Bayesian Alternatives to Statistical Significance

Imad Ali[*] Jonah Gabry[†]

February 10, 2019

Abstract

This paper outlines common misconceptions of Frequentist statistical significance and proposes a more intuitive Bayesian alternative. At a high-level Frequentist methods focus on the distribution of the test statsitic as opposed to the parameter of interest, whereas the methods proposed here allow the researcher to perform inference directly on the parameter of interest.

Contents

1 Introduction **2**
1.1 Misinterpreting statistical significance 2

2 Inference on Sample Parameters **2**
2.1 One Sample t-test 2
2.2 Bayesian alternative to the one sample t-test 4
2.3 Bayesian alternative to the two sample t-test 5

3 Inference on Regression Parameters **6**

4 Conclusion **6**

5 References **7**

6 Appendix A - Binomial test **8**

[*]National Basketball Association
[†]Columbia University

1 Introduction

Some literature reviewy stuff.

In the following section we discuss common misconceptions surrounding test statistics, p-values, and generally speaking the misuse of the term statistical significance. We then show an alternative approach under the Bayesian framework to evaluate significant differences in parameters. Finally, we extend this to how we can deal with the significance of parameters in generalized linear regression models.

1.1 Misinterpreting statistical significance

A quantity is defined to be statistically significant if the test statistic computed with this quantity is unlikely to occur under the null hypothesis, where the null hypothesis assumes that there is no material difference between two quantities. What this means is that if we repeatedly draw samples of data from the population under the null hypothesis, compute the same statistic, and plot the empirical distribution of the statistic; then the statistic we computed from our original data would fall somewhere in the tails (i.e. a low probability region). One way to quantify this is to compute the proportion of the distribution that is in the tails from the computed test statistic. This quantity is known as a p-value and tells you the probability of computing a statistic under the null hypothesis that is as extreme as the one at hand. Lower p-values indicate that the computed test statistic is way out in the tail(s) of the distribution.

In this discussion of statistical significance we have been talking about quantities that we, as applied researchers, are not really interested in; test statistics and p-values. We are more interested in the underlying quantities used to compute these statistics. It would be more appropriate to do inteference on quantities that we are actually interested in compared to statistics derived from these quantities.

2 Inference on Sample Parameters

It is easier to talk about the process of inference in context so the one sample t-test is used as an example. First we will show how statistical significance is determined using the test and then we will outline a Bayesian alternative by modeling the data using **rstan**, the R interface to the Stan probabilistic programming language.

2.1 One Sample t-test

The *one sample t-test* is used to test the whether the mean of a sample of data $\bar{x}$ is significantly different from some hypothesized value μ. The test statistic associated with this test assumes that data x is generated from the normal distribution and is calculated as,

$$\text{t-statistic} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

where σ is the standard deviation of the data x and n is the sample size of the data. This statistic follows the t-distribution with $n-1$ degrees of freedom.

Below we have a sample of ten observations generated from $\mathcal{N}(4.5, 2)$.

```
x <- c(5.820883, 2.667825, 3.332511, 3.388233, 7.976444,
       5.925112, 6.465919, 7.064625, 3.012066, 2.771472)
```

Suppose we want to test whether the mean of these data is statistically different from a hypothesized value $\mu = 4.5$. This is what is known as a two-tailed test since we are interested in determining whether the mean is significantly greater than 4.5 or significantly less than 4.5. Applying the formula to the data gives the following t-statistic,

$$\frac{4.84 - 4.5}{2.01/\sqrt{10}} \approx 0.5394$$

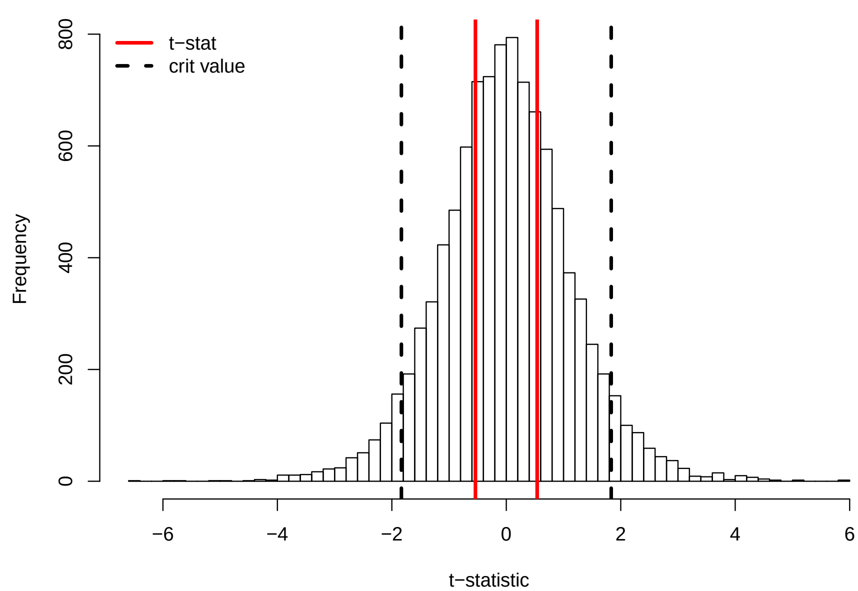
Since we are dealing with a two-tailed test our test statistics of interest are $\mid 0.5394 \mid$ and $- \mid 0.5394 \mid$. As mentioned in the introduction, we want to determine how unlikely it is to compute test statistics more extereme than the ones we have computed. We can do this using the cumulative distribution function for the t-distribution $\Phi(\mu, \nu)$ parameterized by location μ and degrees of freedom ν. Since the t-distribution is symmetric we can compute the probability of calculating a statistic less than or equal to -0.5394 under the null hypotheis and multiply this value by 2. Performing this computation yields,

$$2 \cdot \Phi(0.5394, 10 - 1) \approx 0.6027$$

So there is about a 0.6 probability that a test statistic could occur that is more extereme than the one we computed. In other words, if we repeatedly randomly sample 10 observations under the null hypotheis $x \sim \mathcal{N}(4.5, 2)$ and compute the t-statistic with this data, then 60% of the time we will observe values more extreme than what we calculated with the orginal data sample. That is a high probability and suggests that there may be no material difference between the test statistic computed from the data sample and the test statistics computed under the null hypothesis. Under the Frequentist interpretation, this in turn implies that the mean of our sample is not statistically different from the hypothesized value, 4.5.

The distribution of t-statistics under the null hypothesis can be illustrated by sampling 10 observations from $\mathcal{N}(4.5, 2)$ B times, and plotting a histogram of these B t-statistics. The figure below presents a plot of the distribution where $B = 10000$.

Figure 1: Empirical Distribution of t-Statistic



The red lines indicate the t-statistic computed with the data sample provided above. The dashed black lines indicate the critical values where the proportion of the distribution in the tails from those dashed lines is cumulatively 0.1. If our computed t-statistic was outside the critical values then we could say that there is less than 0.1 probability that we would calculate a statistic as extreme as the one we have calculated, thus indicating that there is a significant difference between the mean from our data and the null hypothesis. But the statistic falls inside the critical value interval, which leads us to conclude that $\bar{\mu}$ is not statistically different from 4.5.

2.2 Bayesian alternative to the one sample t-test

The approach above focused on the test statistic, removing us from the parameter of interest $\bar{\mu}$. The Bayesian alternative to this test involves modeling the data with prior information, and performing inference on the parameter estimate $\bar{\mu}$ directly.

In the context of the t-test assumptions the data is modeled using the normal distribution. The prior distributions on the location and scale parameters are up to the researcher. Here, we define normal distributions on both parameters, which gives us the following model,

$$
\begin{aligned}
x &\sim \mathcal{N}(\mu, \sigma) \\
\mu &\sim \mathcal{N}(0, 3) \\
\sigma &\sim \mathcal{N}^{+}(0, 3)
\end{aligned}
$$

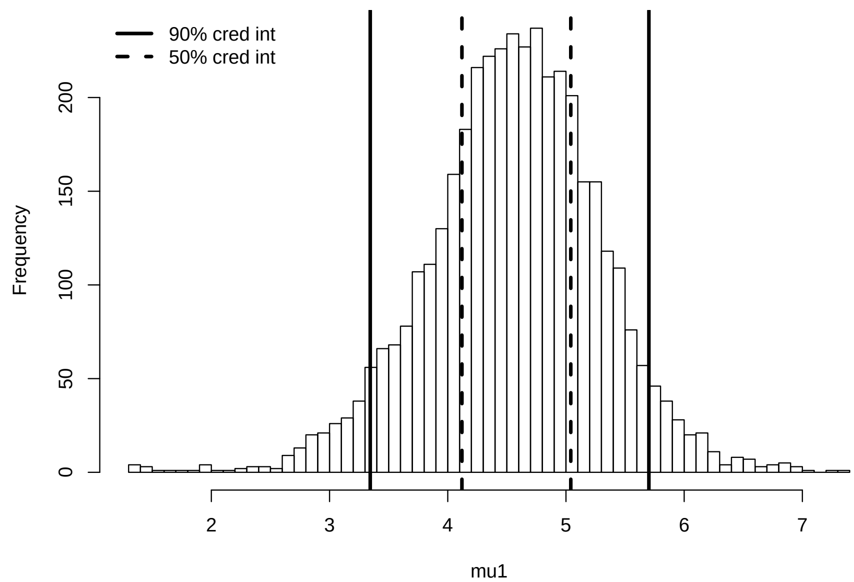
Fitting the model below to the same data sample in rstan gives samples that define the marginal posterior distributions for both μ and σ, along with the posterior predictive distribution for x. This distributions are illustrated in the figure below.

```
data {
  int<lower=0> N;
  vector[N] y;
  real mu_loc;
  real sigma_loc;
  real<lower=0> mu_scale;
  real<lower=0> sigma_scale;
}
parameters {
  real mu;
  real<lower=0> sigma;
}
model {
  target+= normal_lpdf(y | mu, sigma);
  target+= normal_lpdf(mu | mu_loc, mu_scale);
  target+= normal_lpdf(sigma | sigma_loc, sigma_scale);
}
generated quantities {
  real y_hat[N];
  for (n in 1:N)
    y_hat[n] = normal_rng(mu, sigma);
```

```
        }
```

We are still interested in testing if the mean is different from the hypothesized value, 4.5. To do this we can compute sample quantiles using probabilities defined by the researcher. These *uncertainty intervals* enables us to communicate the probability that the estimated mean is between two values conditional on the data and prior information. Looking at the figure below, which illustrates the 50% and 90% credible intervals for the location parameter, we identify that 90% of the marginal posterior is between 3.3 and 5.7. It is at the discretion of the researcher to conclude whether this interval is satisfactory to conclude that the estimated mean is not different from 4.5.

Figure 2: Marginal posterior distribution of μ



Aside from performing inference directly on parameters, an additional benefit with this Bayesian approach is that it enables researchers to quantify their prior information about the parameters in the model. This is particularly useful for small samples compared, where prior information has more influence over the marginal posterior distribution of the location parameter. With larger samples more narrow priors will need to be defined in order for them to have any substantial influence.

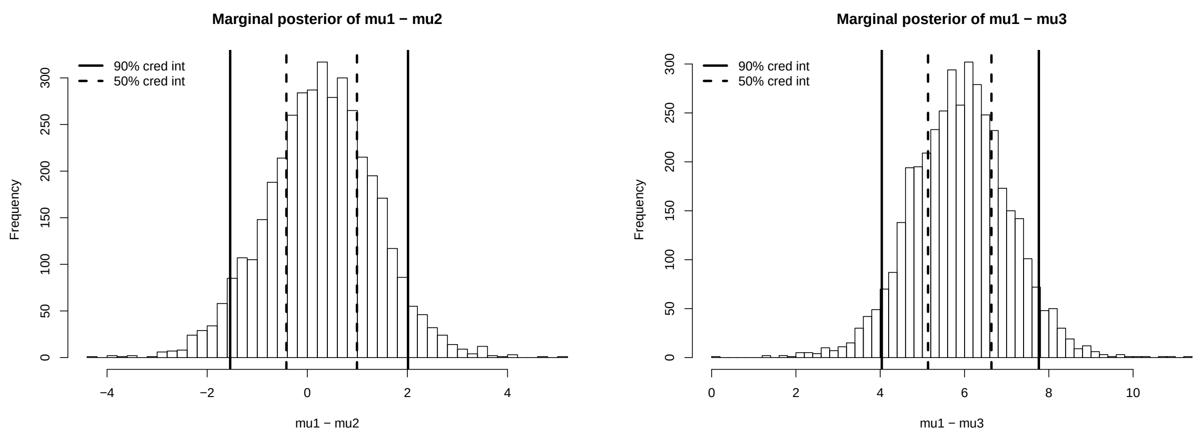
2.3 Bayesian alternative to the two sample t-test

Here we extend the one-sample test defined above to two samples. Suppose we have three data sets (x_1, x_2, and x_3) generated from the normal distribution.

```
x1 <- c(5.820883, 2.667825, 3.332511, 3.388233, 7.976444,
        5.925112, 6.465919, 7.064625, 3.012066, 2.771472)
x2 <- c(6.329095, 4.575028, 1.346890, 6.440587, 7.479409,
        1.736296, 5.091112, 3.865324, 7.728403, 1.792258)
x3 <- c(-2.541337, -2.778205, -1.363428, -4.821875, -2.526202,
        -3.587017, -1.073344, -2.400323,  3.287509,  3.278655)
```

We want to determine whether the mean of x_1 is different from that of x_2 and x_3. Modeling each set of data according to the approach in the previous section gives the marginal posterior distributions for the location parameters (μ_1, μ_2, and μ_3). We can then compute the 50% and 90% quantiles for $\mu_1 - \mu_2$ and $\mu_1 - \mu_3$. For $\mu_1 - \mu_2$ we find that 90% of the difference in location parameters is between -1.5 and 2.0, which means 90% of the time we cannot confidently conclude that the means of the two samples, x_1 and x_2 are different. On the other hand the 90% uncertainty interval for $\mu_1 - \mu_3$ is 4.0 and 7.8 so we can say that 90% of the time there is a difference between the means of x_1 and x_3. The figure below illustrates the uncertainty interval for each of these cases.

Figure 3: Two-sample Inference on Location Parameters



3 Inference on Regression Parameters

4 Conclusion

...

5 References

6 Appendix A - Binomial test

The example outlined above involved the t-test which required normally distributed data. Here we look into inference on count data, specifically data generated from the binomial distribution. In Frequentist statistics the test used to compare the probability of success between some realized data sample and a hypothesized value is the *binomial test*. This is commonly used in **A/B testing** where researchers are interested in the determining whether there is a substantial difference in the click through rate of an online advertising campaign.

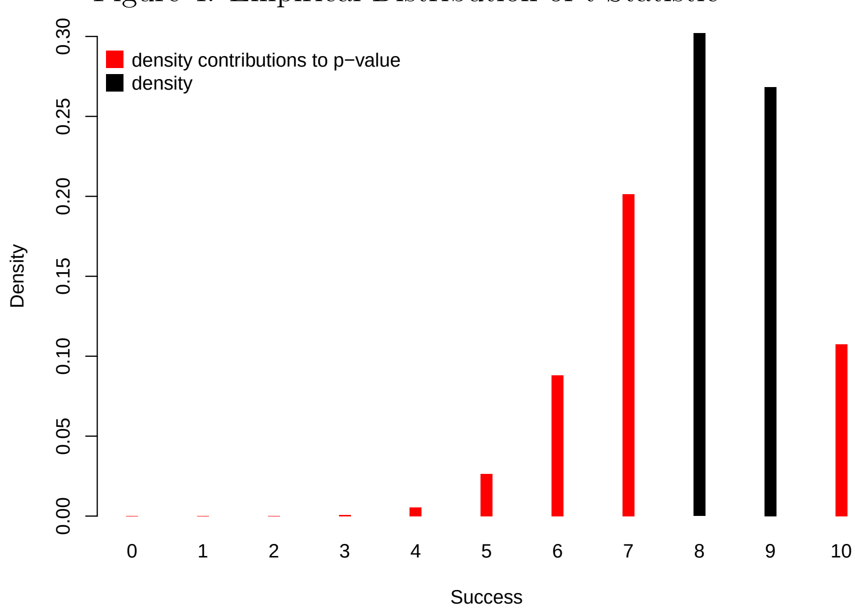
In the context of the binomial test the p-value is calculated as,

$$d = Bin(\hat{x}, N, p_0)$$
$$\text{p-value} = \sum_{x=0}^{N} Bin(k, N, p_0) \text{ s.t. } Bin(x, N, p_0) \leq d$$

where x is the number of successes, $\hat{x}$ is the number of observed successes (data), N is the number of trials, and p_0 is the hypothesized value for the probability of success.

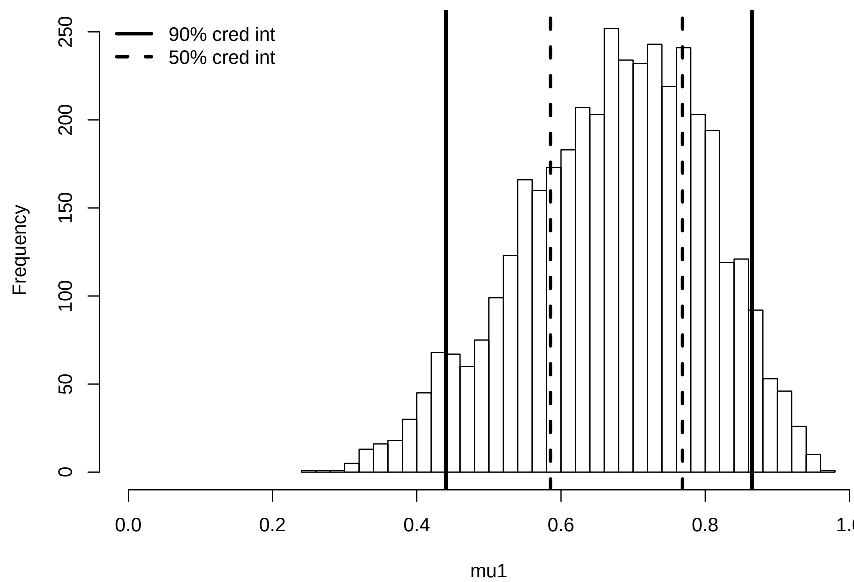
As an example consider $x = 7$, $N = 10$, and $p_0 = 0.8$. Using the method described above we get a p-value of approximately 0.4296, preventing us from rejecting the null hypothesis that the probability of success is different from 0.8. This result is illustrated in the figure below which shows the density evaluated for each success under the hypothesized success rate. The densities that constitute the p-value are in red.

Figure 4: Empirical Distribution of t-Statistic



As mentioned above, the Bayesian approach involves modeling the data with prior information. In this case, the model is $7 \sim Bin(10, p)$ with a prior distribution defined on p (the Beta distribution is typically used). We can then compute the quantiles and determine whether the estimated probability of success p is different from the hypothesized value $p_0 = 0.8$. In this case we cannot conclude that the estimated probability of success is different from 0.8 since 90% the time the parameter is between 0.44 and 0.86, conditional on the data and prior information. This result is illustrated below.

Figure 5: Posterior distribution of p



This approach can easily be extended to multiple observations. The only difference would be to model the vector of successes (instead of a single observation) and a vector of associated trials. The inference on the probability of success parameter would be the same.