012 - p-values

EPIB 607 - FALL 2020

Sahir Rai Bhatnagar
Department of Epidemiology, Biostatistics, and Occupational Health
McGill University

sahir.bhatnagar@mcgill.ca

slides compiled on October 7, 2020

p-values

Applications

Summary

p-values and statistical tests

Definition (p-value)

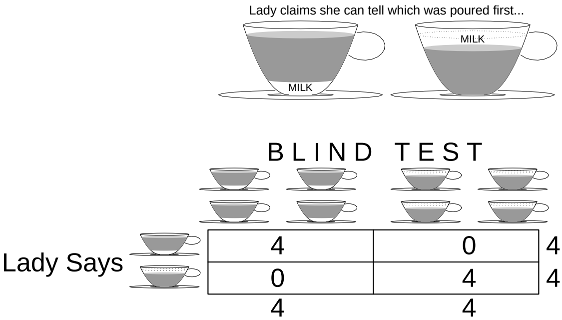
A ***probability concerning the observed data****, calculated under a* ***Null Hypothesis*** *assumption, i.e., assuming that the only factor operating is sampling or measurement variation.*

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.*

Caution: A p-value is NOT the probability that the null 'hypothesis' is true

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

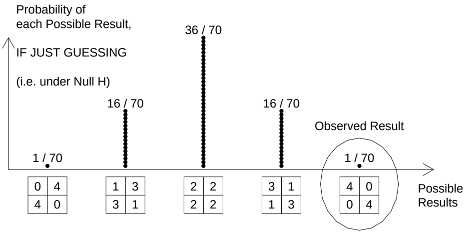


Null Hypothesis (H_{null}): she can not tell them apart, i.e., just guessing.

Alternative Hypothesis (H_{alt}): she can.

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 example, 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'.

p-value via the Normal (Gaussian) distribution.

- When judging extremeness of a sample mean or proportion (or difference between 2 sample means or proportions) calculated from an amount of information that is sufficient for the Central Limit Theorem to apply, one can use 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.

More about the p-value

- The p-value is a **probability concerning data**, **conditional on 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.*

$$\begin{aligned} p_{value} &= P(\text{this or more extreme data}|H_0) \\ &\neq P(H_0|\text{this or more extreme data}). \end{aligned}$$

- Statistical tests are often coded as statistically significant or not according to whether results are extreme or not with respect to a reference (null) distribution. 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 prosecutor's fallacy [1]

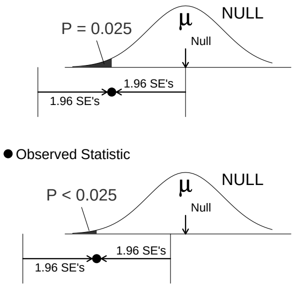
- Let's suppose a defendant has been accused of robbery
- The null hypothesis is that the defendant is innocent. Instructions to juries are quite explicit about this.
- **Prosecutor**: "If the defendant were innocent, wouldn't it be remarkable that the police found him at the scene of the crime with a bag full of money in his hand, a mask on his face, and a getaway car parked outside?" P(innocent | evidence)
- **Jury**: Considers the evidence in light of the presumption of innocence and judges whether the evidence against the defendant would be plausible if the defendant were in fact innocent. P(evidence | innocent)

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

The prosecutor's fallacy in a game of poker

- Imagine the judges were playing a game of poker with the Archbishop of Canterbury.
- If the Archbishop were to deal a royal flush on the first hand, one might suspect him of cheating.
- The probability of the Archbishop dealing a royal flush on any one hand, assuming he is an honest card player, is P(royal flush | innocent) = 1 in 70 000.
- But if the judges were asked whether the Archbishop was honest, given that he had just dealt a royal flush, they would be likely to quote a probability greater than 1 in 70 000 $\rightarrow$ P(innocent | royal flush).

(Intimate) Relationship between *p*-value and CI



- (Upper graph) If upper limit of 95% CI *just touches* null value, then the 2 sided *p*-value is 0.05 (or 1 sided *p*-value is 0.025).
- (Lower graph) If upper limit *excludes* null value, then the 2 sided *p*-value is less than 0.05 (or 1 sided *p*-value is less than 0.025).
- (Graph not shown) If 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 or 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}.

p-values

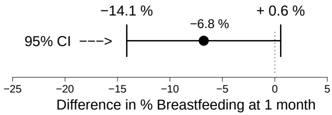
Applications

Summary

Do infant formula samples $\downarrow$ dur$^{n.}$ of breastfeeding?[2]

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	



[2] Bergevin Y, Dougherty C, Kramer MS. Lancet. 1983 1(8334):1148-51

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.
- (message from 1st author:) plan to have **enough statistical power**. his study had only 50% power to detect a difference of 10 percentage points)

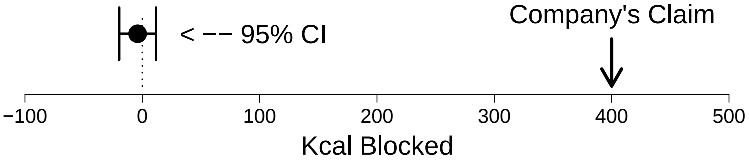
Do starch blockers really block calorie absorption?

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.

p-values

Applications

Summary

SUMMARY - 1

- 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 distribution (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 that there's 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

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

does not necessarily mean that

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

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.