Computational Biology Group

D BSSE

ETH zürich

Introduction to Bayesian Statistics with R

2: Notes - p-values and confidence intervals

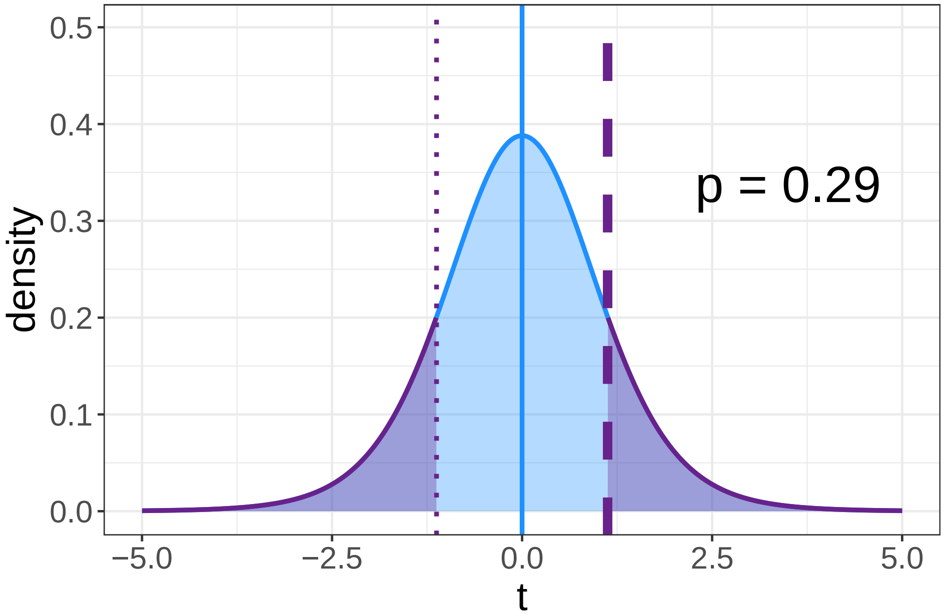
Jack Kuipers

18 February 2025

p-value recap

Hopefully you can remember from the last notes that when we worked out a *p*-value with the *t*-test, we placed a *t*-distribution on the origin corresponding to the null mean value of μ_0, and computed how much of that distribution was more extreme than our observed *t*-statistic

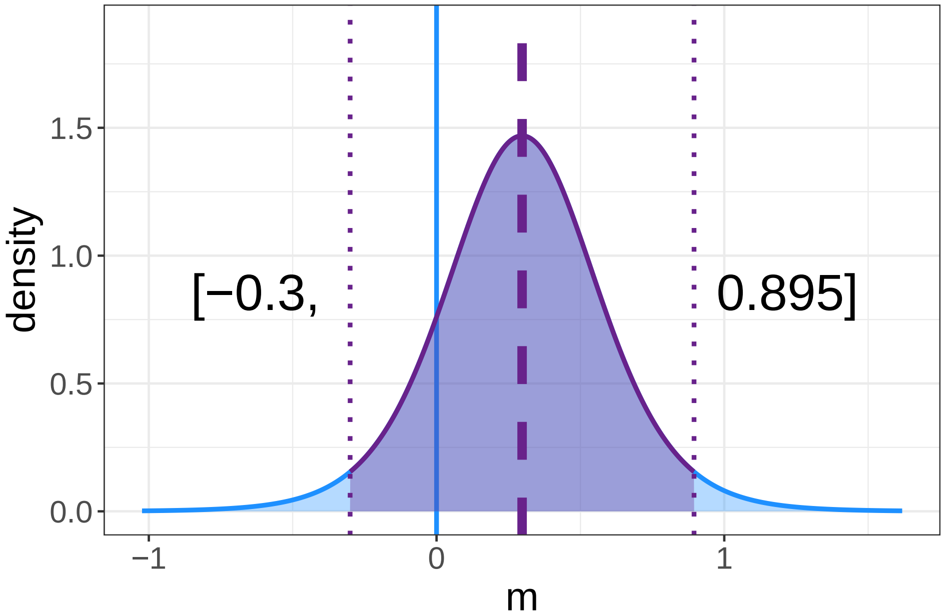
$$t = \frac{m-\mu_0}{s_m} = \sqrt{N}\frac{m-\mu_0}{s}$$



Confidence intervals

We can map back to the original coordinates and get the best estimate of the distribution of the sample mean itself, which turns out to be a rescaled *t*-distribution

$$m \sim \overline{x} + s_m\mathcal{T}_{N-1} = \overline{x} + \frac{s}{\sqrt{N}}\mathcal{T}_{N-1}$$



The **confidence interval** contains the central part of the distribution, usually we include the central 95%, and it is denoted by the lower and upper boundaries. In our case it would be [-0.3, 0.895], and can be extracted from the `t.test(...)` function with `$conf.int`:

```
t.test(local_samples)$conf.int # our sample is stored as local_samples
```

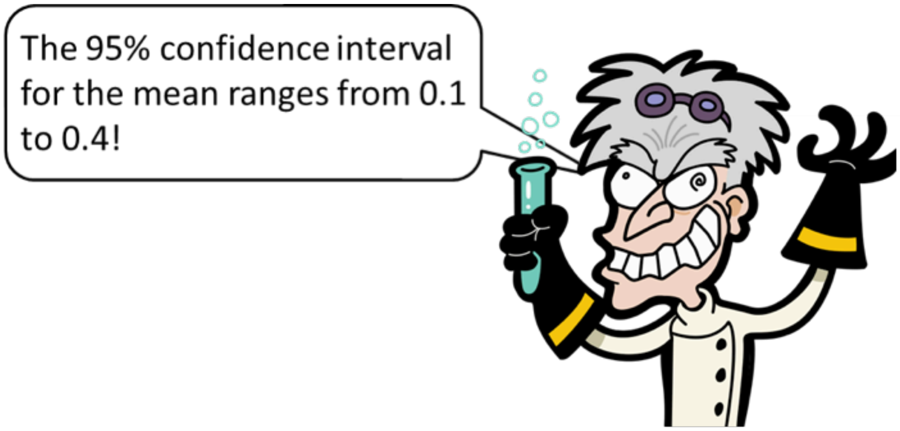
```
## [1] -0.3003473  0.8949409
## attr(,"conf.level")
## [1] 0.95
```

Confidence intervals and p-values are related, for example if the 95%-confidence interval excludes 0, we know our p-value is less than 5% (why?). They are just different numbers extracted from the same underlying distribution and process. One case we centre around 0 (p-values) and compute an area under the curve, the other case (confidence intervals) we centre on the sample mean and extract the borders of an area under the curve.

With this in mind, have a look at the following set of questions about confidence intervals from **Hoekstra et al (2014)** on the next page.

Confidence interval questionnaire

Professor Bumbledorf conducts an experiment, analyzes the data, and reports:


The 95% confidence interval for the mean ranges from 0.1 to 0.4!

Please mark each of the statements on the right as “true” or “false”. False means that the statement does not follow logically from Bumbledorf’s result. Also note that all, several, or none of the statements may be correct.

[1.] The probability that the true mean is greater than 0 is at least 95%.

[2.] The probability that the true mean equals 0 is smaller than 5%.

[3.] The “null hypothesis” that the true mean equals 0 is likely to be incorrect.

[4.] There is a 95% probability that the true mean lies between 0.1 and 0.4.

[5.] We can be 95% confident that the true mean lies between 0.1 and 0.4.

[6.] If we were to repeat the experiment over and over, then 95% of the time the true mean falls between 0.1 and 0.4.

Which statements are TRUE/FALSE?

If you’ve never seen these questions before, please have a good think about them and write your answers down somewhere (no peeking!)

Done, OK, let’s go through them on the next page.

Confidence interval diatribe

All the statements are false.

If that was clear to you, feel free to skip to the next section, otherwise to see why, remember that the sample mean (and its distribution) is a random variable. It depends on the unknown mean, but we cannot simply "invert". What we do know, as a mantra for confidence intervals is:

- the confidence interval $[L, U]$ was computed with a method that is successful in capturing the true mean μ in $(1 - \alpha)$ of cases

where we have L and U as the left and right endpoints of the interval, μ being the true mean value and α is the significance level. This can be written as

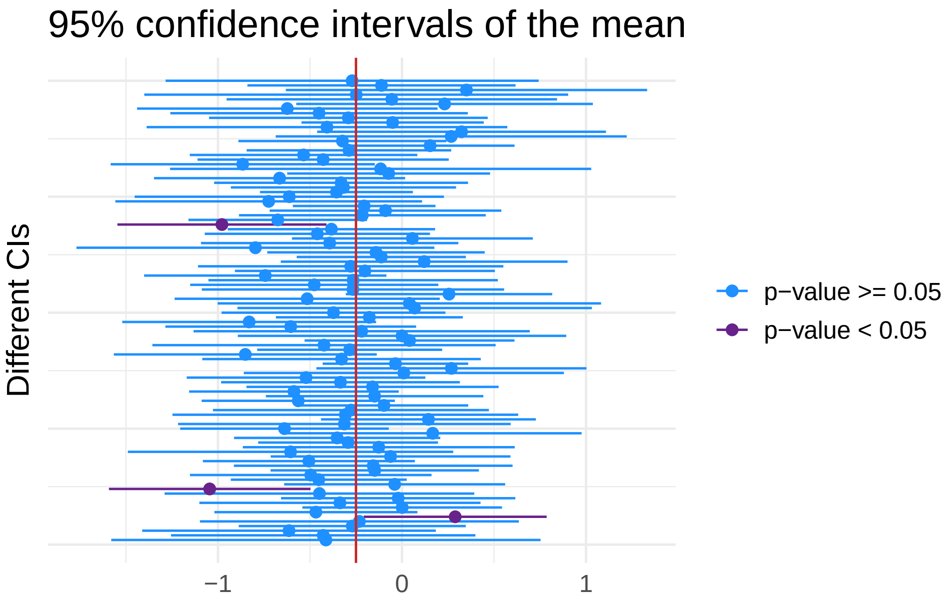
$$P(L \leq \mu \leq U) = 1 - \alpha$$

In that equation, which quantities are random and which are fixed? The true mean μ and the significance level α are fixed, but the confidence interval bounds $[L, U]$ are random since they are constructed from the random sample of data.

Now, keeping $\alpha = 0.05$, it is true that 95% of confidence intervals will contain the true mean on average. We can check this with a basic simulation in R. First we draw 100 times 10 samples from a normal with a standard deviation of 1 and a mean of $-\frac{1}{4}$ and compute confidence intervals, sample means and p-values.

```
conf_ints <- NULL
for (ii in 1:100) {
    sample <- rnorm(10, mean = -0.25)
    tt   <- t.test(sample, mu = -0.25)
    low  <- tt$conf.int[1]
    high <- tt$conf.int[2]
    pval <- tt$p.value
    me   <- tt$estimate
    conf_ints  <- rbind(conf_ints,
                  data.frame(id = ii, low = low, high = high, pval = pval, me = me))
}
```

Then we plot the results. Remembering that for two-sided tests, p-values are related to confidence intervals: if a p-value is significant at a significance level α, the $1 - \alpha$ confidence interval does *not* contain the null population parameter (here $\mu = -\frac{1}{4}$), we colour by significance:



Here 3 were significant so 97 percent of the confidence intervals included the true mean. For a larger number of repetitions, say a million:

```
n_reps <- 1e6
true_mean = -0.25
mean_inside <- rep(NA, n_reps)
for (ii in 1:n_reps) {
    tt <- t.test(rnorm(10, mean = true_mean), mu = true_mean)
    mean_inside[ii] <- tt$conf.int[1] < true_mean && tt$conf.int[2] > true_mean
}
round(100*mean(mean_inside), 2)
```

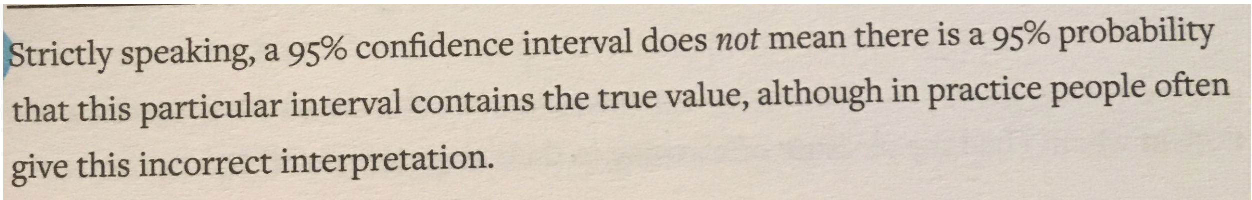
```
## [1] 94.98
```

we indeed get around 95% of the confidence intervals including the true mean of $-\frac{1}{4}$.

Why, then are statements 4 and 5 above false? The pedantic answer is that μ is fixed. Whether μ lies between 0.1 and 0.4 is deterministic and either true or false. Is it not a probabilistic question, and we cannot answer it with a probability. Worse, we only know the sample mean and do not know the true mean, so we cannot answer the question at all.

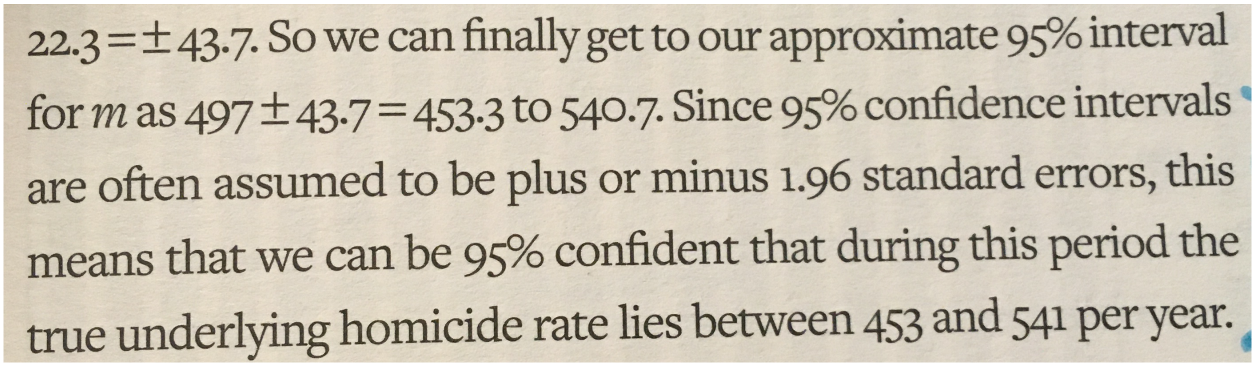
A weak analogy would be rolling a dice (or to be pedantic, a die). Before rolling it, the chance you get a 6 is $\frac{1}{6}$. Once rolled, you get a 1 and it either is a 6 or isn't (not in this case). Before running our experiment and getting the data, the probability the confidence interval containing the true mean is 95%. For the particular random realisation of the confidence interval (and for particular values of $[L, R]$) we cannot say.

Confidence intervals are notoriously misunderstood. Take for example the correct note in a book from a statistics professor (for the public understanding of risk):



Strictly speaking, a 95% confidence interval does *not* mean there is a 95% probability that this particular interval contains the true value, although in practice people often give this incorrect interpretation.

while their own incorrect description 7 pages later of the confidence interval computed from real data, ironically proves their point about misinterpretation:



22.3=±43.7. So we can finally get to our approximate 95% interval for *m* as 497±43.7=453.3 to 540.7. Since 95% confidence intervals are often assumed to be plus or minus 1.96 standard errors, this means that we can be 95% confident that during this period the true underlying homicide rate lies between 453 and 541 per year.

My complaining may seem like pedantic sophistry, but inadvertent misunderstandings of confidence intervals actually go to the very heart of Bayesian statistics.

Back to p-values

Remember that if the 95% confidence interval excludes 0

```
## [1] "95% CI [-0.3, 0.895] does not exclude 0"
```

then then p-value is below 5%

```
## [1] "p = 0.29, not below 5%"
```

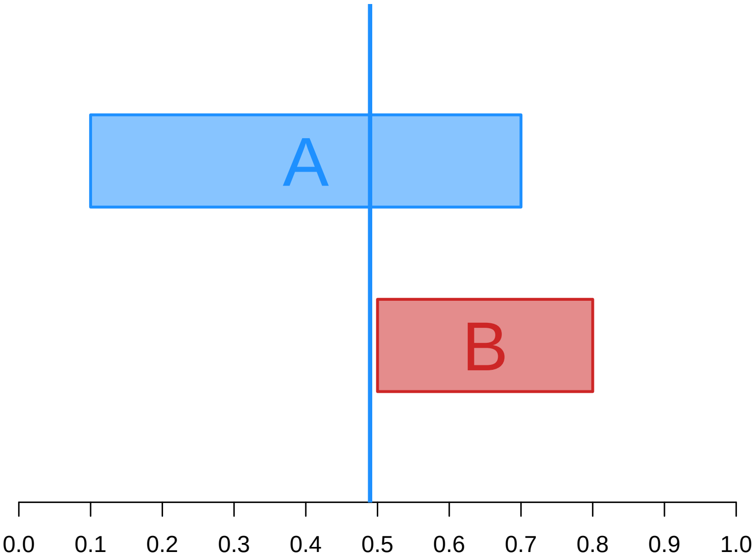
But we know (from last time) that the p-value is the probability of the data (or more extreme) **given** the null hypothesis.

Importantly, and fundamentally, this is NOT the probability of the hypothesis **given** the data.

To make it clearer that these are not the same thing, we now explore the world of conditional probability.

Probability

Let's look at a simple example of a random uniform variable between 0 and 1, denoted by the vertical line, and include two blocks in our plot: A from 0.1 to 0.7, and B from 0.5 to 0.8.

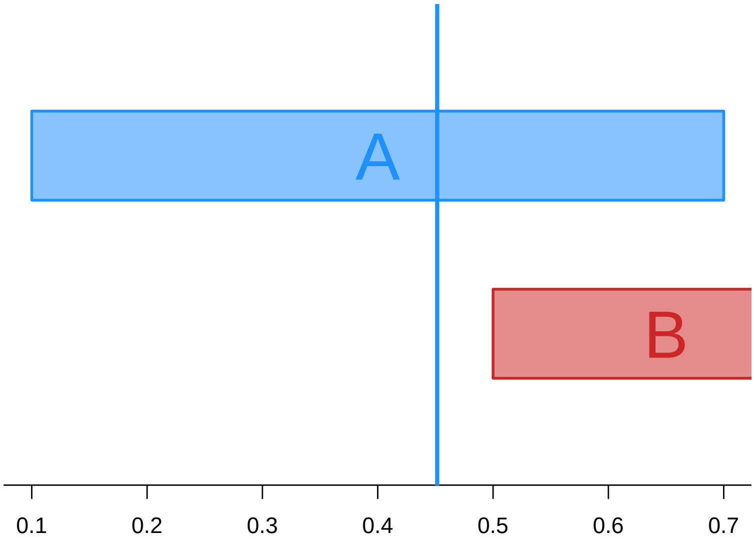


In this setup, we can see that $P(A)$, the probability a uniform random variable hits A, is 0.6 since 60% of the line is covered by A. Likewise $P(B) = 0.3$.

More involved is the probability that the sampled value hits both A and B, denoted by $P(A \cap B)$. What is it's value? Only the region between 0.5 and 0.7 is covered by both A and B, so $P(A \cap B) = 0.2$.

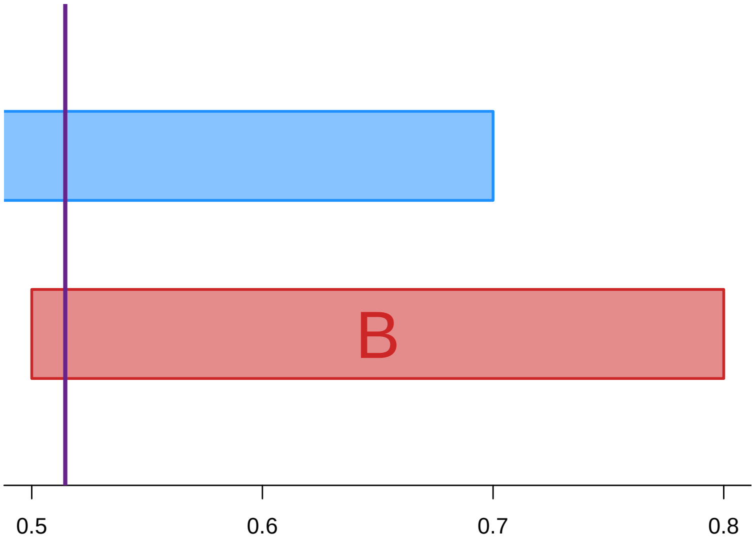
Conditional probability

More interesting are questions of conditional probability. Given that we hit A, say, what is the probability we also hit B. As a first step we can ignore any samples that do not hit A and essentially zoom in only on the region covered by A.



The region of A is now our entire universe once we conditioned upon A happening, so what is $P(B \mid A)$, the probability the sample also hits B, given that it hits A? Of the 0.6 region covered by A, 0.2 is also covered by B, giving a probability of $P(B \mid A) = \frac{0.2}{0.6} = \frac{1}{3}$

For $P(A \mid B)$, we only consider the region covered by B:



From this plot we can see that $P(A \mid B) = \frac{2}{3}$

Bayes theorem

To formalise these example, we have the rule of conditional probability

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

which basically says we look at the intersection of both events, and we rescale and normalise to those events where A occurred. We can also rearrange the equation

$$P(B \mid A)P(A) = P(A \cap B) = P(A \mid B)P(B)$$

and obtain Bayes theorem in its standard form

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}$$

A useful aside for Bayes theorem is that we can also write $P(A)$ in the denominator in terms of conditional probabilities for the different possible values of B:

$$P(A) = P(A \mid B)P(B) + P(A \mid \neg B)P(\neg B)$$

A family example

To get used to Bayes theorem we gave a simple example of entering all families with exactly two children and at least one daughter into a raffle. What is the probability the winning family has two daughters?

The answer is on the next page, so have a think first.

The probability is $\approx \frac{1}{3}$. Since we have removed families with two sons, there are three possibilities for the two children: $\{D, D\}$, $\{D, S\}$, $\{S, D\}$ and only one of those has two daughters.

In terms of Bayes theorem, let's denote by C the condition that a family with two children has at least one daughter, then $P(C) = \frac{3}{4}$ and $P(D, D \cap C) = \frac{1}{4}$ and so

$$P(D, D \mid C) = \frac{P(D, D \cap C)}{P(C)} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}$$

We can also easily check this with simulations:

```
n_fam <- 1e5
first_child <- sample(c("D", "S"), n_fam, replace = TRUE)
second_child <- sample(c("D", "S"), n_fam, replace = TRUE)
families <- cbind(first_child, second_child)
# remove those with two sons
two_sons <- which(rowSums(families == "S") == 2)
selected_families <- families[-two_sons, ]
# work out the fraction with two daughters
p_two_daughters <- mean(rowSums(selected_families == "D") == 2)
signif(p_two_daughters, 3) # print the simulation results
```

```
## [1] 0.332
```

A testing example

This example comes from **Eddy (1982)**, and asked of medical doctors to see if they can get the right ballpark probability in the end:

1% of women at age forty who participate in routine screening have breast cancer.

80% of women with breast cancer will get positive mammographies.

9.6% of women without breast cancer will also get positive mammographies.

A woman in this age group had a positive mammography in a routine screening.

What is the probability that she actually has breast cancer?

Before I give the answer, it is worth having a go yourself if you've never seen this question. Even if you have seen it before, it is worth checking you can still compute the right answer!

We can plug all the information into Bayes theorem, using + to signify a positive test and C to indicate having breast cancer.

$$P(C \mid +) = \frac{P(+ \mid C)P(C)}{P(+)} = \frac{0.8 \times 0.01}{P(+)} = \frac{0.008}{P(+)}$$

To proceed we use the expanded version of the denominator in terms of the two possible cancer states

$$P(+) = P(+ \mid C)P(C) + P(+ \mid \neg C)P(\neg C) = 0.8 \times 0.01 + 0.096 \times 0.99 = 0.10304$$

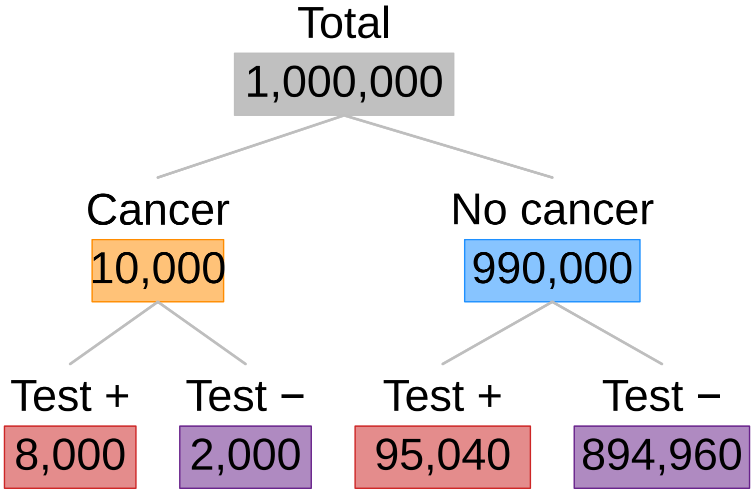
and obtain

$$P(C \mid +) = \frac{0.008}{0.10304} = 0.078$$

around 8%. This is quite far from a lot of people's intuitive answer of around 80% in line with the true positive rate of the test, and illustrates the dangers of "inverting" conditional probabilities in your head.

An easier solution

It can be easier to avoid algebra and think of a large population:



And compare those with a positive test: 8,000 have cancer, while 95,040 don't.

$$P(C \mid +) = \frac{8,000}{8,000 + 95,040} = 0.078$$

Bayes theorem and p-values

We took this detour into conditional probability because our p-values are a conditional probability of the data (or more extreme) **given** the null hypothesis being true

$$p = P(D^+ \mid H_0)$$

where with the + we intend the data D or more extreme. However, we would really like to know the probability of the hypothesis given the data instead: $P(H_0 \mid D)$. We can plug this into Bayes theorem

$$P(H_0 \mid D) = \frac{P(D \mid H_0)P(H_0)}{P(D)}$$

and it is clear to compute this “inversion” we would need to know, in addition to the probability or **likelihood** of the data under the null (which we have or can compute similarly to the p-value), also:

the **prior** probability of the null hypothesis: $P(H_0)$ * the probability of the data $P(D)$

For the second part, we can expand the denominator

$$P(D) = P(D \mid H_0)P(H_0) + P(D \mid \neg H_0)P(\neg H_0)$$

and see we further need the probability of the data under **all** possible alternative hypotheses: $P(D \mid \neg H_0)$.

The p-value does not, can not, and will never include all the information to deduce the probability of any hypothesis **given** the data. Learn the mantra:

- A p-value is the probability of the data (or more extreme) **under** the null hypothesis.

stick to the script, and be skeptical of any deviations from the definition.

Back to confidence intervals

Returning to confidence interval questionnaire, we can look at the logical consequences from statements 4 or 5 above. If the probability that the true mean is between 0.1 and 0.4 were 95%, then the p-value (with the null of 0 mean) would be less than 5%. Also, with 95% probability of the true mean being between 0.1 and 0.4, the probability it is 0 (or to give it some width, that it is between -0.1 and +0.1) must be less than 5%. The probability of the null hypothesis must therefore be below 5%, and we would have deduced the probability of an hypothesis just from an observed sample.

From Bayes theorem, we know this isn’t possible without lots more information not contained in our original sample. Statements 4 and 5 are basically underlying inversions of conditional probabilities, but particularly subtle and insidious.

Summary

We can now summarise our whistlestop recap of t-tests:

p-values are a measure of surprise

- probability of the data statistic (sample mean) or more extreme **given** the null

Commonly misinterpreted

- not the probability of the hypothesis **given** the data!

For normal data the rescaled sample mean

Confidence intervals are the central part of estimates of the distribution of the statistic (sample mean)

- contain similar information to the p-value but even more commonly misinterpreted

To “invert” **conditional** probabilities

- **Bayes theorem**

$$P(H_0 \mid D) = \frac{P(D \mid H_0)P(H_0)}{P(D)}\,, \qquad P(D) = P(D \mid H_0)P(H_0) + P(D \mid \neg H_0)P(\neg H_0)$$

Need information beyond the p-value: the **prior** H_0 and to consider all alternative hypotheses!

→ Exercises 2