

Homework #9

Reading: 9.1-3

Submit in the inbox outside Sequoia 229 before 5:00 PM.

This final problem set doubles as the practice final.

1. *Summary Statistics*

The following data was collected on a pair of random variables. X is the count of instances of the word “statisticians” in 10 articles sampled without replacement from the scientific literature, while Y is the count of the sentence fragment “are extremely attractive”.

	X	Y
[1,]	0	1
[2,]	1	0
[3,]	1	0
[4,]	2	0
[5,]	0	1
[6,]	2	0
[7,]	3	0
[8,]	1	0
[9,]	1	0
[10,]	0	0

- Describe the difference between the population and the sample.
- Calculate the sample mean, median, max, min, standard deviation, and variance of X and Y .
- Calculate the sample covariance and the correlation between X and Y .
- Is it useful to calculate a correlation between discrete variables like counts?
- Based on this sample, are these two random variables linearly independent? Are they statistically independent?
- In general, why might your conclusion about the linear independence of the two RVs in a sample be misleading about the population? Intuitively, how could you reduce the probability of making a bad conclusion about the linear independence of two RVs from finite data? (Hint: is the sample correlation a random variable?)

2. *Classification and Regression*

For the X and Y variables of the previous problem, calculate:

- An estimation equation for X given Y using linear regression.

- (b) An estimation equation for Y given X using linear regression.
- (c) Now define $W = 0$ if $X \leq 1$ and $W = 1$ if $X \geq 2$. This corresponds to a dichotomization of the random variable X . Derive the empirical binary classifier function $P(W = 1|Y = l)$.

3. *Statisticians Do it Discretely*

A discrete random variable $X \in \{1, 2, 3\}$ has the following probability mass function:

$$\begin{aligned} P(X = 1) &= p_1 \\ P(X = 2) &= p_2 \\ P(X = 3) &= p_3 \end{aligned}$$

- (a) Express p_3 in terms of p_1 and p_2 .
- (b) What is the cdf of X ?
- (c) Obtain expressions for the mean and standard deviation.
- (d) What is $P(X = 2|X \neq 3)$?
- (e) What is $E[X|X \neq 3]$? What is $\text{Var}[X|X \neq 3]$?
- (f) The following sample data was collected from this distribution:

[1] 2 2 3 1 2 1 1 3 3 2 2 2 3 2 3 2 3 1 2 3 1 2 3 2 2

- Estimate p_1 and p_2 from their frequency of occurrence in the sample. Do you need a separate equation for p_3 ?
- (g) Use the method of moments to estimate p_1 and p_2 from this data and compare the results to the frequency estimate.

4. *Just a Stats Survivor*

In many medical and engineering studies, it is useful to model the survival time of a person¹ or a component as a random variable T . The following definitions are standard:

- The *survival function* $S(t)$ is defined as the probability² that a component survives beyond time t :

$$S(t) = P(T > t)$$

In some disciplines, this is called the “reliability function” $R(t)$.

¹Indeed, reliability engineering principles are now being applied to the burgeoning field of life extension, or SENS (Strategically Engineered Negligible Senescence). See here for a good review: ieeexplore.ieee.org/ie15/6/29382/01330807.pdf?arnumber=1330807

²Note that because lowercase and uppercase t and T are typographically different, I don’t use the dummy variable u here. I only used that to prevent confusion between X and x in written rather than typeset solutions.

- The *hazard rate* $\lambda(t)$ is proportional to the probability of failure in the time interval $[t, t + dt]$ given that a component has survived to time t :

$$\lambda(t)dt = \frac{P(T \in [t, t + dt])}{P(T > t)}$$

This can be interpreted as the instantaneous failure rate at time t .

- Express the survival function and the hazard rate in terms of the pdf and cdf of T .
- Derive explicit expressions for the hazard rate and survival function for the following distribution:

$$f_T(t) = \frac{\alpha}{\beta} t^{\alpha-1} e^{-t^\alpha/\beta} \quad (0 \leq t < \infty; \alpha > 0; \beta > 0)$$

This is called a *Weibull* distribution with shape parameter α and scale parameter β and is important in modeling failure time distributions in many industrial and medical applications.

- Suppose that $\alpha = 1$ and $\beta = \frac{1}{\lambda}$. Do you recognize this distribution? What is the hazard rate and survival function for this distribution? Why might this not be a good model for human survival times? (Hint: Remember HW6, Problem 2).
- Suppose that human survival times (in years) are modeled by a Weibull with shape parameter $\alpha = 2$ and scale parameter $\beta = 50$. Under this model, what is the probability that a random human survives beyond age 80?

5. Functions of RVs

- Suppose that a random variable X can have each of the seven values $-3, -2, -1, 0, 1, 2, 3$. Determine the pmf of $Y = X^2 - X$.
- Suppose that a random variable $X \sim \mathcal{N}(\mu, \sigma^2)$. Determine the pdf of $Y = X^2$. If $\mu = 0$ and $\sigma = 1$, show that this corresponds to a chi-squared distribution and identify the number of degrees of freedom.

6. Mixture Distributions

Suppose that the random vector $\vec{X} = (X_1, X_2, X_3)$ has the following joint distribution function:

$$f_{X_1, X_2, X_3}(u_1, u_2, u_3) = \begin{cases} cu_1^{1+u_2+u_3}(1-u_1)^{3-u_2-u_3} & \text{for } 0 < x_1 < 1 \text{ and } x_2, x_3 \in \{0, 1\} \\ 0 & \text{else} \end{cases}$$

- Which of the components of this random vector are continuous and which are discrete?
- What is the value of the constant c ?
- Determine $P(X_2, X_3)$.
- Determine $P(X_1|X_2 = 1, X_3 = 1)$.

7. *Covariance/Correlation*

- (a) Suppose that X and Y are negatively correlated. Is $\text{Var}(X + Y)$ larger or smaller than $\text{Var}(X - Y)$?
- (b) Suppose that X_i for $i \in \{1, \dots, n\}$ are random variables such that the variance of each is 1 and the correlation between each pair of random variables is ρ . Determine $\text{Var}(X_1 + \dots + X_n)$. Is this larger or smaller than the variance of the sum assuming independence?

8. *Applied Statistics 101*

- (a) Suppose that a person drinks from a bottle containing 40 ounces of a certain beverage. Suppose that the average size of each sip is 1 ounce with a standard deviation of .25 ounces and that sip volumes are statistically independent. Determine the probability that the bottle will not be empty after 30 sips have occurred. You may assume that sips continue occurring even after the bottle is empty.
- (b) A group of 50 students crowd up against the entrance to a party. Each student must wait a time T_i for the highly trained doorman to confiscate all liquids and lipstick (which of course may be bombs), perform an ID check, and complete a retinal scan. The specific distribution of these T_i is unknown but constant, and it is known that the average waiting time is 1 minute with standard deviation of .5 minutes. Assuming that waiting times occur independently of each other, what is the probability that the last person in line gets into the party after waiting less than 40 minutes?

9. *Stat Happens in Vegas, Stays in Vegas*

You have been hired by the Mirage to ensure that the proportion of times that each roulette wheel comes up red³ is close to $p = 18/38$. To do this you are allowed to spin each wheel as many times as you want.

- (a) Express this as a hypothesis testing problem. What is the null hypothesis? What is the alternative hypothesis?
- (b) Suppose that you observed that 30 of 76 spins came up red. What is the probability of observing as great a deviation or greater from the expected mean under the null distribution? Use a two-tailed test.
- (c) How many spins of the wheel do you need to make to have a 95% chance of detecting deviations of greater than .005 from the true mean?

³Note that this probability is supposed to be 18/38, not .50, or else the casino would not win money on average.