BST 140.651
Problem Set 1

1 Basic probability

1. Given only the simple axiomatic foundations, prove the following.
   a. $P(\emptyset) = 0$.
   b. $P(E) = 1 - P(E^c)$.
   c. If $A \subset B$ then $P(A) \leq P(B)$.
   d. For any A and B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
   e. $P(A \cup B) = 1 - P(A^c \cap B^c)$.
   f. $P(A \cap B^c) = P(A) - P(A \cap B)$.
   g. $P(\cup_{i=1}^n E_i) \leq \sum_{i=1}^n P(E_i)$ (Bonferroni's inequality).
   h. $P(\cup_{i=1}^n E_i) \geq \max_i P(E_i)$.
   i. $P(B \cap A^C) = P(B) - P(B \cap A)$
2. (From Casella and Berger) $P(A) = 1/3$ and $P(B^C) = 1/4$, can A and B be disjoint?
3. Suppose that an influenza epidemic strikes a area. In 17% of two parent families at least one of the parents has contracted the disease. In 12% of the families the father has contracted influenza while in 6% of the families both the mother and father have contracted influenza.
   a. What's the probability that the mother has contracted influenza?
   b. What's the probability that neither the mother nor the father has contracted influenza?
   c. What's the probability that the mother has contracted influenza but the father has not?
   d. What's the probability that the father has contracted influenza but the mother has not?
   e. What event does the probability one minus the probability that both have contracted influenza represent?

2 Random variables, univariate densities

1. Suppose $h(x)$ is such that $h(x) > 0$ for $x = 1, 2, \ldots, I$. Argue that $p(x) = h(x)/\sum_{i=1}^I h(i)$ is a valid pmf.
2. Suppose a function h is such that $h > 0$ and $c = \int_{-\infty}^{\infty} h(x)dx < \infty$. Show that $f(x) = h(x)/c$ is a valid density.
3. Suppose that, for a randomly drawn subject from a particular population, the proportion of a their skin that is covered in freckles follows a density that is constant on $[0, 1]$. (This is called the **uniform density**.) That is, $f(x) = k$ for $0 \leq x \leq 1$.

a. Draw this density. What must k be?

b. Suppose a random variable, X, follows a uniform distribution. What is the probability that X is between .1 and .7? Interpret this probability in the context of the problem.

c. Verify the previous calculation in R. What's the probability that $a < X < b$ for generic values $0 < a < b < 1$?

d. What is the distribution function associated with this density?

e. What is the median of this density? Interpret the median in the context of the problem.

f. What is the 95^{th} percentile? Interpret this percentile in the context of the problem.

g. Do you believe that the proportion of freckles on subjects in a given population could feasibly follow this distribution? (Why or why not.)

4. Consider a randomly drawn leaf from a particular tree population. The proportion of the leaf that is covered in blight follows density that is a right triangle with one vertex at $(0, 0)$, one at $(0, k)$ and one at $(1, 0)$.

a. What is the value of k that makes this a valid density?

b. What is the survival function associated with this density? Plug in $.5$ into the survival function. What is this quantity in the context of the problem?

c. What is the probability of drawing a leaf more than 80% covered in blight?

d. What is the 95^{th} percentile? Interpret this quantile in the context of the problem.

e. Do you believe that the proportion of blight on leaves could feasibly follow this distribution? (Why or why not.)

5. Let $0 \leq \pi \leq 1$ and f_1 and f_2 be two continuous densities with associated distribution functions F_1 and F_2 and survival functions S_1 and S_2. Let $g(x) = \pi f_1(x) + (1 - \pi) f_2(x)$.

a. Show that g is a valid density.

b. Write the distribution function associated with g in the terms of F_1 and F_2.

c. Write the survival function associated with g in the terms of S_1 and S_2.

d. Repeat the previous questions where $\{f_i\}_{i=1}^I$ is a collection of densities and $\{\pi_i\}_{i=1}^I$ is a point on the I dimensional simplex ($\sum_{i=1}^I \pi_i = 1$ and $0 \leq \pi_i \leq 1$) and $g = \sum_{i=1}^I \pi_i f_i$.

6. Radiologists have created cancer risk summary that, for a given population of subjects, follows (a specific instance of) the **logistic** density

$$\frac{e^{-x}}{(1 + e^{-x})^2} \qquad \text{for } -\infty < x < \infty.$$

a. Show that this is a valid density.

b. Calculate the distribution function associated with this density.

c. What value do you get when you plug 0 into the distribution function? Interpret this result in the context of the problem.

d. Define the *odds* an event with probability p as $p/(1-p)$. Prove that the p^{th} quantile from this distribution is $\log\{p/(1-p)\}$; which is the natural log of the odds of an event with probability p.

7. Quality control experts estimate that the time (in years) until a specific electronic part from an assembly line fails follows (a specific instance of) the **Pareto** density

$$\frac{1}{x^2} \qquad \text{for } 1 < x < \infty.$$

a. Show that this is a valid density.

b. What is the survival function associated with this density? Interpret a value (say 10 years) evaluated in the survival function in the context of the problem.

c. Show that the p^{th} quantile from this density is $1/(1-p)$. For $p = .8$ interpret this value in the context of the problem.

8. Suppose that a density is of the form cx^k for some constant $k > 1$ and $0 < x < 1$.

a. Find c.

b. Derive the distribution function for f.

c. Derive a formula for the p^{th} quantile from f.

d. Let $0 \leq a < b \leq 1$. Derive a formula for $P(a < X < b)$.

9. Suppose that the time in days until hospital discharge for a certain patient population follows a density $f(x) = c\exp(-x/10)$ for $x > 0$.

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

b. Find the distribution function for this density.

c. Find the survival function.

d. Calculate the probability that a person takes longer than 11 days to be discharged.

e. What is the median number of days until discharge?

10. The (lower) incomplete gamma function is defined as $\Gamma(k, c) = \int_0^c x^{k-1}\exp(-x)dx$. By convention $\Gamma(k, \infty)$, the complete gamma function, is written $\Gamma(k)$. Consider a density

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

where α is a known number.

a. Argue that this is a valid density.

b. Write out the survival function associated with this density using gamma functions

c. Let β be a known number; argue that

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

is a valid density. This is known as the **gamma density**.

d. Plot the Gamma density for different values of α and β.

11. The **Weibull density** is useful in survival analysis. Its form is given by

$$\frac{\gamma}{\beta} x^{\gamma-1} \exp\left(-x^{\gamma}/\beta\right),$$

for $x > 0$ and γ and β are fixed known numbers.

a. Demonstrate that the Weibull density is a valid density.

b. Calculate the survival function associated with the Weibull density.

c. Calculate the median of the Weibull density.

d. Plot the Weibull density for different values of γ and β.

12. The Beta function is given by $B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}$ for $\alpha > 0$ and $\beta > 0$. It turns out that

$$B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta).$$

The **Beta density** is given by $\frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$ for fixed $\alpha > 0$ and $\beta > 0$. This density is useful for

a. Argue that the Beta density is a valid density.

b. Argue that the uniform density is a special case of the beta density.

c. Plot the beta density for different values of α and β.

13. A famous formula is $e^{\lambda} = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}$ for any value of λ. Assume that the count of the number of people infected with a particular disease per year follows a mass function given by

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad \text{for} \quad x = 0, 1, 2, 3, \ldots$$

where λ is a fixed known number. (This is know as the **Poisson mass function**.)

a. Argue that $\sum_{x=0}^{\infty} P(X = x) = 1$.

14. Consider counting the number of coin flips from an unfair coin with success probability p until a head is obtained, say X. The mass function for this process is given by $P(X = x) = p(1-p)^{x-1}$ for $x = 1, 2, 3, \ldots$. This is called the **geometric mass function**.

a. Argue mathematically that this is a valid probability mass function. Hint, the geometric series is given by $\frac{1}{1-r} = \sum_{k=0}^{\infty} r^k$ for $|r| < 1$.

b. Calculate the survival distribution $P(X > x)$ for the geometric distribution for integer values of x.

15. Let U be a uniform $(0, 1)$ density. Calculate the distribution function and density of U^P where p is a power. What is the name of this density?

16. Let X be an exponential 1 random variable. Calculate the distribution and density for $\log(X)$.

3 Expected values and variances

1. Using the rules of expectations prove that $\mathrm{Var}(X) = E[X^2] - E[X]^2$ where $\mathrm{Var}(X) = E[(X-\mu)^2]$.

2. Let $g(x) = \pi f_1(x) + (1-\pi) f_2(x)$ where f_1 and f_2 are densities with associated means and variances μ_1, σ_1^2 and μ_2, σ_2^2, respectively. You showed already that g is a valid density. What is it's associated mean and variance?

3. Suppose that a density is of the form $(k+1)x^k$ for some constant $k > 1$ and $0 < x < 1$.

   a. What is the mean of this distribution?

   b. What is the variance?

4. Suppose that the time in days until hospital discharge for a certain patient population follows a density $f(x) = \frac{1}{10}\exp(-x/10)$ for $x > 0$.

   a. Find the mean and variance of this distribution?

   b. The general form of this density (the exponential density) is $f(x) = \frac{1}{\beta}\exp(-x/\beta)$ for $x > 0$ for a fixed value of β. Calculate the mean and variance of this density.

5. The Gamma density is given by

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

   for fixed values of α and β.

   a. Derive the mean and variance of the gamma density. You can assume the fact (proved in HW 1) that the density integrates to 1 for any $\alpha > 0$ and $\beta > 0$.

   b. The Chi-squared density is the special case of the Gamma density where $\beta = 2$ and $\alpha = p/2$ for some fixed value of p (called the "degrees of freedon"). Calculate the mean and variance of the Chi-squared density.

6. The Beta density is given by

$$\frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} \quad \text{for} \quad 0 < x < 1$$

   and $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$.

   a. Derive the mean of the beta density. Note the following is useful for simplifying results: $\Gamma(c+1) = c\Gamma(c)$ for $c > 0$.

   b. Derive the variance of the beta density.

7. The Poisson mass function is given by

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad \text{for} \quad x = 0, 1, 2, 3, \ldots$$

a. Derive the mean of this mass function.

b. Derive the variance of this mass function. Hint, consider $E[X(X-1)]$.

8. Suppose that, for a randomly drawn subject from a particular population, the proportion of a their skin that is covered in freckles follows a uniform density (constant between 0 and 1).

   a. What is the expected percentage of a (randomly selected) person's body that is covered in freckles? (Show your work.)

   b. What is the variance? (Show your work.)

9. You have an MP3 player with a total of 4 songs stored on it. Suppose that songs are played randomly *with replacement*. Let X be the number of songs played until you hear a repeated song.

   a. What values can X take, and with what probabilities?

   b. What is the expected value for X?

   c. What is the variance for X?

10. Argue that any density that is symmetric about a point μ, whose first moment exists, has mean μ.

11. Give a proper density that has no mean. Give a proper density that has a mean but no variance.

12. (Adapted from Casella and Berger) Let x_0 be a point. Argue that $g(x) = f(x)/F(X_0)$ for $x < x_0$ is a valid density. (This is the conditional density given that $x < x_0$.) What is the CDF associated with g?

13. Let X be a random variable with mean μ and variance σ^2. Show that $(X-\mu)/\sigma$ has mean 0 and variance 1.

14. You are playing a game with a friend where you flip a coin and if it comes up heads you give him a dollar and if it comes up tails she gives you a dollar. You play the game ten times.

    a. What is the expected total earnings for you? (Show your work; state your assumptions.)

    b. What is the variance of your total earnings? (Show your work; state your assumptions.)

    c. Suppose that the die is biased and you have a $.4$ chance of winning for each flip. repeat the calculations in parts a and b

15. Note that the code

```
temp <- matrix(sample(1 : 6, 1000 * 10, replace = TRUE), 1000)
xBar <- apply(temp, 1, mean)
```

    In R produces $1,000$ averages of 10 die rolls. That is, it's like taking ten dice, rolling them, averaging the results and repeating this $1,000$ times.

    a. Do this in R. Plot histograms of the averages.

b. Take the mean of `xBar`. What should this value be close to? (Explain your reasoning.)

c. Take the standard deviation of `xBar`. What should this value be close to? (Explain your reasoning.)

16. Note that the code

```
xBar <- apply(matrix(runif(1000 * 10), 1000), 1, mean)
```

produces $1,000$ averages of 10 uniforms.

a. Do this in R. Plot histograms of the averages.

b. Take the mean of `xBar`. What should this value be close to? (Explain your reasoning.)

c. Take the standard deviation of `xBar`. What should this value be close to? (Explain your reasoning.)

17. You flip a coin with a probability p of heads. How many times do you have to flip it on average until you get a head? What's the variance of the number of flips to get a head?

18. Let X be a random variable such that $0 \leq X \leq 1$. What's the largest that the variance of X can be? What distribution for X obtains this variance?

4 Random vectors, independence, conditional probabilities

1. When at the free-throw line for two shots, a basketball player makes at least one free throw 90% of the time. 80% of the time, the player makes the first shot, while 70% of the time she makes both shots.

   a. Does it appear that the player's second shot success is independent of the first?

   b. What is the conditional probability that the player makes the second shot given that she made the first? What would it be if she missed the first?

2. Assume that an act of intercourse between an HIV infected person and a non-infected person results in a $1/500$ probability of spreading the infection. How many acts of intercourse would an uninfected person have to have with an infected persons to have a 50% probability of obtaining an infection? State the assumptions of your calculations.

3. You meet a person at the bus stop and strike up a conversation. In the conversation, it is revealed that the person is a parent of two children and that one of the two children is a girl. However, you do not know the gender of the other child, nor whether the daughter she mentioned is the older or younger sibling.

   a. What is the probability that the other sibling is a girl? What assumptions are you making to perform this calculation?

   b. Later in the conversation, it becomes apparent that she was discussing the older sibling. Does this change your probability that the other sibling is a girl?

4. A particularly sadistic warden has three prisoners, A, B and C. He tells prisoner C that the sentences are such that two prisoners will be executed and let one free, though he will not say who has what sentence. Prisoner C convinces the warden to tell him the identity of one of the prisoners to be executed. The warden has the following strategy, which prisoner C is aware of. If C is sentenced to be let free, the warden flips a coin to pick between A and B and tells prisoner C that person's sentence. If C is sentenced to be executed he gives the identity of whichever of A or B is also sentenced to be executed.

   a. Does this new information about one of the other prisoners give prisoner C any more information about his sentence?

   b. The warder offers to let prisoner C switch sentences with the other prisoner whose sentence he has not identified. Should he switch?

5. Derive the mean and variance for the difference in means between a collection of data $\{X_{1i}\}_{i=1}^{I}$ and $\{X_{2i}\}_{i=1^I}$. Where (X_{i1}, X_{i2}) are iid pairs of data (yet possibly dependent). Define any notation that you use. How does the variance change when you assume that the observations within a pair are independent.

6. Quality control experts estimate that the time (in years) until a specific electronic part from an assembly line fails follows (a specific instance of) the **Pareto** density

$$\frac{3}{x^4} \qquad \text{for } 1 < x < \infty.$$

   a. What is the average failure time for components from this density? (Show your work.)

   b. What is the variance? (Show your work.)

   c. The general form of the Pareto density is given by $\frac{\beta\alpha^\beta}{x^{\beta+1}}$ for $0 < \alpha < x$ and $\beta > 0$ (for fixed α and β). Calculate the mean and variance of the general Pareto density.

5 Conditional probabilities, Bayes rule

1. A web site (www.medicine.ox.ac.uk/bandolier/band64/b64-7.html) for home pregnancy tests cites the following:

   > When the subjects using the test were women who collected and tested their own samples, the overall sensitivity was 75%. Specificity was also low, in the range 52% to 75%.

   a. Interpret a positive and negative test result using diagnostic likelihood ratios using both extremes of the specificity.

   b. A woman taking a home pregnancy test has a positive test. Draw a graph of the positive predictive value by the prior probability (prevalence) that the woman is pregnant. Assume the specificity is 63.5%

   c. Repeat the previous question for a negative test and the negative predictive value.

6 Standard distributions

1. (Adapted from Rosner page 135) Suppose that the diasolic 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?
2. 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?
3. 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?

7 Limit theorems and sampling distributions

1. 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?

2. 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?

3. 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?

4. 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.)

5. 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 $100\ mmHg$?

6. 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.)

7. 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.

8. 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.)

9. 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?

10. 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 $\hat{SE}_{\hat{p}_1 - \hat{p}_2}$.

11. 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.)

12.

13. 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.

14. 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.)

15. 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?

8 Likelihood

2. 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.

3. 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.

4. 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.

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

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

   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.

6. 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.

7. 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 $\alpha = 1$).

8. 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 log-normal density is given by

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

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

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

9 Programming

1. Simulate $1,000$ random uniform[0,1] variables variables and do the following:

   a. Calculate their mean, variance and standard deviation.

   b. Plot a histogram.

2. Take your $1,000$ simulated uniform random variables, say the vector U, and create the vector $Y = 2 + 4 * U + E$ where E is $1,000$ simulated standard normal random variables.

   a. Plot Y versus U

   b. Use $lm(Y\ U)$ to find the line that best fits the data.

   c. Overlay the line onto your plot.

3. Write a function in R that takes in a matrix and normalizes it so that every row and column has a mean of 0 and a variance of 1. Note, to normalize a vector, subtract its mean from every element then divide every resulting difference by the standard deviation. Test your function by simulating matrices of random uniforms.

4. Use R to create a graph of an Archimedian spiral.

5. Graph Tupper's self referential formula in R.

6. Create a data frame in R using the Y and U from above. Add a third variable that is $2Y$. Change the names of the variables.

7. Write a function in R that takes in a character vector of first and last names in the format **firstname.lastname** and returns a character vector of the form **lastname.firstname**.

8. Create three random 2×2 matrices consisting of uniform variables. Create a (three element) list of these matrices. Use an lapply statement to create a new list consisting of the matrices squared.