Some notes on Bayesian Statistics

Frequentist vs. Bayesian Inference

Bayesian statistics/inference are presented in opposition to what is called *frequentist* statistics. If you've taken any type of introductory statistics class, you almost certainly learned some type of frequentist statistics. The following three postulates come from Wasserman (2013).

1. Probability refers to limiting relative frequencies. Probabilities are objective properties of the real world.
2. Parameters are fixed, unknown constants. Because they are not fluctuating, no useful probability statements can be made about parameters.
3. Statistical procedures should be designed to have well-defined long run frequency properties. For example, a 95 percent confidence interval should trap the true value of the parameter with limiting frequency at least 95 percent.

These three frequentist postulates can be compared with their Bayesian counterparts.

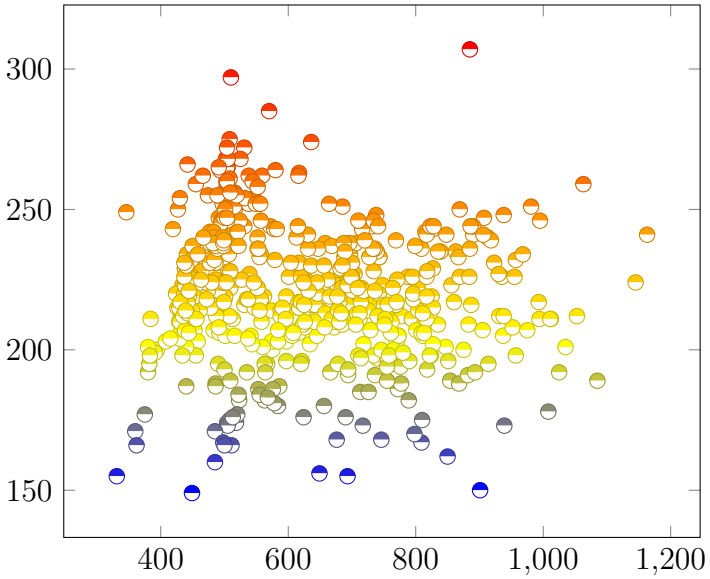
1. Probability describes degree of belief, not limiting frequency. As such, we can make probability statements about lots of things, not just data which are subject to random variation. For example, I might say that "the probability that Albert Einstein drank a cup of tea on August 1, 1948" is .35. This does not refer to any limiting frequency. It reflects my strength of belief that the proposition is true.
2. We can make probability statements about parameters, even though they are fixed constants.
3. We make inferences about a parameter θ by producing a probability distribution for θ. Inferences, such as point estimates and interval estimates, may then be extracted from this distribution.

Linear Modeling

To get an intuition about the distinction between the two types of inference, let's work through an example of linear modeling under both views. Linear models have the following general format:

$$y = \beta_0 + \Sigma_{i=1}^{n} \beta_i x_i$$

Here, y is a predicted/dependent variable while our collection of x values are predictors/independent variables. When we only have one x we define a line (think back to $y = mx + b$ from high school math). The following data come from Hillenbrand et al. (1995). It contains information about vowel production of 48 adult female speakers of English. The x-axis corresponds to F1 while the y-axis corresponds to f0.



It has been claimed that high vowels have an intrinsically higher pitch than low vowels. While it's not super strong, it does seem like this is the case in the plot above. But how do we confirm that there is in fact a negative linear effect? This is where linear modeling comes in to play. We can do linear regresison on this data. Frequentists would say we are trying to find a model of the following type:

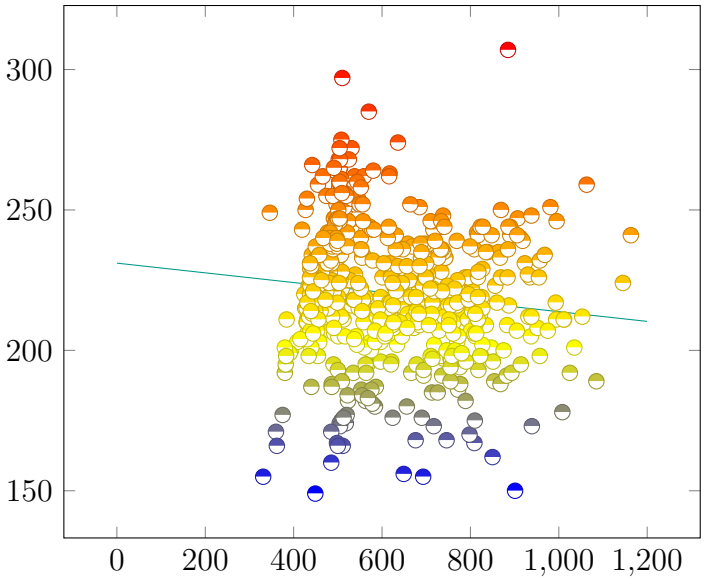
$$y = \beta_0 + \beta_1 x + \epsilon$$

In layman's terms, we want to find a line that goes through our data with some wiggle room for points to exist off of the line. The latter part is what the ϵ captures. The distance between any line and the actual points are called the *residuals*. We say the line that best fits the data is the one that reduces the residual sum of square (also called the least squares method). Here's what that looks like with fancy notation:

$$RSS = \Sigma_{i=1}^{n}(\hat{y}(x_i) - y_i)^2$$

The term $\hat{y}(x_i)$ refers to the predicted value at a given point x. The best line for our observed data has an intercept (β_0) of 231.06476 and a slope (β_1) of -0.01729. This is shown on the same plot from above.

We got some other information that might be of interest to us. For example, we find that the slope parameter significantly varies from 0 with a p-value of 0.003823 and confidence



interval of [-0.02897489, -0.005595637]. Ok, but how should we interpret what this tells us? Frequentist statistics are based on the premise that we are answering the question, "what is the probability of seeing this data given our hypothesis?" A p-value in this instance tells us what is the probability of seeing the data that we did **or something even more extreme** given the null hypothesis that there is no effect of F1 on f0. The only inference we should be making here is that we reject the null hypothesis since that's what we were testing in the first place. Notice in this case we also get a point estimate of the slope value which best fits the line. Let's now compare this to a Bayesain approach.

Here are the three steps that we need to do Bayesian inference.

1. Choose a probaility density $f(\theta)$ – called the **prior distribution** – that expresses our beliefs about a parameter θ before we see any data.
2. Choose a statistical model $f(x \mid \theta)$ that reflects our beliefs about x given θ.
3. After observing data $X_1, \ldots, X_n$, we update our beliefs and calculate the **posterior** distribution $f(\theta \mid X_1, \ldots, X_n)$.

Notice in the final line that this indicates that Bayesian analysis doesn't give us a point estimate and instead informs us of the most likely parameters of the model given the data we have seen. This is all centered around Bayes' theorem. If you've never seen it, it looks like this:

$$\mathbb{P}(\Theta = \theta \mid X = x) = \frac{\mathbb{P}(X = x \mid \Theta = \theta)\mathbb{P}(\Theta = \theta)}{\Sigma_\theta \mathbb{P}(X = x \mid \Theta = \theta)\mathbb{P}(\Theta = \theta)}$$

And also like this:

$$f(\theta \mid x) = \frac{f(x \mid \theta) f(\theta)}{\int f(x \mid \theta) f(\theta) d\theta}$$

You may see $f(x \mid \theta)$ replaced by $\mathfrak{L}(\theta)$ since this is the *likelihood*. [1] You will often also see the denominator removed because it only acts a scaling factor. This allows us to make the following claim:

$$f(\theta \mid x) \propto \mathfrak{L}(\theta) \times f(\theta)$$

This tells us that the posterior ($f(\theta \mid x)$) scales proportionally to the likelihood of the data given the model parameters ($\mathfrak{L}(\theta)$) times the prior probability of the data ($f(\theta)$). When we explain it in words it's very elegant and straightforward. In fact, this method was known long before it became popular. One big problem is that in most cases we don't have an analytical solution and instead have to rely on simulations to compute the posterior distribution. As was the case for the frequentist approach, we won't get into the details and instead I'll just show you what the output of a Bayesian analysis of our data would look like.

The model we'll use to analyze the f0 data is listed below. Note "$\sim$" below can be read as, "is distributed with properties..."

$$\begin{aligned}
y_i &\sim \mathcal{N}(\mu_i, \sigma) \\
\mu_i &= \beta_0 + \beta_1 (x_i - \bar{x}) \\
\beta_0 &\sim \mathcal{N}(220, 20) \\
\beta_1 &\sim \mathcal{N}(0, 1) \\
\sigma &\sim \mathcal{U}(0, 50)
\end{aligned}$$

Ultimately, we are starting with some prior assumptions and seeing how the observation of new data changes our beliefs. Our assumptions are as follows:

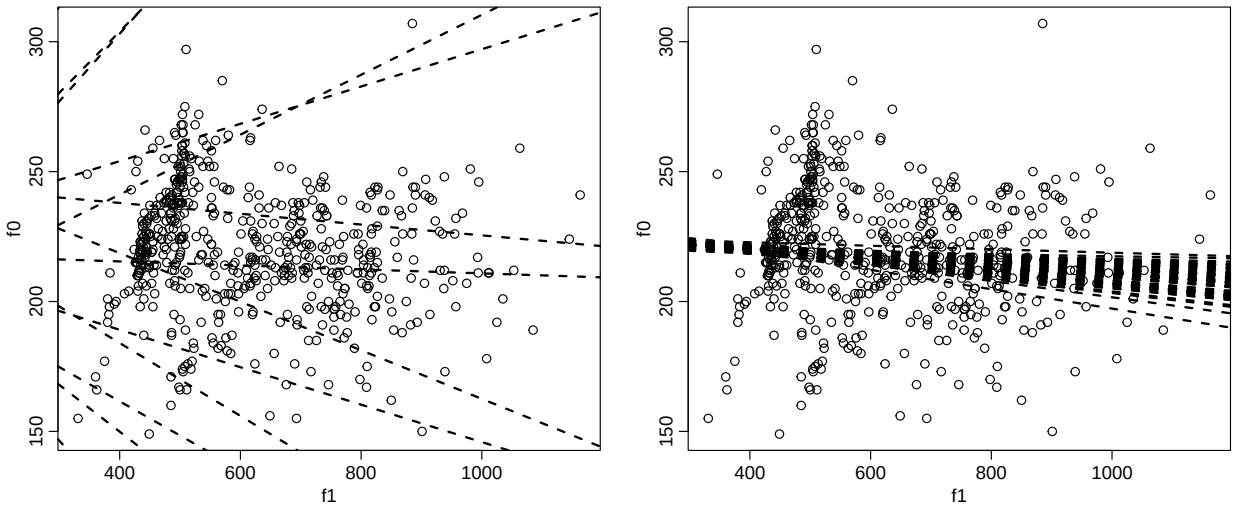
1. Individual f0 (y_i) values are normally distributed based on individual means but constant standard deviation. [This is the likelihood]
2. Individual means are constants based on "slope" and "intercept" values, as well as the observed sample mean of F1 $\bar{x}$.
3. Intercept (mean f0) is normally distributed.
4. Slope is normally distributed around 0 (Prior belief that there is *no* effect of F1 on f0)

[1] I should be a little more careful, technically $\mathfrak{L}_n(\theta) = f(x_1, \ldots, x_n \mid \theta) = \Pi_{i=1}^{n} f(x_i \mid \theta)$

5. Standard deviation is uniformly distributed.

Let's ignore σ for now and just think about how we would predict y. We could do this by using the formula in line 2 with values for β_0 and β_1 drawn from the distributions. I sampled from each distribution 50 times and plotted the lines in the left plot below. Note, fewer than 50 lines appear which means we have some pretty wild prior ideas about what could be a possible relationship between f0 and F1.

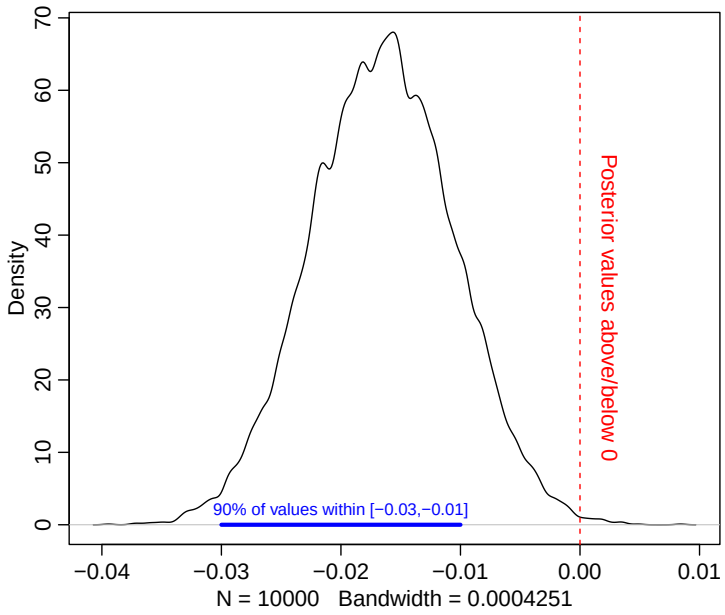
Doing Bayesian inference is now typically done with some type of sampling method. Essentially, we want to take a meaningful sample of our original distribution and then use those values to create a new distribution. If you want to learn more about how this is done in practice, I advise looking at McElreath (2018) or Kruschke (2014). The plot on the right shows the result of sampling values from the new posterior distribution.



Clearly we can see that there is a negative trend given our new distribution. But let's bring this back to regression. We can look at the posterior distribution specifically for the β_1 parameter to get an idea of how f0 and F1 are related. There are two things we might be interested in. First, we might be interested in how many of the posterior β_1 values in our sample are *less than zero*. The method I used created a posterior distribution with 10000 samples and 9,971 (99.71%) of the values were less than zero. This is pretty good evidence that F1 has a negative effect on f0 zero. Additionally, we can look at the 90% "compatibility interval". This shows us the interval where 90% of the posterior distribution for this parameter lies. Here, we see that it is between -0.01 and -0.03 (recall that frequentist linear regression gave us a value of -0.01729).

Notice, in this case we are saying nothing about the validity of any hypothesis. Instead, we are saying *given prior expectations and new evidence, here is where we expect the parameter relating f0 and F1 to be.* There are no p-values, we're simply updating our beliefs about the world based on new evidence. As you can hopefully see, ones prior can really affect the outcome. This is a point of contention with people feeling strongly about it from both

directions. Again, consult the books referenced above for more discussion.



References

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). Acoustic characteristics of american english vowels. *The Journal of the Acoustical society of America*, 97(5):3099–3111.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

McElreath, R. (2018). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.

Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.