Jordan Schupbach
STAT 532
September 11, 2015

Homework # 2

1. - Many books define probabilities using the axioms of probabilities (sometimes stated as Kolmogorov's Axioms). In 'Statistical Inference' by Casella and Berger, the authors introduce probabilities with the following definition.

     Given a sample space S and an associated sigma algebra $\mathcal{B}$, a *probability function* is a function P with domain $\mathcal{B}$ that satisfies

     1. $P(A) \geq 0$ for all $A \in \mathcal{B}$.

     2. $P(S) = 1$.

     3. If $A_1, A_2, \cdots \in \mathcal{B}$ are pairwise disjoint, then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

     In the book 'Bayesian Data Analysis (3rd Edition)' by Gelman et al., the authors write this same definition in somewhat simpler terms: "the mathematical definition of probability: that probabilities are numerical quantities, defined on a set of 'outcomes,' that are nonnegative, additive over mutually exclusive quantities, and sum to 1 over all possible mutually exclusive outcomes."

   - The word 'probability' has shared many meanings over the course of history, and its specific meaning should be made in the context of when it was said and by whom. Uses of the word originate from two separate goals. One, in describing *uncertainty*, and the other in the process of *induction*. Prior to about 1700, probability was limited to describing the cogency of a declaration. After this, the mathematical definition emerged out of the gambling world and eventually became the dominant use of the word. To further convolute the definition, divisions within the discipline of statistics use the word in different ways. For instance, frequentists restrict its usage to describing a long-run frequency, whereas Bayesians and Likelihoodists accept its usage as a measure of certainty for a single outcome.

2. One criticism of the likelihood principle is that it doesn't take into account experimental design. All of the information regarding the parameter is contained in the likelihood. Thus, two completely different experiments could yield the same inference about a parameter despite having different study designs, so long as both have the same likelihood. To illustrate, consider two experiments of flipping a coin, where the researcher's parameter of interest is p, the probability that a single coin flip will result in a "heads". In one experiment, call it E_1, we flip a coin 12 times and record the number of heads. In the other experiment, E_2, we record the number of tails until the third head. We have E_1 associated with the family of $binomial(12, p)$ pmfs and E_2 associated with the family of $nbinom(3, p)$ pmfs. Now, consider two sample points $x_1 = 3$ (3 out of 12 heads in E_1) and $x_2 = 12$ (the third head occurs on the 12th coin

flip in E_2). We have the following likelihood functions for each experiment

$$L(p|x_1 = 12) = \binom{12}{3} p^3(1-p)^{12} \text{ for } E_1$$

and

$$L(p|x_2 = 13) = \binom{11}{2} p^3(1-p)^{12} \text{ for } E_2$$

Before continuing, we will state the formal likelihood principle:

Suppose that we have two experiments, $E_1 = (\boldsymbol{X}_1, \theta, \{f_1(\boldsymbol{x}_1|\theta\})$ and $E_2 = (\boldsymbol{X}_2, \theta, \{f_2(\boldsymbol{x}_2|\theta\})$, where the unknown parameter θ is the same in both experiments. Suppose $\boldsymbol{x}_1^*$ and $\boldsymbol{x}_2^*$ are sample points from E_1 and E_2 respectively, such that

$$L(\theta|\boldsymbol{x}_2^*) = CL(\theta|\boldsymbol{x}_1^*)$$

for all θ and for some constant C that may depend on $\boldsymbol{x}_1^*$ and $\boldsymbol{x}_2^*$ but not θ. Then,

$$Ev(E_1, \boldsymbol{x}_1^*) = Ev(E_2, \boldsymbol{x}_2^*)$$

where $Ev(E, \boldsymbol{x})$ stands for the *evidence about θ arising from E and $\boldsymbol{x}$*.

That is, the formal likelihood principle states that the evidence about p from both experiments with their respective samples should be equal since their likelihoods are proportional. In fact, the MLE in either case is $\hat{p} = \dfrac{3}{12} = .25$. Suppose the following mixed experiment. A researcher has his colleague flip a coin and carry out E_1 if he flips a heads and carry out E_2 if he flips a tails. The researcher, unknowing of which experiment was conducted, receives the data of 3 heads out of 12 flips from his colleague. In either case, the researcher concludes that the probability of getting a heads on a single coin flip is .25, however, he does not know whether 3 or 12 was the fixed value. The former is an illustration of the formal sufficiency principle and the conditionality principle; which, by Birnbaum's theoerem, is both necessary and sufficient for the formal likelihood principle.

Now, let us consider testing the null hypothesis $H_0 : p = .5$ versus the alternative $H_a : p < .5$. Given that the null is true, the probability of observing 7 or fewer heads in E_1 is

$$\begin{aligned} P(X_1 \leq 3|p = .5, n = 12) &= \left(\binom{12}{12} + \binom{12}{11} + \binom{12}{10} + \binom{12}{9} \right) \left(\frac{1}{2} \right)^{12} \\ &= \frac{299}{4096} \\ &\approx .073 \end{aligned}$$

In E_2, we have the probability of needing to conduct 12 or more experiments is

$$\begin{aligned} P(X_2 \geq 12|p = .5, r = 3) &= 1 - P(X_2 \leq 12|p = .5, r = 3) \\ &= 1 - \left(\binom{10}{2} \left(\frac{1}{2}\right)^{11} + \binom{9}{2} \left(\frac{1}{2}\right)^{10} + \cdots + \binom{2}{2} \left(\frac{1}{2}\right)^{3} \right) \\ &= \frac{134}{4096} \\ &\approx .0327 \end{aligned}$$

These calculations are the associated p-values of each experiment. Hence, we would reject the null hypothesis in E_1 at the .05 confidence level and would fail to reject the null in E_2 if we knew which experiment was conducted. Some would say that because of this the likelihood principle is flawed as it doesn't take into account the experimental design. Likelihoodists would likely respond by saying that this simply exemplifies the flaws in significance testing. The arguments within this problem come from 2 sources. The first is in section 6.3.2 in 'Statistical Inference' by Casella and Berger and the second is in the wikipedia page for 'Likelihood.' I am uncertain as to the original source of this problem, however I would suspect that it comes from the debates between Fisher and Neyman.

3. The sampling distribution is related to the likelihood function in that if we fix a sample $\boldsymbol{X}$, then the function of θ defined by

$$L(\theta|\boldsymbol{x}) = f(\boldsymbol{x}|\theta)$$

is the likelihood function. Then, if we consider two values of θ, say θ_1 and θ_2, and find that

$$P_{\theta_1}(\boldsymbol{X} = \boldsymbol{x}) = L(\theta_1|\boldsymbol{x}) > L(\theta_2|\boldsymbol{x}) = P_{\theta_2}(\boldsymbol{X} = \boldsymbol{x})$$

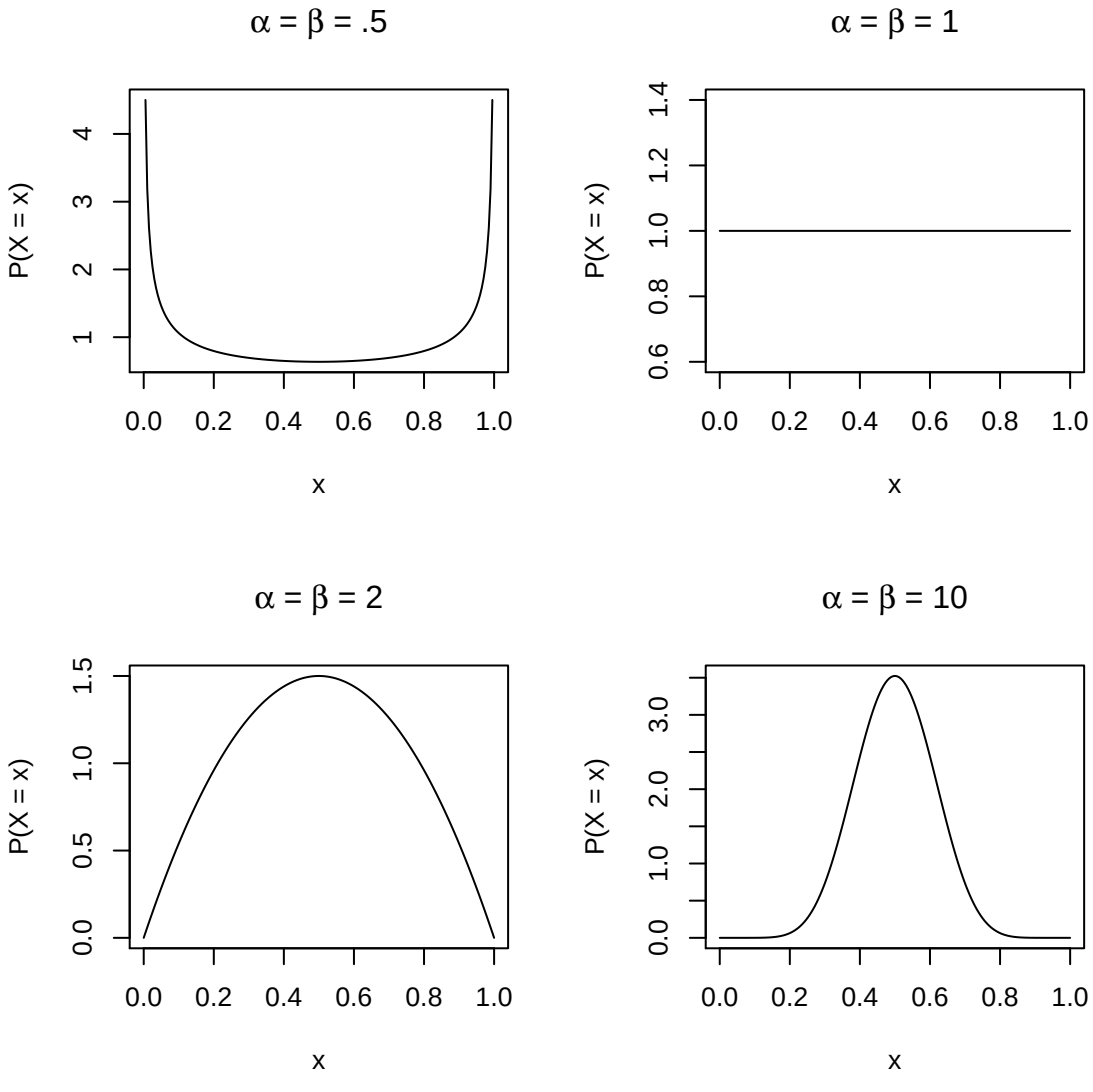
we find that the value observed sample is more likely to have occured if $\theta = \theta_1$ than if $\theta = \theta_2$. Hence, if we maximize $L(\theta|\boldsymbol{x})$, we find the most likely value θ given the sample $\boldsymbol{x}$. Thus, the likelihood function (and by extension, MLE's) is directly related to the sampling distribution through the definition of the likelihood function.

4. The likelihood function is a function of our parameter of interest for a given sample which tells us how 'likely' a given value of our parameter of interest is, given we observe a specific set of data. To be clear though, we do not mean 'likely' in a strict probabalistic sense. In fact, a likelihood function is not a probability function. First, the sum over all possible values of the likelihood function does not have to be 1. That is, in probabilities, when we consider an experiment, the probability of something happening is 1. If we want to make further statements about what that something is, it might be less than 1 (but greater than zero). In contrast, the sum over the likelihood function can be any nonnegative real number (it might even be infinite).

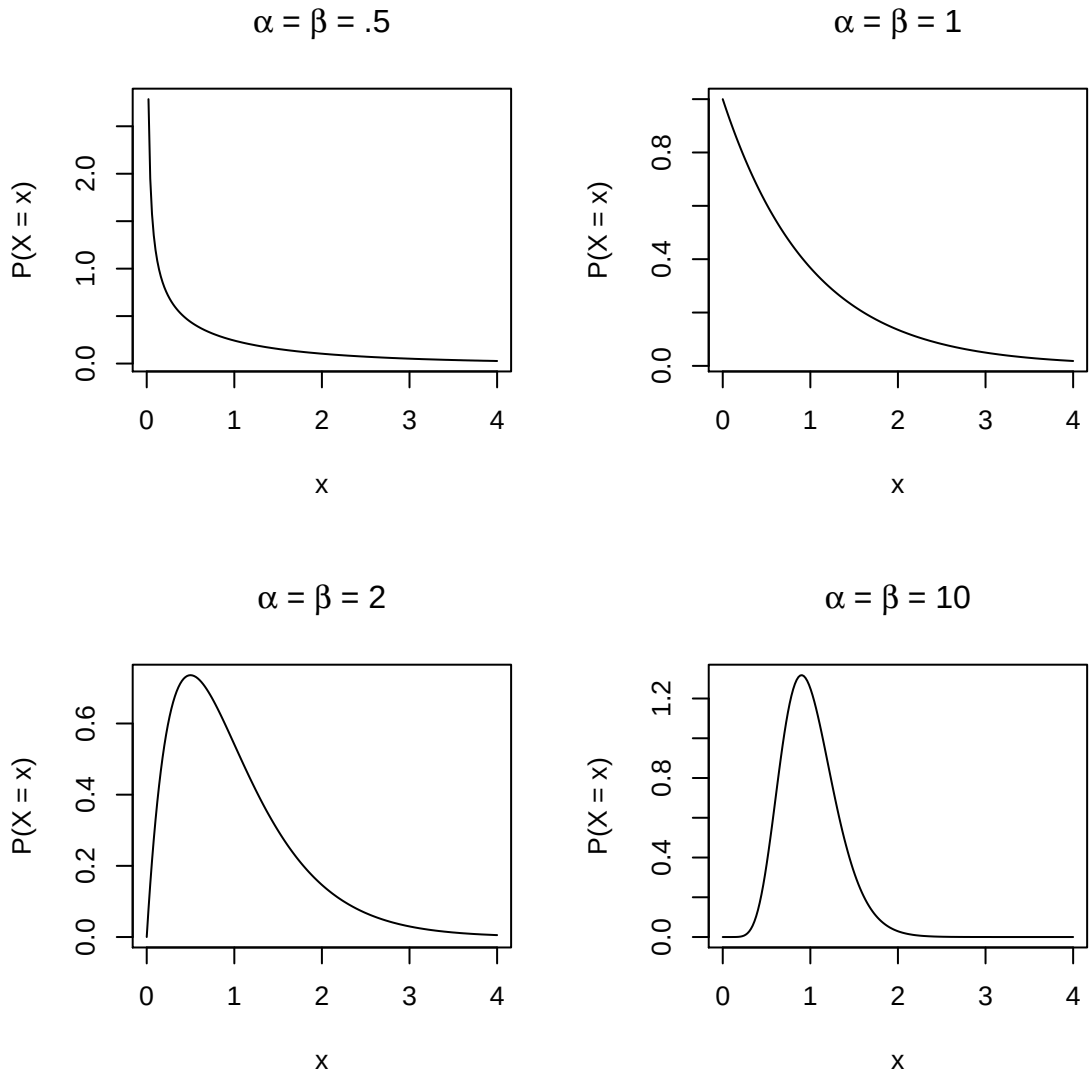
5. If we want to estimate survivability of some birds, and one week, all of the birds survived, we might conclude that the probability of all birds surviving any given week

is 1. This would effectively mean that these birds are invincible, which we know not to be true. If we use Bayesian methods, situations such as these are not a problem, as we can simply restrict our prior distribution so that it is impossible to conclude survivability of 1. As pointed out in class, our choice of prior can in some ways be related to increasing our sample size. For instance, choosing a $beta(1,1)$ prior would be equivalent to adding one survived and one died to the sample and choosing a $beta(.5,.5)$ would be equivalent to adding a half to each group. In effect, based on how likely or unlikely we think high (or low) values of survivabililty are in the samples for various sample sizes, we may want to choose an appropriate prior as to not strongly influence the inferences made from the data observed.

6. We get the following plots of the beta distribution for different values of $\alpha = \beta$.



7. We get the following plots of the gamma distribution for different values of $\alpha = \beta$.



The gamma distribution is often used for the shape parameter for a normal model with mean known and variance unknown as it is the conjugate prior for this distribution.

8. (a)
   (b)

9.

10. (a)
    (b)
    (c)
    (d)
    (e)
    (f)

11.

12.

13.

14.

15.

16. (a)

    (b)

    (c)

17.

18.

R-Code

```
## Require Packages
require(xtable)

x <- rep("NA", 8)
for (i in 1:8) x[i] <- choose(20, 21 - i)
x <- as.numeric(x)
sum(x) * (1/2)^20
```