if it were true – would be important and that which – if it were true – would not. '*n*' an important determinant of precision.
- A lab value in upper 1% of reference dist$^{rn.}$ (of values derived from people without known diseases/conditions) does not mean that there is a 1% chance that person in whom it was measured is healthy; 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, P-value $\neq$ probability that null hypothesis is true.
- The fact that

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

does not necessarily mean that

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

SUMMARY - 2

- Precise estimates distinguish b/w that which – if it were true – would be important and that which – if it were true – would not. '*n*' an important determinant of precision.
- A lab value in upper 1% of reference dist$^{rn.}$ (of values derived from people without known diseases/conditions) does not mean that there is a 1% chance that person in whom it was measured is healthy; 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, P-value $\neq$ probability that null hypothesis is true.
- The fact that

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

does not necessarily mean that

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

SUMMARY - 2

- Precise estimates distinguish b/w that which – if it were true – would be important and that which – if it were true – would not. '*n*' an important determinant of precision.
- A lab value in upper 1% of reference dist$^{rn.}$ (of values derived from people without known diseases/conditions) does not mean that there is a 1% chance that person in whom it was measured is healthy; 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, P-value $\neq$ probability that null hypothesis is true.
- The fact that

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

does not necessarily mean that

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

SUMMARY - 3

- Ultimately, P-values, CI's and other evidence from a study need to be combined with other information bearing on parameter or process.
- Don't treat any one study as last word on the topic.
- Worry also 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 (confounding...) in a comparison.

SUMMARY - 3

- Ultimately, P-values, CI's and other evidence from a study need to be combined with other information bearing on parameter or process.
- Don't treat any one study as last word on the topic.
- Worry also 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 (confounding...) in a comparison.

SUMMARY - 3

- Ultimately, P-values, CI's and other evidence from a study need to be combined with other information bearing on parameter or process.
- Don't treat any one study as last word on the topic.
- Worry also 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 (confounding...) in a comparison.

SUMMARY - 3

- Ultimately, P-values, CI's and other evidence from a study need to be combined with other information bearing on parameter or process.
- Don't treat any one study as last word on the topic.
- Worry also 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 (confounding...) in a comparison.