Parameter estimation and hypothesis testing: a frequentist approach

Michael Franke

We first look at a Bayesian approach of inferring posterior beliefs about a coin's bias, derived from observing outcomes of flips. We constrast this to frequentist approaches to constructing a point-valued estimate and a confidence interval around it. We then look at the definition of a p-value and its use in hypothesis testing.

Parameter inference: Bayesian approach

Posterior beliefs about a coin bias

Suppose grandma gives you her lucky coin. You wonder if it is biased in any way. Since you cannot ask it, you resort to flipping and observing outcomes. The assumption is that the inherent bias $\theta \in [0; 1]$ of the coin probabilistically influences the likelihood of observing a particular outcome. The probability of a single outcome of heads (aka. a *success*) is θ, that of tails (aka. a *failure*) is $1 - \theta$. On the assumption that the coin's bias is θ,[1] when we flip the coin n times the probability of observing k successes is given by the *binomial distribution*:

[1] And assuming that each coin flip is probabilistically independent of the others.

$$\text{Binom}(K = k; n, \theta) = \binom{n}{k} \cdot \theta^k \cdot (1 - \theta)^{n-k}$$

Simplifying the notation, we will write $P(k \mid \theta)$ and call this the *likelihood function*. The likelihood function is parameterized on θ and gives a probability distribution over all possible observable outcomes when flipping the coin a contextually fixed number of times, here written as n.

The goal is to infer the latent and unknown θ from an observed given k. Bayes rule allows us to calculate:

$$P(\theta \mid k) = \frac{P(k \mid \theta) \cdot P(\theta)}{P(k)}$$

The ingredients of this equation are:

- the *posterior distribution* $P(\theta \mid k)$ specifying our beliefs about how likely each value of θ is given fixed k;
- the *likelihood function* $P(k \mid \theta)$ specifying how likely each observation of k is for a fixed θ (here given by the binomial distribution);
- the *prior distribution* $P(\theta)$ specifying our initial (aka. *a priori*) beliefs about how likely each value of θ might be;
- the *marginal likelihood* $P(k) = \int P(k \mid \theta) \cdot P(\theta)\ \mathrm{d}\theta$ specifying how likely an observation of k is under our prior beliefs about θ.

Example 1. We flip grandma's coin $n = 24$ times and observe 7 heads. What should we believe about θ? It depends, of course, on what we believe before we made this observation. Suppose our prior is *flat*, i.e., we assign the same probability to any value of $\theta \in [0; 1]$ (see Figure 1).[2] With these prior

[2] Flat prior beliefs are also called, uninformative priors, uninformed priors, Laplace priors, or, pompously, maximum entropy priors.

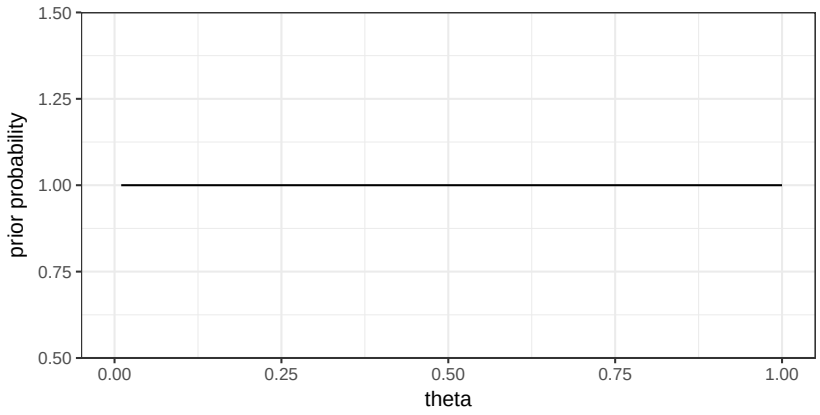

Figure 1: Flat prior belief about coin bias θ.

beliefs, by Bayes rule, our posterior beliefs are given as in Figure 2.[3]

[3] This example is meant to illustrate only how, in general, a Bayesian approach to parameter inference works. We will come back later to computing this example and many other applications of Bayesian parameter inference.

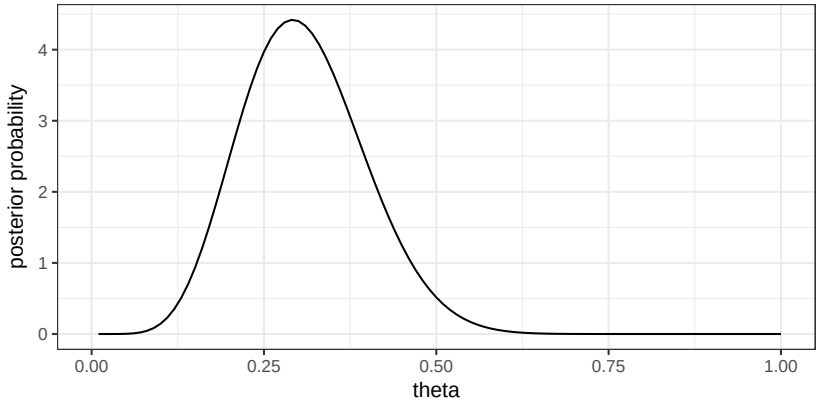

Figure 2: Posterior belief about coin bias θ, based on a flat prior belief, after observing $k = 7$ successes in $n = 24$ flips.

Posterior means and credible intervals

The probability distribution in Figure 2 contains rich information. It specifies how likely each value of θ is, given flat prior beliefs updated by the observed data. Such rich information is difficult to process and communicate in language. It is therefore convenient to have conventional means of summarizing the rich information carried in a probability distribution like in Figure 2. Customarily, we summarize in terms of a point-estimate and/or an interval estimate. The POINT ESTIMATE gives information about a "best value", i.e., a salient point, such as the expectation (in Bayesian approaches) or the most

likely value (in frequentist approaches (see below)). The INTERVAL ESTIMATE gives, usually, an indication of how closely other "good values" are scattered around the "best value".

A common Bayesian point estimate of coin bias parameter θ is the mean of the posterior distribution. It gives the value of θ which we would expect to see, when basing out expectations on the posterior distribution:

$$\mathbb{E}_{P(\theta|k)} = \int \theta \cdot P(\theta \mid k)\, \mathrm{d}\theta$$

If we start with flat beliefs, the expected value of θ after k successes in n flips can be calculated rather easily as $\frac{k+1}{n+2}$.[4] For our example case, we calculate the expected value of θ as $\frac{8}{26} \approx 0.308$ (see also Figure 3).

A common Bayesian interval estimate of coin bias parameter θ is a CREDIBLE INTERVAL.[5] An interval $[l; u]$ is a γ% credible interval for a random variable X if two conditions hold, namely

$$P(l \leq X \leq u) = \frac{\gamma}{100}$$

and, secondly, for every $x \in [l; u]$ and $x' \notin [l; u]$ we have $P(X = x) > P(X = x')$. Intuitively, a 95% credible interval gives the range of values in which we believe with relatively high certainty that the true value resides. Figure 3 indicates the 95% credible interval, based on the posterior distribution $P(\theta \mid D)$ of θ, for the running example.[6]

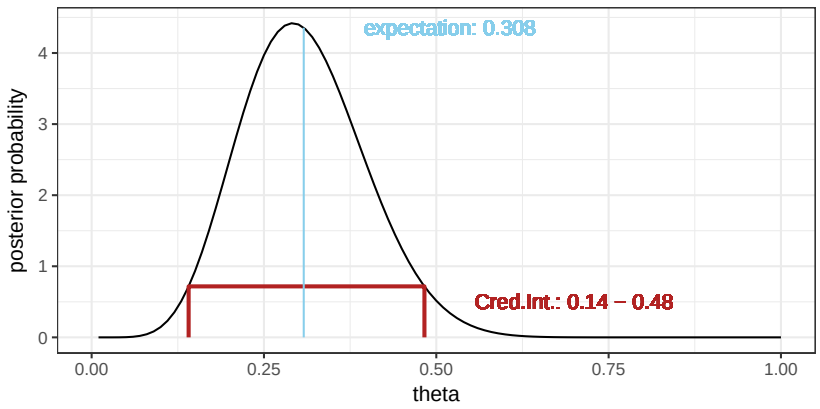

Figure 3: Posterior belief about coin bias θ, based on a flat prior belief, after observing $k = 7$ successes in $n = 24$ flips, including the 95% credible interval.

Subjectivism vs. Frequentism

The posterior distribution shown in Figure 2 is the outcome of rational belief update, applied to the prior beliefs and the observed data. In a frequentist approach to probability, we cannot do this. Frequentists deny that a probability distribution over a latent parameter like θ is meaningful. The only statements about probabilities that are conceptually sound, according to a frequentist

[4] This is also known as *Laplace's rule*, or the *rule of succession*.

[5] Also frequently called "highest-density intervals", even when we are dealing not with density but probability mass.

[6] Not all random variables have a credible interval for a given γ, according to this definition. A bimodal distribution might not, for example. We can therefore generalize the concept to a finite set of disjoint convex CREDIBLE REGIONS, all of which have the second property of the definition above and all of which conjointly are realized with γ% probability. Unfortunately, common parlor uses the term "credible interval" to refer to credible regions as well. The same disaster occurs with alternative terms, such as "γ% highest-density intervals", which also often refers to what should better be called "highest-density regions".

interpretation, are those that derive from intuitions about limiting frequencies when (hypothetically) performing a random process (like throwing a dice or drawing a ball from an urn). Bluntly put, there is no "(thought) experiment" which can be repeated so that its objective results, on average, align with whatever subjective prior beliefs the Bayesian analysis needs. As a result, the frequentist approach to statistical inference needs alternative methods for parameter estimation — methods that do *not* rely on (subjective) priors $P(\theta)$.

To understand the contrast better, think of two different ways of specifying a STATISTICAL MODEL. For the Bayesian analyst, a statistical model of a data-generating process consists of a likelihood function $P(d \mid \theta)$ and a prior $P(\theta)$ over parameter values.[7] The Bayesian analyst can compute rational belief updates because she avails herself of priors $P(\theta)$. In contrast, the frequentist analyst denies that a concept like $P(\theta)$ makes any sense. She therefore cannot appeal to rational belief update in the sense of Bayes rule. Her statistical model consists solely of the likelihood function $P(d \mid \theta)$. She must (and can) do with just that.

[7] Since, in many contexts, the meaning will be clear enough, we follow common practice and write $P(d \mid \theta)$ as a shortcut for $P(D = d \mid \Theta = \theta)$. Here D is the class of all relevant observable data and Θ is the range of a possibly high-dimensional vector of parameter values. (Clearly, many statistical models will want to have several, if not many parameters.)

A frequentist approach to parameter estimation

We pick up the running example of grandma's coin again. It was flipped $n = 24$ times and landed heads $k = 7$ times. Again, the goal is to draw inferences about the underlying bias $\theta \in [0; 1]$. Being a frequentist, we do that based only on likelihood function $P(k \mid \theta)$, given here by the binomial distribution as before. We will derive, from $P(k \mid \theta)$ alone, plausible point and interval estimates.[8]

[8] There are many constructions for both point and interval estimates. We here only look at what seem to be the single most prominent exemplars in each category.

Maximum likelihood estimate

The MAXIMUM LIKELIHOOD ESTIMATE (MLE) is a point estimate based on the likelihood function alone. It specifies the value of θ for which the observed data is most likely. We often use the notation $\hat{\theta}$ to denote the MLE of θ:

$$\hat{\theta} = \arg\max_{\theta} P(d \mid \theta)$$

Example 2. For coin flips, the maximum likelihood estimate is easy to calculate as $\frac{k}{n}$, yielding $\frac{7}{24} \approx 0.292$ for the running example. Figure 4 shows a graph of the non-normalized likelihood function and indicates the maximum likelihood estimate (the value that maximizes the curve).[9]

[9] Notice that the posterior in Figure 2 does indeed have the exact same shape as the likelihood function in Figure 4, but not the same y-axis values. The former is normalized, the latter is not. Posteriors and likelihood functions cease to have identical shapes when prior beliefs are not flat. When they are, the maximum likelihood estimate coincides with the MAXIMUM A POSTERIORI (MAP) value, i.e., the value that maximizes the posterior distribution.

Confidence intervals

The most commonly used interval estimate in frequentist analyses are CONFIDENCE INTERVALS. Although (frequentist) confidence intervals *can* coincide with (subjectivist) credible intervals in specific cases, they generally do not. And even when confidence and credible values yield the same numerical results, these notions are fundamentally different that ought not to be confused.

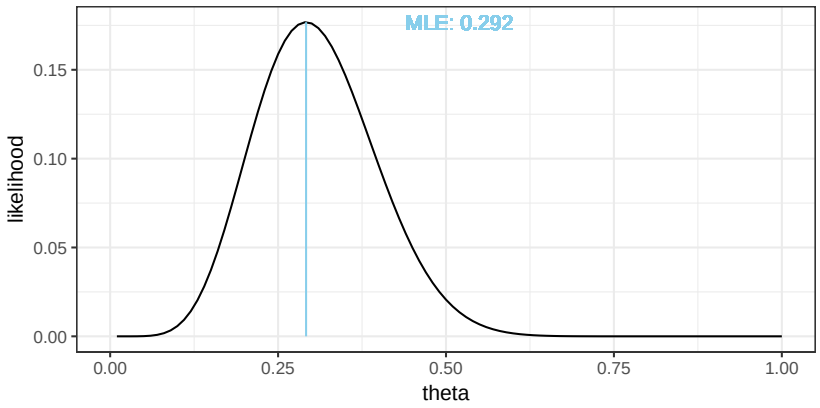

Figure 4: Non-normalized likelihood function for the observation of $k = 7$ successes in $n = 24$ flips, including maximum likelihood estimate.

Let's look at credible intervals to establish the proper contrast. Recall that part of the definition of a credible interval for a posterior distribution over θ, captured here notationally in terms a random variable Θ, was the probability $P(l \leq \Theta \leq u)$ that the value realized by random variable Θ lies in the interval $[l; u]$. This statement makes no sense to the frequentist. There cannot be any non-trivial value for $P(l \leq \Theta \leq u)$. The true value of θ is either in the interval $[l; u]$ or it is not. To speak of a probability that θ is in $[l; u]$ is to appeal to an ill-formed concept of probability which the frequentist denies.

In order to give an interval estimate nonetheless, the frequentist appeals to probabilities that she can accept: probabilities derived from (hypothetical) repetitions of a genuine random event with objectively observable outcomes. Let D be the random variable that captures the probability with which data $D = d$ is realized. We obtain a pair of derived random variables X_l and X_u from a pair of functions $g_{l,u} : d \mapsto \mathbb{R}$. A $\gamma\%$ *confidence interval* for observed data d_{obs} is the interval $[g_l(d_{\text{obs}}), g_u(d_{\text{obs}})]$ whenever functions $g_{l,u}$ are constructed in such a way that

$$P(X_l \leq \theta_{\text{true}} \leq X_u) = \frac{\gamma}{100}$$

where θ_{true} is the unknown but fixed true value of θ. In more intuitive words, a confidence interval is the outcome of a special construction (functions $g_{l,u}$) such that, when applying this procedure repeatedly to outcomes of the assumed data-generating process, the true value of parameter θ will lie inside of the computed confidence interval in exactly $\gamma\%$ of the cases.

In some complex cases, the frequentist analyst relies on functions g_l and g_u that are easy to compute but only approximately satisfy the condition $P(X_l \leq \theta_{\text{true}} \leq X_u) = \frac{\gamma}{100}$. For example, we might use an asymptotically correct calculation, based on the observation that, if n grows to infinity, the binomial distribution approximates a normal distribution.[10] We can then calculate a confidence interval, *as if* our binomial distribution actually was a normal distribution. If n is not large enough, this will be increasingly

[10] See Info Box 1 for a derivation of this asymptotic approximation. This should give you a general idea of the flavor of frequentist approximations, upon which much of the frequentist approach is built.

Aymptotic approximation of a binomial confidence interval

Let X be the random variable that determines the binomial distribution, i.e., the probability of seeing k successes in n flips. For large n, X approximates a normal distribution with a mean $\mu = n \cdot \theta$ and a standard deviation of $\sigma = \sqrt{n \cdot \theta \cdot (1-\theta)}$. The random variable U:

$$U = \frac{X - \mu}{\sigma} = \frac{X - n \cdot \theta}{\sqrt{n \cdot \theta \cdot (1-\theta)}}$$

Let $\hat{P}$ be the random variable that captures the distribution of our maximum likelihood estimates for an observed outcome k:

$$\hat{P} = \frac{X}{n}$$

Since $X = \hat{P} \cdot n$ we obtain:

$$U = \frac{\hat{P} \cdot n - n \cdot \theta}{\sqrt{n \cdot \theta \cdot (1-\theta)}}$$

We now look at the probability that U is realized to lie in a symmetric interval $[-c, c]$, centered around zero — a probability which we require to match our confidence level:

$$P(-c \leq U \leq c) = \frac{\gamma}{100}$$

We now expand the definition of U in terms of $\hat{P}$, equate $\hat{P}$ with the current best estimate $\hat{p} = \frac{k}{n}$ based on the observed k and rearrange terms, yielding the asymptotic approximation of a binomial confidence interval:

$$[\hat{p} - \frac{c}{n} \cdot \sqrt{n \cdot \hat{p} \cdot (1-\hat{p})}; \hat{p} + \frac{c}{n} \cdot \sqrt{n \cdot \hat{p} \cdot (1-\hat{p})}]$$

This approximation is conventionally considered precise enough when the *rule of thumb*

$$n \cdot \hat{p} \cdot (1-\hat{p}) > 9$$

is met.

Info Box 1: Asymptotic approximation of a binomial confidence interval

imprecise. Rules of thumb are used to decide how big n has to be to involve at best a tolerable amount of imprecision (see Info Box 1).

Example 3. For our running example ($k = 7, n = 24$), the rule of thumb recommends *not* using the asymptotic calculation. If we did nonetheless, we would get a confidence interval of $[0.110; 0.474]$. For the binomial distribution also a more reliable calculation exist, which yields $[0.126; 0.511]$ for the running example.[11]

[11] We can use numeric simulation to explore how good/bad a particular approximate calculation is.

Excursion: numeric simulation as a tool to investigate statistical concepts

Philosophical quibbles aside, which concept is better: confidence intervals or credible intervals? If we use confidence intervals for a binomial parameter, how bad would an asymptotic approximation really be as compared to what is referred to as the exact method of calculation?

Numerical simulations can help answer these questions.[12] The idea is simple but immensely powerful. We simulate, repeatedly, a ground-truth and synthetic results for fictitious experiments, and then we apply the statistical tests/procedures to these fictitious data sets. Since we know the ground-truth, we can check which tests/procedures got it right.

[12] Even if the math seems daunting, this method is much more tangible and applicable and requires only basic programming experience.

Example 4. Let's look at a simulation, comparing credible intervals to confidence intervals, the latter of which calculated by asymptotic approximation or the so-called exact method. To do so, we repeatedly sample a ground-truth (e.g., a known coin bias θ_{true}) from a flat distribution over $[0; 1]$.[13]. We then simulate an experiment in a synthetic world with θ_{true}, using a fixed value of n, here taken from the set $n \in \{10, 25, 100, 1000\}$. We then construct a confidence interval (either approximately or precisely) and a 95% credible interval; for each of the three interval estimates. We check whether the ground-truth θ_{true} is *not* included in any given interval estimate. We calculate the mean number of times such non-inclusion (errors!) happen for each kind of interval estimate. Figure 5 shows the results of such a simulation experiment, based on 10,000 samples of θ_{true}.

[13] This is already not innocuous. We are fixing, as it were, an assumption about how likely ground-truths should actually occur in the real world.

Hypothesis testing with p-values

Hypothesis testing is the workhorse of much of frequentist statistics.[14] Researchers hold a research hypothesis H_1. The hypothesis that is tested is the logical negation of H_1, usually referred to as the null-hypothesis H_0.[15] If empirical observations are sufficiently *un*likely from the point of view of the null-hypothesis H_0, this is treated as evidence in favor of the research hypothesis H_1. A measure, perhaps approximate, of how unlikely the data is in the light of H_0 is the p-value. A conventional threshold on p-values governs a categorical decision whether to reject or not reject the null-hypothesis. This threshold is essentially an upper-bound on a particular kind of error,

[14] A more precise but less elegant heading for this section would be: "Null-hypothesis significance testing with controlled α-errors."

[15] It is called null-hypothesis, because it is a baseline, a null point, against which the real beast of interest is compared.

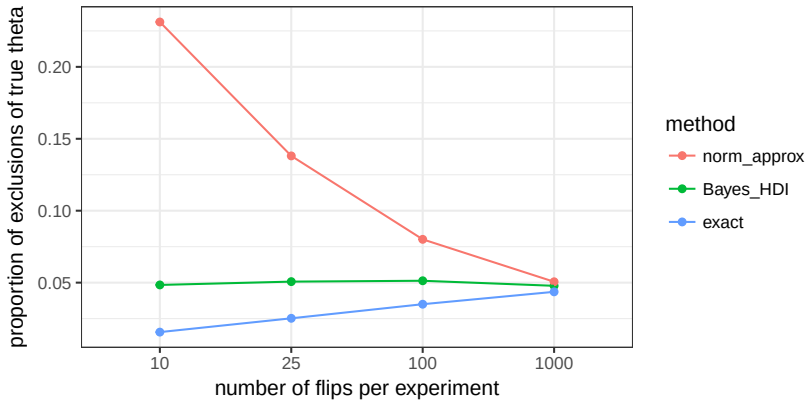

Figure 5: Outcomes of a numerical simulation to compare interval estimates.

namely the error to falsely reject the null-hypothesis when it is in fact true (a so-called type-1 error or α-error). In the following we will fill these general ideas with life.[16]

p-values

As before we fix a parameterized likelihood function $P(d \mid \theta)$, such as the binomial distribution. For our purposes here, we may think of a research hypothesis as a statement about the parameter (vector) θ. Often, the null-hypothesis fixes a parameter of interest to a point-value, but this is not required for this approach to work.[17]

Example 5. A research hypothesis for our running example and its corresponding null-hypothesis are:

H_1: grandma's coin is biased $\theta \neq 0.5$

H_0: grandma's coin is fair $\theta = 0.5$

The null-hypothesis fixes a likelihood function for any kind of data we might perceive on (hypothetical) repetitions of an experiment. We here write $P(D = d \mid H_0)$ or, for short, $P(d \mid H_0)$, e.g., $P(D = d \mid \Theta = 0.5)$ for our grandma-coin example. Let's write D for the random variable that captures the probability distribution over any observation that is conceivable for our experiment, from the point of view of the null-hypothesis. (E.g., when fixing $n = 24$ the space of logically possible outcomes is $K = \{0, 1, \ldots, 24\}$ observations of heads.) Then $T = t(D)$ is a derived random variable, the so-called SAMPLING DISTRIBUTION, with the function $t : d \mapsto t(d) \in \mathbb{R}$ a so-called TEST STATISTIC. The p-value associated with actual data observation d_{obs} is the probability of observing any alternative outcome with a value of the test statistic that is at least as extreme as $t(d_{\text{obs}})$:[18]

$$P(T \geq t(d_{\text{obs}})) = P(D \in \{d \mid t(d) \geq t(d_{\text{obs}})\})$$

[16] Null-hypothesis significance testing is superficially related to Popperian falsificationism. It is, however, quite the opposite when looked at more carefully. Popper famously denied that empirical observation could constitute positive evidence in favor of a research hypothesis. Research hypotheses can only be refuted, viz., when their logically consequences are logically incompatible with the observed data. In a Popperian science, what is refuted are research hypotheses; frequentist statistics instead seeks to refute null-hypotheses and counts successful refutation of a null-hypothesis as evidence in favor of a research hypothesis.

[17] Some ((cheeky) Bayesian) people object that it makes no sense to conceive of a point-value null-hypothesis in the first place. The probability that the coin is exactly fair has no non-trivial measure.

[18] Keep in mind that the null-hypothesis is implicit in the sampling distribution.

A special case obtains when the test statistic is the inverse likelihood function: $t(d) = P(D = d \mid H_0)^{-1}$. We then speak of an EXACT p-VALUE (and an ensuing EXACT TEST). It follows that the exact p-value of observation d_{obs} gives the probability of observing an event which is, from the point of view of H_0, no more likely than d_{obs} (or: an event at least as unlikely (so: at least as extreme)):[19]

$$P(D \in \{d \mid P(d \mid H_0) \leq P(d_{\text{obs}} \mid H_0)\})$$

[19] Using a test statistic that does not yield an exact test can be useful for various reasons. We might want to select a particular aspect of the data that we want to measure the extremes of. Or, we might wish to avoid the complexity of calculating an exact p-test and so resort to a clever test statistic that simplifies computation, yet approximates the exact calculation reasonably well (under certain conditions).

Example 6. Grandma's coin was flipped $n = 24$ times and landed heads $k = 7$ times. To calculate the associated exact p-value under the null-hypothesis $\theta = 0.5$, we consult the likelihood function in Figure 6, note the likelihood of the observed data (on the x-axis at $k = 7$) in comparison to all other conceivable outcomes of the flip-$n = 24$-times experiment, and sum the probabilities of all events which are at least as unlikely as the observed $k = 7$. This yields the value $p \approx 0.06391$.

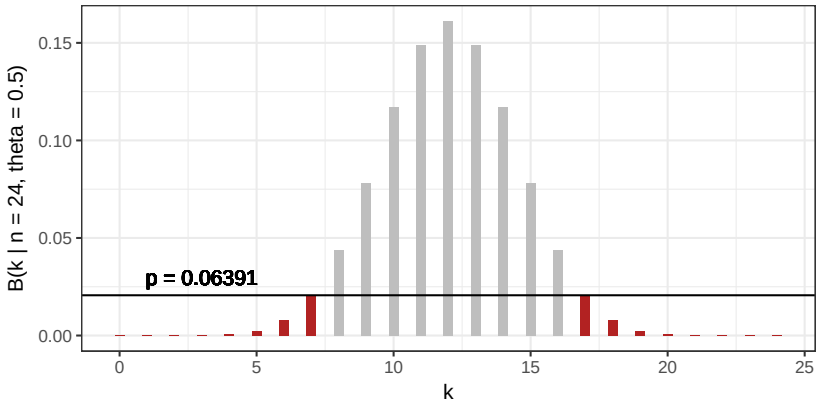

Figure 6: Visualization of the computation of an exact p-value for a binomial likelihood function for a flip-$n = 24$-times experiment, null hypothesis $\theta = 0.5$ and observation $k = 7$.

Statistical significance, decisions & α-error control

Based on a prespecified threshold α, the frequentist analyst makes a simple binary decision. If the p-value of the observed data is no bigger than α, the test result is called (STATISTICALLY) SIGNIFICANT (AT LEVEL α) and the null-hypothesis is rejected. If it is not, the result is not statistically significant and the null-hypothesis is not rejected.[20]

Rejecting a null-hypothesis that is in fact true, is a so-called type-1 error, or α-error. By such an error, the researcher ends up claiming mistaken evidence for the research evidence H_1. To avoid this kind of error, the frequentist approach imposes a strict regime of α-error control. This is achieved by setting the threshold α beneath which a test result is claimed to be significant and therefore a null-hypothesis, to a rather low value. A conventional threshold is $\alpha = 0.05$, but also $\alpha = 0.01$ or $\alpha = 0.001$ are advocated. The

[20] Not rejecting the null-hypothesis is *not* accepting it. The approach sketched so far does not yield evidence *in favor* of the null-hypothesis. Enthusiasts of Bayesian approaches keep repeating that their favorite statistical toy *does* deliver a quantitative notion of evidence which can also turn out in favor of the null-hypothesis.

α level gives an upper-bound on the number of type-1 errors (when each test is performed with significance level α and all tests are executed on their own fresh data set which was collected in the appropriate manner precisely for conducting this test).[21]

[21] These clumsy provisos are essential. Their impracticality is one of the potential reasons for the publication of many false results.

Example 7. By the standard significance threshold $\alpha = 0.05$ we would *not* reject the null hypothesis that grandma's coin is fair in an experiment where we flipped $n = 24$ times and observed $k = 7$ heads. This says nothing about whether or how much we would endorse the null-hypothesis. It just says that we should not brush this possibility off.