2 95% of statistics (AJ Chapters 5-7)

1. (*) A p-value is a statistic $P : X \to [0,1]$ such that under H_0,

$$\mathbb{P}(P \le t) \le t \text{ for } t \in [0,1].$$

The p-value should be designed such that small values indicate a divergence between the data and the null hypothesis; the critical region for a test of size α is then $\{x : P(x) \in [0, \alpha]\}$.

   (a) Show that an independent Uniform(0,1) random variable is always a p-value. What is the power of the test?

   (b) Let X be a continuous random variable with strictly positive p.d.f. $f : \mathbb{R} \to (0, \infty)$. Show that the c.d.f.

$$F(x) = \mathbb{P}(X \le x), \qquad F : \mathbb{R} \to [0,1]$$

   is an invertible function and that the distribution of $F(X)$ is Uniform$(0,1)$.

   (c) Let $X_1, \ldots, X_n \sim N(\theta, 1)$.

      i. Find a p-value for testing $H_0 : \theta = 0$ against $H_1 : \theta < 0$.

      ii. What is the critical region in terms of $X_1, \ldots, X_n$?

      iii. You are working at the 5% significance level. You want the probability of rejecting H_0 to be at least 90% if $\theta < -\epsilon$, where ϵ is a small, positive number. How large does n have to be as a function of ϵ?

   (d) Let $X_1, \ldots, X_n \sim N(\theta, 1)$. Find a p-value for testing $H_0 : \theta = 0$ against $H_1 : \theta \ne 0$.

   (e) Let $X_1, \ldots, X_n \sim N(0, \sigma^2)$. Find a p-value for testing $H_0 : \sigma^2 = 1$ against $H_1 : \sigma^2 < 1$.

2. (*) Suppose $X_{1,} \ldots X_n \sim Poisson(\lambda)$ iidrv so $\mathbb{P}(X_i = k) = e^{-\lambda} \frac{\lambda^k}{k!}$.

   (a) What, roughly speaking, is the distribution of $\bar{X} = \sum_{i=1}^n X_i / n$?

   (b) Given λ, find an interval such that $\bar{X}$ belongs to that interval with probability 95%.

   (c) Re-arrange your answer to part (b) to produce an approximate 95% confidence interval for λ.

   (d) Let $\lambda = 2$ and $n = 5$. Using R, do the following a large number of times: Sample $X_{1,} \ldots X_n \sim Poisson(\lambda)$ and check if your confidence interval includes λ. What fraction of the time does your confidence interval include λ.

   (e) How about when $n = 100$?

3. ~~Problems~~ Opportunities with Hypothesis Testing

(a) The county of Midsomer has a population of 10,001. Or it did; there has been a murder. If we pick someone from the population uniformly at random, there is a one in 10,000 chance they are the murderer.
A genetic test has been devised to test someone's DNA with a DNA sample from the murder weapon. Relatives have similar DNA, so the machine will show a red light, if the test subject is the murderer, or any of the murderer's 99 closest blood relatives. For the other 9,900 inhabitants of Midsomer, the machine will show a green light.
Consider the hypothesis test being applied to a randomly chosen citizen of Midsomer. Consider H_0 : "not a murderer" against H_1 : "a murderer".

   i. The machine show a red light. Should you (i) reject H_0, or (ii) fail to reject H_0 / accept H_1, at a 5% significance level? Interpret your test result.
   ii. What is the probability the test subject is the killer?
   iii. How, in the general context of hypothesis testing, can you reduce the number of false positives a test produces?

(b) Suppose $X_1, \ldots, X_{100} \sim N(\theta, 1)$. Consider a test $H_0 : \theta = 0$ against $H_1 : \theta = 100$. Calculate a critical region for a hypothesis test at a 5% significance level. What is the power of the test? You observe $\bar{X} = 10$. Do you reject H_0 or fail to do so? Compare the values of the likelihood function at $\theta = 0$ and $\theta = 100$. What do you think of the conclusion of the test?

(c) Coins have two sides: heads or tails. They never fall on their side. (Well it did once in Midsomer, but we don't talk about that any more).

   i. A coin might be biased. Consider $H_0 : \mathbb{P}(heads) = 0.6$ against $H_1 : \mathbb{P}(heads) = 0.4$. You toss the coin four times. Work at the 5% significance level.
      A. Should you reject H_0 if you get four heads?
      B. Should you reject H_0 if you get four tails?
   ii. A bag contains one biased coin ($\mathbb{P}(heads) = 0.6$) and 999 biased-the-other-way coins ($\mathbb{P}(heads) = 0.4$). A coin is picked from the bag uniformly at random. You toss it 4 times and get four heads. Using part (i), should you reject H_0. Calculate the probability (conditional on the four tosses) that H_0 describes the coin.

4. The Hard EM algorithm is as its name suggest, much easier than the general EM-algorithm.

   - Split the data $x = (x_o, x_m)$. You have only observed x_o. Start at some $\theta^{(0)}$.
   - Iterate $\theta^{(t)} \to \theta^{(t+1)}$ by:
     - sampling $x_m = x_m(t)$ conditional on $\theta^{(t)}$, and then
     - setting $\theta^{(t+1)}$to be the MLE for $x = (x_0, x_m)$.

   A die has side-probabilities $p_1 + \cdots + p_6 = 1$. Roll the die $2n$ times to get $Y_1, \ldots, Y_{2n} \in \{1, \ldots, 6\}$ independent. You observe

$$x_o = (Y_1 + Y_2, Y_3 + Y_4, \ldots, Y_{2n-1} + Y_{2n})$$

   but you do not observe the whole dataset, i.e.

$$x_m = (Y_1, \ldots, Y_{2n}).$$

   (a) Download the file hard_em.R from the website. It contains x_o for $n = 1000$,

Total	2	3	4	5	6	7	8	9	10	11	12
#	1	11	24	38	82	133	138	166	169	147	91

   and a function for estimating the M.L.E. for the (p_i) using the hard EM algorithm. Run the code to calculate the MLE approximately.

   (b) Try using some other method (perhaps moment generating functions, and the square root operation) to approximate the distribution of the die. Compare with your answer to part (a).

   (c) Reduce n to say 100 (`xo<-xo[1:100]`).

Total	2	3	4	5	6	7	8	9	10	11	12
#	0	1	2	4	11	9	12	17	22	15	7

   Faced with a small amount of data, can you modify the algorithm in some simple way to improve performance.

5. Likelihood ratio test: Let $t \in (0, \infty)$ be some known quantity. Let $X_1, \ldots, X_n \sim$ Exponential(r) with r unknown. Consider the simple test $H_0 : r = 1$ against $H_1 : r = t$. The simple likelihood ratio test has critical region

$$C_\alpha := \left\{ x : \frac{L(r = 1 \mid x)}{L(r = t \mid x)} \leq A_\alpha \right\} \qquad L(\mu \mid x) = \prod_{i=1}^{n} r \exp(-r x_i).$$

   (a) Calculate the critical region as a function of t, and of the test size α, and of sample size n.

   (b) Calculate the power $\beta$of the simple likelihood ratio test.

   (c) Assuming βneeds to be at least 0.9, what is the minimum sample size n as a function of t.

6. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ independent.

   (a) Is there a uniformly most powerful test for $H_0 : p = 1/2$ against $H_1 : p > 1/2$?

   (b) Is there a uniformly most powerful test for $H_0 : p = 1/2$ against $H_1 : p \neq 1/2$?

7. Let $L(\theta \mid X)$ denote the likelihood. Assume regularity conditions are met...
   Wilks' theorm states that when testing a composite null hypothesis $H_0 : \theta \in \Omega_0$ against a composite alternative hypothesis $H_1 : \theta \in \Omega_1 = \Omega \setminus \Omega_0$, the distribution of

$$-2 \log \frac{L(\hat{\theta}_0 \mid X)}{L(\hat{\theta} \mid X)} \sim \chi^2_{\dim \Omega - \dim \Omega_0}$$

$$\text{where} \qquad \hat{\theta} = \arg\max_{\theta \in \Omega} L(\theta \mid X), \qquad \hat{\theta}_0 = \arg\max_{\theta \in \Omega_0} L(\theta \mid X).$$

   The probability distribution when rolling a die is described by 6 numbers, adding to one. You have two dice, and you roll each die n times. Denote the outcomes $X_1, \ldots, X_n$ for die 1, and $Y_1, \ldots, Y_n$ for die 2.

i	1	2	3	4	5	6
X	130	116	109	197	223	225
Y	100	108	128	209	210	245

   Test H_0 : the dice have the same distribution against H_1 : the distributions differ.

8. Read Chapters 6 and 7 of AJ's notes.