Hypothesis testing

We often wish to use our observed data to draw conclusions about the plausibility of various hypotheses. Returning to our example of coin flipping, we might wish to know whether the parameter θ is say less than $1/2$. The Bayesian method allows us to compute this value directly from the posterior distribution:

$$\Pr(\theta < 1/2 \mid x, n, \alpha, \beta) = \int_0^{1/2} p(\theta \mid x, n, \alpha, \beta)\,\mathrm{d}\theta.$$

Classical method

There is a sharp contrast between the simplicity of the approach above and the frequentist method. The classical approach to hypothesis testing uses the likelihood as a way of generating fake datasets of the same size as the observations. The likelihood then serves as a so-called "null hypothesis" that allows us to generate hypothetical datasets under some condition.

From these, we compute *statistics,* which, like estimators, can be any function of the hypothesized data. We then identify some critical set C for this statistic which contains some large portion $(1-\alpha)$ of the values corresponding to the datasets generated by our null hypothesis. If the statistic computed from the observed data falls outside this set, we reject the null hypothesis with "confidence" α. Note that the "rejection" of the null hypothesis in classical hypothesis testing is purely a statement about the observed data (that it looks "unusual"), and not about the plausibility of alternative hypotheses!

Bayesian method

Suppose we are reasoning about a parameter θ in light of data $\mathcal{D}$, and wish to consider a hypothesis of the form $\theta \in \mathcal{H}$, where $\mathcal{H} \subseteq \Theta$ is some set of possible values for this parameter.

We have seen that the Bayesian approach to hypothesis testing is straightforward. We first derive the posterior distribution $p(\theta \mid \mathcal{D})$ and then may compute the probability of the hypothesis directly:

$$\Pr(\theta \in \mathcal{H} \mid \mathcal{D}) = \int_{\mathcal{H}} p(\theta \mid \mathcal{D})\,\mathrm{d}\theta.$$

Coin fairness

Let's consider an explicit example. Suppose we are interested in the unknown bias of a coin $\theta \in (0,1)$, and begin with the uniform prior on the interval $(0,1)$:

$$p(\theta) = \mathcal{U}\bigl(\theta; 0, 1\bigr) = \mathcal{B}(\theta; \alpha = 1, \beta = 1).$$

Let's collect some data to further inform our belief about θ. Suppose we flip the coin independently $n = 50$ times and observe $x = 30$ heads. After gathering this data, we wish to consider the natural question of whether the coin is fair: that is, whether $\theta = 1/2$.

From the results in the previous lecture, we can compute the posterior distribution easily. It is an updated beta distribution:

$$p(\theta \mid \mathcal{D}) = \mathcal{B}(\theta; 31, 21).$$

We may now compute the posterior probability of the hypothesis that the coin is fair:

$$\Pr(\theta = 1/2 \mid \mathcal{D}) = \int_{1/2}^{1/2} p(\theta \mid \mathcal{D})\,\mathrm{d}\theta = 0.$$

The posterior probability of the coin being *exactly* fair is zero! This should not be surprising, as suggesting that we could possibly know the bias of the coin with infinite precision is unfathomable.

We may however relax the question a bit to get some more insight. One option would be to consider a parameterized family of hypotheses of the form

$$\mathcal{H}(\varepsilon) = (1/2 - \varepsilon, 1/2 + \varepsilon).$$

Thus a high probability of the hypothesis $\mathcal{H}(\varepsilon)$ corresponds to the notion that the coin is “near fair” with an allowed error of ε. We may then compute the posterior probability of these hypotheses and consider how they vary as a function of ε. Figure 1 shows the results for the coin-flipping example above. We can see that there’s approximately a 50% posterior probability that the bias of the coin is in the interval $(0.4, 0.6)$, corresponding to $\varepsilon = 0.1$. We also have evidence to conclude $\theta \in (0.25, 0.75)$ with near certainty. These probabilities help constrain exactly how “fair” or “not fair” we believe the coin to be in light of our evidence.

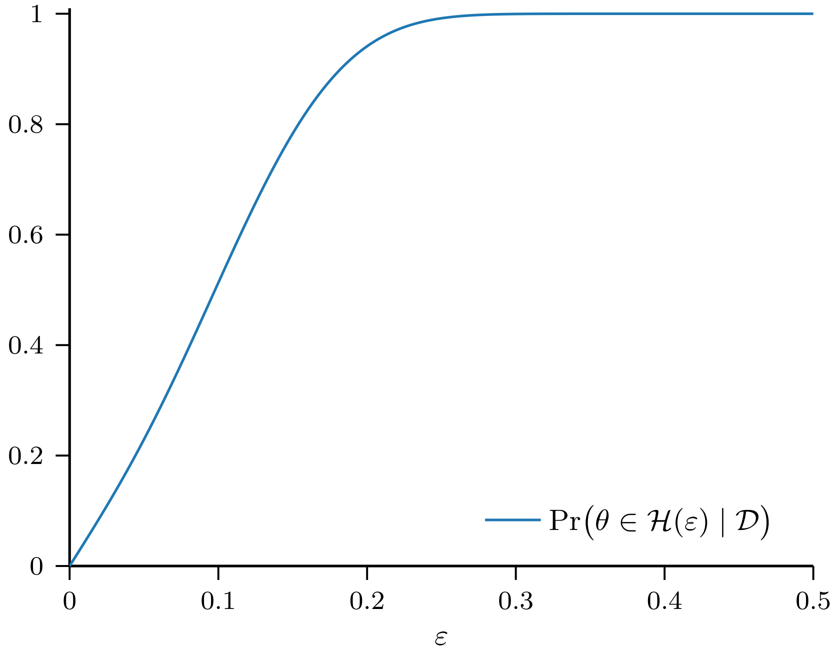

Figure 1: The posterior probability of the hypotheses $\mathcal{H}(\varepsilon)$ for $0 < \varepsilon < 1/2$.

The classical approach to testing this hypothesis is to create a so-called “null hypothesis” $\mathcal{H}_0$ that serves to define what “typical” data may look like assuming that hypothesis. For example, for reasoning about the fairness of a coin, we may choose the natural null hypothesis $\mathcal{H}_0 \colon \theta = 1/2$. Now we can use the likelihood

$$\Pr(x \mid n, \theta = 1/2)$$

to reason about what observed data would look like if this hypothesis were true. This is a critical point: the null hypothesis exists to define what sort of data we would expect to see under an assumed value of θ.

The classical procedure would then define a statistic summarizing a given dataset $s(\mathcal{D})$ in some way. An example for coin flipping would be the sample mean $s(\mathcal{D}) = \hat{\theta} = {}^{x}\!/_{n}$. This happens to be a common estimator for θ as well, but this is a coincidence. We now compute a so-called *critical set* $C(\alpha)$ with the property

$$\Pr\bigl(s(\mathcal{D}) \in C(\alpha) \mid \mathcal{H}_0\bigr) = 1 - \alpha,$$

where α is called the *significance level* of the test. The interpretation of the critical set is that the statistic computed from datasets generated assuming the null hypothesis "usually" have values in this range.

Finally, we compute the statistic for a particular set of observed data and determine whether it falls inside the critical set $C(\alpha)$ we have defined. If so, the dataset appears, according to the statistic, typical for datasets generated from the null hypothesis. If not, the dataset appears unusual, in the sense that data generated assuming the null hypothesis would have such extreme values of the statistic only a small portion of the time ($100\alpha\%$). In this case, you "reject" the null hypothesis with significance $1 - \alpha$.

What is a p-value? It must be the probability that the null hypothesis is true, right? Absolutely not! The null hypothesis cannot be associated with a probability in the classical interpretation of probability. A p-value is actually the minimum α for which you would reject the null hypothesis using this procedure. That is, a p-value is not the probability that the null hypothesis is true, but rather the probability that we would observe results as extreme as those in our dataset, as measured by the chosen statistic, *if the null hypothesis were true!* The p-value is thus only a probability that is well-defined when already assuming the null hypothesis to be true. A p-value does *not* say how extreme our results would appear under alternative hypotheses.

Bayesian model selection will eventually allow us to explicitly quantify the plausibility of a collection of models having generated the observed data.

To interpret the above procedure in the frequency interpretation of probability, the critical sets are constructed by reasoning about the following experiment:

- generate $\mathcal{D}$ assuming $\mathcal{H}_0$;
- compute $s(\mathcal{D})$;
- state $s(\mathcal{D}) \in C(\alpha)$.

In the limit of infinitely many repetitions of this experiment, the final claim will be true exactly $100(1 - \alpha)\%$ of the time. Recall this is the definition of probability in this context: the frequency of occurrence in the limit of infinitely many trials. Note that the experiment we repeat here *includes generating data from the null hypothesis* as its first step! This is not the experiment we are conducting, since we have a dataset in front of us that we want to analyze, which may have been generated in any number of ways.

Summarizing Distributions

In the Bayesian method, the posterior distribution $p(\theta \mid \mathcal{D})$ is the main object of interest and contains all relevant information about θ in light of the observations $\mathcal{D}$. A natural task is to provide a summary of the posterior distribution, for example to efficiently convey its relevant properties.

In the next lecture we will consider point estimation, which is one common summarization method. Another commonly considered problem is **interval summarization,** where we provide an interval (ℓ, u) indicating plausible values of the parameter θ in light of the observed data. Classical interval estimates are known as *confidence intervals,* and we will discuss them in more detail shortly.

The Bayesian approach to interval estimation is straightforward. Again we use the posterior distribution $p(\theta \mid \mathcal{D})$ to guide the construction of an interval summary. If we can find an interval (ℓ, u) such that the posterior probability that $\theta \in (\ell, u)$ is "large" (say, has probability α):

$$\Pr\bigl(\theta \in (\ell, u) \mid \mathcal{D}\bigr) = \int_{\ell}^{u} p(\theta \mid \mathcal{D})\,\mathrm{d}\theta = \alpha,$$

then we call (ℓ, u) an α-**credible interval** for θ. Note the parallel in this definition to our treatment of hypothesis testing above! Effectively, an α-credible interval is simply a hypothesis that has posterior probability equal to α and happens to take the form of an interval.

Examining our coin flipping example from before, we can construct some credible intervals immediately from the data in Figure 1. We have that $\mathcal{H}(\varepsilon = 0.1) = (0.4, 0.6)$ is a 50%-credible interval for the bias of the coin, and $\mathcal{H}(\varepsilon = 0.2) = (0.3, 0.7)$ is a 95%-credible interval. The slightly wider interval $\mathcal{H}(\varepsilon = 0.25) = (0.25, 0.75)$ represents a very high probability credible interval, corresponding to $\alpha > 99\%$.

It is clear from the definition that multiple intervals (in fact, often uncountably many) can serve as a credible interval for a particular value of α. Exactly which interval should we construct to summarize a given distribution? This is a question for which we will need to develop Bayesian decision theory before we can continue, which we will discuss in the next lecture. In short, we will first need to quantify how "desirable" a given credible interval is in some way, then select the one maximizing this measure. For example, we may want to construct the narrowest possible interval, or we may wish it to be centered on a particular point (such as the posterior mean, median, or mode), or we may wish the interval to have some other property.

The classical approach to interval summarization is to construct a so-called confidence interval for the parameter of interest θ. Again a confidence interval is described in terms of repeating a particular experiment infinitely many times. The experiment we consider will proceed as follows. First we are going to define a function $\mathrm{CI}(\mathcal{D})$ that will map a given dataset $\mathcal{D}$ to an interval $(\ell, u) = \mathrm{CI}(\mathcal{D})$. Now we consider repeating the following experiment:

- collect data $\mathcal{D}$
- compute the interval $(\ell, u) = \mathrm{CI}(\mathcal{D})$
- state $\theta \in (\ell, u)$.

In the limit of infinitely many repetitions of this experiment, if the final statement is true with probability α, then the procedure $\mathrm{CI}(\mathcal{D})$ is called an *α-confidence interval procedure,* and we will write $\mathrm{CI}(\mathcal{D}; \alpha)$ to indicate the confidence level α when required.

This might sound like exactly the same definition as a Bayesian credible interval. For example, if we have an α-confidence interval procedure available, then when we plug in a given dataset $\mathcal{D}$, we must have

$$\Pr\bigl(\theta \in \mathrm{CI}(\mathcal{D}; \alpha) \mid \mathcal{D}) = \alpha, \tag{$\star$}$$

right? **No!** This interpretation is widespread, but it is wrong. This conclusion is sometimes known as the **fundamental confidence fallacy,**[1] and confuses the nature of prior information with that of posterior information. Namely, note that the experiment we consider when defining the confidence interval procedure *includes gathering a random dataset* as its first step. All we know is that if we repeat the confidence interval procedure on *infinitely many datasets,* that it will succeed with probability α. However, we usually have only one particular dataset in front of us to analyze that we care about, and we cannot say anything about the interval produced for this dataset in isolation.

Here is a simple example that shows how this erroneous conclusion can go wrong. Suppose we are going to observe two values $x_1, x_2 \in \mathbb{R}$ generated independently from some unknown distribution $p(x)$ and wish to construct a confidence interval for the mean of the distribution generating the data, $\theta = \mathbb{E}[x]$. Consider the following procedure:

$$\mathrm{CI}(\mathcal{D}) = \begin{cases} (-\infty, \infty) & x_1 < x_2 \\ \emptyset & x_2 \geq x_1. \end{cases}$$

Obviously this trivial map is a 50%-confidence interval procedure! Because the values are generated independently, x_1 will be the lesser value exactly 50% of the time. In this case, the absurdly large interval produced will *definitely* contain θ. The other 50% of the time, the interval will be empty, and *definitely will not* contain θ. Therefore the procedure succeeds exactly 50% of the time. However, in half the cases, the posterior probability that θ is inside the interval produced is 100%, and otherwise this probability is 0%. In no case is this probability equal to the confidence level.

Another fallacy in the interpretation of confidence intervals is the so called **precision fallacy,** that shorter confidence intervals indicate the data provide more precise information about θ. A striking illustration of this fallacy is provided by the "lost submarine" example provided by Morey, et al. in the reference given below. I encourage you to read this paper and reflect!

[1] See the following reference for some excellent extended discussion on confidence intervals: Richard D. Morey, et al. (2015). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review* 23(1): 103–123