Inferential statistics I – Hypothesis Testing in the Basic Form of Conditional Probability / Bayes' Rule in R

Absolute Beginner's Stat-o-Sphere

by Steffen Schwerdtfeger

Our first "Stat-o-Sphere" tutorial focuses on the basics of statistical inference, which can be looked at as reflecting three steps of scientific inquiry. Nevertheless, you consider a Bayesian or frequentist approach to (conditional) probability, they both refer to the same formula. The classic frequentist p-value is also just a conditional probability and the only difference to the Bayesian approach is the prior assumptions that are made before gathering or encountering data (see further below).

Below you see different ways to reflect on the steps of scientific inquiry within inferential statistics. In case you feel intimidated by mathematical symbols, *you should never*, as the definition of each formula is just the formula spoken out (see further below). A mere linguistic definition will just be confusing in the long run. Conditional probability is also really easy and intuitive to understand, so avoiding the math does not come with any benefits all, on the contrary. You will also soon learn that most of the people working with stats have a very heuristic approach to it, which explains a lot of confusion around p-values and other concepts in statistics. Our goal is to change that and to show that everybody can be a statistician.

Step 1 Ms. Miranda, hypothesizing in her shelter – they say, she knows-it-owl. A hypothesis for itself can be considered prior (a priori), when the data to evaluate the hypothesis with is *not* given yet. Original photo by Kevin Mueller.

Step. 2 Ms. Miranda, on the way to gather some data – wise enough to always challenge her beliefs. Original photo by Alfred Kenneally.

Step. 3 Ms. Miranda has published her results maximum *open access*, so her insights were spread fast in the community. Ms. Miranda was right with her hypothesis. This is not the case with every hypothesis we belief to be true, keep in mind. Original photo by Ray Hennessy.

I. Step	II. Step	III. Step
formulate hypothesis	gather / encounter data	update hypothesis / model
Prior $P(theta)$	*Joint Probability* $P(theta, data)$	*Posterior* $P(theta \mid data)$
past	present	future

The **prior** is the probability (p) of the $(hypo)theta$, regardless any $data$.

The **joint** is a logical overlapping/conjunction, spoken the probability of $theta$ and (= ",") $data$.

The **posterior** is the probability of the $theta$, given or under the condition or after (= "|") $data$ was observed.

The **classic p-value in frequentist statistics** is the same as the posterior conditional probability, just tilted, and is called the **likelihood** and is denote **p(data|theta)**. The p-value also involves a t-test such that $data$ refers to the difference in the means of, e.g., two groups. Multiplying the likelihood with the prior results in a joint probability:

$$p(data|theta) * p(theta) \; = \; p(data, theta)$$

Given a **uniform prior**, i.e., 50/50 for $theta$ and "not-theta" (denoted $\overline{theta}$), the likelihood becomes equivalent to the posterior probability ($\overline{theta}$ is also referred to as the **null hypothesis**). As the posterior and the likelihood become equivalent, the conditions can be tilted without changing the results (so technically a p-value does not *only* refer to the likelihood). Looking at our three steps above, this circumstance can be looked at as being focused on the present moment only when doing statistical inference. Below you can find the formula for Bayes' rule (which again concerns both frequentist and Bayesian approach).

$$P(theta|data) \; = \; \frac{P(data \mid theta)P(theta)}{P(data)}$$

***Posterior probability**, which includes both unweighted (uniform) and weighted prior*

$$P(theta|data) \; = \; \frac{P(data \mid theta)\cancel{P(theta)}}{\cancel{P(data)}}$$

***Maximum likelihood**, given the special case of a unform prior probability distribution*

$$P(\overline{theta}|data) \; = \; \frac{P(data \mid \overline{theta})\cancel{P(theta)}}{\cancel{P(data)}}$$

***p − value** in the common sense: the data given the null hypothesis* $- \overline{theta}$

An easy way to recall the formula for Bayes' rule is by chanting it clockwise, starting with the posterior on the left (including a small break for the single probabilities $p(data)$ and $p(theta)$):

theta-data || data-theta || theta – data

Further questions? See full tutorial for details!