Stat 200A, Fall 2018
A. Adhikari and Z. Bartha

WEEK 5 EXERCISES

You are expected to do all these problems, but for **Homework 5** please turn in **only Problems 2, 3, 6, and 7** on **Thursday September 27 at the start of lecture**.

1. Warmup

(a) Let H_n be the number of heads in n tosses of a coin and S_n the number of sixes in n rolls of a die. Find $E(H_n)$, $SD(H_n)$, $E(S_n)$, and $SD(S_n)$. Which is bigger: $SD(H_n)$ or $SD(S_n)$? Why?

(b) Continuing Part **a**, what are the distributions of H_n and S_n? For large n, what are the approximate distributions of H_n and S_n?

(c) A population consists of N elements of which G are good. Let R_n be the number of good elements in a sample of size n drawn at random with replacement from the population. Let W_n be the number of good elements in a random sample of size n drawn at random without replacement from the population. Find $E(R_n)$, $SD(R_n)$, $E(W_n)$, and $SD(W_n)$. Which is bigger: $SD(R_n)$ or $SD(W_n)$? Why?

(d) In a city that has over a million voters, 49% of the voters belong to Party A. A simple random sample of 2,500 voters is taken. Approximately what is the chance that the majority of sampled voters belong to Party A? Justify your approximation: which distribution are you approximating, and by what? Why?

2. Random Counts, Part 1

Last week you found the expectations of the random variables below. Now find the variances.

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'll prove it later 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 Part **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.

3. Random Counts, Part 2

(a) In the matching problem 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.

Find $E(M)$ and $Var(M)$. In Week 1 Exercises, you found the approximate distribution of M for large n. Are the expectation and variance consistent with this distribution?

(b) A deck consists of n cards, of which r are red. Cards are dealt at random without replacement till a red card appears. Let X be the number of cards dealt. Find $E(X)$ and $Var(X)$. Use symmetry; there should be no combinatorial terms or factorials in your answers.

(c) A deck consists of n cards, of which r are red and the rest are blue. Cards are dealt at random without replacement till all the red cards have been dealt. Let X be the number of cards dealt. Use symmetry and

Part **b** to find $E(X)$ and $Var(X)$.

4. Correlation

The covariance of random variables X and Y has nasty units: the product of the units of X and the units of Y. Dividing the covariance by the two SDs results in an important pure number.

The *correlation coefficient* of the random variables X and Y is defined as

$$r(X,Y) = \frac{Cov(X,Y)}{SD(X)SD(Y)}$$

It is called the correlation, for short. The definition explains why X and Y are called *uncorrelated* if $Cov(X,Y) = 0$.

(a) Let X^* be X is standard units and let Y^* be Y in standard units. Check that

$$r(X,Y) = E(X^*Y^*)$$

(b) Use the fact that both $(X^* + Y^*)^2$ and $(X^* - Y^*)^2$ are non-negative random variables to show that $-1 \le r(X,Y) \le 1$.

[First find the numerical values of $E(X^*)$ and $E(X^{*2})$. Then find $E(X^* + Y^*)^2$.]

(c) Show that if $Y = aX + b$ where $a \ne 0$, then $r(X,Y)$ is 1 or -1 depending on whether the sign of a is positive or negative.

(d) Consider a sequence of i.i.d. Bernoulli (p) trials. For any positive integer k let X_k be the number of successes in trials 1 through k. **Use bilinearity** to find $Cov(X_n, X_{n+m})$ and hence find $r(X_n, X_{n+m})$.

(e) Fix n and find the limit of your answer to **(d)** as $m \to \infty$. Explain why the limit is consistent with intuition.

5. Relations Between Random Variables

This exercise is about departures from the "independent and identically distributed" (i.i.d.) model, with particular attention to correlation.

(a) Let X_1 and X_2 be the numbers appearing on the first and second rolls of a die. Let $S = X_1 + X_2$ and $D = X_1 - X_2$. Are S and D identically distributed? Are they independent? Are they uncorrelated?

(b) Construct two random variables X and Y such that X and Y are identically distributed and negatively correlated, that is, $Cov(X,Y)$ is negative. You can do this easily on the space of a few tosses of a coin.

(c) Construct two random variables X and Y such that $X \ne Y$, X and Y are identically distributed and positively correlated, that is, $Cov(X,Y)$ is positive. This too can be done on the space of a few tosses of a coin.

6. The "Sample Variance"

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

7. Geometric Mean

(a) Let U have the uniform distribution on the interval $(0, 1)$. Find the cdf of $-\log(U)$. Identify this as the cdf of a well known distribution and provide the relevant parameters.

(b) Let $U_1, U_2, \ldots, U_n$ be i.i.d. uniform on the interval $(0, 1)$, and let $G_n = (U_1 U_2 \cdots U_n)^{1/n}$ be the geometric mean of the sample. Show that there is a constant c such that $G_n \xrightarrow{P} c$ as $n \to \infty$, and find c.

(c) For large n and small $\epsilon > 0$, approximate $P(|G_n - c| < \epsilon)$. Justify your answer.

8. Empty Boxes

There are n balls and $2n$ boxes. Each ball is placed in a box picked uniformly at random, independent of the placement of all other balls. Let W_n be the proportion of empty boxes.

(a) Find $E(W_n)$ and $Var(W_n)$.

(b) Show that there is a constant c such that $W_n \xrightarrow{P} c$ as $n \to \infty$, and find c.

9. Reliability

Let X_n be the number of successes in n i.i.d. Bernoulli (0.9) trials. About how large does n have to be so that the chance of 100 or more successes is about 99%?

Versions of this calculation are used by airlines to work out by how much they will overbook their flights, or by manufacturers who need to get a minimum number of good items using a process that has some chance of producing duds.