course EPIB 607 1980-1990, 2001

SEE PAGES 13&14 FOR MANY WRONG WAYS TO DESCRIBE A CONFIDENCE INTERVAL

SEE PAGES 15-17 FOR MANY WRONG WAYS TO DESCRIBE A P-VALUE

Introduction to Inference*

Lady tasting tea - page 7

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 M&M]

- 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

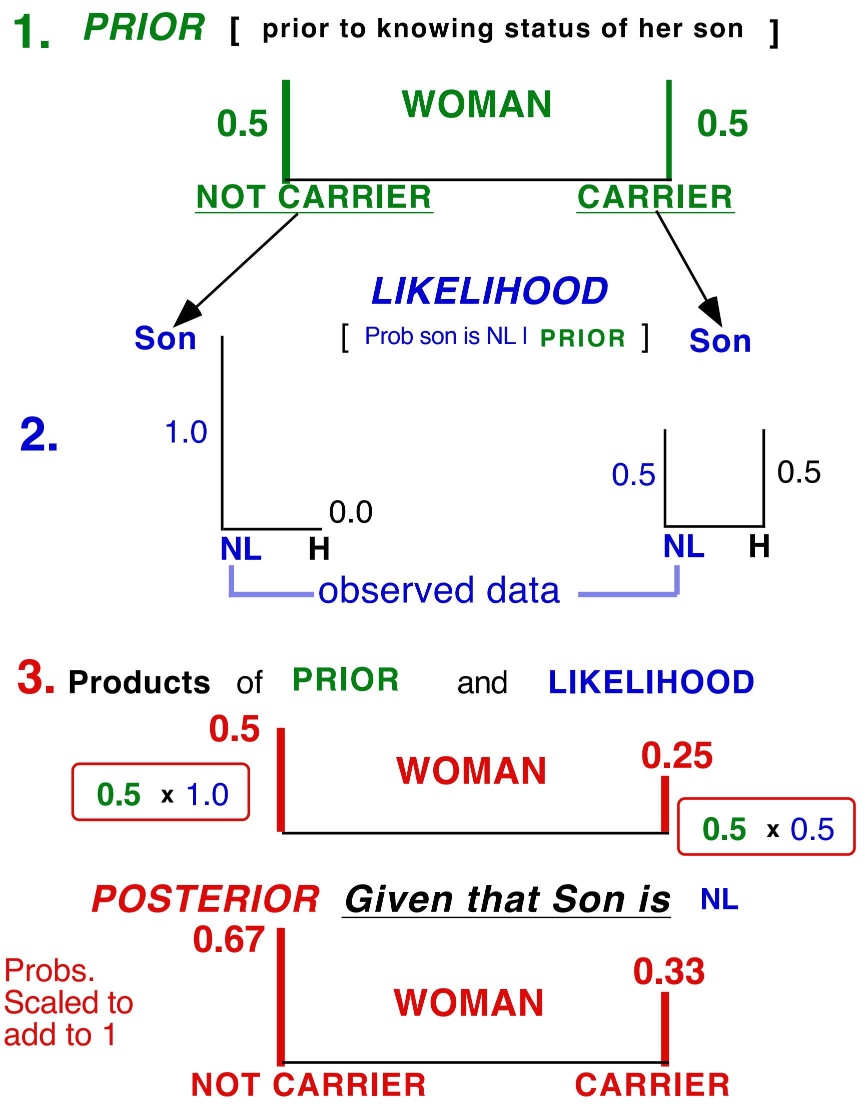
Unlike Bayesian inference, there is no quantified pre-test or pre-data "impression"; the ultimate statements are about data, conditional on an assumed null or other hypothesis.

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.

Bayes Theorem : Haemophilia

Brother has haemophilia => Probability (**WOMAN** is Carrier) = 0.5
New Data: Her **Son** is Normal (NL) .
Update: Prob[Woman is Carrier, given her son is NL] = ??



1. ***PRIOR*** **[prior to knowing status of her son]**

WOMAN
0.5 NOT CARRIER
0.5 CARRIER

LIKELIHOOD
[Prob son is NL | PRIOR]

2.
Son: 1.0 NL, 0.0 H
Son: 0.5 NL, 0.5 H
observed data

3. Products of PRIOR and LIKELIHOOD

WOMAN
0.5 (0.5 x 1.0)
0.25 (0.5 x 0.5)

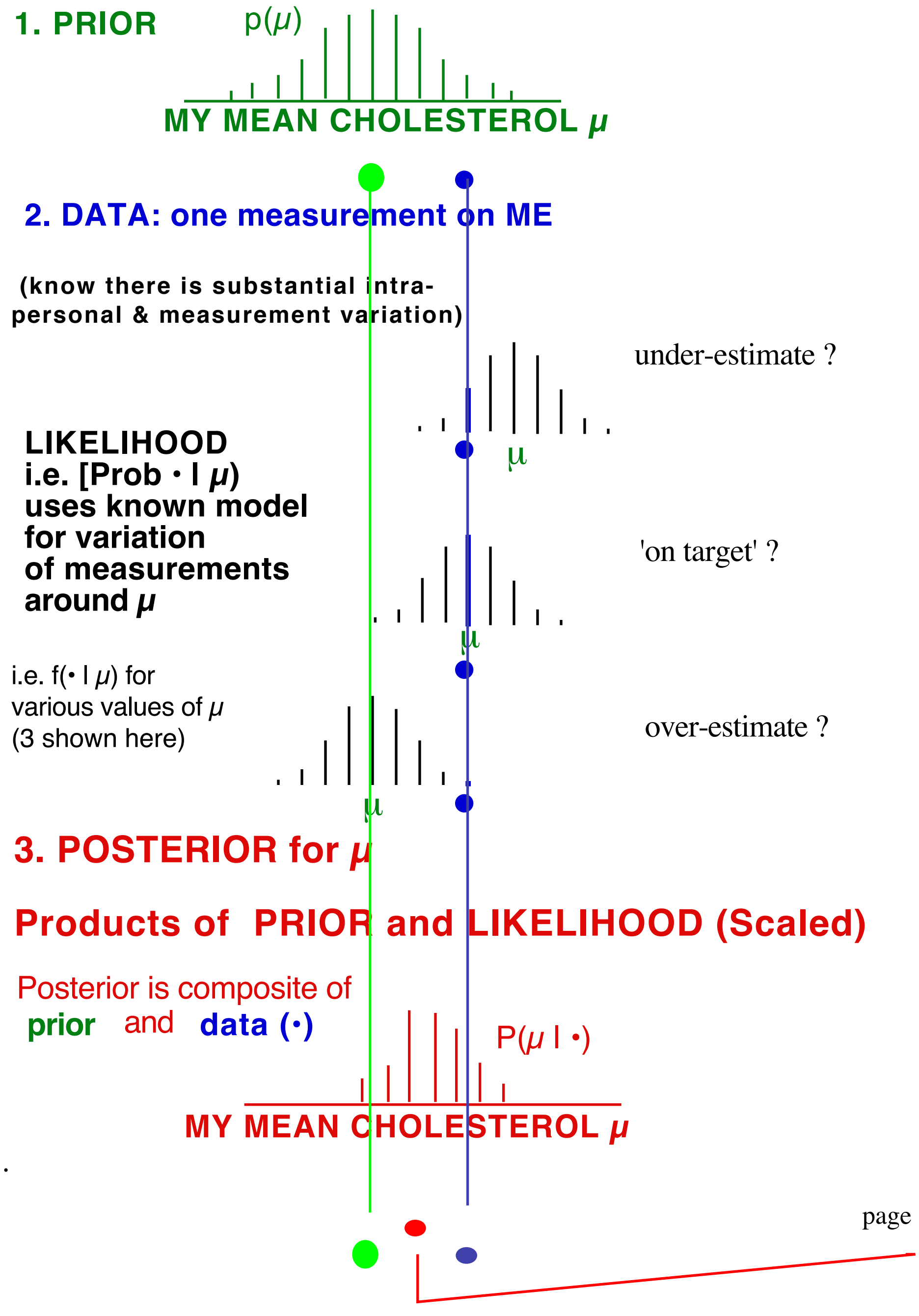
POSTERIOR ***Given that Son is*** **NL**

Probs. Scaled to add to 1

WOMAN
0.67 NOT CARRIER
0.33 CARRIER

Bayesian Inference for a quantitative parameter

E.g. Interpretation of a measurement that is subject to intra-personal (incl. measurement) variation. Say we know the pattern of inter-personal and intra-personal variation. Adapted from Irwig (JAMA 266(12):1678-85, 1991)



1. PRIOR

2. DATA: one measurement on ME

(know there is substantial intra-personal & measurement variation)

under-estimate ?

'on target' ?

over-estimate ?

LIKELIHOOD i.e. [Prob • | μ) uses known model for variation of measurements around μ

i.e. f(• | μ) for various values of μ (3 shown here)

3. POSTERIOR for μ

Products of PRIOR and LIKELIHOOD (Scaled)

Posterior is composite of **prior** and **data (•)**

Bayesian Inference ... in general

- Interest in a parameter θ .
- Have prior information concerning θ in form of a prior distribution with probability density function p(θ). [to distinguish, might use lower case p for prior]
- Obtain new data x (x could be a single piece of information or more complex)

Likelihood of the data for any contemplated value θ is given by

$$L[\, x \mid \theta \,] = \text{prob}[\, x \mid \theta \,]$$

Posterior probability for θ, GIVEN x, is calculated as:

$$P(\, \theta \mid x \,) = \frac{L[\, x \mid \theta \,]\ p(\theta)}{\int L[\, x \mid \theta \,]\, p(\theta)\, d\theta}$$

[To distinguish, might use UPPER CASE P for POSTERIOR]. The denominator is a summation/integration (the ∫ sign) over range of θ and serves as a scaling factor that makes P(θ) sum to 1.

In Bayesian approach, post-data statements of uncertainty about θ are made directly from the function P(θ | x) .

Notice that posterior distribution is not centered on the one measurement, but on a value less extreme, i.e., on a value that is a compromise between the prior and the data

Re: **Previous 2 examples of Bayesian inference**

Haemophilia example

θ = possible status of woman:

θ = "Carrier" or "Not a carrier"

$p(\theta = \text{Carrier}) = 0.5$

$p(\theta = \text{Not a Carrier}) = 0.5$

x = status of son

L[x=Normal | Woman is Carrier] = 0.5

L[x=Normal | Woman is Not Carrier] = 1

$P(\theta = \text{Carrier} \mid x=\text{Normal})$

$$= \frac{L[x=N \mid \theta = C]\ p[\theta = C]}{L[x=N \mid \theta = C]\ p[\theta = C] + L[x=N \mid \theta = \text{Not } C]\ p[\theta = \text{Not } C]}$$

[equation for predictive value of a diagnostic test with binary results]

Cholesterol example

θ = my mean cholesterol level

θ = ?? i.e. $p[\theta]$ = ?

In absence of any knowledge about me, would have to take as a prior the distribution of mean cholesterol levels for population my age and sex

x = one cholesterol measurement on me

Assume that if a person's mean is θ, the variation around θ would be Gaussian with standard deviation σ_w. (Bayesian argument does not insist on Gaussian-ness). So...

L[X=x | my mean is θ] is obtained by evaluating height of Gaussian(θ,σ_w) curve at X = x

$$P[\theta \mid X = x] = \frac{L[\ X = x \mid \theta\ ]\ p[\theta]}{\int L[\ X = x \mid \theta\ ]\ p[\theta]\ d\theta}$$

If intra-individual variation is Gaussian with SD σ_w and the prior is Gaussian with mean μ and SD σ_b [b for between individuals], then the mean of the posterior distribution is a weighted average of μ and x, with weights inversely proportional to the squares of σ_w and σ_b respectively. So, the less the intra-individual and lab variation, the more the posterior is dominated by the measurement x on the individual --- and vice versa.

(Frequentist) Confidence Interval (CI) or Interval Estimate for a parameter θ

Formal definition:

A level 1 - α Confidence Interval for a parameter θ is given by two statistics

U_{pper} and L_{ower}

such that when θ is the true value of the parameter,

$$\textbf{Prob} \; (\text{L}_{\text{ower}} \le \theta \le \text{U}_{\text{pper}}) = 1 - \alpha$$

α	**1 - α**
0.05	0.95
0.01	0.99

- CI is a **statistic**: a quantity calculated from a sample
- usually use α = 0.01 or 0.05 or 0.10, so that the "level of confidence", 1 - α, is 99% or 95% or 90%. We will also use "α" 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 θ in a fraction 1 - α of applications**. If we were not fussy about the semantics, we might say that any particular CI has a 1-α chance of covering θ.
- 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.

Meaning of a CI is often "massacred" in the telling... users slip into Bayesian-like statements without realizing it.

STATISTICAL CORRECTNESS

The Frequentist CI (statistic) is the SUBJECT of the sentence (speak of long-run behaviour of CI's).

In Bayesian parlance, the parameter is the SUBJECT of the sentence [speak of where parameter values lie].

Book "Statistical Inference" by Oakes is good here..

Large-sample CI's

Many large-sample CI's are of the form

$$\hat{\theta} \pm \text{multiple of SE}(\hat{\theta}) \quad \text{or} \quad f^{-1} [\; f\{\hat{\theta}\} \pm \text{multiple of SE}(f\{\hat{\theta}\} \;] ,$$

where f is some function of $\hat{\theta}$ which has close to a Gaussian distribution, and f^{-1} is the inverse function

(other motivation is variance stabilization; cf A&B ch11)

examples of the latter are:

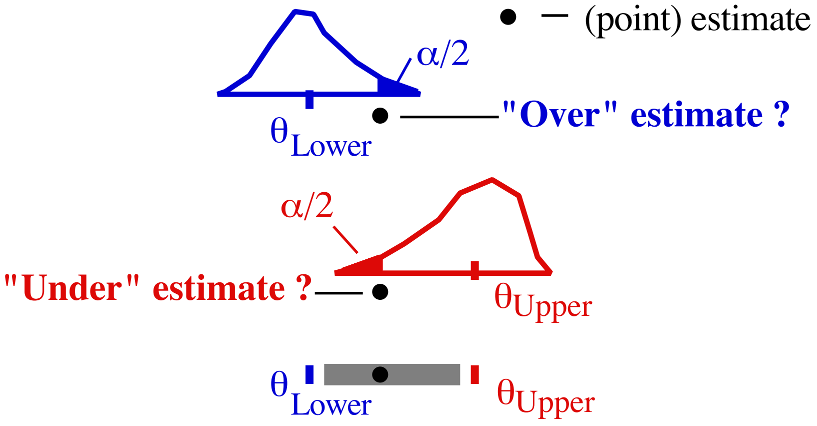
θ = odds ratio

f = *ln* ;	f^{-1} = exp

θ = proportion π

f = arcsine ;	f^{-1} = reverse
f = logit ;	f^{-1} = exp(•)/[1+exp(•)]

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



NB: shapes of distributions may differ at the 2 limits and thus yield asymmetric limits: see e.g. CI for π , based on binomial.
Notice also the use of concept of tail area (p-value) to construct CI.

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

parameter	estimate	SD*(estimate)
θ	$\hat{\theta}$	$SD(\hat{\theta})$
μ_X	$\bar{x}$	$\frac{\sigma_X}{\sqrt{n}}$
$\mu_{\Delta X}$	$\bar{d}$	$\frac{\sigma_d}{\sqrt{n}}$
π	p	$\frac{\sqrt{\pi[1-\pi]}}{\sqrt{n}}$
$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_1$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
$\pi_1 - \pi_2$	$p_1 - p_2$	$\sqrt{\frac{\pi_1[1-\pi_1]}{n_1} + \frac{\pi_2[1-\pi_2]}{n_2}}$

The last two are of the form $\sqrt{SD_1^2 + SD_2^2}$

* **In practice**, measures of individual (unit) variation about θ {e.g. σ_X, $\sqrt{\pi[1-\pi]}$, ...} are replaced by estimates (e.g. s_x, $\sqrt{p[1-p]}$, ... } calculated from the sample data, and if sample sizes are small, adjustments are made to the "multiple" used in the multiple of SD(θ). To denote the "substitution" , some statisticians and texts (e.g., use the term "SE" rather than SD; others (e.g. Colton, Armitage and Berry), use the term SE for the SD of a statistic -- whether one 'plugs in' SD estimates or not. Notice that M&M delay using SE until p504 of Ch 7.

Semantics behind Confidence Interval (e.g.)

parameter	estimate	SD(estimate)
μ_X	$\bar{x}$	$\frac{\sigma_X}{\sqrt{n}}$

Probability is $1-\alpha$ that ...

$\boxed{\bar{x}}$ falls within $\substack{z_{\alpha/2}\\ t_{\alpha/2}}SD(\bar{x})$ of $\boxed{\mu_X}$ (see footnote 1)

Probability is $1-\alpha$ that ..

$\boxed{\mu_X}$ **"falls"** within $\substack{z_{\alpha/2}\\ t_{\alpha/2}}SD(\bar{x})$ of $\boxed{\bar{x}}$ (see footnote 2)

$$Pr\left\{ \boxed{\mu_X} - \substack{z_{\alpha/2}\\ t_{\alpha/2}}SD(\bar{x}) \le \boxed{\bar{x}} \le \boxed{\mu_X} + \substack{z_{\alpha/2}\\ t_{\alpha/2}}SD(\bar{x}) \right\} = 1-\alpha$$

$$Pr\left\{ - \substack{z_{\alpha/2}\\ t_{\alpha/2}}SD(\bar{x}) \le \boxed{\bar{x}} - \boxed{\mu_X} \le + \substack{z_{\alpha/2}\\ t_{\alpha/2}}SD(\bar{x}) \right\} = 1-\alpha$$

$$Pr\left\{ + \substack{z_{\alpha/2}\\ t_{\alpha/2}}SD(\bar{x}) \ge \boxed{\mu_X} - \boxed{\bar{x}} \ge - \substack{z_{\alpha/2}\\ t_{\alpha/2}}SD(\bar{x}) \right\} = 1-\alpha$$

$$Pr\left\{ \boxed{\bar{x}} + \substack{z_{\alpha/2}\\ t_{\alpha/2}}SD(\bar{x}) \ge \boxed{\mu_X} \ge \boxed{\bar{x}} - \substack{z_{\alpha/2}\\ t_{\alpha/2}}SD(\bar{x}) \right\} = 1-\alpha$$

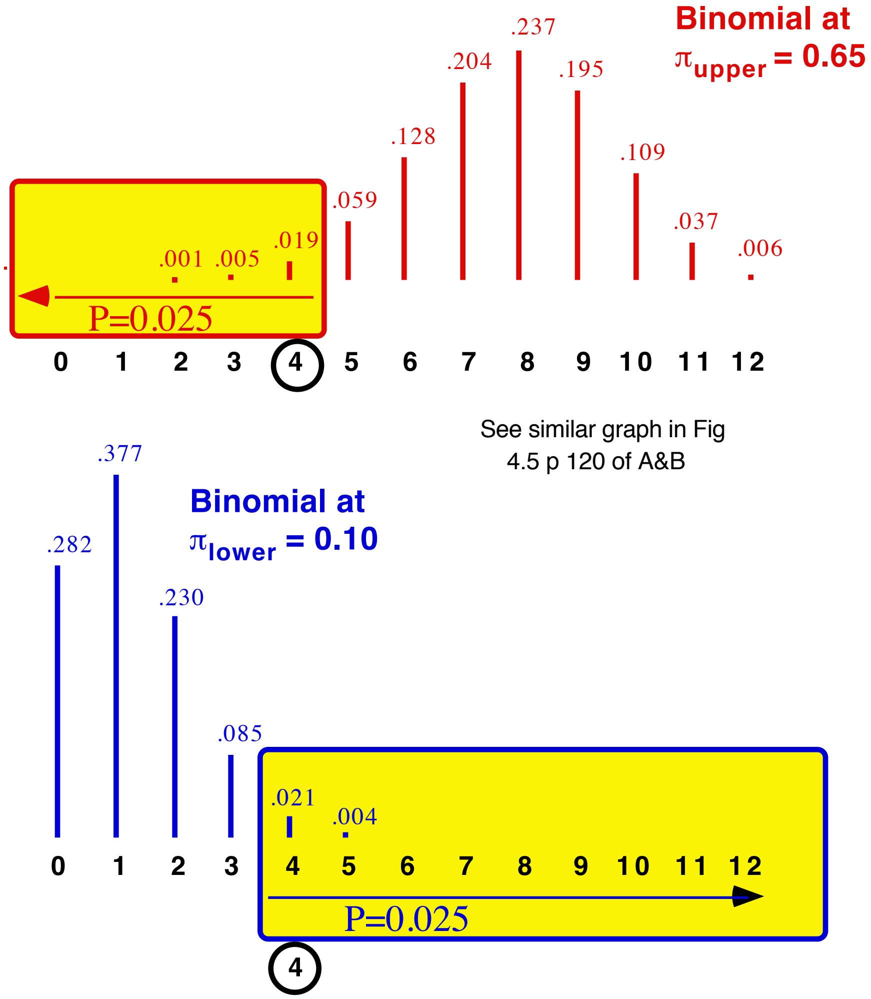
$$Pr\left\{ \boxed{\bar{x}} - \substack{z_{\alpha/2}\\ t_{\alpha/2}}SD(\bar{x}) \le \boxed{\mu_X} \le \boxed{\bar{x}} + \substack{z_{\alpha/2}\\ t_{\alpha/2}}SD(\bar{x}) \right\} = 1-\alpha$$

[1] This is technically correct, because the subject of the sentence is the statistic xbar. Statement is about behaviour of xbar.

[2] This is technically incorrect, because the subject of the sentence is the parameter. μ_X. In the Bayesian approach the parameter is the subject of sentence. In special case of "prior ignorance" [e.g. if had just arrived from Mars], the incorrectly stated frequentist CI is close to a Bayesian statement based on the posterior density function p(μ_X l data).

Technically, we are not allowed to "switch" from one to the other [it is not like saying "because St Lambert is 5 Km from Montreal, thus Montreal is 5Km from St Lambert".] **Here μ_X is regarded as a fixed (but unknowable) constant; it doesn't "fall" or "vary around" any particular spot; in contrast you can think of the statistic xbar "falling" or "varying around" the fixed μ_X.**

Clopper-Pearson 95% CI for π based on observed proportion 4/12



See similar graph in Fig 4.5 p 120 of A&B

Notice that Prob[4] is counted twice, once in each tail .

The use of CI's based on Mid-P values, where Prob[4] is counted only once, is discussed in Miettinen's Theoretical Epidemiology and in §4.7 of Armitage and Berry's text.

Constructing a Confidence Interval for μ

Assumptions & steps (simplified for didactic purposes)

(1) Assume (for now) that we know the sd (σ) of the Y values in the population. If we don't know it, suppose we take a "conservative" or larger than necessary estimate of σ.

(2) Assume also that either

(a) the Y values are normally distributed or

(b) (if not) n is large enough so that the Central Limit Theorem guarantees that the distribution of possible $\bar{y}$'s is well enough approximated by a Gaussian distribution.

(3) Choose the degree of confidence (say 90%).

(4) From a table of the Gaussian Distribution, find the z value such that 90% of the distribution is between -z and +z. Some 5% of the Gaussian distribution is above, or to the right of, $z = 1.645$ and a corresponding 5% is below, or to the left of, $z = -1.645$.

(5) Compute the interval $\bar{y} \pm 1.645\ SD(\bar{y})$, i.e., $\bar{y} \pm 1.645\ \sigma/\sqrt{n}$

Warning: Before observing $\bar{y}$, we can say that there is a 90% probability that the $\bar{y}$ we are about to observe will be within ±1.645 SD($\bar{y}$)'s of μ . But, after observing $\bar{y}$, we cannot reverse this statement and say that there is a 90% probability that μ is in the calculated interval.

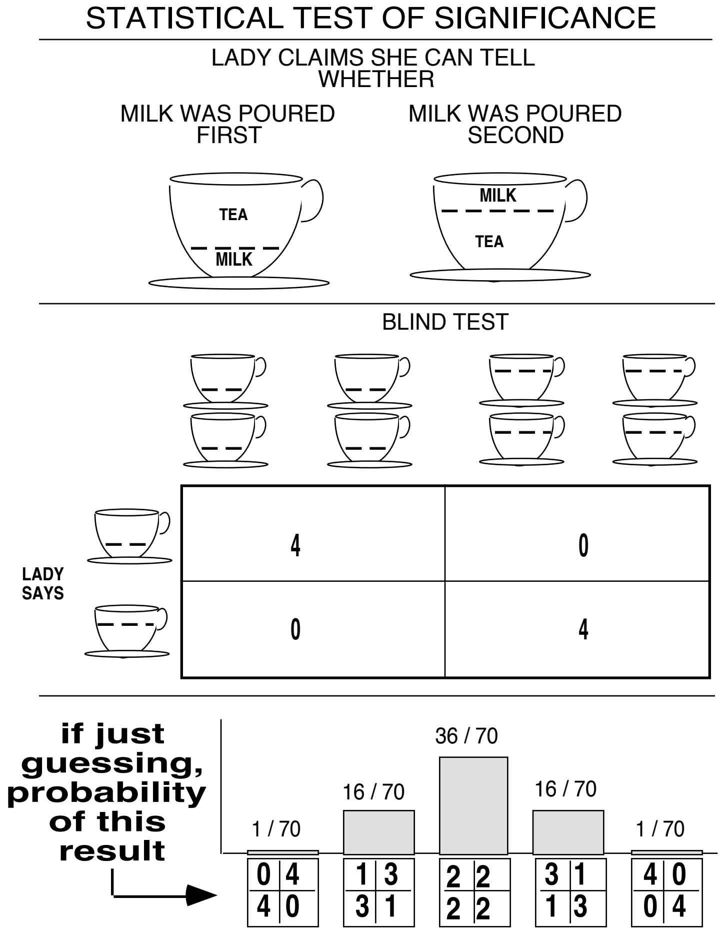
We **can** say that we are **USING A PROCEDURE IN WHICH SIMILARLY CONSTRUCTED CI's "COVER" THE CORRECT VALUE OF THE PARAMETER (μ in our example) 90% OF THE TIME.** The term "confidence" is a statistical legalism to indicate this semantic difference.

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)



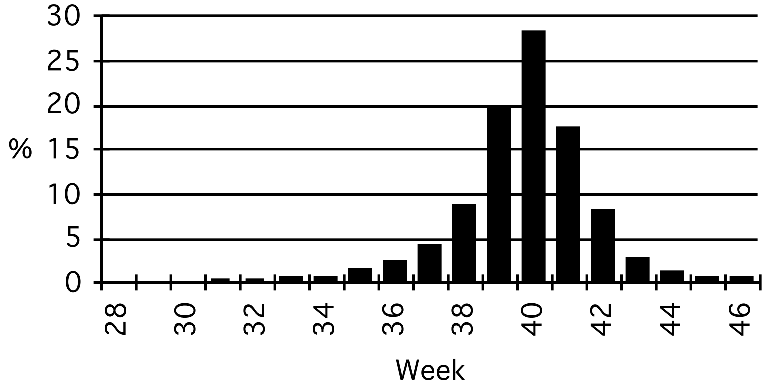
Example 2

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 μ)

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= α 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 - α	α
	H_a	β	1 - β

The quantities (1 - β) and (1 - α) are the "sensitivity (power)" and "specificity" of the statistical test. Statisticians usually speak instead of the complements of these probabilities, the false positive fraction (α) and the false negative fraction (β) 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-α].

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 + I H_a) and specificity = Prob(test - I H_0) but rather, the other way round, of Prob(H_a I test +) and Prob(H_0 I 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 I H) into Prob(H I 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 $\alpha = 0.05$, $\beta = 0.2$.

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.

Below are some previous students' answers to questions from 2nd Edition of Moore and McCabe Chapter 6. For each answer, say whether the statement/explanation is correct and why.

Question 6.2

A New York Times poll on women's issues interviewed 1025 women and 472 men randomly selected from the United States excluding Alaska and Hawaii. The poll found that 47% of the women said they do not get enough time for themselves.

(a) The poll announced a margin of error of ±3 percentage points for 95% confidence in conclusions about women. Explain to someone who knows no statistics why we can't just say that 47% of all adult women do not get enough time for themselves.

(b) Then explain clearly what "95% confidence" means.

(c) The margin of error for results concerning men was ± 4 percentage points. Why is this larger than the margin of error for women?

1	True value will be between 43 & 50% in 95% of repeated samples of same size.	• **No**. Estimate will be between μ – margin & μ + margin in 95% of samples.
2	Pollsters say their survey method has 95% chance of producing a range of percentages that includes π.	• **Good**. Emphasize average performance in repeated applications of method.
3	If this same poll were repeated many times, then 95 of every 100 such polls would give a range that included 47%.	• **No!**. See 1.
4	You're pretty sure that the true percentage π is within 3% of 47% . "95% confidence" means that 95% of the time, a random poll of this size will produce results within 3% of π.	• Bayesians would object (and rightly so!) to this use of the "true parameter" as the subject of the sentence. They would insist you use the statistic as the subject of the sentence and the parameter as object.
5	If took 100 different samples, in 95% of cases, the sample proportion will be between 44% and 50%.	• **NO**! The sample proportion will be between truth – 3% & truth + 3% in 95% of them.
6	With this one sample taken, we are sure 95 times out of 100 that 41-53% of the women surveyed do not get enough time for themselves.	• **NO**. 95/100 times the estimate will be within 3% of π, i.e., estimate will be in interval π – margin to π + margin. Method used gives correct results 95% of time.
7	In 95 of 100 comparable polls, expect 44 - 50% of women will give the same answer.	• **NO**. Same answer? as what?
	Given a parameter, we are 95% sure that the mean of this parameter falls in a certain interval.	**Not** given a parameter (ever) . If we were, wouldn't need this course! Mean of a parameter makes no sense in frequentist inference.
8	"using the poll procedure in which the CI or rather the true percentage is within +/- 3, you cover the true percentage 95% of times it is applied.	• A bit **muddled**... but "correct in 95% of applications" is accurate.
9	Confident that a poll (such) as this one would have measured correctly that the true proportion lies between in 95% .	• **???** [I have trouble parsing this!] In 95% of applications/uses, polls like these come within ± 3% of truth.
10	95% chance that the info is correct for between 44 and 50% of women.	• **???** 95% confidence in the procedure that produced the interval 44-50
11	95% confidence -> 95% of time the proportion given is the good proportion (if we interviewed other groups).	• "Correct in 95% of applications" **Good** to connect the 95% with the long run, not specifically with this one estimate. Always ask yourself: what do I mean by "95% of the time" ? If you substitute "applications" for "time", it becomes clearer.
12	It means that 47% give or take 3% is an accurate estimate of the population mean 19 times out of 20 such samplings.	• **???** 95% of applications of CI give correct answer. How can the same interval 47%±3 be accurate in 19 but not in the other 1?
Q6.4	"This result is trustworthy 19 times out of 20"	• **???** "this" result: Cf. the distinction between "my operation is successful 19 times out of 20 … " and "operations like the one to be done on me are successful 19 times out of 20"
	95% of all samples we could select would give intervals between 8669 and 8811.	• Surely **not**!

Question 6.18

The Gallup Poll asked 1571 adults what they considered to be the most serious problem facing the nation's public schools; 30% said drugs. This sample percent is an estimate of the percent of all adults who think that drugs are the schools' most serious problem. The news article reporting the poll result adds, "The poll has a margin of error -- the measure of its statistical accuracy -- of three percentage points in either direction; aside from this imprecision inherent in using a sample to represent the whole, such practical factors as the wording of questions can affect how closely a poll reflects the opinion of the public in general" (The New York Times, August 31, 1987).

The Gallup Poll uses a complex multistage sample design, but the sample percent has approximately a normal distribution. Moreover, it is standard practice to announce the margin of error for a 95% confidence interval unless a different confidence level is stated.

a **The announced poll result was 30%±3%. Can we be certain that the true population percent falls in this interval?**

b **Explain to someone who knows no statistics what the announced result 30%±3% means.**

c This confidence interval has the same form we have met earlier:

estimate ± $Z^*\sigma_{\text{estimate}}$

(Actually s is estimated from the data, but we ignore this for now.)

What is the standard deviation σ_{estimate} of the estimated percent?

d Does the announced margin of error include errors due to practical problems such as undercoverage and nonresponse?

1	This means that the population result will be between 27% and 33% 19/20 times.	• **NO! Population result is wherever it is and it doesn't move**. Think of it as if it were the speed of light.
2	95% of the time the actual truth will be between 30 ± 3% and 5% it will be false.	• It either is or it isn't … the truth doesn't vary over samplings.
3	If this poll were repeated very many times, then 95 of 100 intervals would include 30% .	• **NO**. 95% of polls give answer within 3% of truth, NOT within 3% of the mean in this sample.
4	Interval of true values ranges b/w 27% + 33%.	• **???** There is only one true value. AND, it isn't 'going' or 'ranging' or 'moving' anywhere!
5	Confident that in repeated samples estimate would fall in this range 95/100 times.	• **NO**. Estimate falls within 3% of π in 95% of applications
6	95% of intervals will contain true parameter value and 5% will not. Cannot know whether result of applying a CI to a particular set of data is correct.	• **GOOD**. Say "Cannot know whether CI derived from a particular set of data is correct." Know about behaviour of procedure! If not from Mars, (i.e. if you use past info) might be able to bet more intelligently on whether it does or not.
7	In 1/20 times, the question will yield answers that do not fall into this interval.	• **No**. In 5% of applications, estimate will be more than 3% away from true answer. See 1,2,3 above.
8	This type of poll will give an estimate of 27 to 33% 19 times out of 20 times.	• **NO**. Won't give 27 ± 3 19/20 times. Estimate will be within ± 3 of truth in 19/20 applications
9	5% risk that μ is not in this interval.	• ??? If an after the fact statement, somewhat inaccurate.
10	95 out 100 times when doing the calculations the result 27-33% would appear.	• **No it wouldn't**. See 1,2,3,7.
11	95% prob computed interval will cover parameter.	• Accurate if viewed as a prediction.
12	The true popl'n mean will fall within the interval 27-33 in 95% of samples drawn.	• **NO**. True popl'n mean will not "fall" anywhere. It's a fixed, unknowable constant. Estimates may fall around it.

Question 6.22

In each of the following situations, a significance test for a population mean μ is called for. **State the null hypothesis H_0 and the alternative hypothesis H_a in each case.**

a Experiments on learning in animals sometimes measure how long it takes a mouse to find its way through a maze. The mean time is 18 seconds for one particular maze. A researcher thinks that a loud noise will cause the mice to complete the maze faster. She measures how long each of 10 mice takes with a noise as stimulus.

a The examinations in a large accounting class are scaled after grading so that the mean score is 50. a self-confident teaching assistant thinks that his students this semester can be considered a sample from the population of all students he might teach, so he compares their mean score with 50.

c A university gives credit in French language courses to students who pass a placement test. The language department wants to know if students who get credit in this way differ in their understanding of spoken French from students who actually take the French courses. Some faculty think the students who test out of the courses are better, but others argue that they are weaker in oral comprehension. Experience has shown that the mean score of students in the courses on a standard listening test is 24. The language department gives the same listening test to a sample of 40 students who passed the credit examination to see if their performance is different.

1	Ho: Is there good evidence against the claim that $\pi_{male} > \pi_{female}$ Ha: Fail to give evidence against the claim that $\pi_{male} > \pi_{female}$.	• **NO**. Hypotheses do not include statements about data or evidence. . This student mixed parameters and statistics/data ... Put Ho, Ha in terms of parameters π_{male} vs π_{female} only; H's have nothing to do with new data; Evidence has to do with p-values, data.
2	$\bar{x}$ = average. time of 10 mice w/ loud noise. Ho: mu - $\bar{x} = 0$ or mu = $\bar{x}$	• **NO**! Ho must be in terms of parameter(s). IT MUST NOT SPEAK OF DATA
3	Ho : a loud noise has no effect on the rapidity of the mouse to find its way through the maze.	• OK if being generic. but not if it makes a prediction about a specific mouse (sounds like this student was talking about a specific mouse. **H is about mean of a population, i.e. about mice (plural). It is not about the 10 mice in the study!**

Question 6.24

A randomized comparative experiment examined whether a calcium supplement in the diet reduces the blood pressure of healthy men. The subjects received either a calcium supplement or a placebo for 12 weeks. The statistical analysis was quite complex, but one conclusion was that "the calcium group had lower seated systolic blood pressure (P=0.008) compared with the placebo group." **Explain this conclusion, especially the P-value, as if you were speaking to a doctor who knows no statistics**. (From R.M. Lyle et al., "Blood pressure and metabolic effects of calcium supplement in normotensive white and black men," Journal of the American Medical Association, 257 (1987), pp. 1772-1776.)

1	The P-value is a probability: "P=0.008" means 0.8% . It is the probability, assuming the null hypothesis is true, that a sample (similar in size and characteristics as in the study) would have an average BP this far (or further) below the placebo group's average BP. In other words, if the null hypothesis is really true, what's the chance 2 group of subjects would have results this different or more different?	• Not bad!
2	Only a 0.008 chance of finding this difference by chance if, in the population there really was no difference between treatment and central groups.	• Good!
3	The p-value of .008 means that the probability of the observed results if there is, in fact, no difference between "calcium" and "placebo" groups is 8/1000 or 1/125.	• Good, but would change to "the observed results or results more extreme"

4	The p-value measures the probability or chance that the calcium supplement had no effect.	• **No**. First, Ho and Ha refer not just to the n subjects studied, but to all subjects like them. They should be stated in the <u>present</u> (or even future) tense. Second, the p-value is about data, under the null H. It is not about the credibility of Ho or Ha.
5	There is strong evidence that Ca supplement in the diet reduces the blood pressure of healthy men. The probability of this being wrong conclusion according to the procedure and data corrected is only 0.008 (i.e. 0.8%) .	• Stick to "official wording" .. IF Ca makes no Δ to average BP, chance of getting ... Notice the <u>illegal</u> attempt to **make the p-value into a predictive value** -- about as illegal as a statistician trying to interpret a medical test that gave a reading in the top percentile of the 'health' population -- without even seeing the patient!
6	Only 0.8% chance that the lower BP in Calcium group is lower than placebo due to chance.	• If Ca++ does nothing, then prob. of obtaining a result ≥ this extreme is ... Wording borders on the illegal.
7	The chance that the supplement made no change or raised the B/P is very slim.	• **NO**! p-value is a <u>conditional</u> statement, predicated (calculated on supposition that) Ca makes no difference to μ. Often stated in present tense. p-value is more 'after the data' in 'past-conditional' tense. Again, wording bordering on illegal.
8	There is 0.8% that this difference is due to chance alone and 99.2% chance that this difference is a true difference.	• **Not really**.. Just like the previous statements, this type of language is crossing over into Bayes-Speak.
9	There has been a significant reduction in the BP of the treated group... there's only a probability of 0,8% that this is due to chance alone.	• **NO**. Cannot talk about the cause... Can say "IF no other cause than chance, then prob. of getting ≥ a difference of this size is ...

<u>Question 6.32</u>

The level of calcium in the blood in healthy young adults varies with mean about 9.5 milligrams per deciliter and standard deviation about $\sigma = 0.4$. A clinic in rural Guatemala measures the blood calcium level of 180 healthy pregnant women at their first visit for prenatal care. The mean is $\bar{x} = 9.57$. Is this an indication that the mean calcium level in the population from which these women come differs from 9.5?

a **State H_0 and H_a.**

b **Carry out the test and give the P-value, assuming that $\sigma = 0.4$ in this population. Report your conclusion.**

c The 95% confidence interval for the mean calcium level μ in <u>this</u> population is obtained from the margin of error, namely $1.96 \times 0.4 / \sqrt{180} = 0.058$. i.e. as 9.57 ± 0.058 or 9. We are confident that μ lies quite close to 9.5. This illustrates the fact that a test based on a large sample (n=180 here) will often declare even a small deviation from H_0 to be statistically significant.

1	95% of the time the mean will be included in the interval of 9.512 to 9.628 and 5% will be missed.	• **No**. See 1,2,3 in Q6.18 above
2	Ho: There is no difference between sample area and the population area: H0: $\mu = \bar{x}$. Ha: There is a significant difference between the sample mean and the population area. **PS** A professor in the dept. of Math and Statistics questioned what we in Epi and Biostat are teaching, after he saw in a grant proposal submitted by one of our doctoral students (now a faculty member!) a statement of the type H0: $\mu = \bar{x}$. **Please do not give our critic any such ammunition! -- JH**	• **NO**. This is quite muddled. Unless one takes a census, there will **always** -- because of sampling variability -- be some <u>non-zero</u> difference between $\bar{x}$ and μ. The question posed is whether "mean calcium level (μ) in the population from which these women come differs from 9.5" **ALSO**: Must state H's in terms of PARAMETERS. Here there is one population. If two populations, identify them by subscripts e.g. Ho: $\mu_{area1} = \mu_{area2}$. "Significant" is used to interpret data. (and can be roughly paraphrased as "<u>evidence</u> that true parameter is non-zero". **Do not use "significant" when stating hypotheses.**

3	95% CI: $\mu \pm 1.96\ \sigma / \sqrt{180}$	**• NO** CI for μ is $\bar{x} \pm 1.96\ \sigma/\sqrt{180}$!!!! If we knew μ, we would say $\mu \pm 0$!! and we wouldn't need a statistics course!
4	Ho : mu = $\bar{x}$ = 9.57 Ha : mu ≠ $\bar{x}$ = 9.57	**• NO**. Cannot use sample values in hypotheses. Must use parameters.
5	μ differs from 9.5 and the probability that this difference is only due to chance is 2%.	• Correct to say that "we found evidence that μ differs from 9.5" In frequentist inference, can speak probabilistically only about data (such as $\bar{x}$). This miss-speak illustrates that we would indeed prefer to speak about μ rather than about the data in a sample. We should indeed start to adopt a formal Bayesian way of speaking, and not 'mix our metaphors' as we commonly do when we try to stay within the confines of the frequentist approach. What does this difference mean? Should not speak about the probability that this difference is only due to chance.
6	Ho : $\mu = 9.5$ Ha : $\mu \neq 9.5$	**• Correct**. Notice that Ho & Ha say nothing about what you will find in your data.
7	Q6.44: Ho: $\bar{x} = 0.5$; Ha: $\bar{x} > 0.5$	**• NO**. Must write H's in term of parameters!
8	There is a 0.96% probability that this difference is due to chance alone.	**• NO**. This shorthand is so short that it misleads. If want to keep it short, say something like "difference is larger than expected under sampling variation alone". Don't get into attribution of cause.

Rather than leave this column blank...

http://www.stat.psu.edu/~resources/Cartoons/

http://watson.hgen.pitt.edu/humor/