BST 140.651
Problem Set 3

Problem 1. Imagine that a person, say his name is Flip, has an oddly deformed coin and tries the following experiment. Flip flips his coin 10 times, 7 of which are heads. You think maybe Flip's coin is biased towards having a greater probability of yielding a head than 50%.

a. What is the maximum likelihood estimate of p, the true probability of heads associated with this coin?

b. Plot the likelihood associated with this experiment. Renormalize the likelihood so that its maximum is one. Does the likelihood suggest that the coin is fair?

c. Whats the probability of seeing 7 or more heads out of ten coin flips if the coin was fair? Does this probability suggest that the coin is fair? Note this number is called a P-value.

d. Suppose that Flip told you that he did not fix the number of trials at 10. Instead, he told you that he had flipped the coin until he obtained 3 tails and it happened to take 10 trials to do so. Therefore, the number 10 was random while the number three 3 fixed. The probability mass function for the number of trials, say y, to obtain 3 tails (called the negative binomial distribution) is

$$\left(\begin{array}{c} y-1 \\ 2 \end{array} \right) (1-p)^3 p^{y-3}$$

for $y = 3, 4, 5, 6, \ldots$. What is the maximum likelihood estimate of p now that we've changed the underlying mass function?

e. Plot the likelihood under this new mass function. Renormalize the likelihood so that its maximum is one. Does the likelihood suggest that the coin is fair?

f. Calculate the probability of requiring 10 or more flips to obtain 3 tails if the coin was fair. (Notice that this is the same as the probability of obtaining 7 or more heads to obtain 3 tails.) This is the Pvalue under the new mass function.

(Aside) This problem highlights a distinction between the likelihood and the P-value. The likelihood and the MLE are the same regardless of the experiment. That is to say, the likelihood only seems to care that you saw 10 coin flips, 7 of which were heads. Flip's intention about when he stopped flipping the coin, either at 10 fixed trials or until he obtained 3 tails, are irrelevant as far as the likelihood is concerned. The P-value, in comparison, does depend on Flip's intentions.

Problem 2. Suppose a researcher is studying the number of sexual acts with an infected person until an uninfected person contracts an sexually transmitted disease. She assumes that each encounter is an independent Bernoulli trial with probability p that the subject becomes infected. This leads to the so-called geometric distribution $P(\text{Person is infected on contact } x) = p(1-p)^{x-1}$ for $x = 1, \ldots$.

a. Suppose that one subject's number of encounters until infection is recorded, say x. Symbolically derive the ML estimate of p.

b. Suppose that the subjects value was 2. Plot and interpret the likelihood for p.

c. Suppose that is often assumed that the probability of transmission, p, is $.01$. The researcher thinks that it is perhaps strange to have a subject get infected after only 2 encounters if the probability of transmission is really on 1%. According to the geometric mass function, what is the probability of a person getting infected in 2 or fewer encounters if p truly is $.01$?

d. Suppose that she follows n subjects and records the number of sexual encounters until infection (assume all subjects became infected) $x_1, \ldots, x_n$. Symbolically derive the ML estimate of p.

e. Suppose that she records values $x_1 = 3$, $x_2 = 5$, $x_3 = 2$. Plot and interpret the likelihood for p.

Problem 3. In a study of aquaporins 6 frog eggs received a protein treatment. If the treatment of the protein is effective, the frog eggs would implode. The experiment resulted in 5 frog eggs imploding. Historically, ten percent of eggs implode without the treatment. Assuming that the results for each egg are independent and identically distributed:

a. What's the probability of getting 5 or more eggs imploding in this experiment if the true probability of implosion is 10%? Interpret this number.

b. What is the maximum likelihood estimate for the probability of implosion?

c. Plot and interpret the likelihood for the probability of implosion.

Problem 4. (Adapted from Rosner page 135) Suppose that the diastolic blood pressures of $35-44$ year old men are normally distributed with mean 80 ($mm\ Hg$) and variance 144. For the same population, the systolic blood pressures are also normally distributed and have a mean of 120 and variance 121.

a. What is the probability that a randomly selected person from this population has a DBP less than 90?

b. What DBP represents the 90^{th}, 95^{th} and 97.5^{th} percentiles of this distribution?

c. What's the probability of a random person from this population having a SBP 1, 2 or 3 standard deviations above 120? What's the corresponding probabilities for having DBPs 1, 2 or 3 standard deviations above 80?

d. Suppose that 10 people are sampled from this population. What's the probability that 50% (5) of them have a SBP larger than 140?

e. Suppose that $1,000$ people are sampled from this population. What's the probability that 50% (500) of them have a SBP larger than 140?

f. If a person's SBP and DBP are independent, what's the probability that a person has a SBP larger than 140 and a DBP greater than 90? Is independence a good assumption?

g. Suppose that an average of 200 people are drawn from this population. What's the probability that this average is smaller than 81.3?

Problem 5. Suppose that IQs in a particular population are normally distributed with a mean of 110 and a standard deviation of 10.

a. What's the probability that a randomly selected person from this population has an IQ between 95 and 115?

b. What's the 65^{th} percentile from this distribution?

c. Suppose that 5 people are sampled from this distribution. What's the probability 4 (80%) or more have IQs above 130?

d. Suppose that 500 people are sampled from this distribution. What's the probability 400 (80%) or more have IQs above 130?

e. Consider the average of 100 people drawn from this distribution. What's the probability that this mean is larger than 112.5?

Problem 6. Suppose that 400 observations are drawn at random from a distribution with mean 0 and standard deviation 40.

a. What's the approximate probability of getting a sample mean larger than 3.5?

b. Was normality of the underlying distribution required for this calculation?

Problem 7. Recall that R's function `runif` generates (by default) random uniform variables that have means $1/2$ and variance $1/12$.

a. Sample $1,000$ observations from this distribution. Take the sample mean and sample variance. What numbers should these estimate and why?

b. Retain the same $1,000$ observations from part a. Plot the sequential sample means by observation number. Hint. If x is a vector containing the simulated uniforms, then the code `y <- cumsum(x) / (1 : length(x))` will create a vector of the sequential sample means. Explain the resulting plot.

c. Plot a histogram of the $1,000$ numbers. Does it look like a uniform density?

d. Now sample $1,000$ *sample means* from this distribution, each comprised of 100 observations. What numbers should the average and variance of these $1,000$ numbers be equal to and why? Hint. The command

```
x <- matrix(runif(1000 * 100), nrow = 1000)
```

creates a matrix of size $1,000 \times 100$ filled with random uniforms. The command `y<-apply(x,1,mean)` takes the sample mean of each row.

e. Plot a histogram of the $1,000$ sample means appropriately normalized. What does it look like and why?

f. Now sample $1,000$ *sample variances* from this distribution, each comprised of 100 observations. Take the average of these $1,000$ variances. What property does this illustrate and why?

Problem 8. Note that R's function `rexp` generates random exponential variables. The exponential distribution with rate 1 (the default) has a theoretical mean of 1 and variance of 1.

a. Sample $1,000$ observations from this distribution. Take the sample mean and sample variance. What numbers should these estimate and why?

b. Retain the same $1,000$ observations from part a. Plot the sequential sample means by observation number. Explain the resulting plot.

c. Plot a histogram of the $1,000$ numbers. Does it look like a exponential density?

d. Now sample $1,000$ *sample means* from this distribution, each comprised of 100 observations. What numbers should the average and variance of these $1,000$ numbers be equal to and why?

e. Plot a histogram of the $1,000$ sample means appropriately normalized. What does it look like and why?

f. Now sample $1,000$ *sample variances* from this distribution, each comprised of 100 observations. Take the average of these $1,000$ variances. What property does this illustrate and why?

Problem 9. Consider the distribution of a fair coin flip (i.e. a random variable that takes the values 0 and 1 with probability $1/2$ each.)

a. Sample $1,000$ observations from this distribution. Take the sample mean and sample variance. What numbers should these estimate and why?

b. Retain the same $1,000$ observations from part a. Plot the sequential sample means by observation number. Explain the resulting plot.

c. Plot a histogram of the $1,000$ numbers. Does it look like it places equal probability on 0 and 1?

d. Now sample $1,000$ *sample means* from this distribution, each comprised of 100 observations. What numbers should the average and variance of these $1,000$ numbers be equal to and why?

e. Plot a histogram of the $1,000$ sample means appropriately normalized. What does it look like and why?

f. Now sample $1,000$ *sample variances* from this distribution, each comprised of 100 observations. Take the average of these $1,000$ variances. What property does this illustrate and why?

Problem 10. Consider a density for the proportion of a person's body that is covered in freckles, X, given by $f(x) = cx$ for $0 \leq x \leq 1$ and some constant c.

a. What value of c makes this function a valid density?

b. What is the mean and variance of this density?

c. You simulated $100,000$ sample means, each comprised of 100 draws from this density. You then took the variance of those $100,000$ numbers. Approximately what number did you obtain? (Explain.)

Problem 11. Suppose that DBPs drawn from a certain population are normally distributed with a mean of $90\ mmHg$ and standard deviation of $5\ mmHg$. Suppose that $1,000$ people are drawn from this population.

a. If you had to guess the number of people people in having DBPs less than 80 $mmHg$ what would you guess?

b. You draw 25 people from this population. What's the probability tha the sample average is larger than $92\ mmHg$?

c. You select 5 people from this population. What's the probability that 4 or more of them have a DBP larger than $100\ mmHg$?

Problem 12. You need to calculate the probability that a *standard normal* is larger than 2.20, but have nothing available other than a regular coin. Describe how you could estimate this probability using only your coin. (Do not actually carry out the experiment, just describe how you would do it.)

Problem 13. Let X_1, X_2 be independent, identically distributed coin flips (taking values $0 =$ failure or $1 =$ success) having success probability π. Give and interpret the likelihood ratio comparing the hypothesis that $\pi = .5$ (the coin is fair) versus $\pi = 1$ (the coin always gives successes) when both coin flips result in successes.

Problem 14. The density for the population of increases in wages for assistant professors being promoted to associates ($1 =$ no increase, $2 =$ salary has doubled) is uniform on the range from 1 to 2.

a. What's the mean and variance of this density?

b. Suppose that the sample variance of 10 observations from this density was sampled say $10,000$ times. What number would we expect the average value from these $10,000$ variances to be near? (Explain your answer briefly.)

Problem 15. Suppose that the US intelligence quotients (IQs) are normally distributed with mean 100 and standard deviation 16.

a. What IQ score represents the 5^{th} percentile? (Explain your calculation.)

b. Consider the previous question. Note that 116 is the 84^{th} percentile from this distribution. Suppose now that $1,000$ subjects are drawn at random from this population. Use the central limit theorem to write the probability that less than 82% of the sample has an IQ below 116 as a standard normal probability. Note, you do not need to solve for the final number. (Show your work.)

c. Consider the previous two questions. Suppose now that a sample of 100 subjects are drawn from a *new* population and that 60 of the sampled subjects had an IQs below 116. Give a 95% confidence interval estimate of the true probability of drawing a subject from this population with an IQ below 116. Does this proportion appear to be different than the 84% for the population from questions 1 and 2?

Problem 16. Let X be binomial with success probability p_1 and n_1 trials and Y be an independent binomial with success probability p_2 and n_2 trials. Let $\hat{p}_1 = X/n_1$ and $\hat{p}_2 = Y/n_2$ be the associated sample proportions. What would be an estimate for the standard error for $\hat{p}_1 - \hat{p}_2$? To have consistent notation with the next problem, label this value $\hat{SE}_{\hat{p}_1 - \hat{p}_2}$.

Problem 17. You are in desperate need to simulate standard normal random variables yet do not have a computer available. You do, however, have ten standard six sided dice. Knowing that the mean of a single die roll is 3.5 and the standard deviation is 1.71, describe how you could use the dice to approximately simulate standard normal random variables. (Be precise.)

Problem 18. In a sample of 40 United States men contained 25% smokers. Let p be the true prevalence of smoking amongst males in the United States. Write out and draw and interpret the likelihood for p. Is $p = .35$ or $p = .15$ better supported given the data (why, and by how much)? What value of p is best supported (just give the number, do not derive)?

Problem 19. Consider three sample variances, S_1^2, S_2^2 and S_3^2. Suppose that the sample variances are comprised of n_1, n_2 and n_3 iid draws from normal populations $N(\mu_1, \sigma^2)$, $N(\mu_2, \sigma^2)$ and $N(\mu_3, \sigma^2)$, respectively. Argue that

$$\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + (n_3 - 1)S_3^2}{n_1 + n_2 + n_3 - 3}$$

is an unbiased estimate of σ^2.

Problem 20. You need to calculate the probability that a normally distributed random variable is less than 1.25 standard deviations below the mean. However, you only have an oddly shaped coin with a known probability of heads of $.6$. Describe how you could estimate this probability using this coin. (Do not actually carry out the experiment, just describe how you would do it.)

Problem 21. The next three questions (A., B., C.) deal with the following setting. Forced expiratory volume, FEV_1, is a measure of lung function that is often expressed as a proportion of lung capacity called forced vital capacity, FVC. Suppose that the population distribution of FEV_1/FVC of asthmatics adults in the US has mean of $.55$ and standard deviation of $.10$.

A. Suppose a random sample of 100 people are drawn from this population. What is the probability that their average FEV_1/FVC is larger than $.565$?

B. Suppose the population of non-asthmatics adults in the US have a mean FEV_1/FVC of $.8$ and a standard deviation of $.05$. You sample 100 people from the asthmatic population and 100 people from the non-asthmatic population and take the difference in sample means. You repeat this process $10,000$ times to obtain $10,000$ differences in sample means. What would you guess the mean and standard deviation of these $10,000$ numbers would be?

C. Moderate or severe lung dysfunction is defined as $FEV_1/FVC \leq .40$. A colleague tells you that 60% of asthmatics in the US have moderate or severe lung dysfunction. To verify this, you take a random sample of 5 subjects, only one of which has moderate or severe lung dysfunction. What is the probability of obtaining only one or fewer if your friend's assertion is correct? What does your result suggest about their assertion?

Problem 22. Consider a sample of n iid draws from an exponential density

$$\frac{1}{\beta}\exp(-x/\beta) \text{ for } \beta > 0.$$

A. Derive the maximum likelihood estimate for β.

B. Suppose that in your experiment, you obtained five observations

```
1.590 0.109 0.155 0.281 0.453
```

plot the likelihood for β. Put in reference lines at 1/8 and 1/16.

Problem 23. Often infection rates per time at risk are modelled as Poisson random variables. Let X be the number of infections and let t be the person days at risk. Consider the Poisson mass function $(t\lambda)^x \exp(-t\lambda)/x!$ The parameter λ is called the population incident rate.

A. Derive the ML estimate for λ.

B. Suppose that 5 infections are recorded per 1000 person-days at risk. Plot the likelihood.

C. Suppose that five independent hospitals are monitored and that the infection rate (λ) is assumed to be the same at all five. Let X_i, t_i be the count of the number of infections and person days at risk for hospital i. Derive the ML estimate of λ.

Problem 24. Consider n iid draws from a gamma density where α is known

$$\frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1}\exp(-x/\beta) \quad \text{for} \quad \beta > 0, x > 0, \alpha > 0.$$

A. Derive the ML estimate of β.

B. Suppose that $n = 5$ observations were obtained: 0.015, 0.962, 0.613, 0.061, 0.617. Draw a likelihood plot for β (still assume that $\alpha = 1$).

Problem 25. Let $Y_1, \ldots, Y_N$ be iid random variabels from a Lognormal distribution with parameters μ and σ^2. Note $Y \sim$ Lognormal(μ, σ^2) if and only if $\log Y \sim N(\mu, \sigma^2)$. The lognormal density is given by

$$(2\pi\sigma^2)^{-1/2}\exp[-\{\log(y) - \mu\}^2/2\sigma^2]/y \text{ for } y > 0$$

A. Show that the ML estimate of μ is $\hat{\mu} = \frac{1}{N}\sum_{i=1}^N \log(Y_i)$. (The mean of the log of the observations. This is called the "geometric mean".)

B. Show that the ML estimate of σ^2 is then the biased variance estimate based on the log observation

$$\frac{1}{N}\sum_{i=1}^{N}(\log(y_i) - \hat{\mu})^2$$