

P-Values and Statistical 'Tests'

"P-Value"

Defⁿ. 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

Lady claims she can tell which was poured first...



BLIND TEST

| | | | | | |
|-----------|---|---|---|--|---|
| |  |  |  |  | |
| |  |  |  |  | |
| Lady Says |  | | | | |
| |  | | | | |
| | | 4 | | 0 | 4 |
| | | 0 | | 4 | 4 |
| | | 4 | | 4 | |

“Null Hypothesis” (H_{null}): she can not tell them apart.

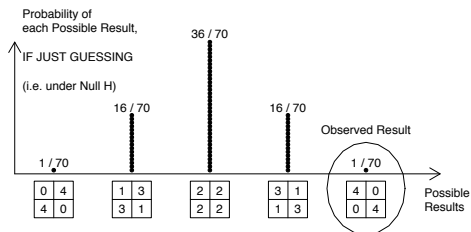
Blind test is equivalent to being asked to say **which 4** of the following 8 Gaelic words are the **correctly spelled** ones. You are told that **4 are correctly spelled & 4 are not**.

| | | | | | | | |
|-------|---------|---------|-------|-------|---------|------|--------|
| 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.



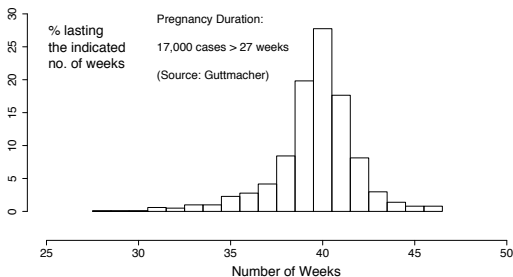
In this e.g., observed result is the most extreme, so

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

- 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^{tn}.
- 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) distⁿ. Medical tests also often coded as '+ve' or '-ve' according to whether results are extreme or not with respect to a ref. (healthy) distⁿ. 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*

$$P_{value} = Prob[data | H_0] \neq Prob[H_0 | data].$$

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} .