Contents

1 Put It to the Test

1.1 A tale of two hypotheses

There are two hypotheses that we want to compare, a null hypothesis h_0 and an alternative hypothesis h_1. Prior to collecting the data that is relevant to the hypotheses we have some beliefs $P(h)$ about which hypotheses are true. We then obtain data d. Unlike frequentist statistics Bayesian statistics does allow us to talk about the probability that the null hypothesis is true, or that the alternative hypothesis is true, for that matter. Better yet, it allows us to calculate the ***posterior probability of the null hypothesis***, using Bayes' rule:

$$P(h_0 \mid d) = \frac{P(d \mid h_0)P(h_0)}{P(d)} \tag{1}$$

Similarly we can, and have, calculated the ***posterior probability of the alternative hypothesis***, using Bayes' rule:

$$P(h_1 \mid d) = \frac{P(d \mid h_1)P(h_1)}{P(d)} \tag{2}$$

We convert these two calculations into one in order to make the comparison.

$$\underbrace{\frac{P(h_1 \mid d)}{P(h_0 \mid d)}}_{\text{Posterior Odds}} = \underbrace{\frac{P(d \mid h_1)}{P(d \mid h_0)}}_{\text{Bayes Factor}} \times \underbrace{\frac{P(h_1)}{P(h_0)}}_{\text{Prior Odds}} \tag{3}$$

There are three terms we can explore.

1. On the left hand side are the deduced the ***posterior odds*** which records the relative compatibility of the hypotheses with the data.

2. On the far right hand side are the ***prior odds***, which indicates what we believed about the two hypotheses *before* applying the data.

3. On the right hand side of the equation and in the middle, is the ***Bayes Factor*** (***BF***), which the relative evidence, that is, the likelihood of data given each hypothesis. This factor quantifies the strength of evidence provided by the data.

Often only the Bayes factor is reported. This is because different analysts might have different views and thus different prior odds surrounding the hypotheses.

There are two examples coming up to put these ideas to work.

1.2 Falsity of a hypothesis

Is *falsity* even a word? TYpical this means some quality of being untrue, even incorrect. In the case of plausibility analysis of a conjecture we do not know if anything is virtually true or false. We do know if a conjecture is consistent with the data, observed, and possibly also supposed (as in priors). So to talk about hypothesis falsification (a much stronger word), or falsity is probably an incorrect way of characterizing the plausible state of a model, a hypothesis, an idea, a conjecture. The question of whether a hypothesis is acceptable as true or not seems to be binary: it's one way or the other. So we then seem to also need to invent some sort of superstructure that encompasses this question of hypothetical alternatives.

Be that as it may, inference is oftem framed in binary terms.

1. An hypothesis is a binary outcome: either the theory is true or it is false.
2. We get a statistical hint, indicator, or cue about the probable falsity of the hypothesis.
3. Our goal is to deduce the impact of the hint or cue on the probable status of the hypothesis. If this frame makes any sense at all, and let's suppose that it does, we (should) use Bayes' theorem to deduce, that is, use a a principled and logical way, the impact of the cue, the indicator, on the status of the hypothesis.

We need all three steps to complete the analysis. We often, if not always, do not complete the last, that third, step.

Example 1.1. Our organization regularly conducts reviews of its research and development activities. Innovation is the key to the success of providing relevant services to our customers. When we observe that our innovations actually meet customer needs, our services reach an ability to service 95% of the customer base. When we do not meet their needs we serve only 5%. The rate at which innovations actually meet customer needs, as customers view them, is only 1%. Yet we persevere! What is the probability of an innovation meeting customer needs given high service levels of 95%?

Here we go.

- Suppose the probability of a positive finding *pos*, that is, the customer service rate, when an hypothesis, that is when we meet customer needs, is true, is $Pr(pos \mid true) = 0.95$. That's more often than not called the ***power of the test***.
- Suppose that the probability of a positive finding *pos*, when an hypothesis is false, is $Pr(pos \mid false) = 0.05$. That's what researchers would call the ***false-positive rate*** .
- The so-called **base rate** is the probability of a true hypothesis, that is, an innovation meets a customer's needs on the customers own terms. We suppose that for 1 in every 100 hypotheses, that is, innovations in our example, turns out to be true. Then $Pr(true) = 0.01$. Does anyone really know what this probability is? The history of science, innovation, and successful start-ups suggests it's small.[1]

Let's compute the posterior.

[1] First, not all ideas, conjectures, innovations are reported, usually only the ones which survive. Even so, on about 15% of new business start ups survive more than 10 years in the US.https://www.bls.gov/bdm/bdmage.htm Second, FDA drug trial rates of success of 13% provide another indicator. Taking both factors jointly into account indicates that under 2% of innovations might make it to market. Third, we might have it wrong as we mix pharmaceuticals with general business start-ups, but at least it is a beginning and so we posit a conservative 1% base rate.

$$
\begin{aligned}
Pr(true \mid pos) &= \frac{Pr(pos \mid true)Pr(true)}{Pr(pos)} && (4)\\
&= \frac{Pr(pos \mid true)Pr(true)}{Pr(pos \mid true)Pr(true) + Pr(pos \mid false)Pr(false)} && (5)\\
&= \frac{Pr(pos \mid true)Pr(true)}{Pr(pos)} && (6)\\
&= \frac{(0.95)(0.01)}{0.01(0.95) + 0.99(0.05)} && (7)\\
&= \frac{0.0095}{0.059} && (8)\\
Pr(true \mid pos) &= 0.16 && (9)
\end{aligned}
$$

The denominator always gets us. It is the weighted average, the expectation, that the likely answer is *pos*, a positive finding. We substitute the relevant values to get this approximate $Pr(true \mid pos) = 0.16$.

So a positive finding corresponds to a 16% chance that the hypothesis is true. This is the same low base-rate phenomenon that applies in medical (and economic policy) testing. You can shrink the false-positive rate to 1% and get this posterior probability up to 0.5, only as good as a coin flip. The most important thing to do is to improve the base rate, $Pr(true)$, and that requires thinking, stewing over, reflection and many iterations of each, not testing.

1.3 Interpreting parameter estimates

A common error in interpretation of parameter estimates is to suppose that because one parameter is sufficiently far from a target level of an indicator (what we might call *significant*) and another parameter estimate (what we might label *not significant*) that the difference between the parameters is also significant. But this is not necessarily so. This isn't just an issue for non-Bayesian analysis: If you want to know the distribution of a difference, then you must compute that difference, a contrast. It isn't enough to just observe, for example, that a slope among males overlaps a lot with zero while the same slope among females is reliably above zero. We must compute the posterior distribution of the difference in slope between males and females.

For example, suppose we have posterior distributions for two parameters, β_f and β_m. We find that β_f's mean adjusted for standard deviation is 0.15 ± 0.02, and β_m's is 0.02 ± 0.10.

So while β_f is reliably different from zero (*significant*) and β_m does not seem to be, the difference between the two (assuming they are uncorrelated) is

$$(0.15 - 0.02) \pm \sqrt{0.02^2 + 0.10^2} = 0.13 \pm 0.10.$$

The distribution of the difference overlaps a lot with zero. In other words, we can be confident that β_f is far from zero, but you cannot be sure that the difference between β_f and β_m is far from zero.

In the context of non-Bayesian significance testing, this phenomenon arises from the fact that statistical significance is inferentially powerful in one way: difference from the null, from the status quo. However, when β_m overlaps with zero, it may also overlap with values very far from zero. Thus its value is highly uncertain. So when we then compare β_m to β_f , that comparison is also uncertain, which shows up in the width of the posterior distribution of the difference $\beta_f - \beta_m$.

Lurking underneath this example is a more fundamental mistake in interpreting statistical significance: the mistake of accepting the null hypothesis as true, at all. Whenever an article, academic or journalistic, or book says something like "we found no difference" or "no effect," this usually means that some parameter was not significantly different from zero, and so the authors might be tempted to adopt zero as the estimate. This is both illogical and extremely common.

To us, one of the biggest advantages to the Bayesian approach is that it answers the right questions, in fact, all of the relevant questions at that. Within the Bayesian framework, it is perfectly sensible and allowable to refer to "the probability that a hypothesis (any including the status quo) is true." We can even try to calculate this probability. Ultimately, isn't that what we *want* our statistical tests to tell ud? To a human being, this would seem to be the whole point of doing statistics: to determine what is true and what isn't, probably, plausibly. Any time that we aren't exactly sure about what the truth is, we should use the language of probability theory to make statements like "there is an 80% chance that Theory A is true, but a 20% chance that Theory B is true instead".

This seems so obvious to a human, yet it is explicitly forbidden within the conventional, framework. To a frequentist, such statements are nonsensical because "the theory is true" is not a repeatable event. A theory is true or it is not, and no probabilistic statements are allowed, no matter how much we might want to make them. It is a one-time binomial outcome with no recourse. We can not interpret the p-value as the probability that the null hypothesis is true. There's a reason why almost every textbook on statistics is forced to repeat that warning. It's because people desperately *want* that to be the correct interpretation.

Frequentist dogma notwithstanding, decades of teaching and consulting suggests to me that most of us think that "the probability that the hypothesis is true" is not only meaningful, it's the thing we care *most* about. It's such an appealing idea that even trained statisticians fall prey to the mistake of trying to interpret a p-value this way. For example, here is a quote from a technical note from the U.S. Bureau of Labor Statistics regarding confidence intervals.

> Statistics based on the household and establishment surveys are subject to both sampling and nonsampling error. When a sample, rather than the entire population,is surveyed, there is a chance that the sample estimates may differ from the true population values they represent. The component of this difference that occurs because samples differ by chance is known as sampling error, and its variability is measured by the standard error of the estimate. There is about a 90-percent chance, or level of confidence, that an estimate based on a sample will differ by no more than 1.6 standard errors from the true population value because of sampling error. BLS analyses are generally conducted at the 90-percent level of confidence.

Also BLS provides this example.

> For example, the confidence interval for the monthly change in total nonfarm employment from the establishment survey is on the order of plus or minus 110,000. Suppose the estimate of nonfarm employment increases by 50,000 from one month to the next. The 90-percent confidence interval on the monthly change would range from -60,000 to +160,000 (50,000 +/- 110,000). These figures do not mean that the sample results are off by these magnitudes, but rather that ***there is about a 90-percent chance that the true over-the-month change lies within this interval.*** Since this range includes values of less than zero, we could not say with confidence that nonfarm employment had, in fact, increased that month. If, however, the reported nonfarm employment rise was 250,000, then all of the values within the 90-percent confidence interval would be greater than zero. In this case, ***it is likely (at least a 90-percent chance) that nonfarm employment had, in fact, risen that month.***

The emphasis is mine in these quotes. This is *not* what a 95% confidence means to a frequentist statistician. The bolded section is wrong. Orthodox methods cannot tell you that "there is a 95% chance that a real change has occurred", because this is not the kind of event to which frequentist probabilities may be assigned. To an ideologically bent frequentist, this sentence should be meaningless.

For hypothesis testing the use of a p-value is often misinterpreted as the probability that the null hypothesis is true. We assume that H_0 is true when we calculate the p-value. Because of this assumption the p-value simply cannot provide information regarding whether H_0 is in fact true.

1. This argument also shows that first, p also cannot be the probability that the alternative hypothesis is true.

2. Second, the p-value is highly sensitive to the sample size.

3. Third, it is not true that the p-value is the probability that any observed difference is simply attributable to the chance selection of observations from the target population.

The $p-value$ is calculated based on an assumption that chance is the only reason for observing any difference. Thus it cannot provide evidence for the truth of any statement.

On the other hand, let's suppose we are Bayesians. Although the bolded passage is the wrong interpretation of a frequentist confidence interval, it's exactly what a Bayesian means when they say that the posterior probability of the finding a parameter value between an lower and an upper value is 89%, suitably redubbed a probability interval. If the Bayesian posterior is actually thing you *want* to report, and what the consumer of your analytical product requires, why are you even trying to use orthodox frequentist methods? If you want to make Bayesian claims, all you have to do is be a Bayesian and use Bayesian tools.

Speaking for myself, I found this to be a the most liberating thing about switching to the Bayesian view many years ago, not ini academia, but in professional consultation with decision makers in complex organizations. Once you've made the jump, you no longer have to wrap your head around counterinuitive definitions of p-values or confidence intervals. You don't have to bother remembering why you can't say that you're 95% confident that the true mean lies within some interval. All we have to do is be honest about what we believed to be true and experienced, including the objective statement – we do not really know!, before we run the study, and then report what we learned from doing it. Sounds nice, doesn't it? To me, this is the big promise of the Bayesian approach: we do the analysis we really want to do, and express what we really believe the data are telling us, all in a principled logically consistent manner.

2 Evidentiary standards you might believe?

> *If [p] is below .02 it is strongly indicated that the [null] hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at .05 and consider that [smaller values of p] indicate a real discrepancy.*
> – Sir Ronald @Fisher1925

Consider the quote above by Sir Ronald Fisher, one of the founders of what has become the orthodox approach to statistics. If anyone has ever been entitled to express an opinion about the intended function of p-values, it's Fisher. In this passage, taken from his classic guide *Statistical Methods for Research Workers*, he's pretty clear about what it means to reject a null hypothesis at $p < .05$. In his opinion, if we take $p < .05$ to mean there is "a real effect", then "we shall not often be astray." This view is hardly unusual: in my experience, most practitioners express views very similar to Fisher's. In essence, the $p < .05$ convention is assumed to represent a fairly stringent evidentiary standard.

How true is this a fairly stringent standard of evidence? One way to approach this question is to try to convert p-values to Bayes factors, and see how the two compare. It's not an easy thing to do because a p-value is a fundamentally different kind of calculation than a Bayes factor. They simply do not measure the same thing at all. However, there have been some attempts to work out the relationship between the two, and it's somewhat surprising. For example, @Johnson2013 presents a compelling case that (for t-tests at least) the $p < 0.05$ threshold rejecting the null corresponds roughly to a Bayes factor of somewhere between 3:1 and 5:1 rejecting the null, and we are allowed to say also accepting the alternative hypothesis. If that's right, then Fisher's claim is a bit of a stretch.

Let's suppose that the null hypothesis is true about half the time (i.e., the prior probability of H_0 is 0.5), and we use those numbers to work out the posterior probability of the null hypothesis given that it has been rejected at $p < .05$. Using the data from @Johnson2013, we see that if you reject the null at $p < .05$, you'll be correct about 80% of the time. I don't know about you, but in my opinion an evidentiary standard that ensures you'll be wrong on 20% of your decisions isn't good enough. The fact remains that, quite contrary to Fisher's claim, if you reject at $p < 0.05$ you shall quite often go astray. It's not a very stringent evidentiary threshold at all.

3 Frequentist vs. Bayesian Inference

3.1 A possibly illuminating example

dWe will solve a simple inference problem using both frequentist and Bayesian approaches. Then we will compare our results based on decisions made with the two methods, to see whether we get the same answer or not. If we do not, we will discuss why that might happen.

Example 3.1. Our organization has a contract to train workers in safe materials handling practices. The materials are highly toxic to people, animals, and plants, and spread toxins both in the air and through water media. Some executives claim that 10% of worker-teams have enough experience not to require further training, while other executives believe that 20% of the worker-teams have the requisite experience. Your department must provide a convincing analysis supporting one or the other claim.

Let's assume these conditions.

- We are being asked to support a decision, and there are associated payoff/losses that we should consider. If we make the correct decision, your gives you a bonus. On the other hand, if we make the wrong decision, we might very well lose our jobs, the stakes are so high.
- However, management knows about the highly likely uncertainty of results and will allow us to be wrong about the 10% rate in 1 in 20 samplings.
- We have funds to draw a random sample of workers from the population. This is no easy task, sinced to sample workers means to observe their practices in person at their work sites. Accessing workers requires expensive travel and lodging expense into very remote regions where physical security may easily be compromised. A 5 person fact-finding team over a 15 day period will entail a consultant cost of $200,000 plus another $300,000 for guides, travel arrangement, licenses, insurance, and other fees for a total of $\500,000, just for this one sampling.

The cost of making a wrong decision is high financially and reputationally, and physically for the worker-teams, the data analytics team and the organization. At the same time, though, data collection is also very costly, with an unknown need for sample size.

3.2 Let's be a frequentist

Let's start with the frequentist inference.

1. We first set out the two opposing hypotheses. One is the **null** hypothesis that only 10% of worker teams are experienced enough in hazardous materials handling practices. The other is that there are more than 10% of worker teams so experienced.

- Hypothesis: H_0 is 10% experienced workers and H_1 is $>$ 10% experienced workers.

2. We then set a tolerance level for being wrong about selecting the null hypothesis as true, when in fact it turns out is is false. A commonly chosen level of signficance is a 1 in 20 time rate of observing that we are wrong about choosing the null hypothesis over the alternative.

- Significance level: wrong about supporting the null hypothesis in only 1 in 20 samplings for an $\alpha = 0.05$.

3. We sample the 5 different worker-teams and find 1 experienced team.

- Observed data: $x = 1$ experienced worker team in $n = 5$ teams.

4. We calculate the probability that the data associated with the alternative hypothesis, namely, that there are 10% or more experienced worker-teams, given the null hypothesis. This is called the $p-value$, *not to be confused with p = the proportion of successes, also known as the probability of a single success.* With the null hypothesis the probability of no experienced worker-teams, in a single sample, is $1 - 0.10 = 0.90$.

- $p - value$: $Pr(x \geq 1 \mid n = 5, p = 0.10) = 1 - Pr(x = 0 \mid n = 5, p = 0.10) = 1 - 0.90^5 \approx 0.41$

We note the clever use of the complement to calculate a bit more easily the 5 sample both-and sequence of no successes at all with each no success 90% probable under the null hypothesis. The $p - value$ is just the probability of observing more extreme outcomes given that the null hypothesis is true. We recall the many admonitions to resist the idea that this is the probability of accepting as true the null hypothesis, or, for that matter, that this is the probability of accepting as false the alternative hypothesis.

We find that the data fails to reject H_0 and we seem to conclude that the data do (plural *data*, singular *datum*) not provide evidence that the proportion of experienced worker-teams is greater than 10%. This means that if we had to pick between 10% and 20% for the proportion of experienced worker-teamss, even though this hypothesis testing procedure does not actually in any way at all confirm the null hypothesis, we would likely stick with 10% since we couldn't find evidence that the proportion of experienced worker-teams is greater than 10%. That this evidence is convincing will depend on whether this conclusion is a statement of the justified true belief of whomever consumes the results of this analysis.

3.3 Now we are Bayesians

The Bayesian inference works differently from the frequentist approach as we illustrate below.

1. We construct a grid of alternative, mutually-exclusive hypotheses. For example, we can start with 5% and work in increments of 5% up to, for example, 50% unobserved proportions of experienced worker-teams in the population. This example would have us list 10 equally spaced vertices (nodes) with 9 edges (intervals). For our immediate purposes we will just use two alternatives to build our comparison: 10% and 20%. This step differs from the null-alternative hypothesis step in that neither of these alternatives is the null, they are agnostically alternative, that's it.

- Unobserved data: H_1 is $p_1 = 10\%$, and H_2 is $p_2 = 20\%$ worker-teams.

2. Initially, we do not have any preference, or expectations for that matter, about the plausibility of one hypothesis over the other. We thus let them be equally probable.

- Prior experience: $Pr(H_1) = Pr(H_2) = 0.50$

3. We sample the 5 different worker-teams and find 1 experienced team. This is exactly the same step as in the frequentist approach.

- Observed data: $x = 1$ experienced worker team in $n = 5$ teams.

4. We will also use the same binomial model as with the frequentist approach. We can justify the use of this model with the notion that we can only observe the binary outcomes of experienced or not experienced (either E or $\overline{E}$ is true) in the repeated sampling (n = 5 here) of worker-teams. In repeated independent samplings of binary data, the appropriate theoretical distribution of successes is the binomial model. Again n is the number of independent samples.

$$E \sim \text{Binomial}(n, p) \tag{10}$$

$$p \sim \text{Uniform(0, 1)} \tag{11}$$

We will simplify this further below to have $\Pr(p) = 0.5$, the mean of the uniform distribution with 0 minimum and 1 maximum.

- The likelihood of observing $x = 1$ experienced worker-team in $n = 5$ samplings. Across hypotheses, likelihoods are mutually exclusive but do not add up to 1.

$$\Pr(x = 1 | H_1 : p_1 = 0.10) = \binom{5}{1} (0.10)^1 \, (0.90)^4 \approx 0.33 \tag{12}$$

$$\Pr(x = 1 | H_2) : p_2 = 0.20) = \binom{5}{1} (0.20)^1 \, (0.80)^4 \approx 0.41 \tag{13}$$

- Posterior $\Pr(H_i : p_i \mid n, x)$ for each of $i = 1, 2$ to infer plausibility of each hypothesis. To force likelihoods into a distribution we would have to normalize them by adding up the likelihoods and compute the contribution of each likelihood to the overall distribution of probability mass across alternative hypotheses. Effectively this is the denominator of the Bayesian posterior distribution formula. We are, after all, solving for the probability that each hypothesis, given the sample data, is true.

$$\begin{align}
\Pr(H_1 : p_1 = 0.10 \mid x = 1) &= \frac{\Pr(H_1)\Pr(x = 1 \mid H_1 : p_1 = 0.10)}{\Pr(x = 1)} \tag{14}\\
&= \frac{(0.50)(0.33)}{(0.50)(0.33) + (0.50)(0.41)} \tag{15}\\
&= 0.44 \tag{16}
\end{align}$$

and, again using the clever complement trick, we have

$$\begin{align}
\Pr(H_2 : p_2 = 0.20 \mid x = 1) &= 1 - 0.45 \tag{17}\\
&= 0.56 \tag{18}
\end{align}$$

The posterior probabilities of whether H_1 or H_2 is correct are close to each other. As a result, with equal priors and a low sample size, it is difficult to make a decision with an unwavering strength of analytical resolve. given the observed data. However, H_2 does have a higher posterior probability than H_1, so if we had to make a decision at this point, we are somewhat justified to choose H_2, i.e., that the proportion of experienced worker-teams is $p_2 = 20\%$ is true is more plausible than the alternative. This choice directly contradicts the decision based on the frequentist approach.

Table 1 summarizes what the results would look like if we had chosen larger sample sizes. Under each of these scenarios, the frequentist method yields a higher p-value than our significance level, so we would fail to reject the null hypothesis with any of these samples. On the other hand, the Bayesian method always yields a higher posterior for the second model where p is equal to 0.20. So the decisions that we would make are contradictory to each other.

```r
options( digits = 2, scipen = 999999)
# we first build two vector lists of observed data
n <- seq(5, 10 , length.out = 2 )
x <- c( 1, 3)
# we specify the false negative rate
null <- 0.1
# here is an example calculation of the p-value interpreted as a probability if in a Bayesian setting
pr_gt_null <- 1 - (1-null)^n
# next the various conjectures as hypotheses about the proportion of successes
nodes <- 2
p_grid <- seq(.1, .2, length.out = nodes)
# equally plausible hypotheses, aka the prior
prior <- 1/nodes
# likelihood <- dbinom(1, n[1], p_grid) * prior
# posterior <- likelihood / sum(likelihood)
post_calc <- function(x, n, p_grid){
  likelihood <- dbinom(x, n, p_grid) * prior
  posterior <- likelihood / sum(likelihood)
  posterior # the last result is what the function returns
}
```

Table 1: Frequentist and Bayesian probabilities for various sample sizes

probability	H	x=1, n=5	x=1, n=10	x=3, n=5	x=3, n=10
Pr(10% experienced | n, x)	0.1	0.44	0.59	0.14	0.22
Pr(20% experienced | n, x)	0.2	0.56	0.41	0.86	0.78
Pr(x or more | 10% experienced)	0.1	0.41	0.65	0.01	0.07

```r
# use tidyr function to make a grid of all combinations of x and n and p_grid
data_grid <- expand_grid( x, n, p_grid )
# first, compose a key and calculate the likelihood
post_table <- data_grid |>
  mutate(
    key = paste0( "x=", x, ", ", "n=", n),
    likelihood = post_calc( x, n, p_grid )
  )
# next by unique data scenario, the key, compute the posterior
post_table <- post_table |>
  group_by( key ) |>
  mutate(
    posterior = likelihood / sum(likelihood)
  ) |>
  select( key, p_grid, posterior ) |>
  pivot_wider( names_from = key, values_from = posterior )
# finally append the null hypothesis row p-values
# finally append the null hypothesis row p-values
data_null <- expand_grid( x, n)
null_table <- data_null |>
  mutate(
    p_value = 1-pbinom( x-1, n, null)
  )
post_table <- post_table |>
  rbind( c( null, t(null_table)[3,] ))
colnames( post_table )[1] <- c("H")
```

However, if we had set up our framework differently in the frequentist method and set our null hypothesis to be $p = 0.20$ and our alternative to be $p < 0.20$, we might obtain different results. What a wonderful exercise to instantiate this claim!

We would have simply exposed the degree of sensitivity of the the frequentist method to the null hypothesis. On the other hand, with the Bayesian method our results would be the same regardless of which order we evaluate our models.

3.4 Exercising our wits

Exercise 3.1. Redo this entire exercise using the idea that results might be different with a frequentist null hypothesis of $H_0 : p_0 = 0.20$. How and why are our conclusions any different? [**Hint**: Reuse the code above. Which variable(s) would you alter?]

Exercise 3.2. Let's try something different than data that uses the binomial distribution. We observe these monthly (successful!) launches of SpaceX rockets: 2, 5, 9. Again, as with the binomial example, we are initially completely agnostic about the existing dietribution of potential launches. Instead of using the binomial distribution, we use a distribution appropriate for count data in mashing hypotheses with data. As frequestists, we want to test the belief that there are 3 launches on average with a significance level of 1 in 20 times that we are wrong about this hypothesis. As a Bayesian we test this belief again against an alternative hypothesis that launch patterns are drawn from a population with mean 6 launches per month. What are the odds that the alternative wins the launch control center's bet that they can beat 3 launches per month? [**Hint:** we might consider the Poisson distribution.]