## Introduction to Statistical Inference*

**Inference is about Parameters (Populations) or general mechanisms -- or future observations. It is not about data (samples) *per se*, although it uses data from samples. Might think of inference as statements about a universe most of which one did not observe.**

**Two main schools or approaches:**

**Bayesian [ not even mentioned by Fletcher]**

- Makes direct statements about parameters and future observations
- Uses previous impressions plus new data to update impressions about parameter(s)

e.g.
Everyday life
Medical tests: Pre- and post-test impressions

**Frequentist**

- Makes statements about observed data (or statistics from data) (used indirectly [but often incorrectly] to assess evidence against certain values of parameter)
- Does not use previous impressions or data outside of current study (meta-analysis is changing this)

e.g.

• Statistical Quality Control procedures [for Decisions]
• Sample survey organizations: Confidence intervals
• Statistical Tests of Hypotheses

Unlike Bayesian inference, there is no quantified pre-test or pre-data "impression"; the ultimate statements are about data, conditional on an assumed null or other hypothesis.

Thus, an explanation of a p-value must start with the conditional "IF the parameter is ... the probability that the data would ..."

Book "Statistical Inference" by Michael W. Oakes is an excellent introduction to this topic and the limitations of frequentist inference.

## (Frequentist) Confidence Interval (CI) or Interval Estimate for a parameter $\theta$

**Formal definition:**

**A level 1 - $\alpha$ Confidence Interval for a parameter $\theta$ is given by two statistics**

$$\mathbf{U_{pper}\ and\ L_{ower}}$$

**such that when $\theta$ is the true value of the parameter,**

$$\mathbf{Prob\ (\ L_{ower} \le \theta \le U_{pper}\ ) = 1 - \alpha}$$

| $\alpha$ | **1 - $\alpha$** |
|---|---|
| 0.05 | 0.95 |
| 0.01 | 0.99 |

- CI is a **statistic**: a quantity calculated from a sample
- usually use $\alpha$ = 0.01 or 0.05 or 0.10, so that the "level of confidence", 1 - $\alpha$, is 99% or 95% or 90%. We will also use "$\alpha$" for tests of significance (there is a direct correspondence between confidence intervals and tests of significance)
- technically, we should say that we are **using a procedure which is guaranteed to cover the true $\theta$ in a fraction 1 - $\alpha$ of applications**. If we were not fussy about the semantics, we might say that any particular CI has a 1-$\alpha$ chance of covering $\theta$.
- for a given amount of sample data] the narrower the interval from L to U, the lower the degree of confidence in the interval and vice versa.

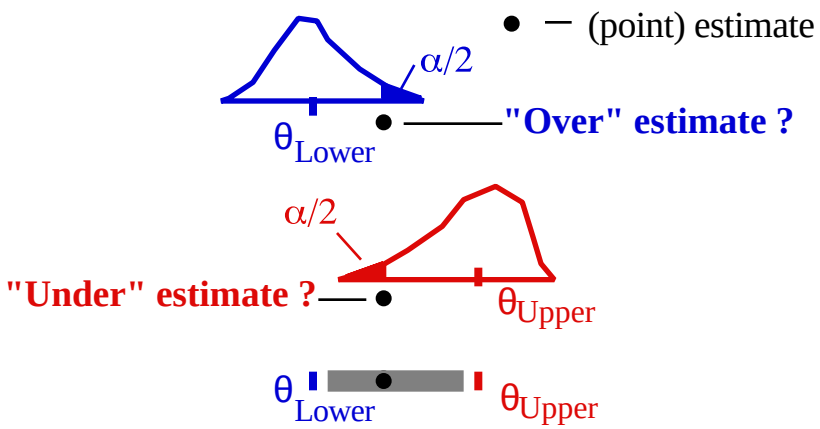
### Large-sample CI's

Many large-sample CI's are of the form

$\hat{\theta} \pm$ multiple of SE($\hat{\theta}$) or $f^{-1}$ [ $f\{\hat{\theta}\} \pm$ multiple of SE($f\{\hat{\theta}\}$ ] ,
where f is some function of $\hat{\theta}$ which has close to a Gaussian distribution, and $f^{-1}$ is the inverse function

example of latter : $\theta$ = odds ratio $f = ln$ ; $f^{-1} = \exp$

## Method of Constructing a 100(1 - α)% CI (in general):




## SD's* for "Large Sample" CI's for specific parameters

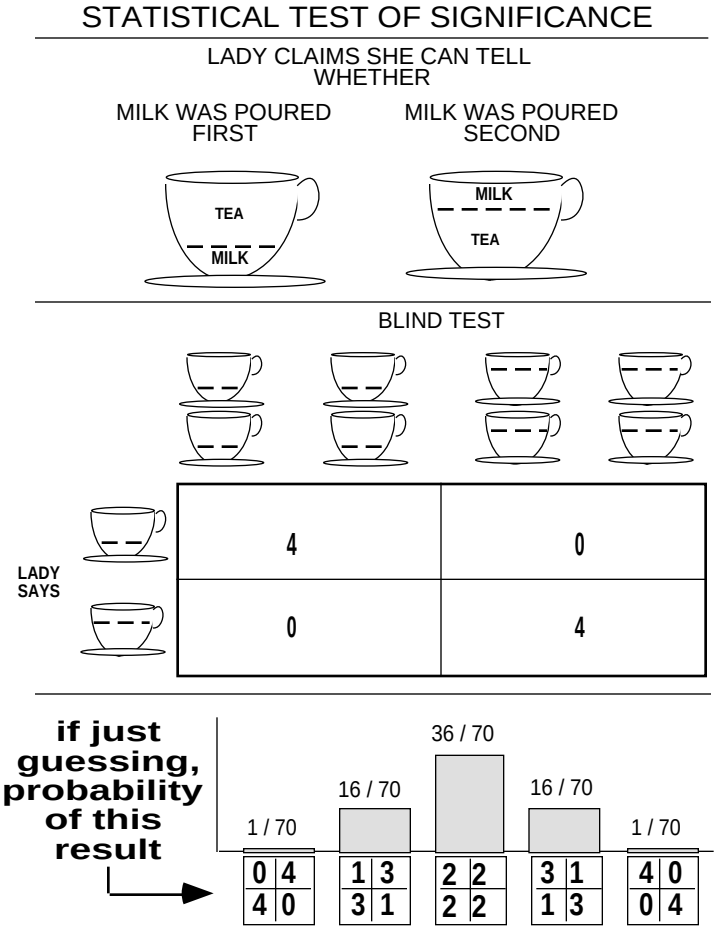
| parameter | | estimate | SD*(estimate) |
|---|---|---|---|
| | $\theta$ | $\hat{\theta}$ | $SD(\hat{\theta})$ |
| mean | $\mu_x$ | $\bar{x}$ | $\frac{\sigma_x}{\sqrt{n}}$ |
| prop. | $\pi$ | $p$ | $\frac{\sqrt{\pi[1-\pi]}}{\sqrt{n}}$ |
| | $\mu_1 - \mu_2$ | $\bar{x}_1 - \bar{x}_1$ | $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ |
| | $\pi_1 - \pi_2$ | $p_1 - p_2$ | $\sqrt{\frac{\pi_1[1-\pi_1]}{n_1} + \frac{\pi_2[1-\pi_2]}{n_2}}$ |

Polling companies who say "polls of this size are accurate to within so many percentage points 19 times out of 20" are being statistically correct -- they emphasize the **procedure** rather than what has happened in this specific instance. Polling companies (or reporters) who say "this poll is accurate .. 19 times out of 20" are talking statistical nonsense -- this specific poll is either "right" or "wrong"!. On average 19 polls out of 20 are "correct ". But ***this*** poll **cannot** be right on average 19 times out of 20!

## (Frequentist) Tests of Significance

Use: To assess the evidence provided by sample data in favour of a pre-specified claim or 'hypothesis' concerning some parameter(s) or data-generating process. As with confidence intervals, tests of significance make use of the concept of a sampling distribution.

Example 1 (see R. A Fisher, Design of Experiments Chapter 2)



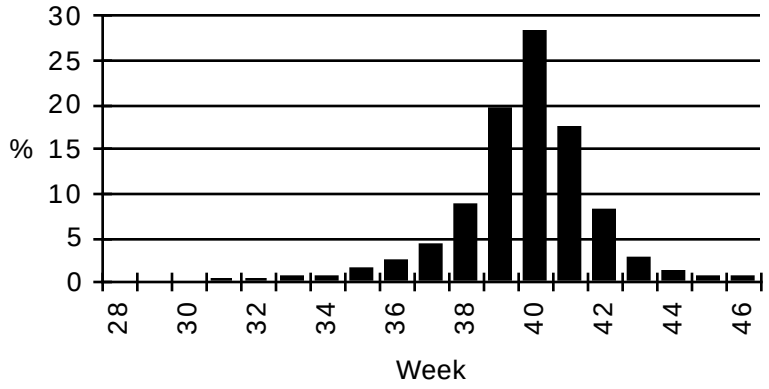

Example 2

In 1949 a divorce case was heard in which the sole evidence of adultery was that a baby was born almost 50 weeks after the husband had gone abroad on military service.

[Preston-Jones vs. Preston-Jones, English House of Lords]

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

Pregnancy Duration: 17000 cases > 27 weeks
(quoted in Guttmacher's book)




In U.S., [Lockwood vs. Lockwood, 19??], a 355-day pregnancy was found to be 'legitimate'.

Other Examples:

3. Quality Control (it has given us terminology)
4 Taste-tests (see exercises )
5. Adding water to milk.. see M&M2 Example 6.6 p448
6. Water divining.. see M&M2 exercise 6.44 p471
7. Randomness of U.S. Draft Lottery of 1970.. see M&M2 Example 6.6 p105-107, and 447-
8. Births in New York City after the "Great Blackout"
9 John Arbuthnot's "argument for divine providence"
10 US Presidential elections: Taller vs. Shorter Candidate.

## Elements of a Statistical Test

The ingredients and the methods of procedure in a statistical test are:

1. A claim about a parameter (or about the shape of a distribution, or the way a lottery works, etc.). Note that the null and alternative hypotheses are usually stated using Greek letters, i.e. in terms of population parameters, and in advance of (and indeed without any regard for) sample data. [ *Some have been known to write hypotheses of the form H: $\bar{y}$ = ... , thereby ignoring the fact that the whole point of statistical inference is to say something about the population in general, and not about the sample one happens to study. It is worth remembering that statistical inference is about the individuals one DID NOT study , not about the ones one did. This point is brought out in the absurdity of a null hypothesis that states that in a triangle taste test, exactly p=0.333.. of the n = 10 individuals to be studied will correctly identify the one of the three test items that is different from the two others.]*

2. A probability model (in its simplest form, a set of assumptions) which allows one to predict how a relevant statistic from a sample of data might be expected to behave under $H_0$.

3. A probability level or threshold or dividing point below which (i.e. close to a probability of zero) one considers that an event with this low probability 'is unlikely' or 'is not supposed to happen with a single trial' or 'just doesn't happen'. This pre-established limit of extreme-ness is referred to as the "α (alpha) level" of the test.

## Elements of a Statistical Test (Preston-Jones case)

1. Parameter (unknown) : DATE OF CONCEPTION

   Claim about parameter

   $H_0$ DATE ≤ HUSBAND LEFT (use = as 'best case')

   $H_a$ DATE > HUSBAND LEFT

2. A probability model for statistic ?Gaussian ?? Empirical?

3. A probability level or threshold
   (a priori ) "limit of extreme-ness" relative to $H_0$
   - for judge to decide

   Note extreme-ness measured as conditional probability, not in days

## Elements of a Statistical Test ...

4. A sample of data, which under $H_0$ is expected to follow the probability laws in (2).

5. The most relevant statistic (e.g. $\bar{y}$ if interested in inference about the parameter $\mu$)

6. The probability of observing a value of the statistic as extreme or more extreme (relative to that hypothesized under $H_0$) than we observed. This is used to judge whether the value obtained is either 'close to' i.e. 'compatible with' or 'far away from' i.e. 'incompatible with', $H_0$. The 'distance from what is expected under $H_0$' is usually measured as a tail area or probability and is referred to as the "P-value" of the statistic in relation to $H_0$.

7. A comparison of this "extreme-ness" or "unusualness" or "amount of evidence against $H_0$ " or P-value with the agreed-on "threshold of extreme-ness". If it is beyond the limit, $H_0$ is said to be "rejected". If it is not-too-small, $H_0$ is "not rejected". These two possible decisions about the claim are reported as "the null hypothesis is rejected at the P= $\alpha$ significance level" or "the null hypothesis is not rejected at a significance level of 5%".

## Elements of a Statistical Test (Preston-Jones case)

4. data: date of delivery.

5. The most relevant statistic (date of delivery; same as raw data: n=1)

6. The probability of observing a value of the statistic as extreme or more extreme (relative to that hypothesized under $H_0$) than we observed

   P-value = Upper tail area : Prob[ 349 or 350 or 351 ...] : quite small

7. A comparison of this "extreme-ness" or "unusualness" or "amount of evidence against $H_0$ " or P-value with the agreed-on "threshold of extreme-ness". Judge didn't tell us his threshold, but it must have been smaller than that calculated in 6.

   Note: the p-value does not take into account any other 'facts', prior beliefs, testimonials, etc.. in the case. But the judge probably used them in his overall decision (just like the jury did in the OJ case).

   .

## "Operating" Characteristics of a Statistical Test

As with diagnostic tests, there are 2 ways statistical test can be wrong:

1) **The null hypothesis was in fact correct but the sample was genuinely extreme and the null hypothesis was therefore (wrongly) rejected.**

2) **The alternative hypothesis was in fact correct but the sample was not incompatible with the null hypothesis and so it was not ruled out.**

The probabilities of the various test results can be put in the same type of 2x2 table used to show the characteristics of a diagnostic test.

| | | **Result of Statistical Test** | |
|---|---|---|---|
| | | "**Negative**" (do not reject $H_0$) | "**Positive**" (reject $H_0$ in favour of $H_a$) |
| **TRUTH** | **$H_0$** | $1 - \alpha$ | $\alpha$ |
| | **$H_a$** | $\beta$ | $1 - \beta$ |

The quantities $(1 - \beta)$ and $(1 - \alpha)$ are the "sensitivity (power)" and "specificity" of the statistical test. Statisticians usually speak instead of the complements of these probabilities, the false positive fraction ($\alpha$ ) and the false negative fraction ($\beta$) as "Type I" and "Type II" errors respectively [It is interesting that those involved in diagnostic tests emphasize the correctness of the test results, whereas statisticians seem to dwell on the errors of the tests; they have no term for $1-\alpha$ ].

Note that all of the probabilities start with (i.e. are conditional on knowing) the truth. This is exactly analogous to the use of sensitivity and specificity of diagnostic tests to describe the performance of the tests, conditional on (i.e. given) the truth. As such, they describe performance in a "what if" or artificial situation, just as sensitivity and specificity are determined under 'lab' conditions.

So just as we cannot interpret the result of a Dx test simply on basis of sensitivity and specificity, likewise we cannot interpret the result of a statistical test in isolation from what one already thinks about the null/alternative hypotheses.

## Interpretation of a "positive statistical test"

It should be interpreted n the same way as a "positive diagnostic test" i.e. in the light of the characteristics of the subject being examined. The lower the prevalence of disease, the lower is the post-test probability that a positive diagnostic test is a "true positive". Similarly with statistical tests. We are now no longer speaking of sensitivity = Prob( test + | $H_a$ ) and specificity = Prob( test - | $H_0$ ) but rather, the other way round, of Prob( $H_a$ | test + ) and Prob( $H_0$ | test - ), i.e. of positive and negative predictive values, both of which involve the "background" from which the sample came.

**A Popular Misapprehension:** It is not uncommon to see or hear seemingly knowledgeable people state that

> "the P-value (or alpha) is the probability of being wrong if, upon observing a statistically significant difference, we assert that a true difference exists"

Glantz (in his otherwise excellent text) and Brown (Am J Dis Child 137: 586-591, 1983 -- on reserve) are two authors who have made statements like this. For example, Brown, in an otherwise helpful article, says (italics and strike through by JH) :

> ~~*"In practical terms, the alpha of .05 means that the researcher, during the course of many such decisions, accepts being wrong one in about every 20 times that he thinks he has found an important difference between two sets of observations"*~~ [1]

But if one follows the analogy with diagnostic tests, this statement is like saying that

> *"1-minus-specificity is the probability of being wrong if, upon observing a positive test, we assert that the person is diseased".*

We know [from dealing with diagnostic tests] that we cannot turn Prob( test | H ) into Prob( H | test ) without some knowledge about the unconditional or a-priori Prob( H ) ' s.

The influence of "background" is easily understood if one considers an example such as a testing program for potential chemotherapeutic agents. Assume a certain proportion P are truly active and that statistical testing of them uses type I and Type II errors of α and β respectively. A certain proportion of all the agents will test positive, but what fraction of these "positives" are truly positive? It obviously depends on α and β, but it also depends in a big way on P, as is shown below for the case of α = 0.05, β = 0.2.

| P --> | 0.001 | .01 | .1 | .5 |
|---|---|---|---|---|
| TP = P(1- β) --> | .00080 | .0080 | .080 | .400 |
| FP = (1 - P)(α)-> | .04995 | .0495 | .045 | .025 |
| Ratio TP : FP --> | ≈ 1 : 62 | ≈ 1: 6 | ≈ 2 : 1 | ≈ 16 : 1 |

Note that the post-test odds TP:FP is

$$P(1-\beta) : (1-P)(\alpha) = \{ P : (1-P) \} \times \left[ \frac{1-\beta}{\alpha} \right]$$

PRIOR × function of TEST's characteristics

i.e. it has the form of a "**prior odds**" P : (1 - P), the "background" of the study, multiplied by a "**likelihood ratio positive**" which depends only on the characteristics of the statistical test. Text by Oakes helpful here

[1] [Incidentally, there is a second error in this statement : it has to do with equating a "statistically significant" difference with an important one... minute differences in the means of large samples will be statistically significant ]

## "SIGNIFICANCE"

*notes prepared by FDK Liddell, ~1970*

And then, even if the cure should be performed, how can he be sure that this was not because the illness had reached its term, or a result of chance, or the effect of something else he had eaten or drunk or touched that day, or the merit of his grandmother's prayers? Moreover, even if this proof had been perfect, how many times was the experiment repeated? How many times was the long string of chances and coincidences strung again for a rule to be derived from it?

*Michel de Montaigne 1533-1592*

The same arguments which explode the Notion of Luck may, on the other side, be useful in some Cases to establish a due comparison between Chance and Design. We may imagine Chance and Design to be as it were in Competition with each other for the production of some sorts of Events, and may calculate what Probability there is, that those Events should be rather owing to one than to the other... From this last Consideration we may learn in many Cases how to distinguish the Events which are the effect of Chance, from those which are produced by Design.

*Abraham de Moivre: 'Doctrine of Chances' (1719)*

If we... agree that an event which would occur by chance only once in (so many) trials is decidedly 'significant', in the statistical sense, we thereby admit that no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon; for the 'one chance in a million' will undoubtedly occur, with no less and no more than its appropriate frequency, however surprised we may be that it should occur to us.

*R A Fisher 'The Design of Experiments'*
*(First published 1935)*

The difference between two treatments is 'statistically significant' if it is sufficiently large that it is unlikely to have risen by chance alone. The level of significance is the probability of such a large difference arising in a trial when there really is no difference in the effects of the treatments. (But the lower the probability, the less likely is it that the difference is due to chance, and so the more highly significant is the finding.)

- Statistical significance does not imply clinical importance.
- Even a very unlikely (i.e. highly significant) difference may be unimportant.
- Non-significance does not mean no real difference exists.
- A significant difference is not necessarily reliable.
- Statistical significance is not proof that a real difference exists.
- There is no 'God-given' level of significance. What level would you require before being convinced:
  - a to use a drug (without side effects) in the treatment of lung cancer?
  - b that effects on the foetus are excluded in a drug which depresses nausea in pregnancy?
  - c to go on a second stage of a series of experiments with rats?
- Each statistical test (i.e. calculation of level of significance, or unlikelihood of observed difference) must be strictly independent of every other such test. Otherwise, the calculated probabilities will not be valid. This rule is often ignored by those who:
  - measure more than on response in each subject
  - have more than two treatment groups to compare
  - stop the experiment at a favourable point.

**"Definitive Negative" Studies?** Starch blockers--their effect on calorie absorption from a high-starch meal.

**Abstract**

It has been known for more than 25 years that certain plant foods, such as kidney beans and wheat, contain a substance that inhibits the activity of salivary and pancreatic amylase. More recently, this antiamylase has been purified and marketed for use in weight control under the generic name "starch blockers." Although this approach to weight control is highly popular, it has never been shown whether starch-blocker tablets actually reduce the absorption of calories from starch. Using a one-day calorie-balance technique and a high-starch (100 g) meal (spaghetti, tomato sauce, and bread), we measured the excretion of fecal calories after normal subjects had taken either placebo or starch-blocker tablets. If the starch-blocker tablets had prevented the digestion of starch, fecal calorie excretion should have increased by 400 kcal. However, fecal reduce the absorption of calories from starch. Using a one-day calorie-balance technique and a high-starch (100 g) meal (spaghetti, tomato sauce, and bread), we measured the excretion of fecal calories after normal subjects had taken either placebo or starch-blocker tablets. **If the starch-blocker tablets had prevented the digestion of starch, fecal calorie excretion should have increased by 400 kcal. However, fecal calorie excretion was the same on the two test days (mean ± S.E.M., 80 ± 4 as compared with 78 ± 2). We conclude that starch-blocker tablets do not inhibit the digestion and absorption of starch calories in human beings.**

Bo-Linn GW. et al New England Journal of Medicine. 307(23):1413-6, 1982 Dec 2

[Overview of Methods: The one-day calorie-balance technique begins with a preparatory washout in which the entire gastrointestinal tract is cleansed of all food and fecal material by lavage with a special calorie-free, electrolyte-containing solution. The subject then eats the test meal, which includes $^{51}CrCl_3$ as a non absorbable marker. After 14 hours, the intestine is cleansed again by a final washout. The rectal effluent is combined with any stool (usually none) that has been excreted since the meal was eaten. The energy content of the ingested meal and of the rectal effluent is determined by bomb calorimetry. The completeness of stool collection is evaluated by recovery of the non absorbable marker.]

---

For an **good paper on topic of 'negative' studies**, see Powell-Tuck J "A defence of the small clinical trial: evaluation of three gastroenterological studies." British Medical Journal Clinical Research Ed..292(6520):599-602, 1986 Mar 1. (Resources for Ch 7)

**Table 1. Standard Test Meal.**

Ingredients

| | |
|---|---|
| Spaghetti (dry weight)* | 100 g |
| Tomato sauce | 112 g |
| White bread | 50 g |
| Margarine | 10 g |
| Water | 250 g |
| $^{51}CrCl_3$ | 4 µCi |

Dietary constituents†

| | |
|---|---|
| Protein | 19 g |
| Fat | 14 g |
| Carbohydrate (starch) | 108 g (97 g) |

•Boiled for seven minutes in 1 liter of water.
† Determined by adding food-table contents of each item

**Table 2. Results in Five Normal Subjects on Days of Placebo and Starch-Blocker Tests.**

| | Placebo Test Day | | | Starch-Blocker test Day | | |
|---|---|---|---|---|---|---|
| | DUPLICATE TEST MEAL* | RECTAL EFFLUENT | MARKER RECOVERY | DUPLICATE TEST MEAL | RECTAL EFFLUENT | MARKER RECOVERY |
| | *kcal* | *kcal* | *%* | *kcal* | *kcal* | *%* |
| 1 | 664 | 81 | 97.8 | 665 | 76 | 96.6 |
| 2 | 675 | 84 | 95.2 | 672 | 84 | 98.3 |
| 3 | 682 | 80 | 97.4 | 681 | 73 | 94.4 |
| 4 | 686 | 67 | 95.5 | 675 | 75 | 103.6 |
| 5 | 676 | 89 | 96.3 | 687 | 83 | 106.9 |
| Means | 677 | 80 | 96.4 | 676 | 78 | 100 |
| ±S.E.M. | ±4 | ±4 | ±0.5 | ±4 | ±2 | ±2 |

*Does not include calories contained in three placebo tablets (each tablet, 1.2±0.1 kcal) or in three Carbo-Lite tablets (each tablet, 2.8±0.1 kcal) that were ingested with each test meal.