Homework_07

October 24, 2019

1 Homework 7

This week, your homework **only involves written work; there are no coding parts**. It is your responsibility to ensure that your homework is submitted completely and properly to Gradescope. Refer to the bottom of the notebook for submission instructions.

1.0.1 1. Varying Parameters and Independence Assumptions

You can leave answers as numerical formulas. You don't have to find decimal values unless they are obvious.

(a) Find the expectation and SD of the number of heads in 26 tosses of a fair coin.

(b) Find the expectation and SD of the number of red cards in 26 cards dealt at random from a standard deck. Compare the answers with those in Part **(a)**: Are the expectations different, and if so, which one is bigger? Are the SDs different, and if so, which one is bigger?

(c) Find the expectation and SD of the number of sixes in 26 rolls of a die. Compare the SD with that in Part **(a)**: Without calculation, say which one is bigger, and why.

1.0.2 Solution

(a) Binomial (26, 1/2); expectation 13, SD $\sqrt{26 \cdot \frac{1}{2} \cdot \frac{1}{2}}$

(b) Hypergeometric (52, 26, 26); expectation 13, SD $\sqrt{26 \cdot \frac{1}{2} \cdot \frac{1}{2}}\sqrt{\frac{26}{51}}$. The expectations are the same and the SD is **(a)** is bigger because the fpc is less than 1.

(c) Binomial (26, 1/6); expectation 26/6, SD $\sqrt{26 \cdot \frac{1}{6} \cdot \frac{5}{6}}$, The SD in **(a)** is bigger because $\sqrt{p(1-p)}$ is largest at $p = 1/2$.

#newpage

1.0.3 2. Collecting Distinct Values

In Homework 4 you found the expectation of each of the random variables below. Go back and see how you did that, and then find the variance of each one.

For one part you will need the fact that the SD of a geometric (p) random variable is $\frac{\sqrt{q}}{p}$ where $q = 1 - p$. We haven't proved that as the algebra takes a bit of work. We will prove it later in the course by conditioning.

(a) A die is rolled n times. Find the variance of number of faces that *do not* appear.

(b) Use your answer to (a) to find the variance of the number of distinct faces that *do* appear in n rolls of a die.

(c) Find the variance of the number of times you have to roll a die till you have seen all of the faces.

1.0.4 Solution

(a) Let X be the number of faces that don't appear in the n rolls. Then $X = \sum_{j=1}^{6} I_j$ where I_j is the indicator of the event that Face j doesn't appear.

By symmetry and the formula for the variance of a sum, $Var(X) = 6Var(I_1) + 6 \cdot 5Cov(I_1, I_2)$. Plug in $Var(I_1) = p(1-p)$ where $p = (5/6)^n$, and $Cov(I_1, I_2) = (4/6)^n - (5/6)^n(5/6)^n$.

(b) If Y is the number of faces that do appear, then $Y = 6 - X$ for X defined in (a). So $Var(Y) = Var(X)$.

(c) If T is the required number of rolls, then $T = 1 + W_2 + W_3 + W_4 + W_5 + W_6$ where W_i is the number of rolls to get the ith new face after the $(i-1)$st new face has appeared. The W_i's are independent and for each i the distribution of W_i is geometric with parameter $p_i = (6-(i-1))/6$. By the addition rule for the variance of a sum of independent random variables,

$$Var(T) = \sum_{i=2}^{6} q_i/p_i^2 \quad \text{where } p_i = (7-i)/6 \text{ and } q_i = 1 - p_i$$

#newpage

1.0.5 3. Bounds

A random variable X, not necessarily non-negative, has $E(X) = 20$ and $SD(X) = 4$. In each part below **find the best bounds you can** based on the information given.

(a) Find upper and lower bounds for $P(0 < X < 40)$.

(b) Find upper and lower bounds for $P(10 < X < 40)$.

(c) Find an upper bound for $P(X \geq 40)$.

(d) Find an upper bound for $P(X^2 \geq 900)$.

1.0.6 Solution

(a) $P(0 < X < 40) = P(|X-20| < 20) = 1 - P(|X-20| \geq 20)$

By Chebyshev's inequality, $P(|X-20| \geq 20) = P(|X-20| \geq 5SD(X)) \leq \frac{1}{25}$.

So $P(0 < X < 40) \geq 1 - \frac{1}{25} = 0.96$

Lower bound 0.96, upper bound 1

(b) $P(10 < X < 40) \geq P(10 < X < 30) = P(|X-20| \leq 10) = P(|X-20| \leq 2.5SD(X)) \geq 1 - \frac{1}{2.5^2} = 0.84$

Lower bound 0.84, upper bound 1

(c) $P(X \geq 40) \leq P(X \leq 0) + P(X \geq 40) = P(|X-20| \geq 20) \leq \frac{1}{25} = 0.04$ by Part (a).

(d) By Chebyshev:

$P(X^2 \geq 900) = P(X \leq -30) + P(X \geq 30) \leq P(X \leq 10 \text{ or } X \geq 30) \leq \frac{1}{2.5^2} = 0.16$ as in 2b.

By Markov:

$E(X^2) = 400 + 16 = 416$

Since $X^2 > 0$ we can use Markov's Inequality:

$P(X^2 \geq 900) \leq \frac{416}{900} \approx 0.462$ which is much bigger than Chebyshev's upper bound.

So the bound to use is Chebyshev's: 0.16.

#newpage

1.0.7 4. The "Sample Variance"

Let $X_1, X_2, \dots, X_n$ be i.i.d., each with mean μ and SD σ. Let $\bar{X} = \frac{1}{n}\sum_{i=1} X_i$ be the sample mean.

(a) Find $E(\bar{X})$ and $SD(\bar{X})$.

(b) For each i, find $Cov(X_i, \bar{X})$. [Plug in the definition of $\bar{X}$ and use bilinearity.]

(c) For each i in the range 1 through n, define the *ith deviation in the sample* as $D_i = X_i - \bar{X}$. Find $E(D_i)$ and $Var(D_i)$. [Write the variance as $Cov(D_i, D_i)$, plug in the definition of D_i, and use bilinearity.]

(d) Define the random variable $\hat{\sigma}^2$ as

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} D_i^2$$

Find $E(\hat{\sigma}^2)$.

For this random variable, the notation $\hat{\sigma}^2$ is pretty standard in statistics. Just think of $\hat{\sigma}^2$ as a symbol; it doesn't help to start thinking about the random variable that is its square root.

(e) Use Part **d** to construct a random variable denoted S^2 that is an unbiased estimator of σ^2. This random variable S^2 is called the *sample variance* and is frequently used in inference.

1.0.8 Solution

(a) Let $S_n = \sum_{i=1}^{n} X_i$. Then $E(\bar{X}) = E(\frac{S_n}{n}) = \frac{n\mu}{n} = \mu$

$Var(\bar{X}) = \frac{1}{n^2} Var(S_n) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$. So $SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$.

(b) $Cov(X_i, \bar{X}) = Cov(X_i, \frac{1}{n}\sum_{j=1}^{n} X_j) = \frac{1}{n}\sum_{j=1}^{n} Cov(X_i, X_j) = \frac{1}{n}Cov(X_i, X_i)$ because $Cov(X_i, X_j) = 0$ for $i \neq j$ by independence.

So $Cov(X_i, \bar{X}) = \frac{\sigma^2}{n}$

(c) $E(D_i) = E(X_i - \bar{X}) = E(X_i) - E(\bar{X}) = 0$

$Var(D_i) = Cov(D_i, D_i) = Cov(X_i - \bar{X}, X_i - \bar{X}) = Cov(X_i, X_i) - Cov(X_i, \bar{X}) - Cov(\bar{X}, X_i) + Cov(\bar{X}, \bar{X}) = \sigma^2 - 2\frac{\sigma^2}{n} + \frac{\sigma^2}{n} = \sigma^2\frac{n-1}{n}$

(d) $E(D_i^2) = Var(D_i)$ because $E(D_i) = 0$. So $E(D_i^2) = \sigma^2\frac{n-1}{n}$ for all i.

$E(\hat{\sigma}^2) = \frac{1}{n}\sum_{i=1}^{n} E(D_i^2) = \frac{1}{n} n\sigma^2\frac{n-1}{n} = \sigma^2\frac{n-1}{n}$

(e) Let $S^2 = \frac{n}{n-1}\hat{\sigma}^2$. Then $E(S^2) = \sigma^2$ so S^2 is an unbiased estimator of σ^2. Note that $S^2 = \frac{1}{n-1}\sum_{i=1}^{n} D_i^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

#newpage

1.0.9 5. Fun with Indicators: The Matching Problem

In the familiar setting of the matching problems, there are n letters labeled 1 through n and n envelopes labeled 1 through n. The letters are distributed at random into the envelopes, one letter per envelope, such that all $n!$ permutations are equally likely.

Let M be the number of letters that fall into envelopes with the corresponding label. That is, M is the number of "matches" or fixed points of the permutation.

(a) Fill in the blank:

$$M = I_1 + I_2 + \ldots + I_n$$

where for each j in the range 1 through n, $I_j = 1$ if ________________________________, and $I_j = 0$ otherwise.

(b) Use Part **(a)** to show that $E(M)$ has the same numerical value for all n.

(c) Use Part **(a)** to show that $Var(M)$ has the same numerical value for all n.

(d) In Lab 3 you found the approximate distribtion of M when n is large. Are your answers to Parts **(c)** and **(d)** consistent with that approximation?

1.0.10 Solution

(a) "where for each j in the range 1 through n, $I_j = 1$ **if letter j is placed in envelope j**, and $I_j = 0$ otherwise."

(b) $E(I_j) = \frac{1}{n}$ for all j so $E(M) = 1$.

(c) By symmetry and the formula for the variance of a sum,

$Var(M) = nVar(I_1) + n(n-1)Cov(I_1, I_2) = n \cdot \frac{1}{n} \cdot \frac{n-1}{n} + n(n-1)\left(\frac{1}{n} \cdot \frac{1}{n-1} - \frac{1}{n} \cdot \frac{1}{n}\right) = \frac{n-1}{n} + 1 - \frac{n-1}{n} = 1$

(d) Recall from Lab 2 that you found M is approximately Poisson (1) for large n. The Poisson (1) distribution has expectation 1 and variance 1, so these results are consistent.

#newpage

1.1 Submission Instructions

Many assignments throughout the course will have a written portion and a code portion. Please follow the directions below to properly submit both portions.

1.1.1 Written Portion

- Scan all the pages into a PDF. You can use any scanner or a phone using an application. Please **DO NOT** simply take pictures using your phone.
- Please start a new page for each question. If you have already written multiple questions on the same page, you can crop the image or fold your page over (the old-fashioned way). This helps expedite grading.
- It is your responsibility to check that all the work on all the scanned pages is legible.

1.1.2 Submitting

- Combine the PDFs from the written and code portions into one PDF. Here is a useful tool for doing so.
- Submit the assignment to Homework 7 on Gradescope.
- **Make sure to assign each page of your pdf to the correct question.**
- **It is your responsibility to verify that all of your work shows up in your final PDF submission.**

1.1.3 We will not grade assignments which do not have pages selected for each question.

```
In [ ]:
```