# 18.05 Problem Set 6, Spring 2018

## (due outside 4-174 Monday, Apr. 9 at 9:30 AM)

**Problem 1.** (20 pts.) **Hypotheses and data.**
Each of the following scenarios describes a setup where data is generated by an unknown distribution. We know enough to be able to form hypotheses about the distribution. For each scenario:

(i) Use a full English sentence to describe the possible hypotheses.

(ii) Use a full English sentence to describe the data in terms of a (theoretically) repeatable experiment.

(iii) Define any notation necessary and give the prior and likelihood using tables or paramaterized pmf or pdf—whichever is appropriate. If the prior is not fully specified choose a reasonable prior and explain your choice.

Explanations do not need to be lengthy. Some pieces of information, e.g. the number of trials, may not be given. In that case you will need to give it a name and give your answer in terms of it.

**(a)** I have four coins in a drawer. One is fair and three have probability 0.4 of heads. I pick one at random. In order to inform my guess about which type of coin I chose, I toss it 3 times getting 2 heads.

**(b)** I'm testing an experimental drug on patients. I will classify the results into two categories, success and failure. I run trials with $n$ patients. I start the trials with no belief about how effective the drug will be. (Since no numbers are given in this problem you will need to assign symbols to the relevant values and give your answer in terms of these.)

**(c)** The drug in part (b) turns out to be effective, being successful in about 75% of patients. I develop a new drug in the same class and run a trial experiment with 30 subjects.

**(d)** The length of time the drug can be detected in the bloodstream follows an exponential distribution with unknown rate parameter $\lambda$. Based on previous experience our belief about $\lambda$ follows a normal distribution with mean $\theta_0$ and variance $\sigma_0^2$.

**Problem 2.** (30 pts.) **Spun gold.**
When spun on edge 250 times a certain coin came up heads 140 times and tails 110.

**(a)** Compute the probability that a fair coin would come up 140 or more times heads. Give an exact formula and use the R function `pbinom` to give a numerical answer.

**(b)** In class and in the notes we considered the odds and Bayes factor for a hypothesis $H$ vs. $H^c$. Given data $D$ we have the posterior odds of $H$ vs. $H^c$ are

$$\text{odds} = \frac{P(H|D)}{P(H^c|D)} = \frac{P(D|H)P(H)}{P(D|H^c)P(H^c)}$$

The factor $\frac{P(D|H)}{P(D|H^c)}$ is the Bayes factor and the factor $\frac{P(H)}{P(H^c)}$ is the prior odds. We have the same definitions for comparing any two hypotheses $H_1$ vs. $H_2$

$$\text{odds} = \frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)P(H_1)}{P(D|H_2)P(H_2)}; \quad \text{Bayes factor} = \frac{P(D|H_1)}{P(D|H_2)}; \quad \text{prior odds} = \frac{P(H_1)}{P(H_2)}.$$

Consider the hypotheses $H_0$ = 'probability of heads is 1/2' and $H_1$ = 'probability of heads is $\frac{140}{250}$'. Using the data of 140 heads and 110 tails, what is the Bayes factor for $H_0$ versus $H_1$?

It shouldn't surprise you that the Bayes factor is much less than 1. Give a short explanation of why you would expect this.

**(c)** Let $\theta$ be the probability that the spun coin comes up heads. Below are five priors, use R to graph them all on a single plot using a different color for each prior. You can find some sample R plotting code in the 'Sample plotting' tutorial at the usual link from our class web site.

Which of the priors do you find most reasonable and why?

(i) $f(\theta)$ is Beta(1,1), or equivalently uniform(0, 1).

(ii) $f(\theta)$ is Beta(10, 10).

(iii) $f(\theta)$ is Beta(50, 50).

(iv) $f(\theta)$ is Beta(500, 500).

(v) $f(\theta)$ is Beta(30, 70).

**(d)** Without computing, rank these priors (lowest to highest) by the *prior probability that the coin is biased toward heads.* Explain your ranking, including ties.

**(e)** Let $x$ be the data of 140 heads and 110 tails. The posterior distribution $f(\theta|x)$ depends on our choice for prior distribution $f(\theta)$. Find the posterior for each of the above priors and graph all them on a single plot using the same colors as in part (b).

Considering the mean and shape of each of the priors, give a short explanation for the positions of each of the posteriors in each graph.

**(f)** Use the R function `pbeta(x,a,b)` to compute the posterior probability of a bias in favor of heads for each of the priors. Rank the posteriors from most biased to least biased towards heads. Is this consistent with your plot in part (d)?

**(g)** Note that $140/250 = 0.56$. Let $H_1$ = '$0.55 \le \theta \le 0.57$' and let $H_0$ = '$0.49 \le \theta \le 0.51$'. Use the following steps to estimate the posterior odds of $H_1$ versus $H_0$, first for prior (i) and then for prior (iv).

Step 1. Use the posterior densities and the fact that $H_0$ and $H_1$ cover small intervals to estimate the posterior probabilities of $H_0$ and $H_1$. (No integral is needed. Leave the normalizing constant uncomputed.)

Step 2. Use the posterior probabilities to give the posterior odds.

Explain the difference between the results for priors (i) and (iv).

**Problem 3.** (20 pts.) **Bayes at the movies.**

**(a)** A local theater employs two ticket sellers, Alice and Bob, although only one of them works on any given day. The number of tickets $X$ that a ticket seller can sell in an hour is modeled by a distribution which has mean $\lambda$, and probability mass function

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

for $k = 0, 1, 2, \ldots$. (This distribution is called a Poisson distribution.)

Suppose that Alice sells, on average, 10 tickets an hour and Bob sells, on average, 15 tickets an hour. One day the manager stays home sick. He knows Bob is supposed to work that day, but thinks there are 1 to 10 odds that Alice is filling in for Bob (based on Bob's prior history of taking advantage of Alice's generous nature when the manager is away). The suspicious manager monitors ticket sales online and observes that over the span of 5 hours there are 12, 10, 11, 4, and 11 tickets sold. What are the manager's posterior odds that Alice is filling in for Bob?

**(b)** This is an entertaining problem, so I kept it as originally written. But the central premise (that the activity of different ticket sellers is reasonably modeled by Poisson distributions with different rate parameters) seems very unlikely to me. Give an argument for why it's unlikely, and describe a different job where it might be a reasonable model.

**Problem 4.** (20 pts.) **Normal is the new normal.**

Your friend transmits an unknown value $\theta$ to you over a noisy channel. The noise is normally distributed with mean 0 and a known variance 9. so the value $x$ that you receive is modeled by $N(\theta, 3^2)$. Based on previous communications, your prior on $\theta$ is $N(5, 4^2)$.

**(a)** Suppose your friend transmits a value to you that you receive as $x = 6$. Show that the posterior pdf for $\theta$ is $N(141/25, 144/25) = N(5.64, (2.4)^2)$. The reading gave formulas for normal-normal updating, but for this problem carry out the calculations from scratch using our usual Bayesian update table and a good dose of algebra.

**(b)** Suppose your friend transmits the same value $\theta$ to you 4 times. You receive the data, i.e. signal plus noise, as $x_1, \ldots, x_4$ with sample mean $\bar{x} = 6$. Using the same prior and known variance $\sigma^2$ as in part (a), show that the posterior on $\theta$ is $N(5.88, 1.97)$. Plot the prior and posterior on the same graph. Describe how the data changes your belief about the true value of $\theta$.

In the reading we gave normal-normal Bayesian update formulas which apply in this case. For this problem you can use these formulas, which we provide here:

$$a = \frac{1}{\sigma^2_{\text{prior}}}, \qquad b = \frac{n}{\sigma^2}, \qquad \mu_{\text{post}} = \frac{a\mu_{\text{prior}} + b\overline{x}}{a+b}, \qquad \sigma^2_{\text{post}} = \frac{1}{a+b}. \tag{1}$$

**(c)** How do the mean and variance of the posterior change as more data is received? What is gained by sending the same signal multiple times? Here we want you to interpret the equations (1).

**Problem 5.** (10 pts.) **Censored data.**
Sometimes data is not reported in full. This can mean only reporting values in a certain range or not reporting exact values. We call such data censored.

We have a 4-sided die and a 6-sided die. One of them is chosen at random and rolled 5 times. Instead of reporting the number of spots on a roll we censor the data and just report

1 if the roll is a 1;   0 if the roll is not a 1.

**(a)** Suppose the data for the five rolls is 1, 0, 1, 1, 1. Starting from a flat prior on the choice of die, update in sequence and report, after each roll, the posterior odds that the chosen die is 4-sided die.

**(b)** A censored value of 1 is evidence in favor of which die? What about 0? How is this reflected in the posterior odds in part (a)?