P-Values and Statistical 'Tests'

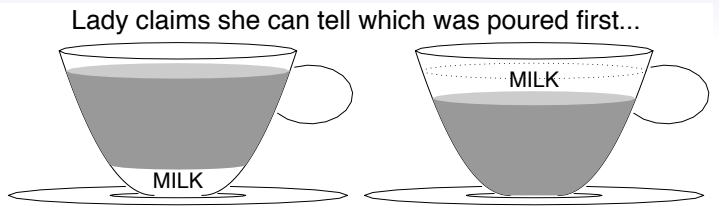
"P-Value"

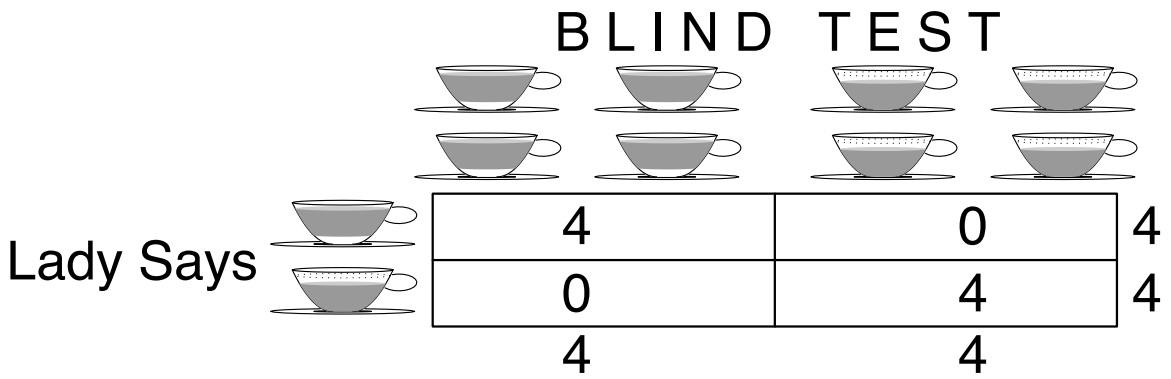
Def$^{n.}$ A **probability concerning the observed data**, calculated under a **Null Hypothesis** assumption, i.e., assuming that the only factor operating is sampling or measurement variation.

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.

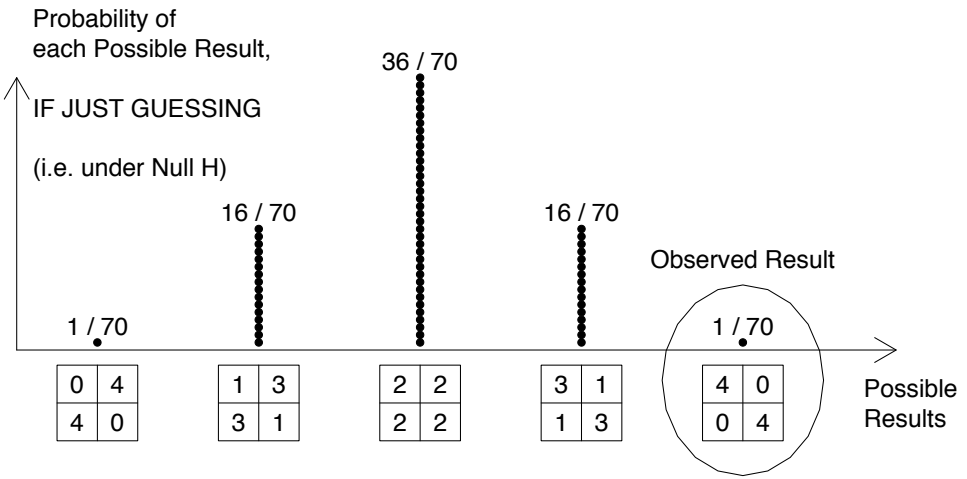
Blind test is equivalent to being asked to say **which 4** of the following 8 Gaelic words are the **correctly spelled** ones. You are told that **4 are correctly spelled & 4 are not**.

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

"Alternative" Hypothesis (H_{alt}): she can (can you think of another "H" ?).

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.



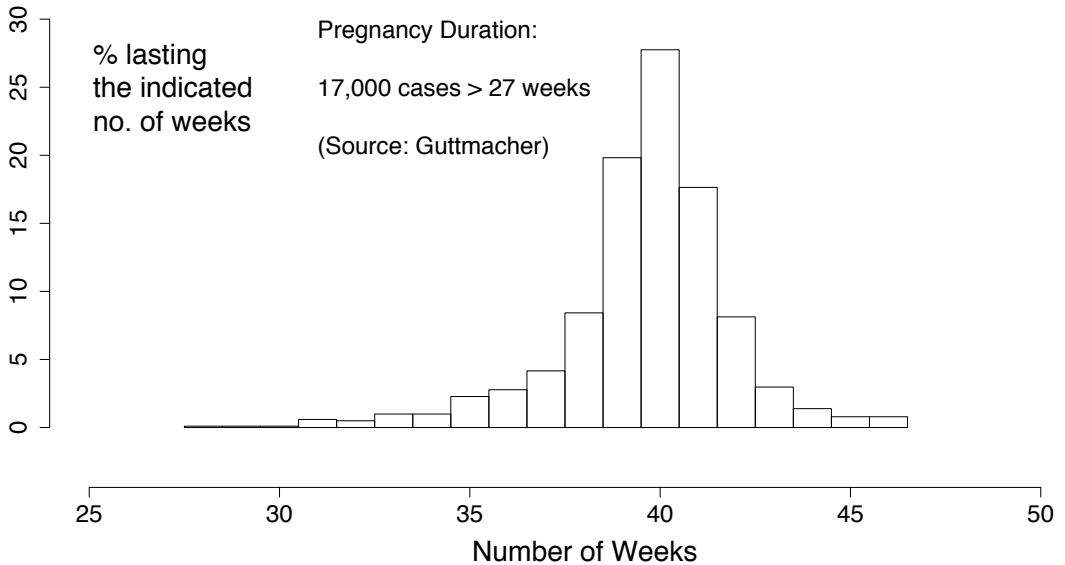
In this e.g., observed result is the most extreme, so

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

- 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.

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.

Don't be overly-impressed by P-values

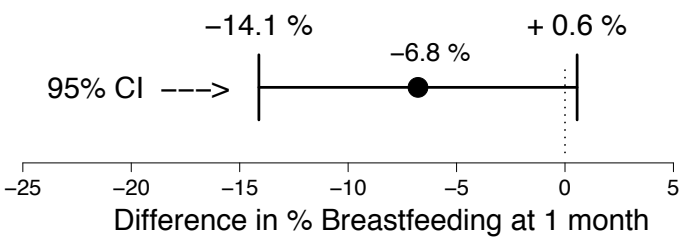
- P-values and 'significance tests' widely misunderstood and misused.
- Very large or very small *n*'s can influence what is / is not 'statistically significant.'
- Use CI's instead.
- *Pre study* power calculations (the chance that results will be 'statistically significant', as a function of the true 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}.

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)

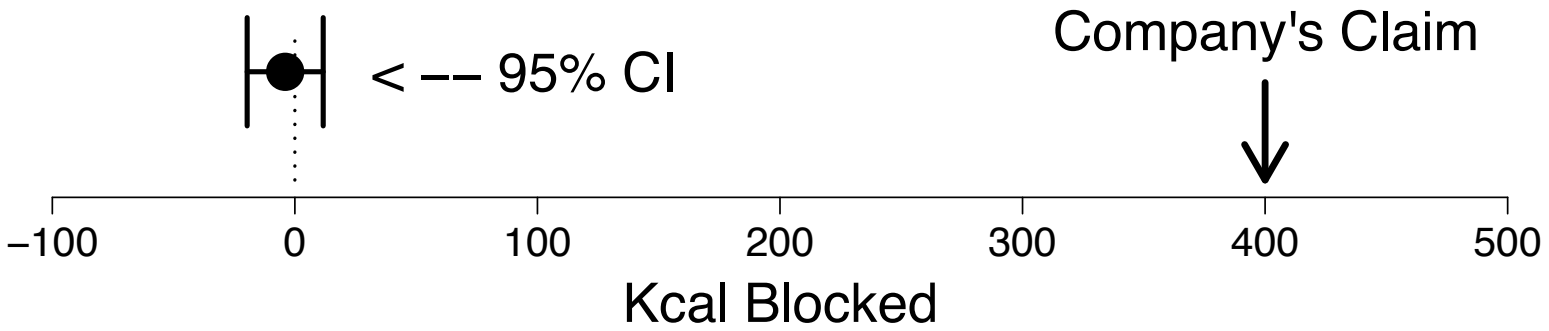
Do starch blockers really block calorie absorption?

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.

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 - 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 in a comparison.