Let’s experiment . . .

This is an introduction to statistical analysis. Please make sure you review all documentation below. If this knowledge does not resonate with you yet, make sure you keep a backup of all your work because you will most likely need them in your career as a computing professional.

Comparing Two Groups with a Continuous or Binary Outcome

Statistical inference is a process of learning from incomplete or imperfect (error-contaminated) data. Can account for this "imperfection" using either a sampling model or a measurement error model. Comparing two datasets.

Statistical hypothesis testing

The de facto standard for statistical inference is statistical hypothesis testing. The goal of hypothesis testing is to evaluate a null hypothesis. There are two possible outcomes:

- reject the null hypothesis  
- fail to reject the null hypothesis

Rejection occurs when a chosen test statistic is higher than some pre-specified threshold value; non-rejection occurs otherwise. Notice that neither outcome says anything about the quantity of interest, the research hypothesis.

Setting up a statistical test involves several subjective choices by the user that are rarely justified based on the problem or decision at hand.

- statistical test to use  
- null hypothesis to test  
- significance level

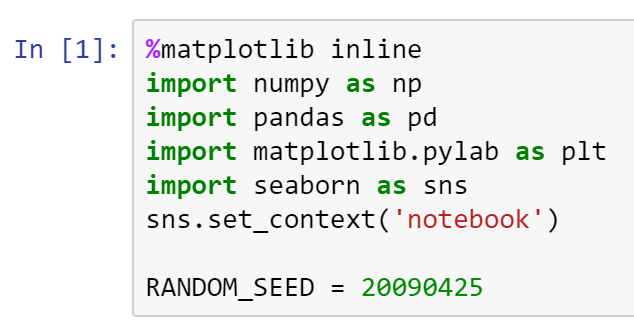
Choices are often based on arbitrary criteria, including "statistical tradition" (Johnson 1999). The resulting evidence is indirect, incomplete, and typically overstates the evidence against the null hypothesis (Goodman 1999).

Most importantly to applied users, the results of statistical hypothesis tests are very easy to misinterpret.

Estimation

Instead of testing, a more informative and effective approach for inference is based on estimation (be it frequentist or Bayesian). That is, rather than testing whether two groups are different, we instead pursue an estimate of how different they are, which is fundamentally more information.

Additionally, we include an estimate of uncertainty associated with that difference which includes uncertainty due to our lack of knowledge of the model parameters (epistemic uncertainty).



An Introduction to Bayesian Statistical Analysis

Practical methods for making inferences from data using "probability models" for quantities we observe and about which we wish to learn (Gelman et a., 2013).

A Bayesian model is described by parameters, uncertainty in those parameters is described using probability distributions. All conclusions from Bayesian statistical procedures are stated in terms of probability statements.

p(o|y) Where: o is unknowns  
 y is observations

What this means is that what is the probability of o given the y. In other words, what is the probability of o happening given the following observations.

This confers several benefits to the analyst, including:

- ease of interpretation, summarization of uncertainty  
- can incorporate uncertainty in parent parameters  
- easy to calculate summary statistics

Bayesian vs Frequentist Statistics: What's the difference?

Any statistical inference paradigm, Bayesian or otherwise, involves at least the following:

1. Some unknown quantities about which we are interested in learning or testing. We call these parameters.
2. Some data which have been observed, and hopefully contain information about.
3. One or more models that relate the data to the parameters, and is the instrument that is used to learn.

The Frequentist World View (or classical statistics)

- All data is random  
- The data that have been observed are considered random, because they are realizations of random processes, and hence will vary each time one goes to observe the system.  
- Model parameters are considered fixed. A parameter's true value is unknown and fixed, and so we condition on them.

In mathematical notation, this implies a (very) general model of the following form:

f(y|o) Where: y is data values (random)  
 o is conditional values (fixed)

Here, the model f accepts data values y as an argument, conditional on particular values of o.

Frequentist inference typically involves deriving estimators for the unknown parameters. Estimators are formulae that return estimates for particular estimands, as a function of data. They are selected based on some chosen optimality criterion, such as unbiasedness, variance minimization, or efficiency.

For example, let's say that we have collected some data on the prevalence of autism spectrum disorder (ASD) in some defined population. Our sample includes n sampled children, y of them having been diagnosed with autism. A frequentist estimator of the prevalence p is:

p = y / n

Why this particular function? Because it can be shown to be unbiased and minimum-variance.

It is important to not that, in a frequentist world, new estimators need to be derived for every estimated that is introduced.

The Bayesian World View

- Data are considered fixed. They used to be random, but once they were written into your lab notebook/spreadsheet they do not change.  
- Model parameters themselves may not be random, but Bayesians use probability distributions to describe their uncertainty in parameter values, and are therefore treated as random. In some cases, it is useful to consider parameters as having been sampled from probability distributions.

This implies that following form:

p(o|y)

This formulation used to be referred to as inverse probability, because it infers from observations to parameters, or from effects to causes.

Bayesians do not seek new estimators for every estimation problem they encounter. There is only one estimator for Bayesian inference: Bayes' Formula.

Priors

Once considered a controversial aspect of Bayesian analysis, the prior distribution characterizes what is known about an unknown quantity before observing the data from the present study. Thus, it represents the information state of that parameter. It can be used to reflect the information obtained in previous studies, to constrain the parameter to plausible values, or to represent the population of possible parameter values, of which the current study's parameter value can be considered a sample.

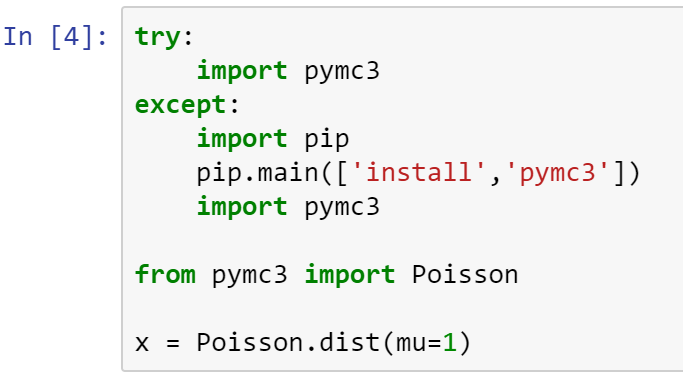
Bayesian Inference, in 3 Easy Steps

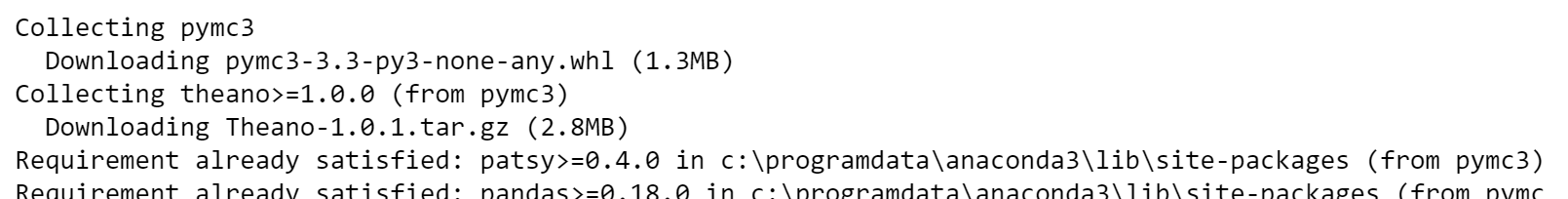
Step 1: Specify a probability model

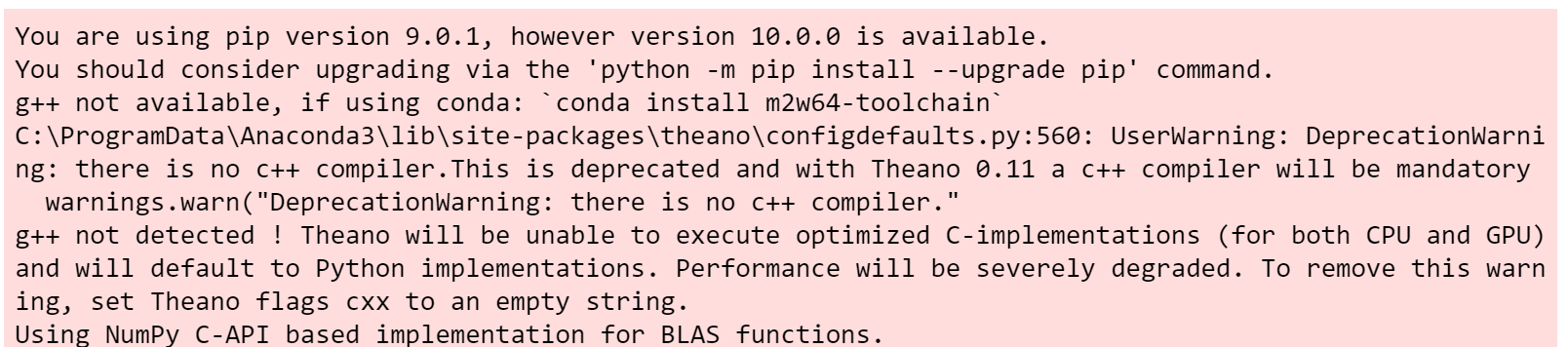
As was noted above, Bayesian statistics involves using probability models to sove problems. So, the first task is to completely specify the model in terms of probability distributions. This includes everything: unknown parameters, data, covariates, missing data, predictions. All must be assigned some probability density. This step involves making choices.

- What is the form of the sampling distribution of the data?

- What form best describes our uncertainty in the unknown parameters?







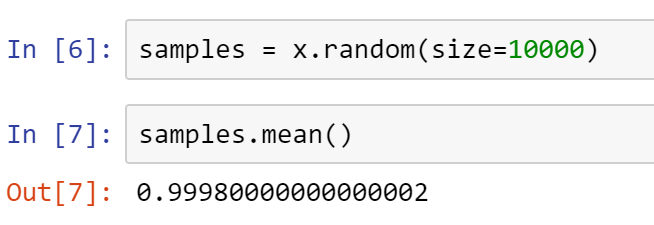
pymc3 is a Python package for Bayesian statistical modeling and Probabilistic Machine Learning.

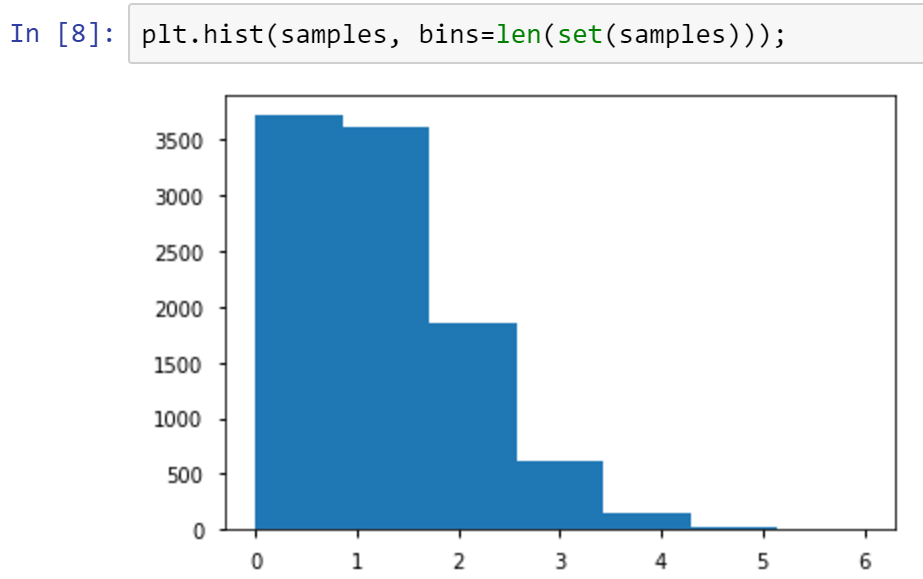
Visit https://docs.pymc.io for documentation

Poisson distribution is a discrete frequency distribution that gives the probability of a number of independent events occurring in a fixed time.

In the code above, mu is the mean of 1.

Size of sample is 10,000.





Above is an example histogram (a distribution) of a randomly generated 10,000 samples.

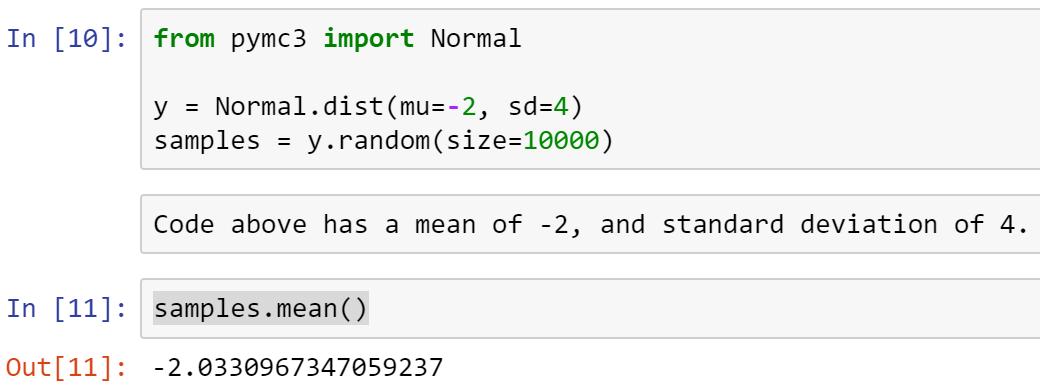
Continuous Random Variables -- an interval between zero and one, or from pos to neg infinity.

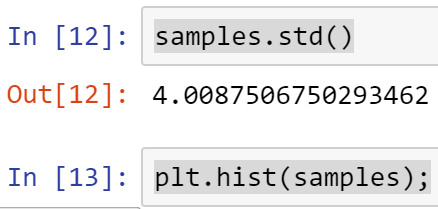
X E [0, 1] Where: E is less than or equal to

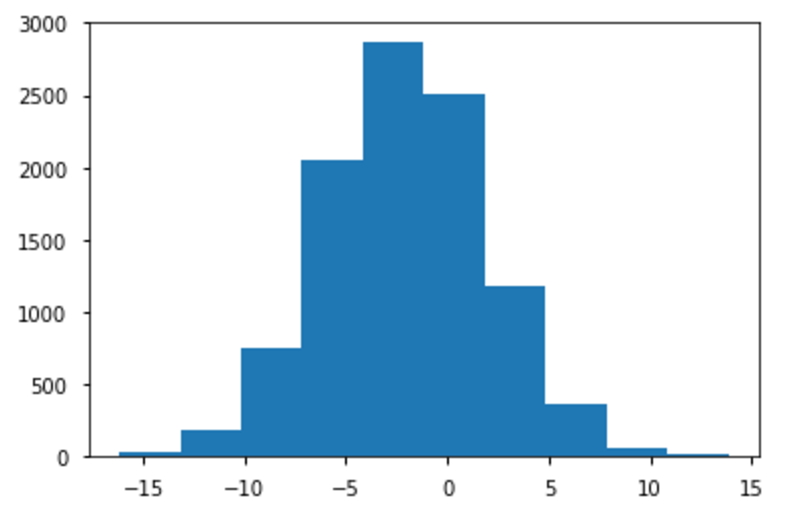
Ranges from 0 to 1

Y E (-infinity, +infinity)

The most famous of these is the normal distribution.







Note the standard Bell Curve generated from a normal distribution.

Step 2: Calculate a posterior distribution

The mathematical form p(o|y) that we associated with the Bayesian approach is referred to as a posterior distribution.

Posterior means later in time; subsequent.

Why posterior? Because it tells us what we know about the unknown "o" after having observed y.

This posterior distribution is formulated as a function of the probability model that was specified in Step 1. Usually, we can write it down but we cannot calculate it analytically. In fact, the difficulty inherent in calculating the posterior distribution for most models of interest is perhaps the major contributing factor for the lack of widespread adoption of Bayesian methods for data analysis.

But, once the posterior distribution is calculated, you get a lot for free:

- point estimates

- credible intervals

- quantiles

- predictions

Step 3: Check your model

Though frequently ignored in practice, it is critical that the model and its outputs be assessed before using the outputs for inference. Models are specified based on assumptions that are largely unverifiable, so the least we can do is examine the output in detail, relative to the specified model and the data that were used to fit the model.

Specifically, we must ask:

- does the model fit data?

- are the conclusions reasonable?

- are the outputs sensitive to changes in model structure?

Let's get into it . . .

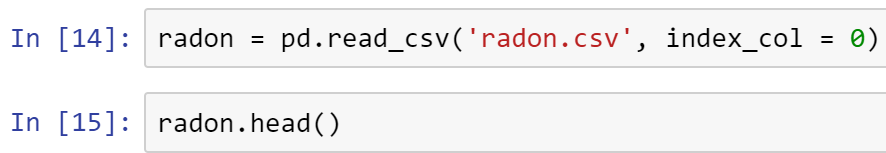
Estimation for one group

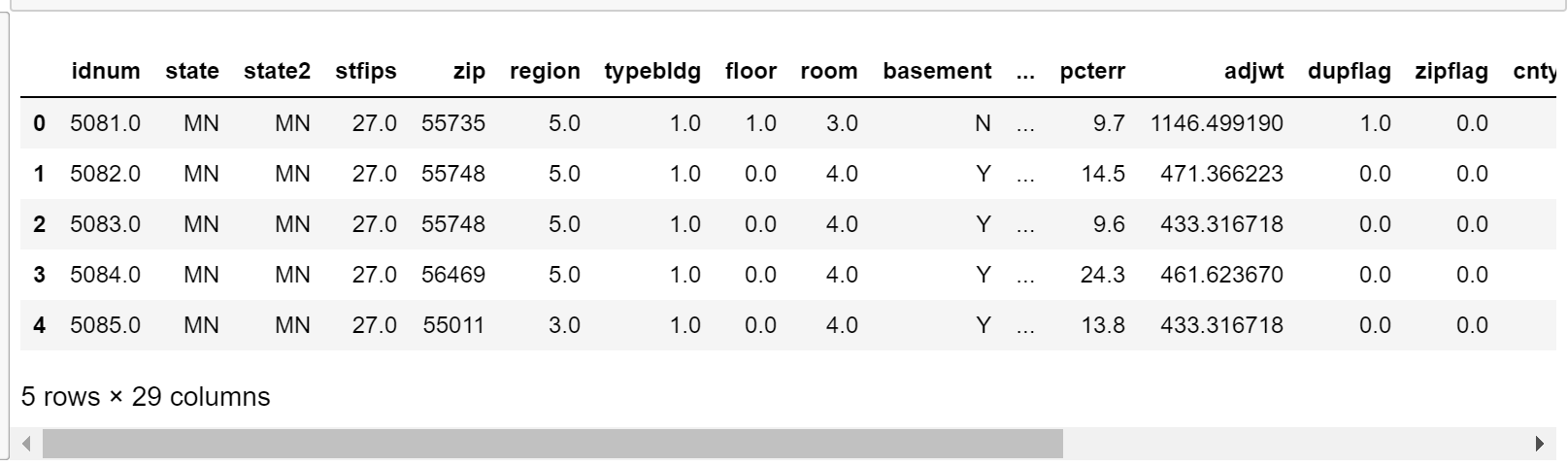
Before we compare two groups using Bayesian analysis, let's start with an even simpler scenario: statistical inference for one group.

For this we will use Gelman et al.'s (2007) radon dataset. In this dataset the amount of the radioactive gas radon has been measured among different households in all counties of several states. Radon gas is known to be the highest cause of lung cancer in non-smokers. It is believed to be more strongly present in households containing a basement and to differ in amount present among types of soil.

The US EPA has set an action level of 4 pCi/L. At or above this level of radon, the EPA recommends you take corrective measures to reduce your exposure to radon gas.

Try the code below, radon.csv datafile is in Canvas. These are real radon readings from houses in Minnesota.



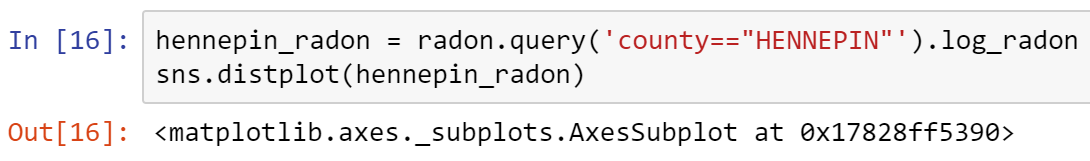


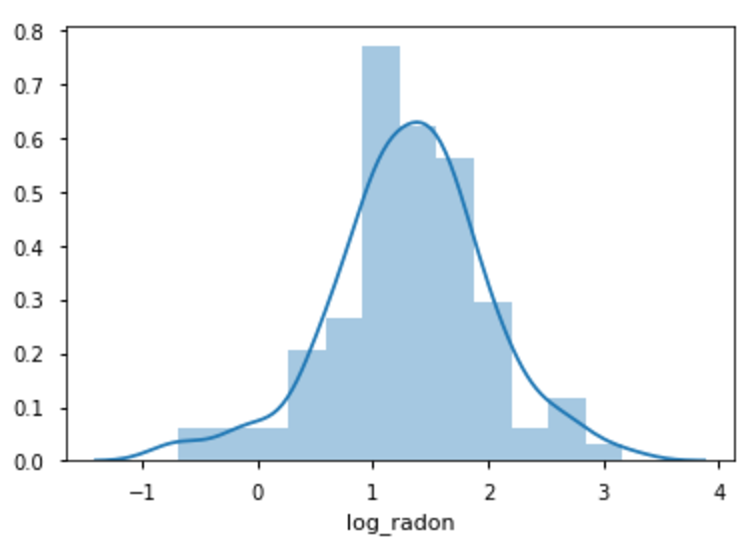
Let's focus on the (log) radon levels measured in a single county (Hennepin).

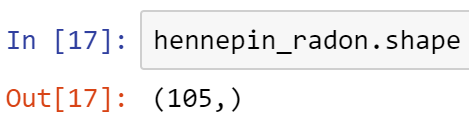
Suppose we are interested in:

- whether the mean log-radon value is greater than 4 pCi/L in Hennepin county

- the probability that any randomly-chosen household in Hennepin county has a reading of > 4







105 observations

The model

Recall that the first step in Bayesian inference is specifying a full probability model for the problem.

This consists of:

- a likelihood function(s) for the observations

- priors for all unknown quantities

The measurements look approximately normal, so let's start by assuming a normal distribution as the sampling distribution (likelihood) for the data.

This implies that we have 2 unknowns in the model; the mean and standard deviation of the distribution.

Prior choice

How do we choose distributions to use as priors for these parameters?

There are several considerations:

- discrete vs continuous values

- the support of the variable

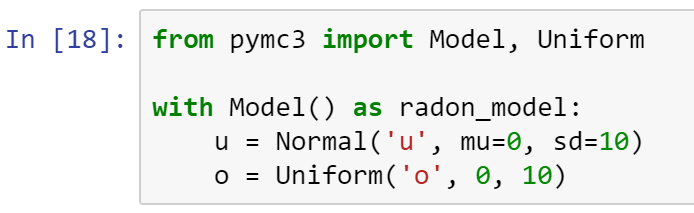
- the available prior information

While there may likely be prior information about the distribution of random values, we will assume no prior knowledge, and specify a diffuse prior for each parameter.

Since the mean can take any real value (since it is on the log scale), we will use another normal distribution here, and specify a large variance to allow the possibility of very large or very small values.

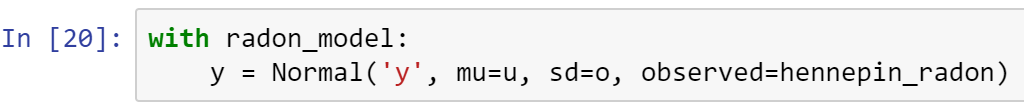
For the standard deviation, we know that the true value must be positive (no negative variances!). I will choose a uniform prior bounded from below at zero and from above at a value that is sure to be higher than any plausible value the true standard deviation (on the log scale) could take.

We can encode these in a Python model, using the PyMC3 package, as follows:



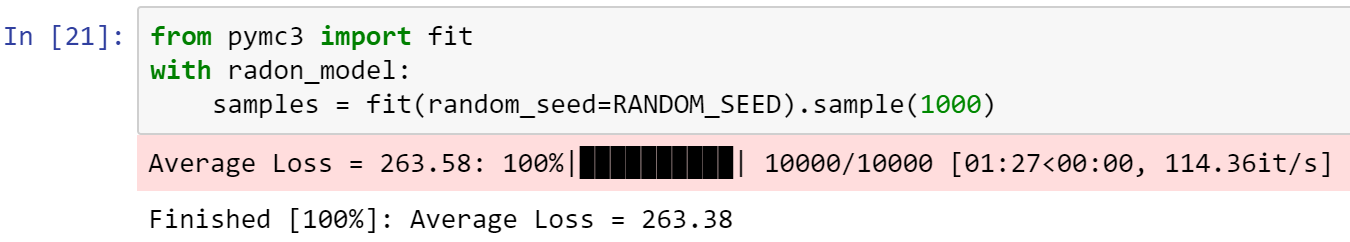
Today there is an array of software choices for Bayesians, including both open source software (e.g., Stan, PyMC, JAGS, emcee) and commercial (e.g., SAS, Stata). These examples can be replicated in any of these environments.

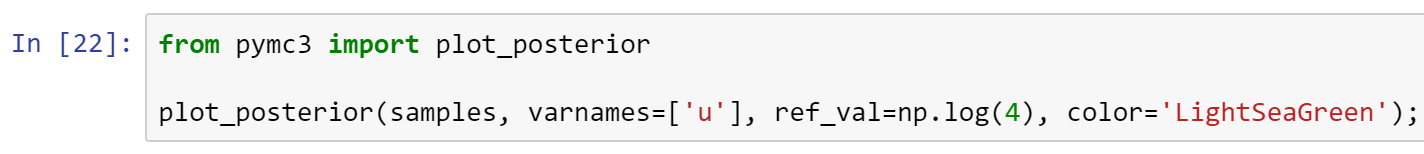
All that remains is to add the likelihood, which takes u and o as parameters, and the log-radon values as the set of observations:

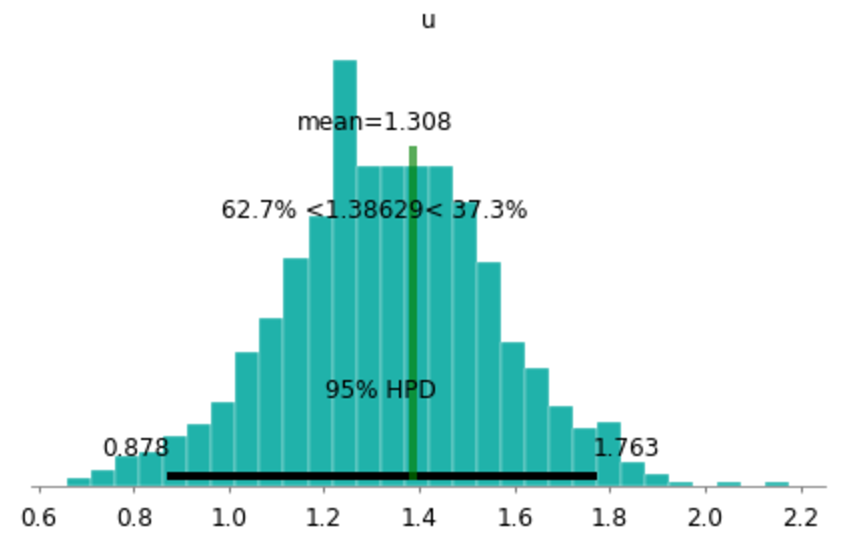


Now, we will fit the model using a numerical approach called variational inference. This will estimate the posterior distribution using an optimized approximation, and then draw samples from it.

Execute the code below and wait for a minute for it to complete.







The plot shows the posterior distribution of u, along with an estimate of the 95% posterior credible interval.

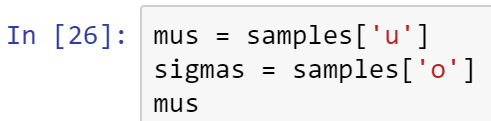
The output 83.1% < 1.38629 < 16.9%

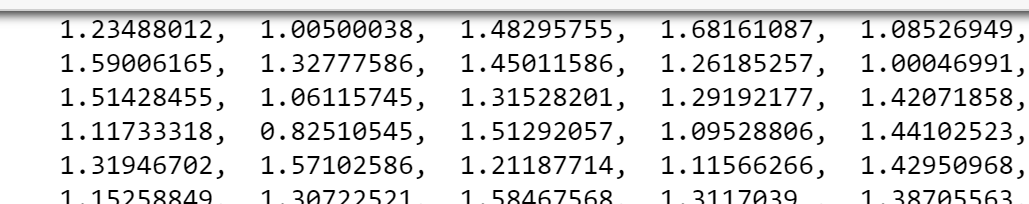
informs us that the probability of u being less than log(4) is 83.1% and the corresponding probability of being greater than log(4) is 16.9%.

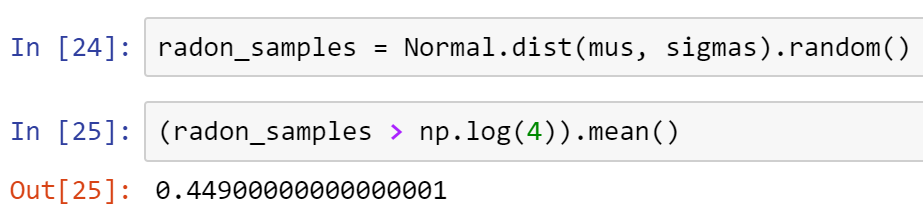
The posterior probability that the mean level of household radon in Henneprin County is greater than 4 pCi/L is 0.17.

Prediction

What is the probability that a given household has a log-radon measurement larger than one? To answer this, we make use of the posterior predictive distribution.







The posterior probability that a randomly-selected household in Hennepin County contains radon levels in excess of 4 pCi/L is 0.45.

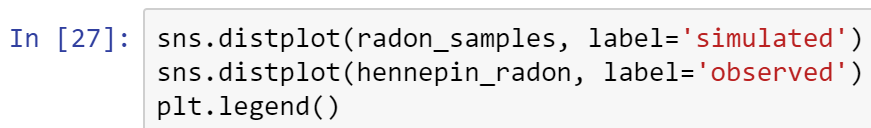
Model checking

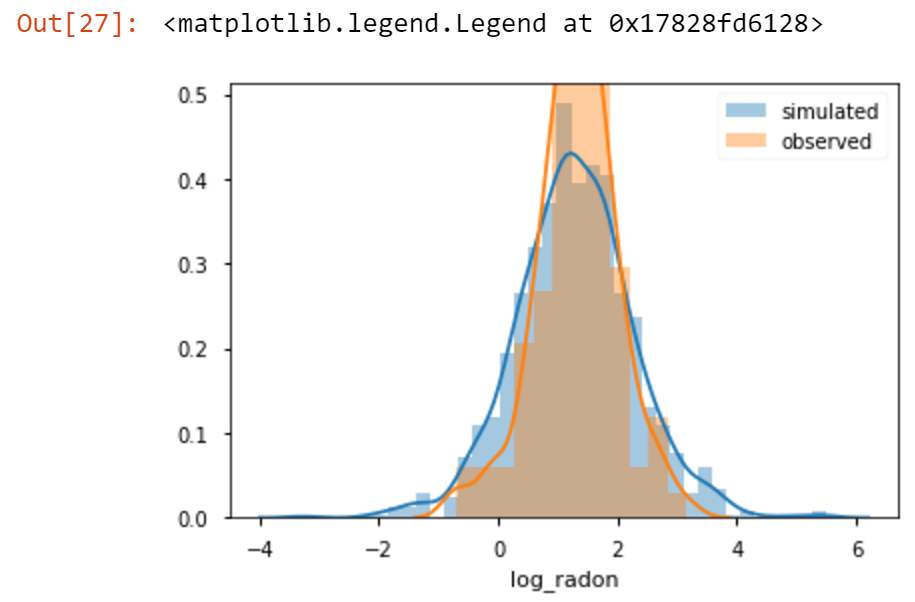
But, how do we know this model is any good?

It's important to check the fit of the model, to see if its assumptions are reasonable. One way to do this is to perform posterior predictive checks. This involves generating simulated data using the model that you built, and comparing that data to the observed data.

One can choose a particular statistic to compare, such as tail probabilities or quartiles, but here it is useful to compare them graphically.

We already have these simulations from the previous exercise!



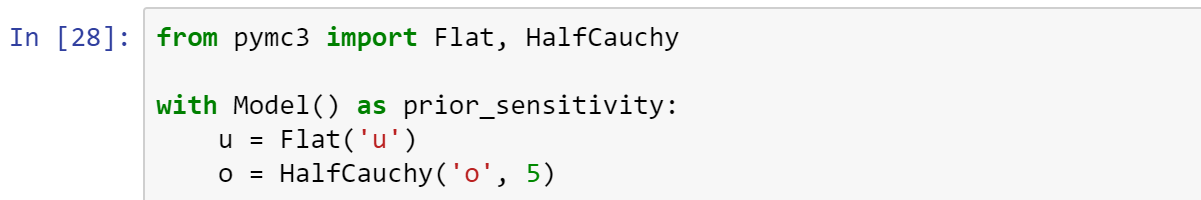


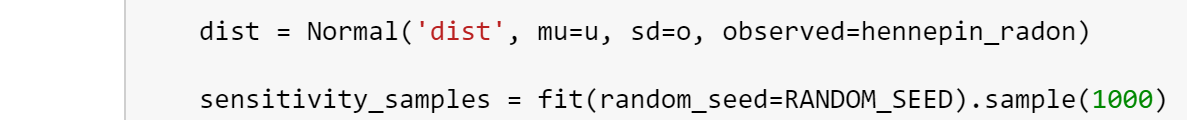
Prior sensitivity

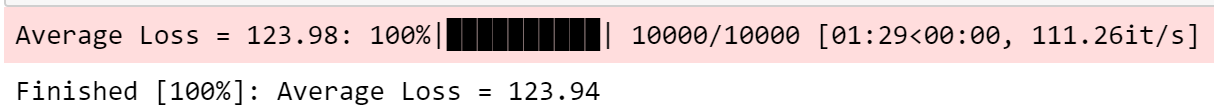
It's also important to check the sensitivity of your choice of priors to the resulting inference.

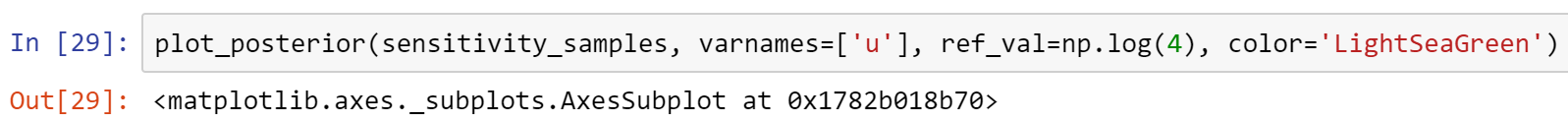
Here is the same model, but with drastically different (though still uninformative) priors specified.

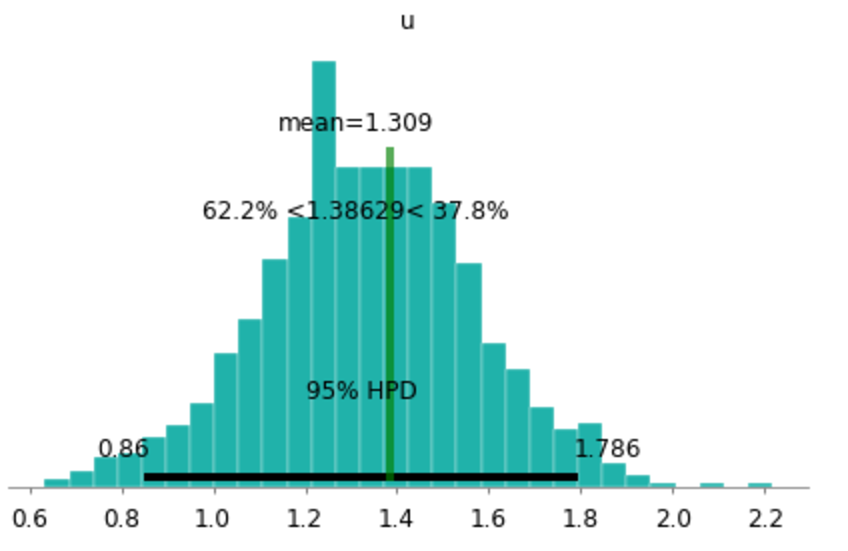
Execute code below and wait for a few seconds.



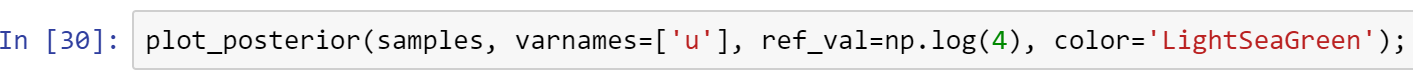


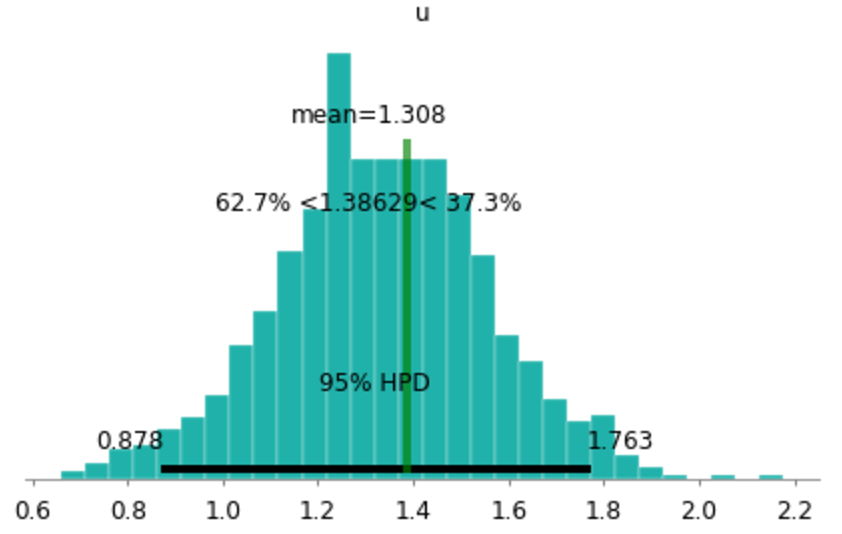






Here is the original model for comparison.

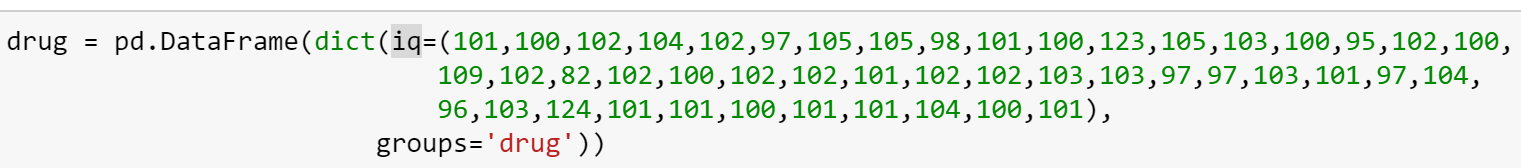




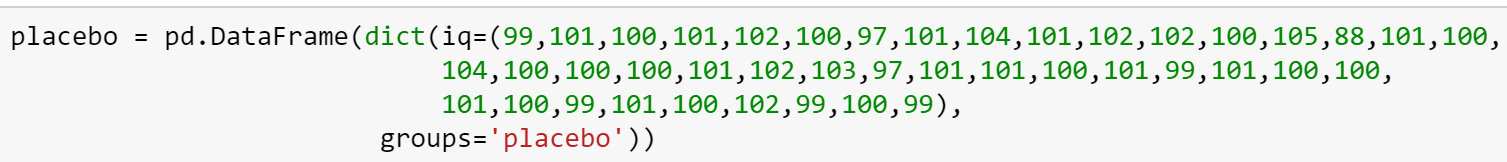
Two Groups with Continuous Outcome

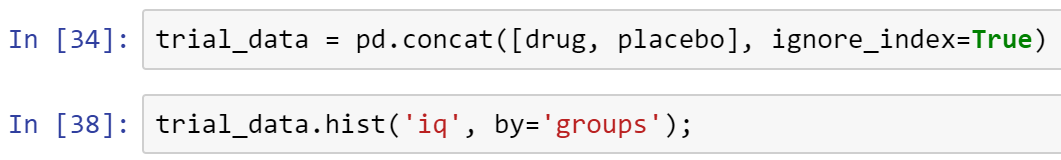
To illustrate how this Bayesian estimation approach works in practice, we will use a fictitious example from Kruschke (2012) concerning the evaluation of a clinical trial for drug evaluation. The trial aims to evaluate the efficacy of a "smart drug" that is supposed to increase intelligence by comparing IQ scores of individuals in a treatment arm (those receiving the drug) to those in a control arm (those receiving a placebo). There are 47 individuals and 42 individuals in the treatment and control arms, respectively.

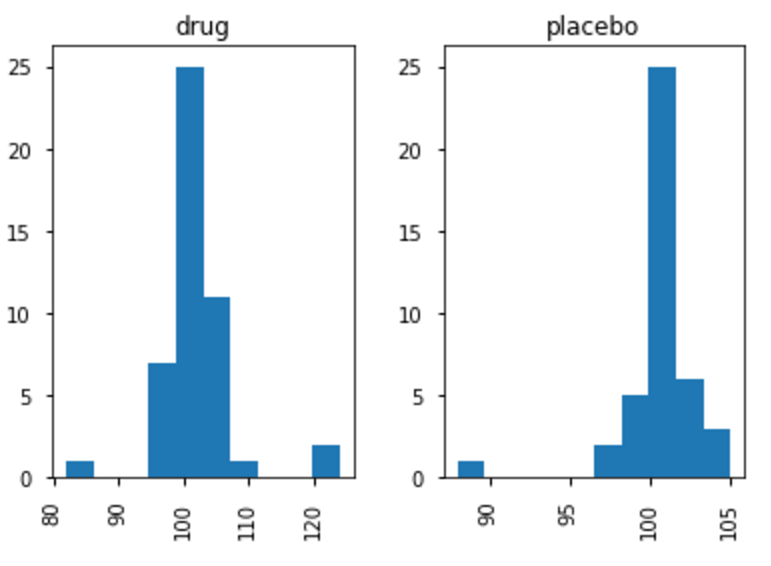












Since there appear to be extreme ("outlier") values in the data, we will choose a Student-t distribution to describe the distributions of the scores in each group. This sampling distribution adds robustness to the analysis, as a T distribution is less sensitive to outlier observations, relative to a normal distribution.

The T distributions have fatter tails, and more flatter than normal distribution.

The three-parameter Student-t distribution allows for the specification of a mean u, a precision (inverse-variance) t, and a degrees-of-freedom parameter u.

The degrees-of-freedom parameter essentially specifies the "normality" of the data, since larger values of v make the distribution converge to a normal distribution, while small values (close to zero) result in heavier tails.

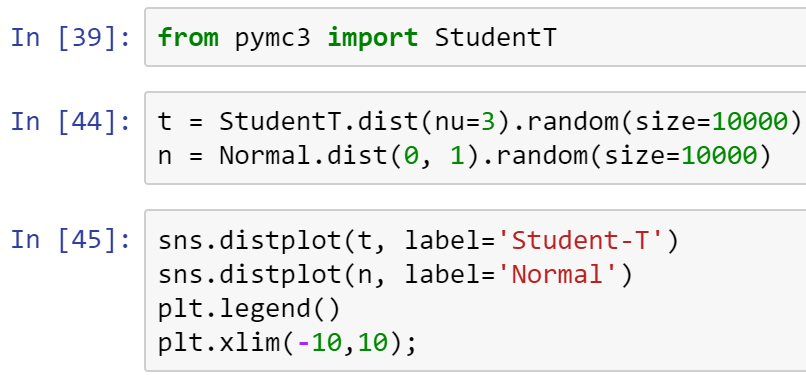
Thus, the likelihood functions of our model are specified as follows:

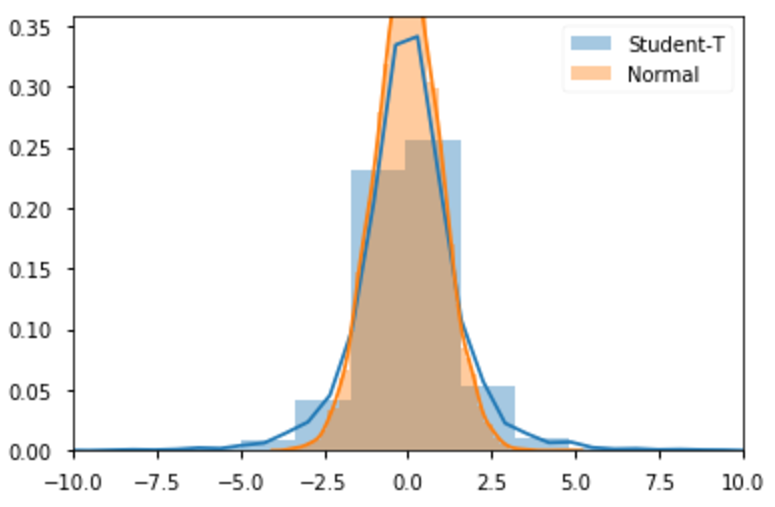
y = T(v,u,o)

As a simplifying assumption, we will assume that the degree of normality u is the same for both groups.

Exercise

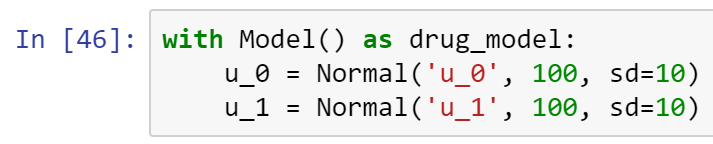
Draw 10000 samples from a Student-T distribution (StudentT in PyMC3) with parameter nu=3 and compare the distribution of these values to a similar number of draws from a Normal distribution with parameters mu=0 and sd=1.



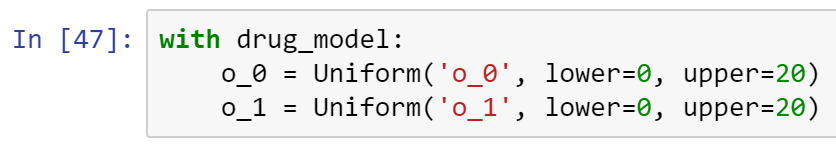


Prior choice

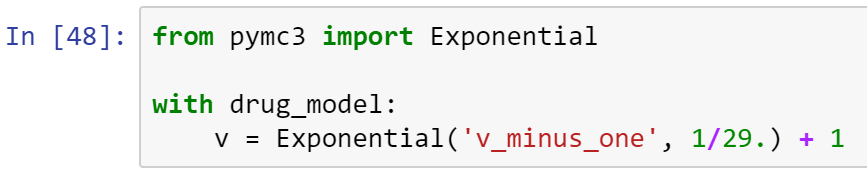
Since the means are real-valued, we will apply normal priors. Since we know something about the population distribution of IQ values, we will center the priors at 100, and use a standard deviation that is more than wide enough to account for plausible deviations from this population mean:

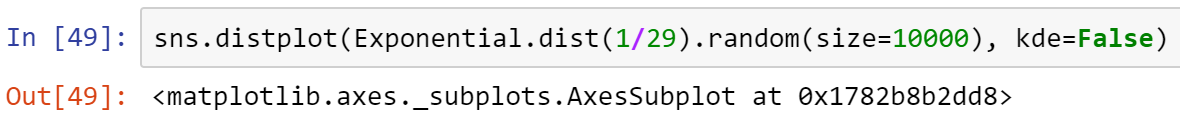


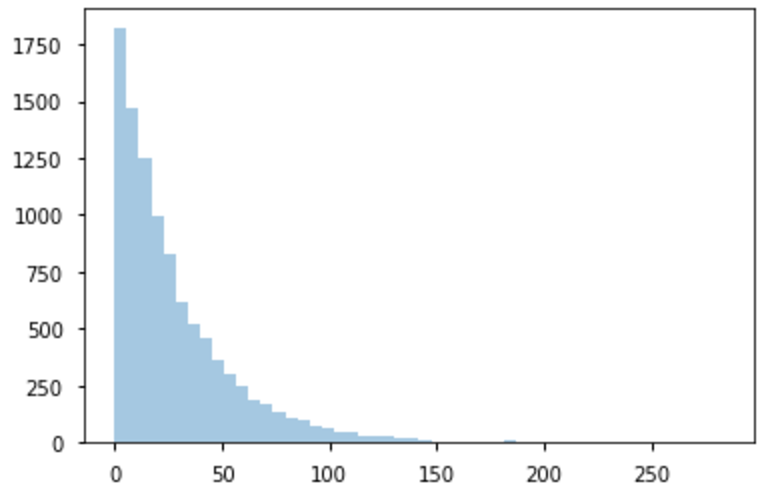
Similarly, we will use a uniform prior for the standard deviations, with an upper bound of 20.



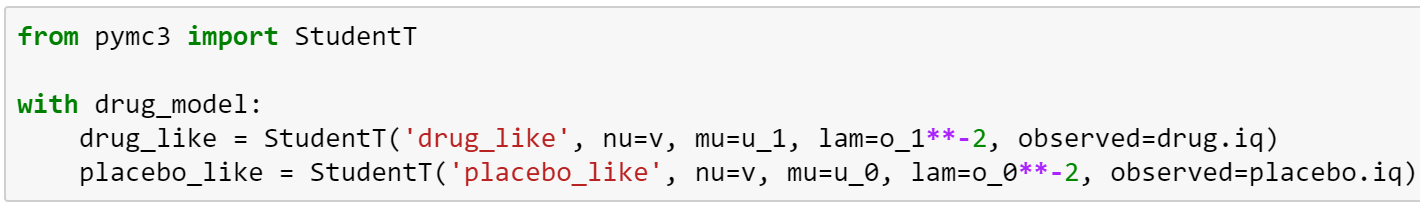
For the degrees-of-freedom parameter u, we will use an exponential distribution with a mean of 30; this allocates high prior probability over the regions of the parameter that describe the range from normal to heavy-tailed data under the Student-T distribution.





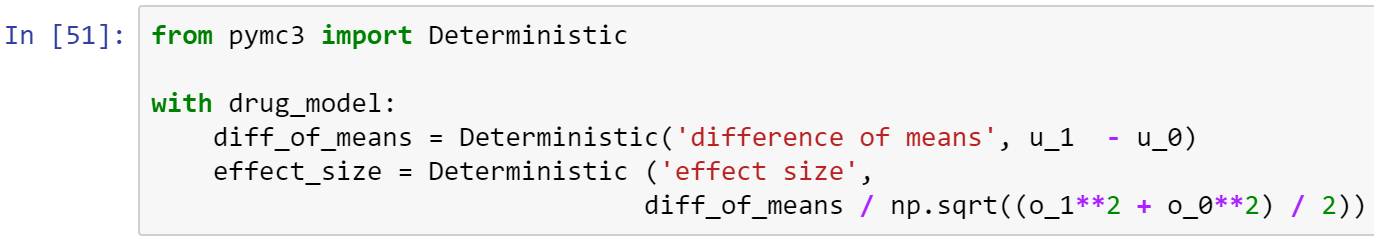


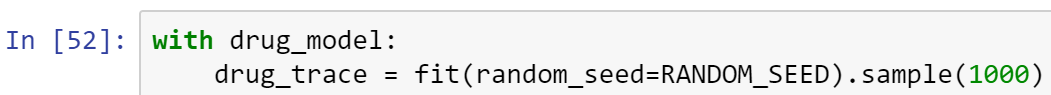


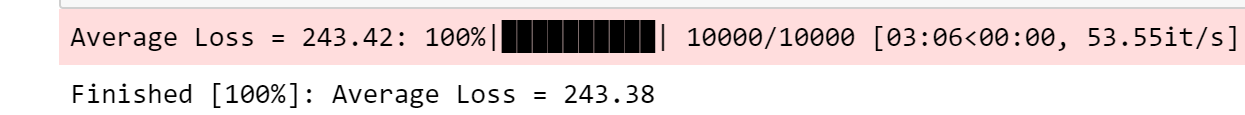


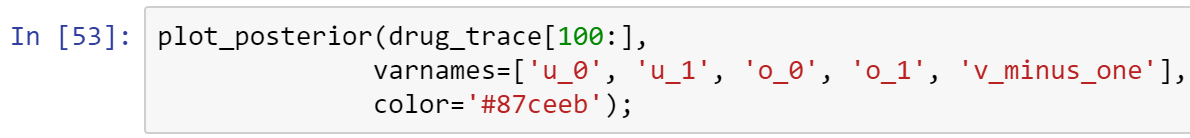
Now that model is fully specified, we can turn our attention to tracking the posterior quantities of interest. Namely, we can calculate the difference in means between the drug and placebo groups.

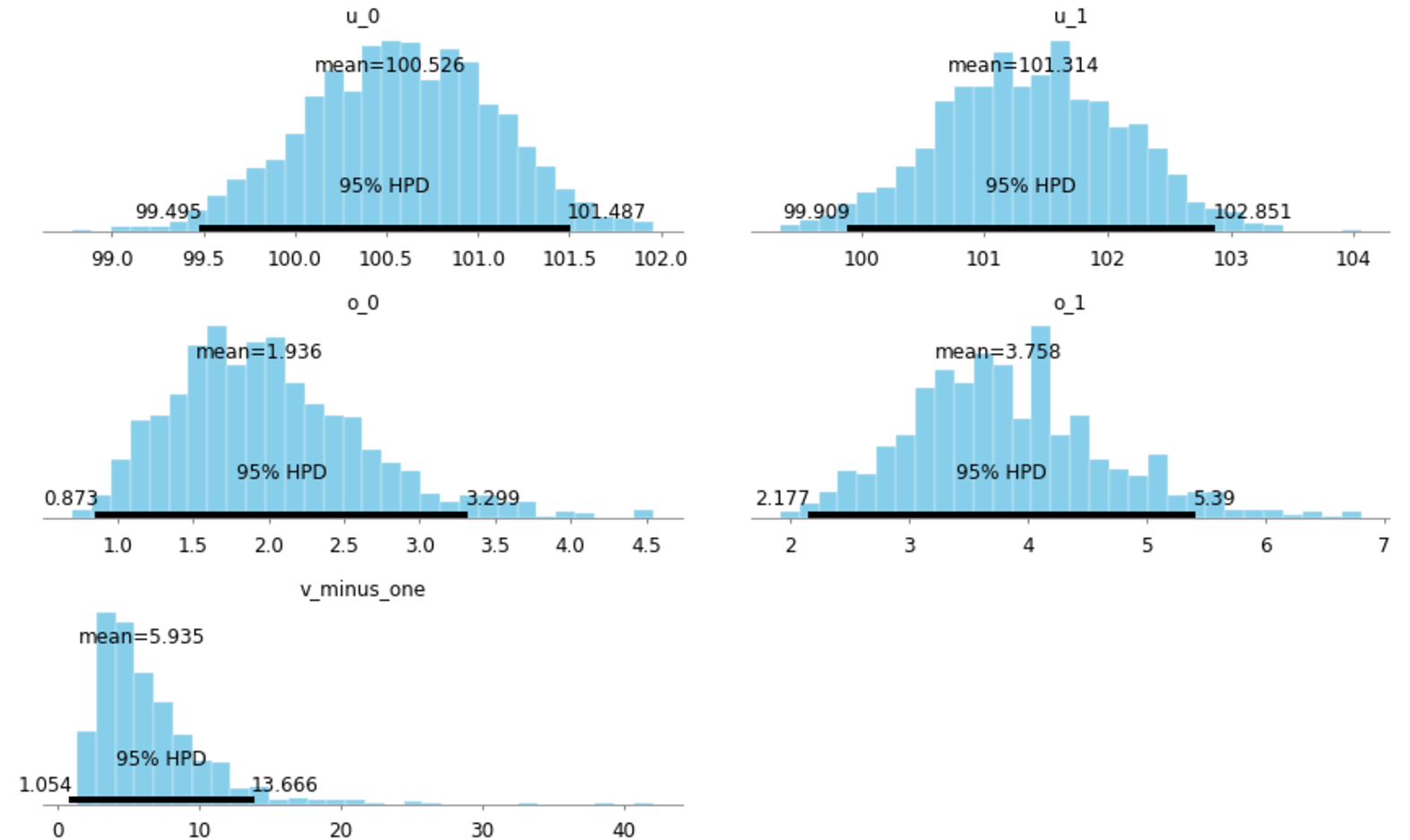
As a joint measure of the groups, we will also estimate the "effect size", which is the difference in means scaled by the pooled estimates of standard deviation. This quantity can be harder to interpret, since it is no longer in the same units as our data, but it is a function of all four estimated parameters. After executing the second set of code below, wait for a few seconds for it to complete.

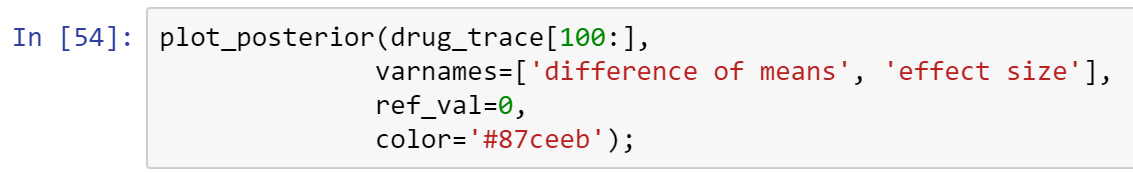


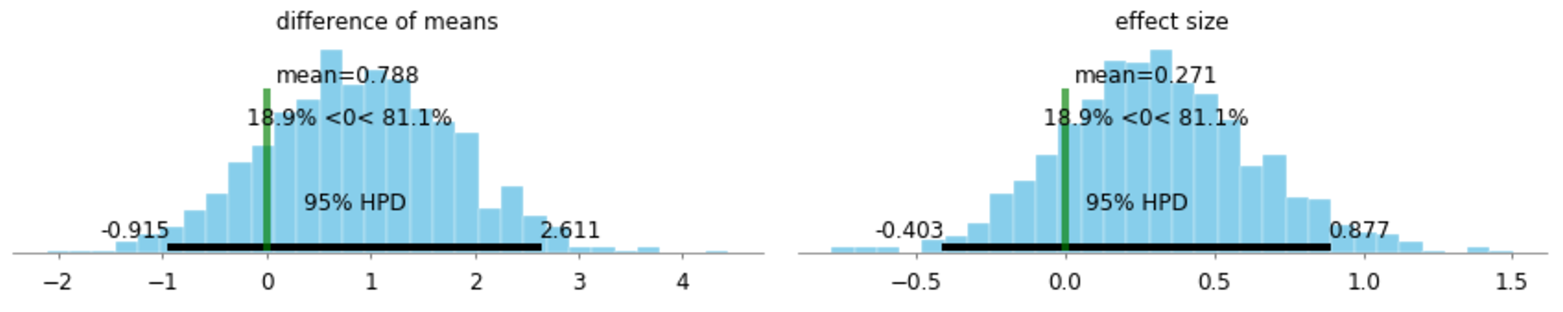












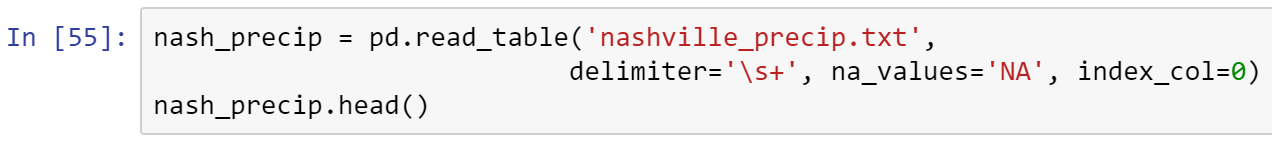
The posterior probability that the mean IQ of subjects in the treatment group is greater than that of the control group is 0.99.

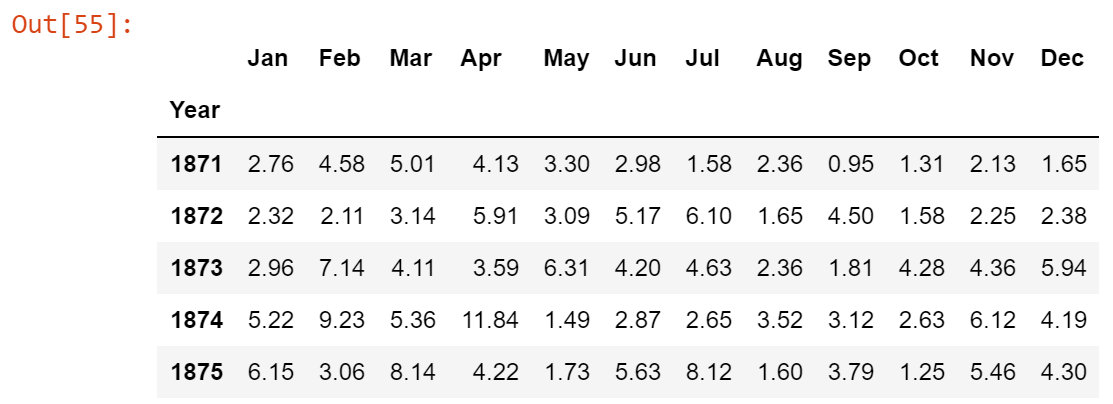
Exercise

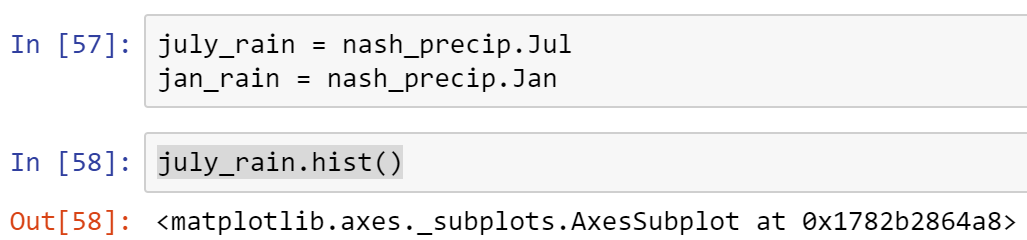
Load the nashville\_precip.txt dataset. Build a model to compare rainfall in January and July.

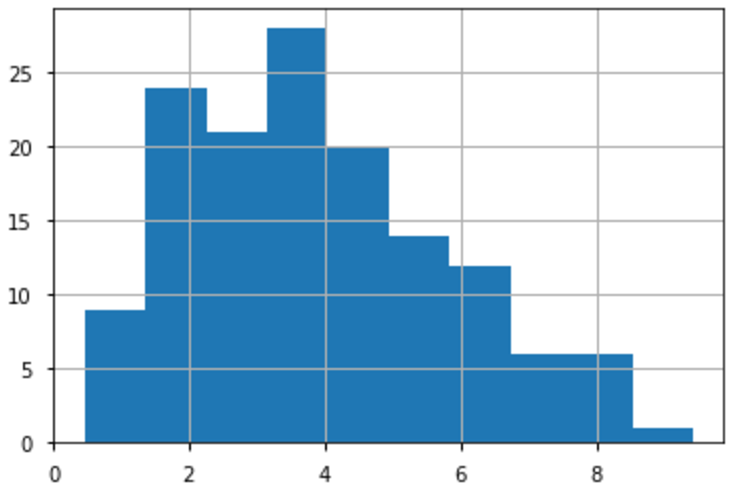
- What's the probability that the expected rainfall in January is larger than in July?

- What's the probability that January rainfall exceeds July rainfall in a given year?





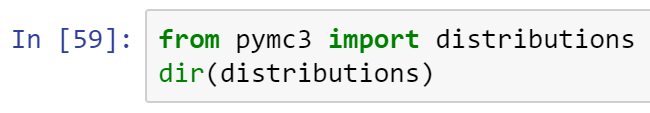


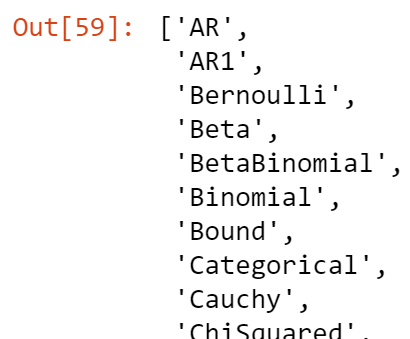


The above distribution is continuous values, but positive

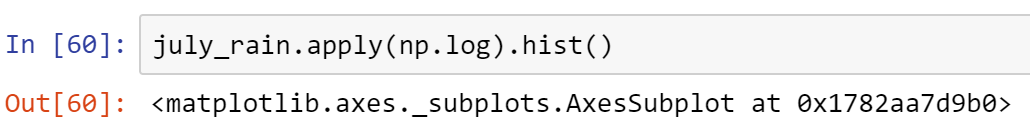
Different types of distributions: https://en.wikipedia.org/wiki/List\_of\_probability\_distributions

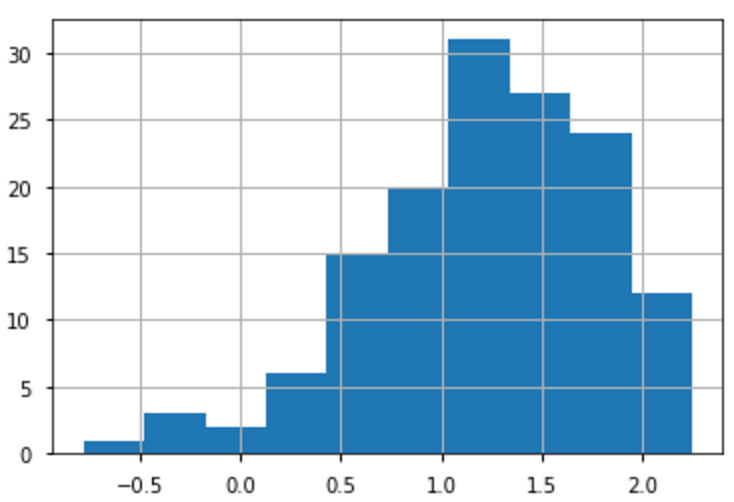
Below are the different distributions in pymc3





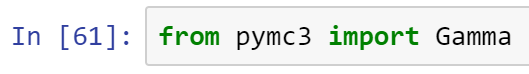
Read the list output generated, and determine which distribution models you recognize.

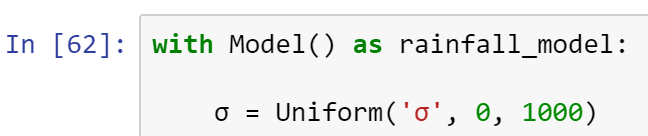


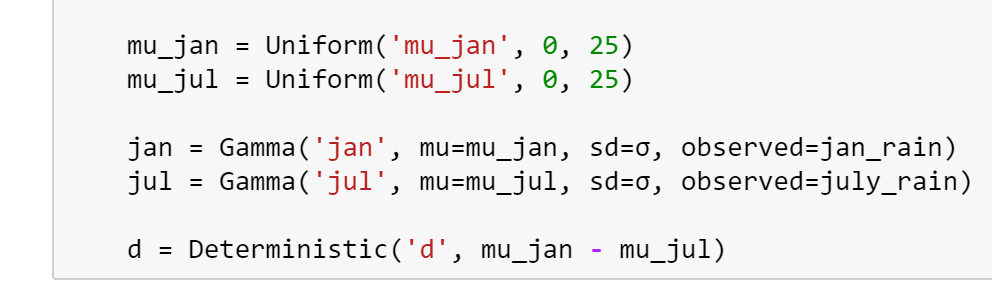


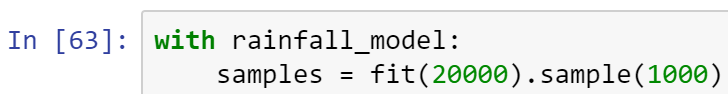
You can also use the Gamma distribution.

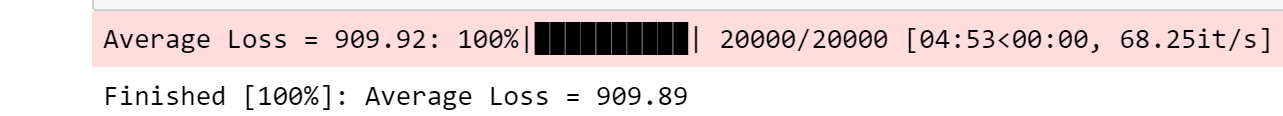
In jupyter notebook, type "\sigma" then hit tab, it will produce the sigma symbol.

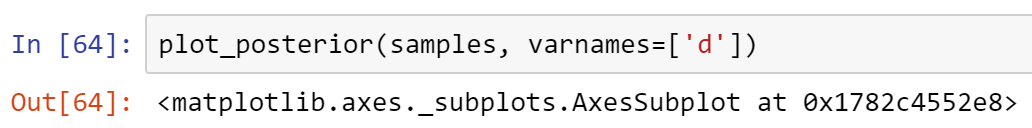


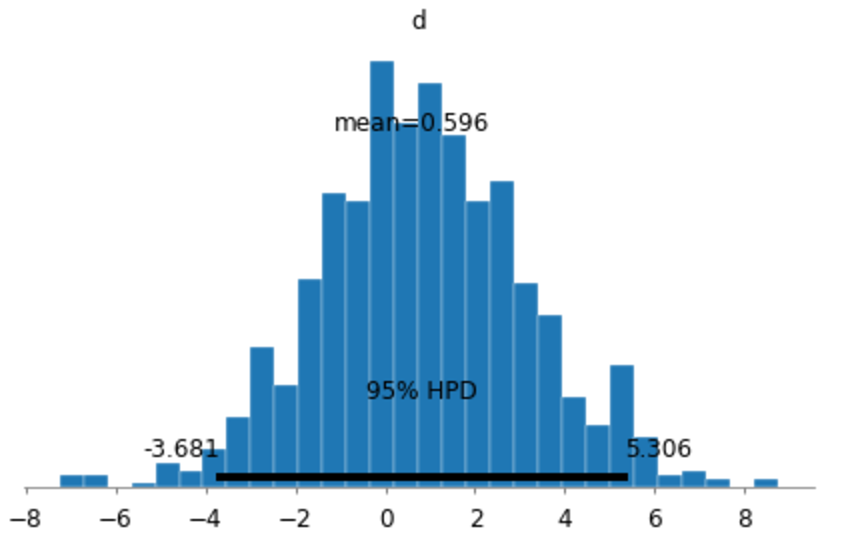












Therefore, in Nashville you got a 60% chance of having rain in July.

Two Groups with Binary Outcome

Now that we have seen how to generalize normally-distributed data to another distribution, we are equipped to analyze other data types. Binary outcomes are common in clinical research:

- survival/death

- true/false

- presence/absence

- positive/negative

Never, ever dichotomize continuous or ordinal variables prior to statistical analysis.

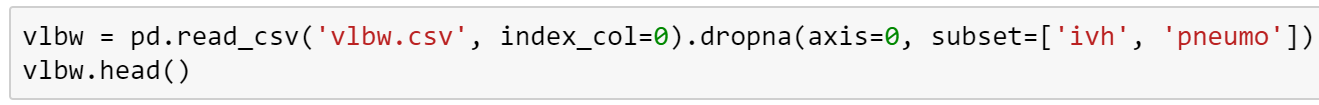
In practice, binary outcomes are encoded as ones (for event occurrences) and zeros (for non-occurrence). A single binary variable is distributed as a Bernoulli random variable.

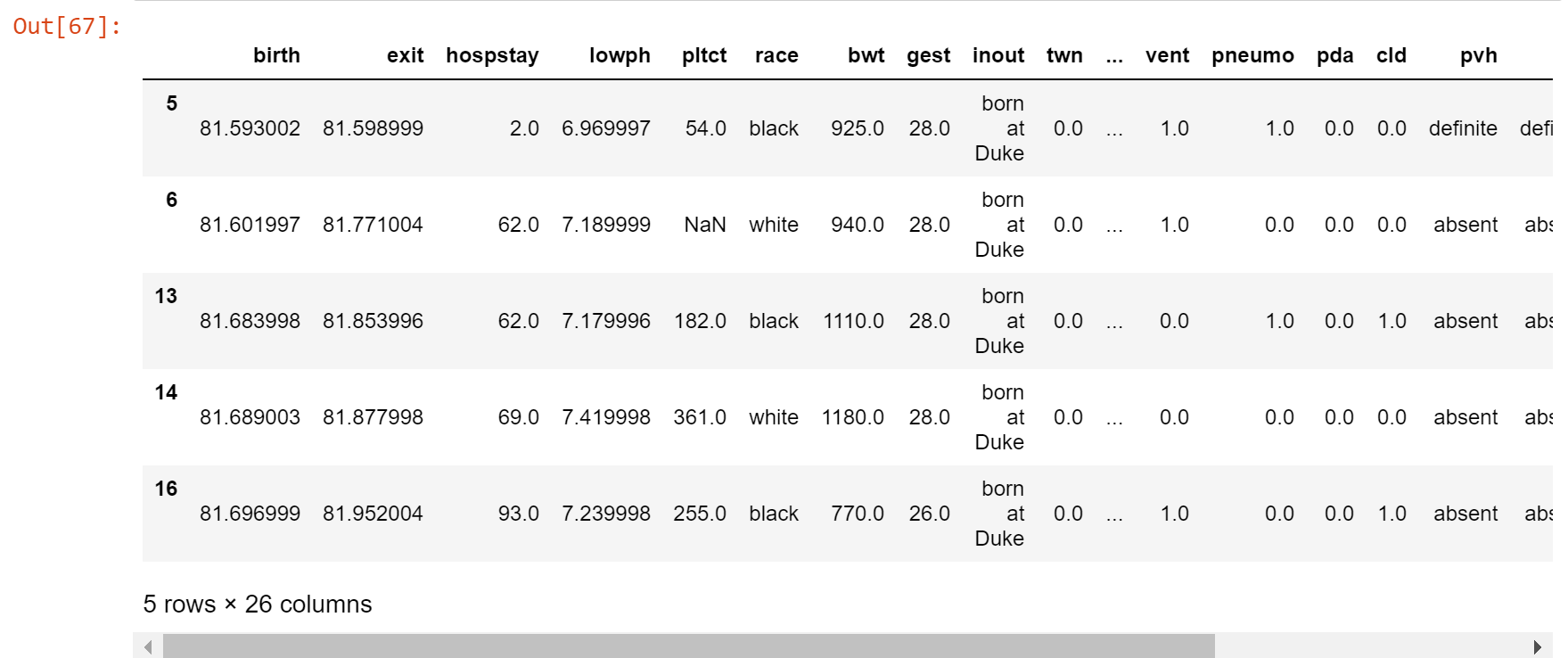
Such events are sometimes reported as sums of individual events, such as the number of individuals in a group who test positive for a condition of interest. Sums of Bernoulli events are distributed as binomial random variables.

The parameter in both models is p, the probability of the occurrence of an event. In terms of inference, we are typically interested in whether p is larger or smaller in one group relative to another.

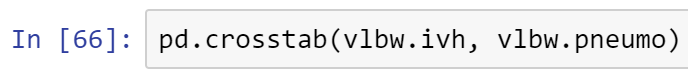
To demonstrate the comparison of two groups with binary outcomes using Bayesian inference, we will use a sample pediatric dataset. Data on 671 infants with very low (<1600 grams) birth weight from 1981-87 were collected at Duke University Medical Center. Of interest is the relationship between the outcome intraventricular hemorrhage (IVH) and predictor such as birth weight, gestational age, presence of pneumothorax and mode of delivery.

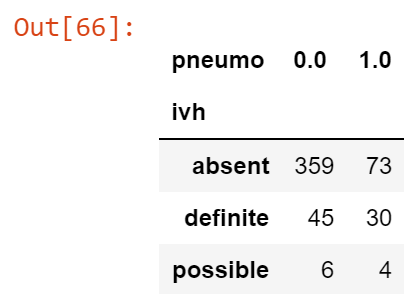




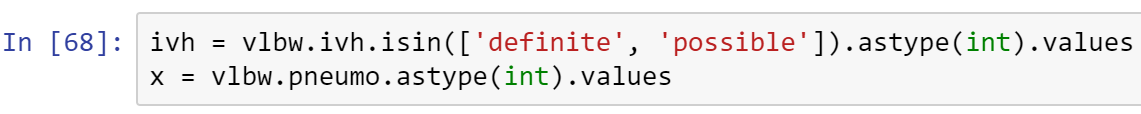


To demonstrate binary data analysis, we will try to estimate the difference between the probability of an intraventricular hemorrhage for infants with a pneumothorax.





We will create a binary outcome by combining "definite" and "possible" into a single outcome.



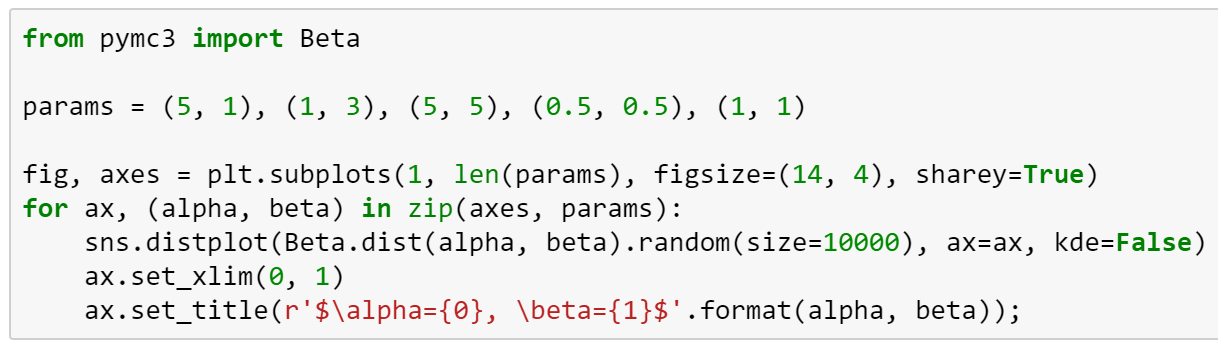
Prior choice

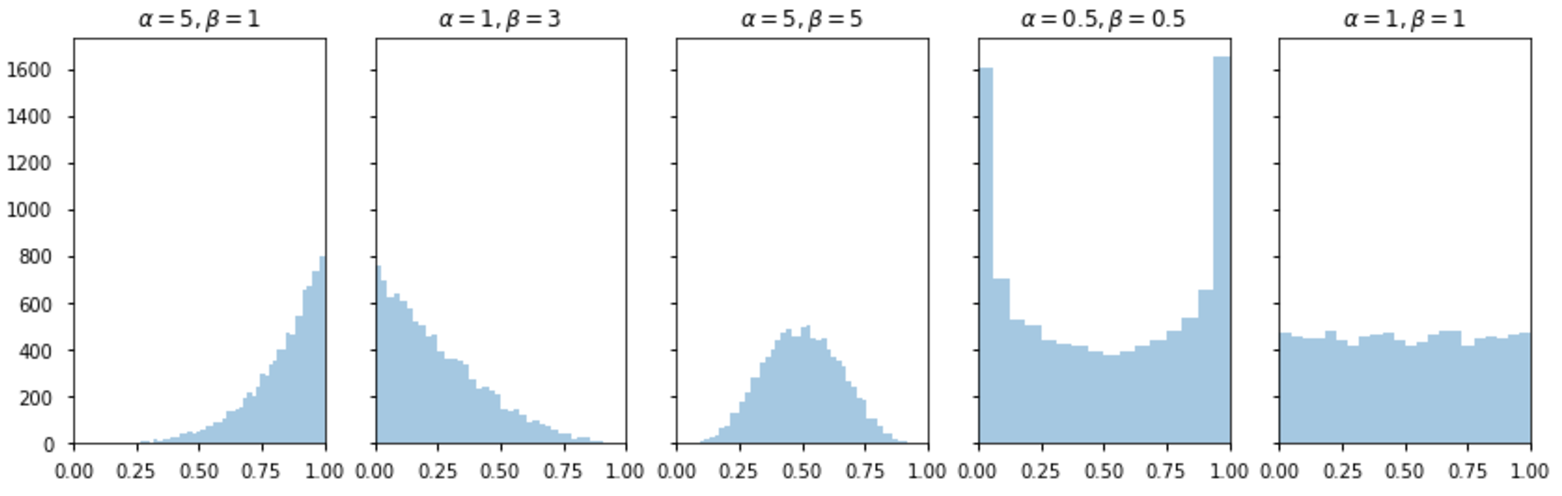
What should we choose as a prior distribution for p?

We could stick with a normal distribution, but note that the value of p is considered by the laws of probability. Namely, we cannot have values smaller than zero nor larger than one. So, choosing a normal distribution will result in ascribing positive probability to unsupported values of the parameter. In many cases, this will still work in practice, but will be inefficient for calculating the posterior and will not accurately represent the prior information about the parameter.

A common choice in this context is the beta distribution, a continuous distribution with 2 parameters and whose support is on the unit interval.



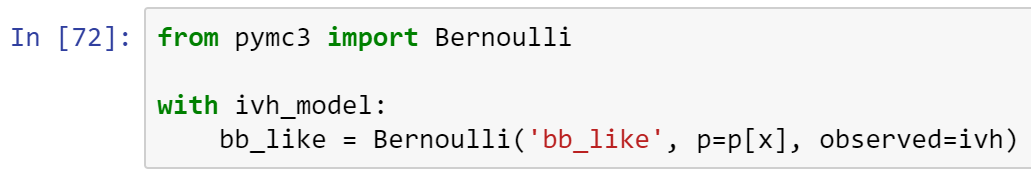




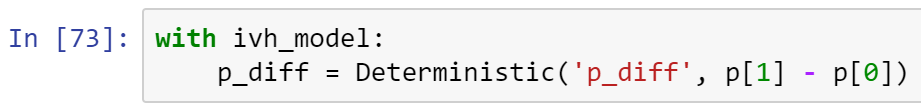
So let's use a beta distribution to model our prior knowledge of the probabilities for both groups. Setting alpha = beta = 1 will result in a uniform distribution of prior mass.

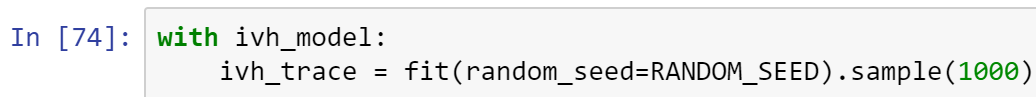


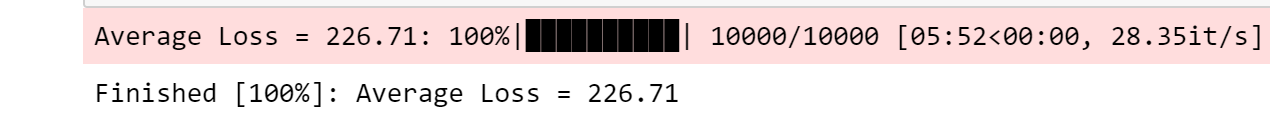
We can now use p as the parameter of our Bernoulli likelihood. Here, x is a vector of zeros as ones, which will extract the appropriate group probability for each subject.

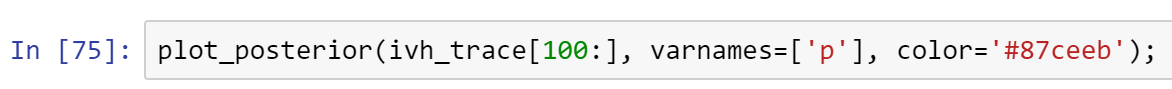


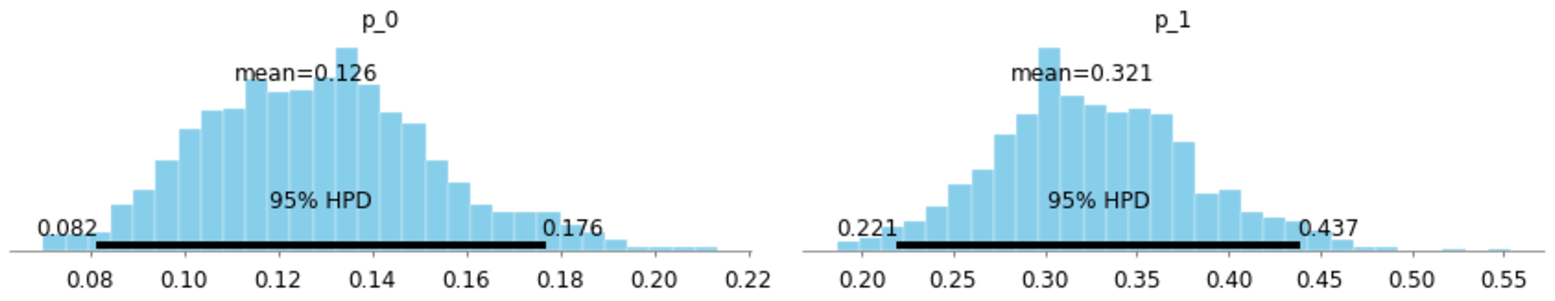
Finally, since we are interested in the difference between the probabilities, we will keep track of this difference.



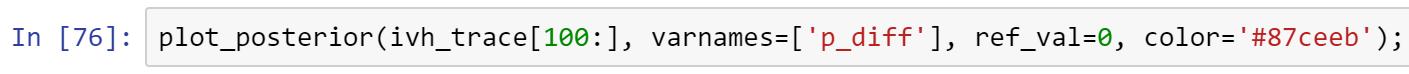


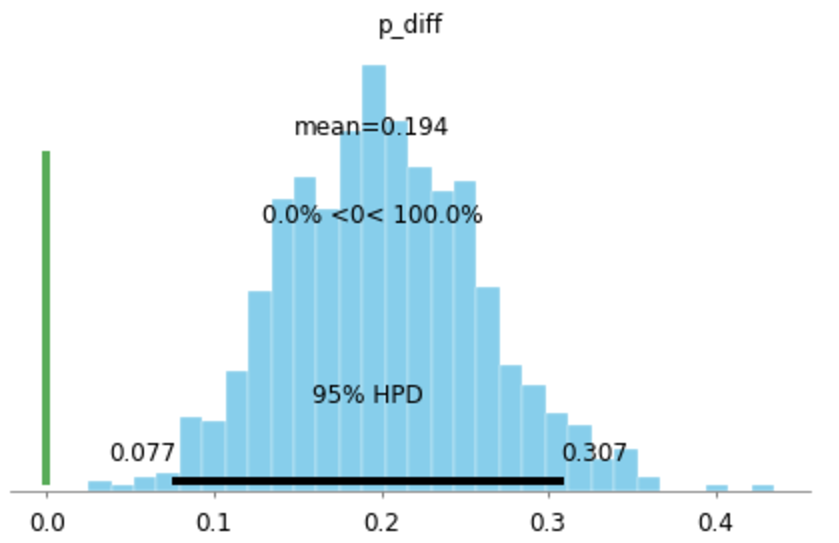






We can see that the probability that p is larger for the pneumothorax with probability one.





**Reflect on this exercise . . .**

What do you gain?

Where can you use this knowledge?

And in what way?

Share your answer to these three questions in the discussion board.

* Copy all your code into a Word doc, place your name on it, and submit in Canvas.