Big Data 2018: Exercise Set 2

Where indicated, exercises have been taken from the following book:

Hodges and Lehmann: J.L. Hodges, Jr. and E.L. Lehmann, Basic Concepts of Probability and Statistics (Second Edition), Holden-Day, 1970.

Exercises

1. (Hodges and Lehmann) In rolls with three dice, the sum $T = 9$ can be produced in six ways, namely as

   (1) 1+2+6

   (2) 1+3+5

   (3) 1+4+4

   (4) 2+2+5

   (5) 2+3+4

   (6) 3+3+3

   Answer the following questions:

   (a) Determine the number of ways in which the sum $T = 10$ can be produced.

   (b) Would you conclude that $\mathbb{P}(T = 9) = \mathbb{P}(T = 10)$?

2. (Hodges and Lehmann) Assume that the $6^3 = 216$ different outcomes of rolls with three dice: (1,1,1),(1,1,2),...,(6,6,6) are all equally likely. Under this assumption, compute $\mathbb{P}(T = 9)$ and $\mathbb{P}(T = 10)$, and discuss the difference between this result and that suggested by the preceding problem.

3. (Hodges and Lehmann) In a study of the performance in course A and B, a department finds that the following model suitably describes the distribution of grades (measured on a 5-point scale) for students completing both courses:

B A	0	1	2	3	4
0	.00	.01	.00	.00	.00
1	.03	.05	.04	.00	.00
2	.01	.04	.26	.05	.00
3	.00	.03	.11	.15	.03
4	.00	.00	.03	.07	.09

Make a table showing the probabilities of

(a) The number U of grade points in course A taking on its various possible values 0,1,2,3,4.

(b) The number V of grade points in course B taking on its various possible values 0,1,2,3,4.

(c) The total number W of grade points in the two courses combined taking on its various possible values.

(d) The difference D between the grade points in course B and those in course A taking on its various possible values.

(e) The number U of grade points in course A taking on its various possible values given that the number V of grade points obtained for course B is 2 (round to 2 decimals).

4. In a collection of song lyrics, 80% of the songs belongs to the Metal genre, and 20% to the Rap genre. In 60% of the Rap lyrics, the word “bitch” occurs at least once. For Metal lyrics this is only 5%. We draw a song at random from the collection and observe that it does contain the word “bitch”. What is the probability that we have drawn a Rap song?

5. Suppose that we have three coloured boxes R (red), B (blue), and G (green). Box R contains 3 apples, 4 oranges, and 3 limes, box B contains 1 apple, 1 orange, and 0 limes, and box G contains 3 apples, 3 oranges, and 4 limes. If a box is chosen at random with probabilities $\mathbb{P}(R) = 0.2, \mathbb{P}(B) = 0.2, \mathbb{P}(G) = 0.6$, and a piece of fruit is removed from the box (with equal probability of selecting any of the items in the box), then what is the probability of selecting an apple? If we observe that the selected fruit is in fact an orange, what is the probability that it came from the green box?

6. The joint probability of gender and admission to the master program of a department of the university is given in the table below:

Gender	Admission	
	Yes	No
Male	0.34	0.16
Female	0.16	0.34

(a) Compute the admission probability for males and females.

(b) Is admission independent of gender?

It turns out that the table above originated from two master programs, A and B. The probability that a student applies to program A is 0.5, and likewise for program B. Each student applies to exactly one of the two programs. The two conditional distributions $\mathbb{P}(\text{gender}, \text{admission} \,|\, \text{program})$ are:

Program	Gender	Admission	
		Yes	No
A	Male	0.04	0.16
	Female	0.16	0.64
B	Male	0.64	0.16
	Female	0.16	0.04

(c) Is admission independent of gender for program A?

(d) Is admission independent of gender for program B?

(e) Demystify the apparent contradiction.

7. Consider the following population of size $N = 8$.

Unit	1	2	3	4	5	6	7	8
$\mathcal{X}$	4	8	8	8	9	9	9	9

Here $\mathcal{X}$ denotes a population variable with mean

$$\mu = \frac{1}{N}\sum_{i=1}^{N} \mathcal{X}_i,$$

and variance

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (\mathcal{X}_i - \mu)^2.$$

(a) Compute the population mean μ and population variance σ^2 of $\mathcal{X}$.

(b) Let X denote the random variable obtained by making a random draw from this population and observing the value of $\mathcal{X}$. Determine the probability distribution $p(x) = \mathbb{P}(X = x)$, the expected value $\mathbb{E}(X)$ and variance $\mathrm{Var}(X)$.

(c) Consider simple random samples (with replacement) of size $n = 2$ from this population. Let X_i denote the observed value of $\mathcal{X}$ on the i-th draw. Determine the distribution $p(x_1, x_2) = \mathbb{P}(X_1 = x_1, X_2 = x_2)$ of the sample.

(d) Use your answer to (c) to determine the probability distribution of the sample mean

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2}{2}$$

(e) Use the probability distribution obtained under (d) to compute $\mathbb{E}(\bar{X})$. Is it equal to μ?

(f) Determine the probability distribution of the sample variance

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n} = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2}{2}$$

(g) Use the probability distribution obtained under (f) to compute $\mathbb{E}(\hat{\sigma}^2)$. Is it equal to σ^2?

(h) Determine the probability distribution of the sample variance

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2}{1}$$

(i) Use the probability distribution obtained under (h) to compute $\mathbb{E}(s^2)$. Is it equal to σ^2?

8. Show that the sample fraction of successes

$$F = \frac{1}{n}\sum_{i=1}^{n} X_i$$

in a binomial sample with success probability π and n trials, has the following properties:

(a) Its expected value is: $\mathbb{E}(F) = \pi$

(b) Its variance is:

$$\text{Var}(F) = \frac{\pi(1-\pi)}{n}$$

9. Use the properties of expectation and variance to show that if X is a random variable with expected value $\mathbb{E}(X) = \mu$, and variance $\text{Var}(X) = \sigma^2$, then

$$Z = \frac{X - \mu}{\sigma}$$

has expected value $\mathbb{E}(Z) = 0$ and variance $\text{Var}(Z) = 1$.

10. Let X be a random variable with expected value $\mathbb{E}(X) = \mu$, and variance $\text{Var}(X) = \sigma^2$. Let

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

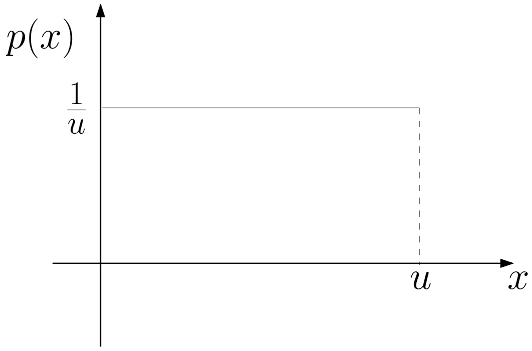
denote the mean value of X in a random sample of size n. Show that $\bar{X}$ is an unbiased estimator of μ (that is, $\mathbb{E}(\bar{X}) = \mu$), and has variance

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

11. Let X be a random variable with probability density function

$$p(x) = \begin{cases} \frac{1}{u} & 0 \le x \le u \\ 0 & \text{otherwise} \end{cases}$$

In words: X has a uniform distribution on the interval $[0, u]$. In a picture:



Answer the following questions:

(a) Determine $\mathbb{E}(X)$.

(b) Determine $\text{Var}(X)$.

The upper bound u is unknown, and we would like to estimate it from a random sample of observations $X_1, \ldots, X_n$. The maximum likelihood estimator of u is the estimator that maximizes the probability of the observed data.
For an estimator $\hat{u}$ of u, we define the following quality criteria:

(1) Bias: $\mathbb{E}[\hat{u}] - u$. An estimator is called unbiased if $\mathbb{E}[\hat{u}] = u$.

(2) Variance: $\mathbb{E}[(\hat{u} - \mathbb{E}(\hat{u}))^2]$. Spread of the estimator around its mean.

(3) Mean Square Error: $\mathbb{E}[(\hat{u} - u)^2]$. Overall quality measure:
how far on average is $\hat{u}$ removed from u?

Here expectation is taken with respect to repeated samples (of some fixed size n) from the population. Note that we have the decomposition "mean square error is squared bias plus variance".

Answer the following questions:

(c) Give a formula for the maximum likelihood estimator of u as a function of the sample $X_1, \ldots, X_n$. You are not required to give a formal proof that the proposed estimator is indeed the maximum likelihood estimator; a good argument is sufficient.

(d) Is the maximum likelihood estimator unbiased? Explain your answer. Again a formal proof is not required.

(e) Give a formal proof, showing that $2\bar{X}$, where $\bar{X}$ is the sample mean, is an unbiased estimator of u.

(f) Determine $\text{Var}(2\bar{X})$.

(g) Write a function in `R` to compare the performance of the maximum likelihood estimator and the estimator given under (d). Your function has the parameters `u`, `n`, and `m`. `u` is the maximum of the uniform distribution, the number we want to estimate. `n` is the sample size, and `m` is the number of samples (each of size `n`) we draw. By drawing `m` samples, we mimic repeated sampling from $U(0, u)$ on the computer. Compute the two estimators for each of the `m` samples of size `n`. Compare their bias, variance and mean squared error. Make a histogram of the sampling distribution of both estimators. Which estimator is preferred? Hint: there is a built-in function `runif` in `R` for drawing samples from a uniform distribution.