Bayesian Inference Assignment 1

1 Poisson Processes with Outliers

Suppose that you are about to collect N independent observations on a count variable. Of those, N_1 are generated by a Poisson distribution with expectation μ_1 and N_2 are generated by a Poisson distribution with expectation $\mu_2 \neq \mu_1$, such that $N = N_1 + N_2$. However, you are assuming all N observations are generated according to the *same* Poisson distribution with expectation μ, in which case you could think of the N_2 observations as unrecognized outliers (with a particular form).

1.1 Sampling Distribution of the Sample Mean

What is the asymptotic distribution — i.e. as $N_1 \uparrow \infty$ and $N_2 \uparrow \infty$ at the same rate — of the sample mean (before you collect the data)?

First, we need to derive the Probability Mass Function (PMF) for the N observations, $\Pr\left(y_1, y_2, \dots y_N \mid \mu_1, \mu_2, N_1, N_2\right)$. We can order the observations however we want without affecting the probability, so assume the first N_1 observations have expectation μ_1 and then the next N_2 observations have expectation μ_2. Treated separately, each group of observations has the same joint distribution that we derived in class. For the first group,

$$\Pr\left(y_1, y_2, \dots y_{N_1} \mid \mu_1\right) = \prod_{n=1}^{N_1} \frac{\mu_1^{y_n} e^{-\mu_1}}{y_n!} = e^{-N_1\mu_1} \mu_1^{\sum_{n=1}^{N_1} y_n} \prod_{n=1}^{N_1} \frac{1}{y_n!} = \frac{e^{-N_1\mu_1}\mu_1^{S_1}}{S_1!}$$

where $S_1 = \sum_{n=1}^{N_1} y_n$. Similarly, for the second group,

$$\Pr\left(y_{N_1+1}, y_{N_1+2}, \dots y_{N_1+N_2} \mid \mu_2\right) = \frac{e^{-N_2\mu_2}\mu_2^{S_2}}{S_2!}$$

where $S_2 = \sum_{n=N_1+1}^{N_1+N_2} y_n$. So, the joint PMF of all N random variables can be written as a product of Poisson PMFs for the two sums:

$$\Pr\left(y_1, y_2, \dots y_N \mid \mu_1, \mu_2, N_1, N_2\right) = \frac{e^{-N_1\mu_1}\mu_1^{S_1}}{S_1!} \times \frac{e^{-N_2\mu_2}\mu_2^{S_2}}{S_2!}$$

However, the sum of two Poisson random variables with different expectations is also Poisson with expectation equal to the sum of the two constituent expectations, which is asserted without proof on Wikipedia but can be proven in various ways if you wanted. For example, in terms of the notation above,

$$\Pr\left(S_1 + S_2 = k \mid N_1\mu_1, N_2\mu_2\right) = \sum_{i=0}^{k} \Pr\left(S_1 + S_2 = k, S_1 = i \mid N_1\mu_1, N_2\mu_2\right) =$$

$$\sum_{i=0}^{k} \Pr\left(S_2 = k - i \mid N_2\mu_2\right) \times \Pr\left(S_1 = i \mid N_1\mu_1\right) = \sum_{i=0}^{k} \frac{e^{-N_1\mu_1}\left(N_1\mu_1\right)^{k-i}}{(k-i)!} \frac{e^{-N_2\mu_2}\left(N_2\mu_2\right)^{i}}{i!} =$$

$$e^{-N_1\mu_1 - N_2\mu_2} \sum_{i=0}^{k} \frac{\left(N_1\mu_1\right)^{k-i}\left(N_2\mu_2\right)^{i}}{i!\,(k-i)!} = \frac{e^{-N_1\mu_1 - N_2\mu_2}}{k!} \sum_{i=0}^{k} \frac{k!}{i!\,(k-i)!} \left(N_1\mu_1\right)^{k-i}\left(N_2\mu_2\right)^{i} =$$

$$\frac{e^{-N_1\mu_1 - N_2\mu_2}}{k!} \sum_{i=0}^{k} \binom{k}{i} \left(N_1\mu_1\right)^{k-i}\left(N_2\mu_2\right)^{i} = \frac{e^{-N_1\mu_1 - N_2\mu_2}}{k!} \left(N_1\mu_1 + N_2\mu_2\right)^{k}$$

where the last step follows from the Binomial Theorem.

Thus, if we let $S = S_1 + S_2$, then S is distributed Poisson with expectation (and variance) $N_1\mu_1 + N_2\mu_2$ for any finite N_1 and N_2. Moreover, as $N_1 \uparrow \infty$ and $N_2 \uparrow \infty$ at the same rate, this expectation approaches infinity and the Poisson distribution converges to the normal, as we indicated during class. Therefore, the sample mean $\overline{y} = \frac{S}{N_1+N_2}$ asymptotically converges to a normal distribution with expectation $\frac{N_1\mu_1+N_2\mu_2}{N}$ and variance $\frac{N_1\mu_1+N_2\mu_2}{N^2}$.

1.2 Power of the Test

A consistent test is one whose power to reject a false null hypothesis approaches 1 asymptotically. Suppose you assume a value for μ to use for a null hypothesis and it is the case that $\mu = \mu_1$. Under the true data generating process outlined above with $\mu_1 \neq \mu_2$ is the test consistent?

If $\overline{y} = \frac{S}{N_1+N_2}$ is asymptotically normal distribution with expectation $\frac{N_1\mu_1+N_2\mu_2}{N}$, then $\overline{y} - \mu_1$ is asymptotically normal with expectation $\frac{N_1\mu_1+N_2\mu_2}{N} - \mu_1 = \frac{(N_1-N)\mu_1+N_2\mu_2}{N} = \frac{N_2(\mu_2-\mu_1)}{N} \neq 0$. Under the null hypothesis, the variance of $\overline{y}$ is $\frac{\mu_1}{N^2}$ so it is not necessary to *estimate* the variance from the data as in a conventional t-test. However, the sample variance is a consistent estimator of the population variance $(N_1\mu_1 + N_2\mu_2)$ so you could utilize that fact to obtain essentially the same conclusion as what follows. Under the null hypothesis, the test statistic $\sqrt{\frac{N}{\mu_1}}\left(\overline{y} - \mu_1\right)$ is asymptotically standard normal, but the null hypothesis is false so $\sqrt{\frac{N}{\mu_1}}\left(\overline{y} - \mu_1\right)$ is asymptotically normal with non-zero expectation $\frac{N_2(\mu_2-\mu_1)}{N}\sqrt{\frac{N}{\mu_1}} = \frac{N_2(\mu_2-\mu_1)}{\sqrt{N\mu_1}}$ and finite variance $\frac{N_1\mu_1+N_2\mu_2}{N^2}\frac{N}{\mu_1} = \frac{N_1\mu_1+N_2\mu_2}{N\mu_1}$. As $N_1 \uparrow \infty$ and $N_2 \uparrow \infty$ at the same rate, this expectation diverges if $\mu_1 \neq \mu_2$ because N_2 increases at a faster rate than does $\sqrt{N_1 + N_2}$. Thus, the probability that the magnitude of the test statistic exceeds 1.96 approaches one and the test is consistent. However, in finite samples, the power of the test may be rather poor if N_2 is small and / or $\mu_1 \approx \mu_2$.

1.3 Posterior Distribution

Now suppose your hypothesized value for μ in the previous subproblem is given by $\frac{a}{b}$ where a is the shape parameter of a Gamma prior distribution and b is the rate parameter of a Gamma prior distribution. What is the posterior distribution conditional on the N observations that are generating according to the process outlined above, but you assume it to be a homogenous Poisson data-generating process with a single expectation?

The PDF of the Gamma prior is

$$f\left(\mu \mid a, b\right) \propto \mu^{a-1}e^{-b\mu}$$

Under an assumed homogenous data-generating process, the *ex-ante* probability of observing the data is the same as we found during class

$$\Pr\left(y_1, y_2, \ldots, y_N \mid \mu\right) = \frac{\mu^S e^{-N\mu}}{S!}$$

The posterior PDF of μ is proportional to the product of these two terms:

$$f\left(\mu \mid a, b, y_1, y_2, \ldots y_N\right) \propto \mu^{a-1}e^{-b\mu}\mu^S e^{-N\mu} = \mu^{a+S-1}e^{-(b+N)\mu} = \mu^{a^*}e^{-b^*\mu}$$

where $a^* = a + S$ and $b^* = b + N$. Thus, the posterior PDF has the same kernel as the Gamma prior but with updated shape a^* and rate b^*. In other words, the posterior distribution is Gamma with shape a^* and rate b^*.

1.4 Summary

Based on your answers to the previous subproblems, what would you conclude about the effect of (these) outliers to Frequentist and Bayesian methods in this particular case?

In the Frequentist case, we will asymptotically reject the null hypothesis that the expectation is μ_1 and the sample mean is a consistent and asymptotically normal estimator of the true expectation $\frac{N_1}{N}\mu_1 + \frac{N_2}{N}\mu_2$ whose variance is $\frac{\frac{N_1}{N}\mu_1 + \frac{N_2}{N}\mu_2}{N^2}$

In the Bayesian case, the posterior distribution is Gamma with shape a^* and rate b^*, which has an expectation of $\frac{a^*}{b^*} = \frac{a+S}{b+N}$ and a variance of $\frac{a^*}{(b^*)^2} = \frac{a+S}{(b+N)^2}$ As $N \uparrow \infty$, the Gamma distribution converges to a normal with expectation $\overline{y}$ and variance $\frac{\overline{y}}{N}$, so the posterior distribution is asymptotically the same as the sampling distribution of the sample mean.

It does not follow that Frequentists should have *beliefs* about the population expectation that are normal with expectation $\overline{y}$ and variance $\frac{\overline{y}}{N}$. All that is strictly justified is concluding that the population expectation is not μ_1. However, in practice, if an applied researcher believed so after collecting the data, it would not be too costly since that is equivalent in this case to a Bayesian analysis with an (improper) Gamma prior with shape $a = 0$ and rate $b = 0$. But, in most other situations, there is not as neat of an equivalence between the sampling distribution of a Frequentist estimator and the posterior distribution for some prior and likelihood.

Conversely, the posterior expectation has good Frequentist properties over datasets of size N that have been randomly sampled from a population. Some people feel that this is relevant, although Bayesians have no reason to care about the properties of some point estimator obtained from a posterior distribution across datasets since the entire posterior distribution provides a complete characterization of beliefs conditional on the one dataset that has actually been collected. Nevertheless, it is not the case that posterior expectations always have good Frequentist properties, although it can be shown that they do in a reasonably wide number of situations.

Thus, the presence of outliers does not have that much effect on Frequentist or Bayesian conclusions about the population expectation in this case. However, the predictions of future data would be underdispersed relative to the actual future data from this mixture of Poisson distributions.

2 Current Population Survey

2.1 Frequentist Perspective

If one wanted to make the case for applying conventional Frequentist techniques to the unemployment rate, it might go something like this:

- Each month, the Bureau of Labor Statistics conducts the Current Population Survey (CPS) to estimate the unemployment rate, which is defined as the ratio of people who are unemployed divided by the number of people in the labor force. In turn, the number of people in the labor force is defined as the number of people who have jobs plus the number of people who do not have jobs but have actively looked for a job in the past few weeks.
- The CPS is *not* a simple random sample from the population of adults who live in the United States, but with complicated weighting schemes it is possible to make the samples (collectively) representative of the population.
- If the unemployment rate in the population (U_t) in month t is at (U^*) — the non-accelerating inflation rate of unemployment (NAIRU) — then the distribution of the estimated unemployment rate in each month's CPS is binomial with probability (U^*) and size equal to the number of people in the labor force in the CPS.

- Thus, in any particular month, one could perform a test of the null hypothesis that $U_t = U^*$ using the `binom.test` function in R with `x` being the number of (weighted) unemployed people in the CPS, `n` being the number of people in the labor force in the CPS, and `p` being U^*
- We would expect such a test to fail to reject the null hypothesis in 19 out of 20 months when $U_t = U^*$ at the $\alpha = 0.05$ level.

For some research area that you are familiar with, make the best case you can for applying conventional Frequentist practice to some other statistic besides the unemployment rate. Then, criticize your arguments from an applied statistics perspective.

The unemployment rate example is more amenable to Frequentist analysis than most situations where Frequentist analysis is applied. The sampling process should be repeated (or at least repeatable) and someone has to care about the distribution of a statistic over datasets that have been randomly sampled from a population. In this case, there is a well-defined population (all adults in the United States), a sampling mechanism that produces a different sample each month, and the Federal Reserve Board takes the unemployment rate into account (along with inflation) when deciding on monetary policy. Nevertheless, even in this case, there are several factors that preclude such a Frequentist from being useful and it does not seem to actually be conducted.

First, in order to test a null hypothesis that $U_t = U^*$, you have to know what value U^* takes. But we do not know what the NAIRU is, and economists have to estimate it. These estimates come from models that are disputed, and even if there were a consensus about what model to use, the NAIRU can change over time. So, at best we have an estimate of what the NAIRU was at time t that comes from the same data, which makes it complicated to derive the null distribution.

A related point is that we do not know whether the economy is generally or ever in at state where $U_t = U^*$. It could be the case that U_t oscillates between being less than the NAIRU for a stretch of months and then crosses over to being greater than the NAIRU. In other words, the NAIRU may not be a stable equilibrium, in which case what good is having a test that is expected to fail to reject the null hypothesis that $U_t = U^*$ in 19 out of 20 months if that null hypothesis is typically (at least slightly) false?

Third, the estimated unemployment rates are revised three times as more data become available. Thus, you could easily (fail to) reject the null hypothesis based on the initial data and later reach the opposite conclusion with the revised data even if the (estimated) NAIRU remains the same. The Frequentist interpretation of probability — as the limit of the ratio of the number of times something occurs to the number of random trials — is problematic when you do not know for sure what the numerator is due to measurement error, missing data, etc. Even if you believe the revised data eventually gets the unemployment rate correct, it may be too late to conduct monetary policy effectively.

Fourth, the (revised) unemployment rate "data" are not really the raw data the Frequentist thought experiment imagines. On one hand, the Federal Reserve Board is more interested in the "seasonally adjusted unemployment rate", which is a somewhat smoothed estimate of what the unemployment rate will be for the next year. Every December retail stores hire a lot of short-term workers which nominally causes the raw unemployment rate to plummet. And every January those short-term workers are let go which nominally causes the raw unemployment rate to spike. But there is often no change in the trend of the economy from December to January and the observed data is largely an artifact of the timing of the holiday season. The seasonal adjustment weights also have to be estimated from models and thus are subject to dispute and changes over time. The final revised unemployment rates are "benchmarked" using other imprecisely measured data such as Gross Domestic Product, which is ostensibly produced by employed workers.

Fifth, even if you reject the false null hypothesis that $U_t = U^*$, it does not do the Federal Reserve Board any good to conclude that the unemployment rate is "not at the NAIRU". The Federal Reserve Board needs to know things like

1. Is the unemployment rate less than or greater than the NAIRU?
2. How much does the unemployment rate differ from the NAIRU?

3. For how long has the unemployment rate been this way?
4. What is the difference between the unemployment rate and the NAIRU predicted to be in the future?

The Frequentist approach does not provide an answer to any of these questions. Remember, just because you reject the null hypothesis that $U_t = U^*$ in favor of the alternative hypothesis that $U_t \neq U^*$ does not mean you have any justification for *believing* the unemployment rate is U_t. Indeed, U_t is just a single number rather than a probability distribution that would characterize uncertain beliefs about the unemployment rate.

As far as I know, the Federal Reserve Board does not bother to test a null hypothesis that $U_t = U^*$. None of the articles that are written in the financial press and blogs discuss the unemployment rate in Frequentist terms because none of the people in their target audience utilize the Frequentist definition of probability. Rather, all of the discussion centers around someone's degree of belief that the unemployment rate is too high or too low relative to the NAIRU and how the Federal Reserve Board should respond. Alan Greenspan, who was the Chair of the Federal Reserve Board from 1987 to 2006 has said explicitly in 2004

> As a consequence, the conduct of monetary policy in the United States has come to involve, at its core, crucial elements of risk management. This conceptual framework emphasizes understanding as much as possible the many sources of risk and uncertainty that policymakers face, quantifying those risks when possible, and assessing the costs associated with each of the risks. In essence, the risk management approach to monetary policymaking is an application of Bayesian decisionmaking.

Other chairs of the Federal Reserve Board have not used the word "Bayesian" but have made statements that are consistent with Greenspan's. Thus, the Federal Reserve Board seems to not approve of a Frequentist approach to analyzing the unemployment rate.

2.2 Getting CPS Data

For the rest of this problem, you are going to need to download some CPS data. The easiest way to do that is to go to

http://ceprdata.org/cps-uniform-data-extracts/cps-outgoing-rotation-group/cps-org-data/

and download the compressed file for the year that you were born to the same directory as your RMarkdown file. Then, unzip that file in the same directory to produce a Stata formatted file, which will have a dta file extension. The Stata formatted file can be loaded into R properly with something like

```r
library(haven)
CPS <- as_factor(read_dta(dir(pattern = "^cepr_.*dta$")))
# as_factor changes the categorical variables in Stata to R factors
```

Finally, filter the `CPS` data.frame down to the month that you were born using the `month` variable so that it is not too big.

A brief description of the variables in `CPS` and the values they take (if categorical) can be obtained by

```r
defs <- sapply(CPS, FUN = attr, which = "label")
vals <- sapply(CPS, FUN = attr, which = "levels")
```

Additional documentation of these variables can be found at http://ceprdata.org/cps-uniform-data-extracts/cps-basic-programs/cps-basic-documentation/ or the from the links on that page but note that `CPS` does not include the household-level variables and recodes / combines / renames some of the individual-level variables. If you are familiar with Stata, it might be helpful to look at the dofiles that create the dataset which can be found at http://ceprdata.org/cps-uniform-data-extracts/cps-outgoing-rotation-group/cps-org-programs/ .

The rest of this analysis will use the December 2018 CPS data:

```r
CPS <- CPS[CPS$month == 12, ]
```

2.3 Selection

One of the criticisms labor microeconometricians make of models of whether an individual is (un)employed is that they take whether the person is in the labor force as an exogenous known instead of attempting to develop a model that explains both why some people are (not) in the labor force and why some people in the labor force are (not) employed. Treating whether someone is in the labor force as an exogenous known is harmful if any of the unmeasured variables that affect whether someone is in the labor force also affect the probability that such people could find jobs if they sought them.

Write a short (about 30 lines) function in the Stan language that draws from the prior predictive distribution for a model that avoids this criticism. It should return a matrix with S rows (the number of simulated draws) and N columns (the number of individuals) where each cell is one of the following values:

- -1 if the person is simulated to be not in the labor force
- 0 if the person is simulated to be in the labor force but simulated to not have a job
- 1 if the person is simulated to be in the labor force and simulated to have a job

In doing so, you should assume that the two error terms — one that affects the probability that the person is (not) in the labor force and the other that affects the probability that the person does (not) have a job conditional on them being in the labor force — are distributed bivariate standard normal with exogenous unknown correlation ρ (but $\mu_1 = 0 = \mu_2$ and $\sigma_1 = 1 = \sigma_2$). There was a Stan function to draw from a bivariate normal distribution in Week 1. You can use whatever variables you like from the `CPS` dataset to form the linear predictors and you can use whatever priors you like for the intercept and slope(s) that comprise those linear predictors.

For each iteration s up to the S that you specify, you should draw a realization of the exogenous unknowns from their respective prior distributions and use those realizations to form a linear predictor for (not) being in the labor force and another linear predictor for (not) being employed if that person *were* in the labor force. Note that these two linear predictor vectors are both of size N because each person has both propensities irrespective of whether or not they actually *are* in the labor force. Then, on iteration s, for each person n up to N (the number of people in the CPS that month), you should draw two error terms from the bivariate standard normal distribution with correlation $\tilde{\rho}$ and add these two error terms to the respective linear predictors for person n. If the sum of the first linear predictor for person n and their error realization is less than zero, then the person is simulated to be not in the labor force and you can just put a -1 into the s-th row and n-th column of the output matrix and move onto the next person. Conversely, if the sum of the first linear for person n and their error realization is greater than zero, then person is simulated to be in the labor force in which case you need to consider the sum of the second linear predictor for person n and the second error term. If that is less than zero, then the person is simulated to be unemployed, and if that is greater than zero, then the person is simulated to have a job. You are going to need `if` and `else` clauses inside the double loop of your Stan program to handle this, but those work the same way in Stan as in R and almost all other computer languages.

Be sure to upload your Stan program along with your RMarkdown file and the HTML or PDF file it generates when you turn in your homework.

Here is a progression of increasingly complex Stan programs that is intended to illustrate good practice. But you only needed to turn in your final model for the prior predictive distribution.

First, I copied the `binormal_rng` function from Week 1 and modified it to return a row vector of size two, rather than a matrix with S rows and two columns.

```stan
functions {
  // modified from Week 1 to only return a row vector of size 2
  row_vector binormal_rng(real mu_X, real mu_Y, real sigma_X, real sigma_Y, real rho) {
    real beta = rho * sigma_Y / sigma_X;
    real sigma = sigma_Y * sqrt(1 - square(rho));
    real x = normal_rng(mu_X, sigma_X);
    real y = normal_rng(mu_Y + beta * (x - mu_X), sigma);
    return [x, y];
  }

  // selection_rng function must follow binormal_rng in order to call binormal_rng
}
```

Then, I wrote a `selection_rng` whose only unknowns are the intercept for the decision to be in the labor force, and if so, the intercept for having a job. For now, I assume that the errors in the two parts of the model are uncorrelated.

```
  matrix selection_rng(int S, int N, vector loc, vector scal) {
    matrix[S, N] draws;
    for (s in 1:S) {
      real alpha_1 = normal_rng(loc[1], scal[1]);
      real alpha_2 = normal_rng(loc[2], scal[2]);
      for (n in 1:N) {
        row_vector[2] epsilon = binormal_rng(0, 0, 1, 1, 0);
        if ((alpha_1 + epsilon[1]) < 0) draws[s, n] = -1;
        else draws[s, n] = (alpha_2 + epsilon[2]) > 0;
      }
    }
    return draws;
  }
```

```r
rstan::expose_stan_functions("selection1.stan")
```

Then, the question becomes what values should I assume for the (normal) priors on these two intercepts? In order to obtain a labor force participation rate of about two thirds, the prior mean for the first intercept should be about

```r
qnorm(2 / 3)
```

```
## [1] 0.4307273
```

Similarly, in order to obtain an unemployment rate of about 5% (among people in the labor force), the prior mean for the second intercept should be about

```r
qnorm(19 / 20)
```

```
## [1] 1.644854
```

The prior standard deviations can be rather small, since I am pretty sure that these two intercepts are about right (even though the unemployment rate has been slightly below 4% in recent times). We can now draw S times from this prior predictive distribution:

```r
S <- 10000
```

```r
draws <- selection_rng(S = S, N = nrow(CPS),
                       loc = qnorm(c(2 / 3, 19 / 20)), scal = c(0.1, 0.1))
mean(draws != -1)                    # labor force participation rate
```

```
## [1] 0.6662408
```

```r
mean(draws == 0) / mean(draws != -1) # unemployment rate
```

```
## [1] 0.05091385
```

Next, I added positively correlated errors to the previous model. Since the error term primarily reflects the productivity of the individual, all economists believe the correlation should be positive but perhaps disagree about its magnidue. Nevertheless, the correlation should not be especially close to 1, so I use a uniform prior between 0 and 0.9.

```
  matrix selection_rng(int S, int N, vector loc, vector scal) {
    matrix[S, N] draws;
    for (s in 1:S) {
      real alpha_1 = normal_rng(loc[1], scal[1]);
      real alpha_2 = normal_rng(loc[2], scal[2]);
      real rho = uniform_rng(0, 0.9);
      for (n in 1:N) {
        row_vector[2] epsilon = binormal_rng(0, 0, 1, 1, rho);
        if ((alpha_1 + epsilon[1]) < 0) draws[s, n] = -1;
        else draws[s, n] = (alpha_2 + epsilon[2]) > 0;
      }
    }
    return draws;
  }
```

```
rstan::expose_stan_functions("selection2.stan")
```

```
draws <- selection_rng(S = S, N = nrow(CPS),
                       loc = qnorm(c(2 / 3, 19 / 20)), scal = c(0.1, 0.1))
mean(draws != -1)                    # labor force participation rate
```

```
## [1] 0.6657471
```

```
mean(draws == 0) / mean(draws != -1) # oops, now unemployment rate is 2%
```

```
## [1] 0.0234138
```

Due to the positive correlation in the errors, the predicted unemployment rate is much too low. Thus, I need to tweak the prior mean and perhaps the prior standard deviation for the second intercept:

```
draws <- selection_rng(S = S, N = nrow(CPS),
                       loc = qnorm(c(2 / 3, 18 / 20)), scal = c(0.1, 0.15))
mean(draws != -1)                    # labor force participation rate
```

```
## [1] 0.6660062
```

```
mean(draws == 0) / mean(draws != -1) # unemployment rate is back near 5%
```

```
## [1] 0.05339524
```

Then, I add a predictor based on the gender of the individual, which has an unknown “slope” even though it only has two possible values in the CPS.

```
  matrix selection_rng(int S, int N, vector female,
                       vector loc, vector scal) {
    matrix[S, N] draws;
    for (s in 1:S) {
      real alpha_1 = normal_rng(loc[1], scal[1]);
      real alpha_2 = normal_rng(loc[2], scal[2]);
      real beta_1  = normal_rng(loc[3], scal[3]);
      real beta_2  = normal_rng(loc[4], scal[4]);
      real rho = uniform_rng(0, 0.9);
      for (n in 1:N) {
        row_vector[2] epsilon = binormal_rng(0, 0, 1, 1, rho);
        real eta_1 = alpha_1 + female[n] * beta_1 + epsilon[1];
        if (eta_1 < 0) draws[s, n] = -1;
        else {
          real eta_2 = alpha_2 + female[n] * beta_2 + epsilon[2];
          draws[s, n] = eta_2 > 0;
        }
      }
    }
    return draws;
  }
```

```
rstan::expose_stan_functions("selection3.stan")
```

If the `female` variable were passed to `selection_rng` without centering, then the interpretation of `alpha_1` and `alpha_2` would change to be the intercepts for men, which are shifted by `beta_1` and `beta_2` respectively. That would be fine, although we would have to adjust the priors. Instead, I pass a centered version of `female`, which now implies `alpha_1` and `alpha_2` are the conditional expectations of the linear predictions for someone of "average" gender. This should not be interpreted as someone who does not identify with one of the binary gender categories but rather as an approximately equally weighted mixture of males and females. Nevertheless, the prior for these *conditional* expected linear predictors should not be that different from the prior for the *marginal* expected linear predictors we were using previously, although we could perhaps shrink the prior standard deviations a small amount. The "average" person in the labor force has a very small chance of being unemployed at a particular time, while people whose productivity is well below average have a probability of being unemployed that is much higher than 0.05.

Although it might differ in particular industries (or other countries), most companies seem to be indifferent between hiring equally qualified men and women, so I use a prior mean of zero for β_2 with a small standard deviation. However, a substantial number of well-qualified women are not in the labor force due to parental demands on their time. Thus, I use a prior mean of -0.4 on β_1. After a bit of tweaking, I arrived at

```
draws <- selection_rng(S = S, female = with(CPS, female - mean(female)),
                       loc = c( qnorm(c(2 / 3, 18.25 / 20)), beta_1 = -0.4, beta_2 = 0),
                       scal = c(alpha_1 = 0.1, alpha_2 = 0.15, beta_1 = 0.1, beta_2 = 0.1))
mean(draws != -1)                    # overall labor force participation rate
```

```
## [1] 0.663037
```

```
mean(draws == 0) / mean(draws != -1) # overall unemployment rate
```

```
## [1] 0.04728041
```

Furthermore, we can investigate the labor force participation rate and the unemployment rate by gender:

```
mean(draws[, CPS$female == 1] >= 0)  # labor force participation rate for women
```

```
## [1] 0.5943666
```

```
mean(draws[, CPS$female == 0] >= 0)  # labor force participation rate for men
```

```
## [1] 0.7376249
```

But this model is not quite right because it is primarily women with children who are married (to relatively high-wage spouses) that face a decision of whether to drop out of the labor force. Let's create an indicator variable for that situation in the `CPS` data.frame:

```
CPS$fmwc <- with(CPS, female == 1 & married == 1 & ownchild > 0)
CPS <- CPS[!is.na(CPS$fmwc), ] # was missing for "non-primary" survey respondents
```

We can use the same Stan function as before (although `fmwc` is being passed rather than `female`), but the magnitudes of the priors need to be adjusted again.

```
draws <- selection_rng(S = S, female = with(CPS, fmwc - mean(fmwc)),
                       loc = c( qnorm(c(2 / 3, 18.1 / 20)), beta_1 = -1, beta_2 = 0),
                       scal = c(alpha_1 = 0.1, alpha_2 = 0.15, beta_1 = 0.1, beta_2 = 0.1))
mean(draws != -1)                    # overall labor force participation rate
```

```
## [1] 0.663334
```

```
mean(draws == 0) / mean(draws != -1) # overall unemployment rate
```

```
## [1] 0.05208016
```

```
mean(draws[, CPS$fmwc == 1] >= 0)  # labor force participation rate for married women w/ children
```

```
## [1] 0.3211583
```

```
mean(draws[, CPS$fmwc == 0] >= 0)  # labor force participation rate for everyone else
```

```
## [1] 0.7006901
```

Next, we can consider the educational attainment of the person, which positively affects the probability that they will be in the labor force and have a job, and thereby decreases the error correlation. Using people who have less than a high school education as the reference category, the CPS implies the following categories for the `educ` variable:

1. (HS): High school graduate but no college
2. (SC): Some college experience
3. (CG): College graduate but not advanced degree
4. (GD): Graduate degree

The coefficients on these variables should be positive and increasing and would seem to be larger for the employment part of the model than the labor force participation part, although there is some uncertainty and perhaps an interaction with the `fmwc` variable that I will ignore for now.

```
  matrix selection_rng(int S, vector fmwc,
                       vector HS, vector SC, vector CG, vector GD,
                       vector loc, vector scal) {
    int N = rows(fmwc);
    matrix[S, N] draws;
    for (s in 1:S) {
      real alpha_1 = normal_rng(loc[1], scal[1]);
      real alpha_2 = normal_rng(loc[2], scal[2]);
      real beta_1  = normal_rng(loc[3], scal[3]);
      real beta_2  = normal_rng(loc[4], scal[4]);
      real gamma_HS_1 = normal_rng(loc[5], scal[5]);
      real gamma_HS_2 = normal_rng(loc[6], scal[6]);
      real gamma_SC_1 = normal_rng(loc[7], scal[7]);
      real gamma_SC_2 = normal_rng(loc[8], scal[8]);
      real gamma_CG_1 = normal_rng(loc[9], scal[9]);
      real gamma_CG_2 = normal_rng(loc[10], scal[10]);
      real gamma_GD_1 = normal_rng(loc[11], scal[11]);
      real gamma_GD_2 = normal_rng(loc[12], scal[12]);
      real rho = uniform_rng(0, 0.5);
      for (n in 1:N) {
        row_vector[2] epsilon = binormal_rng(0, 0, 1, 1, rho);
        real eta_1 = alpha_1
                   + fmwc[n] * beta_1
                   + HS[n] * gamma_HS_1
                   + SC[n] * gamma_SC_1
                   + CG[n] * gamma_CG_1
                   + GD[n] * gamma_GD_1
                   + epsilon[1];
        if (eta_1 < 0) draws[s, n] = -1;
        else {
          real eta_2 = alpha_2
                     + fmwc[n] * beta_2
                     + HS[n] * gamma_HS_2
                     + SC[n] * gamma_SC_2
                     + CG[n] * gamma_CG_2
                     + GD[n] * gamma_GD_2
                     + epsilon[2];
          draws[s, n] = eta_2 > 0;
        }
      }
    }
    return draws;
  }
```

```
rstan::expose_stan_functions("selection4.stan")
```

We need to create and then center some dummy variables for the education categories in the CPS with something like

```
E <- model.matrix(~ educ, data = CPS)[ , -1]
E <- sweep(E, MARGIN = 2, STATS = colMeans(E), FUN = `-`)
```

And then we can call our selection function.

```
draws <- selection_rng(S = S, fmwc = with(CPS, fmwc - mean(fmwc)),
                       HS = E[ , 1], SC = E[ , 2], CG = E[ , 3], GD = E[ , 4],
                       loc = c( qnorm(c(2 / 3, 18.75 / 20)), beta_1 = -1, beta_2 = 0,
                                gamma_HS_1 = 0.1,  gamma_HS_2 = 0.2,
                                gamma_SC_1 = 0.15, gamma_SC_2 = 0.3,
                                gamma_CG_1 = 0.2,  gamma_CG_2 = 0.4,
                                gamma_GD_1 = 0.25, gamma_GD_2 = 0.5),
                       scal = c(alpha_1 = 0.1, alpha_2 = 0.15,
                                beta_1 = 0.1, beta_2 = 0.1,
                                educ = rep(0.25, 8)))
mean(draws != -1)                    # overall labor force participation rate
```

```
## [1] 0.6597917
```

```
mean(draws == 0) / mean(draws != -1) # overall unemployment rate
```

```
## [1] 0.05208188
```

Finally, we introduce the age variable as a predictor. In most situations, age should be measured in decades rather than years because the expected difference in the linear predictor for two otherwise identical people who are one year apart is too negligible for most outcomes (possibly excluding models for childhood development). So, we rescale the units of age to decades:

```
CPS$age <- CPS$age / 10
```

The effect of age could have a somewhat different effect on the probability of being in the labor force than on the probability of having a job given that you are in the labor force, but it has a much different effect for women who are married with children than other people. In general, a quadratic relationship is possible with the coefficient on the quadratic term being negative and the coefficient on the linear term being positive.

For women who are married with children, the probability of being in the labor force is mostly increasing with age or perhaps more specifically with the age of the youngest child. It probably peaks around five decades, which can be used to obtain the coefficient on the quadratic term given a draw for the coefficient on the linear term. For people who are not women that are married with children, the probability of being in the labor force presumably is still quadratic but may have a different trajectory. The probability of having a job given that you are in the labor force would seem to follow a similar trajectories but is likely more muted. The calculations are a bit simpler if we do not center the age variable.

```
  matrix selection_rng(int S, vector fmwc, vector age,
                       vector HS, vector SC, vector CG, vector GD,
                       vector loc, vector scal) {
    int N = rows(fmwc);
    matrix[S, N] draws;
    vector[N] age2 = square(age);
    for (s in 1:S) {
      real alpha_1 = normal_rng(loc[1], scal[1]);
      real alpha_2 = normal_rng(loc[2], scal[2]);
      real beta_1  = normal_rng(loc[3], scal[3]);
      real beta_2  = normal_rng(loc[4], scal[4]);
      real gamma_HS_1 = normal_rng(loc[5], scal[5]);
      real gamma_HS_2 = normal_rng(loc[6], scal[6]);
      real gamma_SC_1 = normal_rng(loc[7], scal[7]);
      real gamma_SC_2 = normal_rng(loc[8], scal[8]);
      real gamma_CG_1 = normal_rng(loc[9], scal[9]);
      real gamma_CG_2 = normal_rng(loc[10], scal[10]);
      real gamma_GD_1 = normal_rng(loc[11], scal[11]);
      real gamma_GD_2 = normal_rng(loc[12], scal[12]);
      real lambda_fmwc_1 = exponential_rng(1) * loc[13];
      real lambda_fmwc_2 = exponential_rng(1) * loc[14];
      real theta_fmwc_1 = -lambda_fmwc_1 / 10;
      real theta_fmwc_2 = -lambda_fmwc_2 / 10;
      real lambda_other_1 = exponential_rng(1) * loc[15];
      real lambda_other_2 = exponential_rng(1) * loc[16];
      real theta_other_1 = -lambda_other_1 / 10;
      real theta_other_2 = -lambda_other_2 / 10;
      real rho = uniform_rng(0, 0.4);
      for (n in 1:N) {
        row_vector[2] epsilon = binormal_rng(0, 0, 1, 1, rho);
        real eta_1 = alpha_1
                   + fmwc[n] * beta_1
                   + HS[n] * gamma_HS_1
                   + SC[n] * gamma_SC_1
                   + CG[n] * gamma_CG_1
                   + GD[n] * gamma_GD_1
                   + (fmwc[n] > 0 ? theta_fmwc_1  * age2[n] + lambda_fmwc_1  * age[n]
                                  : theta_other_1 * age2[n] + lambda_other_1 * age[n])
                   + epsilon[1];
        if (eta_1 < 0) draws[s, n] = -1;
        else {
          real eta_2 = alpha_2
                     + fmwc[n] * beta_2
                     + HS[n] * gamma_HS_2
                     + SC[n] * gamma_SC_2
                     + CG[n] * gamma_CG_2
                     + GD[n] * gamma_GD_2
                     + (fmwc[n] > 0 ? lambda_fmwc_2  * age2[n] + theta_fmwc_2  * age[n]
                                    : lambda_other_2 * age2[n] + theta_other_2 * age[n])
                     + epsilon[2];
          draws[s, n] = eta_2 > 0;
        }
      }
    }
    return draws;
  }
```

2.4 Prior Predictive Checking

Use the `expose_stan_functions` function in the **rstan** package to translate the Stan function you wrote in the previous subproblem to C++, compile it, and make it available to R. Call your function to produce the $S \times N$ matrix of -1, 0, and 1 values and check that it is reasonable in various respects. For example, about two-thirds of adults should be in the labor force and among those, about five percent should be unemployed.

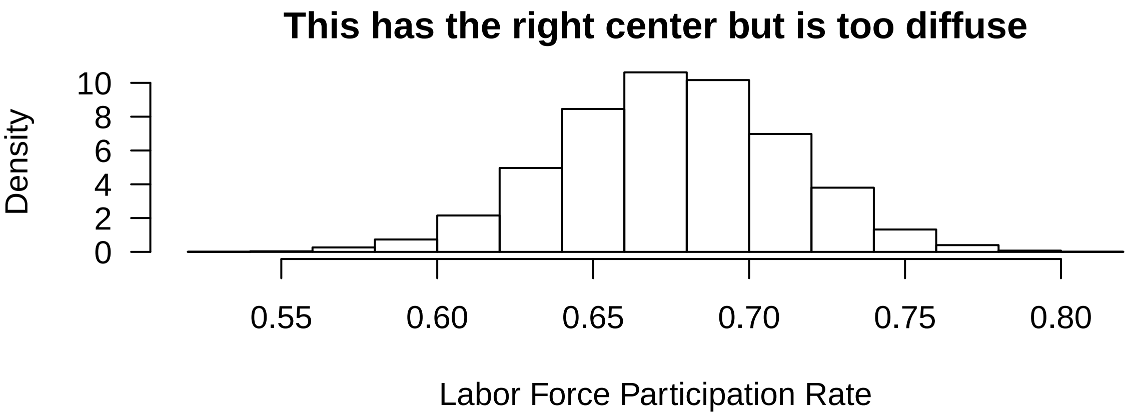
Using the last model from above,

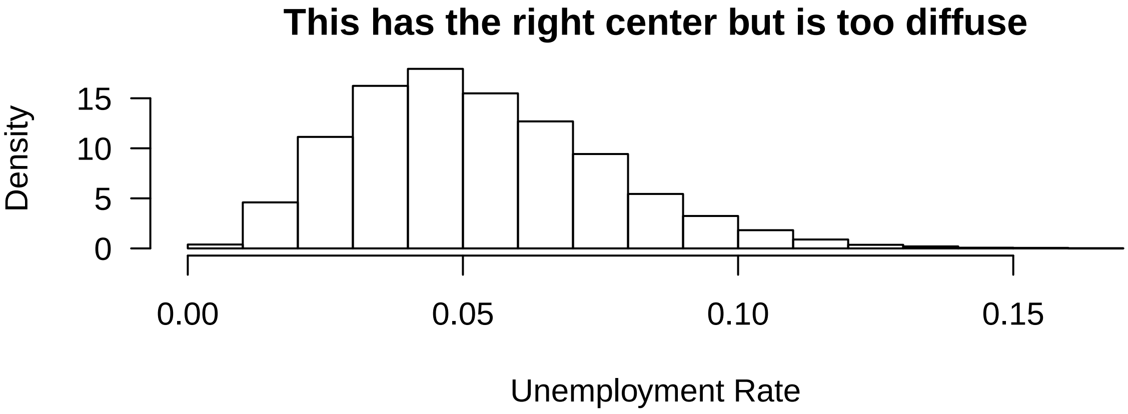
```r
rstan::expose_stan_functions("selection.stan")
```

```r
draws <- selection_rng(S = S, fmwc = with(CPS, fmwc - mean(fmwc)), age = CPS$age,
                       HS = E[ , 1], SC = E[ , 2], CG = E[ , 3], GD = E[ , 4],
                       loc = c( qnorm(c(2 / 3, 18.4 / 20)), beta_1 = -1, beta_2 = 0,
                                gamma_HS_1 = 0.1,  gamma_HS_2 = 0.2,
                                gamma_SC_1 = 0.15, gamma_SC_2 = 0.3,
                                gamma_CG_1 = 0.2,  gamma_CG_2 = 0.4,
                                gamma_GD_1 = 0.25, gamma_GD_2 = 0.5,
                                lambda_fmwc_1 = 0.04, lambda_fmwc_2 = 0.025,
                                lambda_other_1 = 0.02, lambda_other_2 = 0.01),
                       scal = c(alpha_1 = 0.1, alpha_2 = 0.15,
                                beta_1 = 0.1, beta_2 = 0.1,
                                educ = rep(0.25, 8)))
```

More specifically, we can plot the distribution of S realizations of the labor force participation rate and the unemployment rate:

```r
par(mfrow = 2:1, mar = c(4, 4, 2, 2) + .1, las = 1)
hist(rowMeans(draws != -1), prob = TRUE,
     main = "This has the right center but is too diffuse",
     xlab = "Labor Force Participation Rate")
hist(rowMeans(draws == 0) / rowMeans(draws != -1), prob = TRUE,
     main = "This has the right center but is too diffuse",
     xlab = "Unemployment Rate")
```



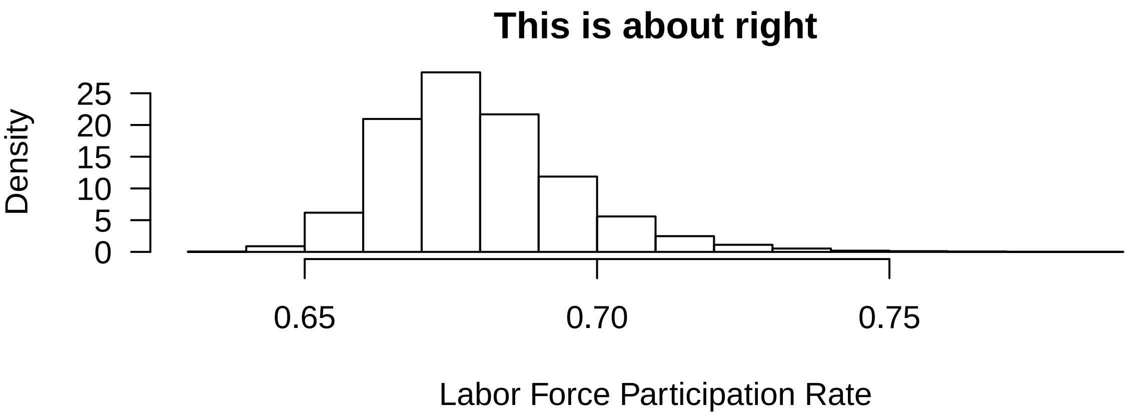


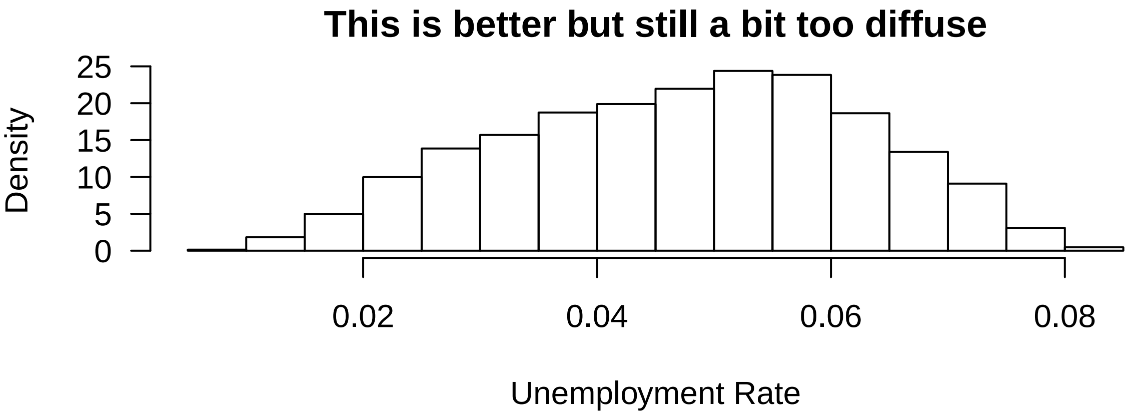
Since these put too much probability on extreme events, I draw again with smaller values of the scale hyperparameters:

```
draws <- selection_rng(S = S, fmwc = with(CPS, fmwc - mean(fmwc)), age = CPS$age,
                       HS = E[ , 1], SC = E[ , 2], CG = E[ , 3], GD = E[ , 4],
                       loc = c( qnorm(c(2 / 3, 18.4 / 20)), beta_1 = -1, beta_2 = 0,
                                gamma_HS_1 = 0.1,  gamma_HS_2 = 0.2,
                                gamma_SC_1 = 0.15, gamma_SC_2 = 0.3,
                                gamma_CG_1 = 0.2,  gamma_CG_2 = 0.4,
                                gamma_GD_1 = 0.25, gamma_GD_2 = 0.5,
                                lambda_fmwc_1 = 0.04, lambda_fmwc_2 = 0.025,
                                lambda_other_1 = 0.02, lambda_other_2 = 0.01),
                       scal = c(alpha_1 = 0.025, alpha_2 = 0.01,
                                beta_1 = 0.025, beta_2 = 0.01,
                                educ = rep(c(0.05, 0.02), times = 4)))
```

Then, the plots come out somewhat better:

```
par(mfrow = 2:1, mar = c(4, 4, 2, 2) + .1, las = 1)
hist(rowMeans(draws != -1), prob = TRUE,
     main = "This is about right",
     xlab = "Labor Force Participation Rate")
hist(rowMeans(draws == 0) / rowMeans(draws != -1), prob = TRUE,
     main = "This is better but still a bit too diffuse",
     xlab = "Unemployment Rate")
```





2.5 Estimating a Probit Model

For this subproblem, we are going to treat the individual's decision to be not in the labor force (`nilf`) as an exogenous known — i.e. do what labor microeconometricans criticize — because the `stan_glm` function is only capable of estimating univariate outcome models, which in this case is for the `empl` variable. Call `stan_glm` in the **rstanarm** package with following arguments:

`data = CPS, family = binomial(link = "probit"), subset = nilf == 0, QR = FALSE`

You should specify the `formula`, `prior`, and `prior_intercept` values to be like those you used in the previous subproblem for the "second" part of the model for (not) being employed conditional on being in the labor force. In this problem, you presumably want to call the functions you use for `prior` and `prior_intercept` with `autoscale = FALSE` so that they are interpreted in their raw units rather than in terms of standard deviations.

```r
library(rstanarm)
options(mc.cores = parallel::detectCores())
```

```r
post <- stan_glm(empl ~ fmwc + educ + age + I(age ^ 2), data = CPS,
                 family = binomial(link = "probit"), subset = nilf == 0, QR = FALSE,
                 prior_intercept = normal(location = 18.4 / 20, scale = 0.01,
                                          autoscale = FALSE),
                 prior = normal(location = c(0, 0.2, 0.3, 0.4, 0.5, 0.02, -0.01),
                                scale = c(0.03, 0.03, 0.03, 0.03, 0.03, 0.03, 0.01),
                                autoscale = FALSE))
```

2.6 Inference

Call the as.matrix function on the result of stan_glm in the previous subproblem in order to obtain a $S \times K$ matrix where K is the number of parameters estimated including the intercept (relative to uncentered predictors). For each parameter, estimate the posterior probability that the parameter is positive from its proportion of positive draws. How does this differ from a frequentist test of a null hypothesis that a parameter is zero against the one-sided alternative hypothesis that it is positive?

```
colMeans(as.matrix(post) > 0)
```

```
##      (Intercept)         fmwcTRUE           educHS educSome college 
##          1.00000          0.58875          1.00000          1.00000 
##      educCollege     educAdvanced              age         I(age^2) 
##          1.00000          1.00000          0.99350          0.04375
```

In this case, the probability that the intercept and the coefficients on education are positive is essentially one. More precisely, the probability they are negative is less than one in 4000, which is the default number of draws from the posterior distribution. The coefficient on the linear age term is also almost definitely positive, although there were a few negative draws. The posterior distribution for the coefficient for women who are married with children has almost equal mass on both sides of zero, while the coefficient on the squared age term has less than a 5% chance of being positive under the posterior distribution.

All of these differ fundamentally from Frequentist p-values, which are conditional probabilities but not posterior probabilities. Specifically, a p-value is the probability of obtaining a value of an even more extreme test statistic from another dataset that is randomly sampled from the same population, conditional on the null hypothesis (in this case that a a coefficient is zero in the population) being true. By convention, if this p-value is sufficiently small, then you reject the null hypothesis in favor of the alternative hypothesis (in this case that the coefficient is positive in the population). But nothing tells you the probability that the coefficient is positive in the population because that is not even a coherent Frequentist question. Coefficients and parameters more generally are not random variables, so you cannot make probabilistic statements about them. You can only make probabilistic statements about *estimators* of those parameters, which are random variables if they are calculated from a sample that has been randomly drawn from a population, and those probabilistic statements are conditional on the value of the parameter you are estimating.

A Bayesian posterior probability, in contrast, is simply the subjective probability that something is the case — such as a parameter being positive — in the data that you condition on. There is no notion of repeated sampling of data because the probabilities are conditional on the data rather than conditional on the parameters.