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

WEEK 6 EXERCISES

You are expected to do all these problems, but for **Homework 6** please turn in **only Problems 2, 3, and 5** on **Thursday October 4 at the start of lecture**.

1. Linear Combinations of Independent Normals

Let Z_1, Z_2, Z_3, Z_4 be i.i.d. standard normal variables. Find the following without integration. You can leave answers in terms of the standard normal cdf Φ if necessary.

(a) $P(3Z_1 + 2Z_2 > Z_3 + 4Z_4)$

$$= \mathbb{P}(3Z_1 + 2Z_2 - Z_3 - 4Z_4 > 0) = \mathbb{P}\left(\frac{3Z_1 + 2Z_2 - Z_3 - 4Z_4}{\sqrt{30}} > 0\right) = 1 - \Phi(0) = \frac{1}{2},$$

since $3Z_1 + 2Z_2 - Z_3 - 4Z_4 \sim N(0, 30)$, so $\frac{3Z_1+2Z_2-Z_3-4Z_4}{\sqrt{30}} \sim N(0,1)$.

(b) $P(Z_1 + Z_2 > Z_3 + Z_4 + 1)$

$$= \mathbb{P}(Z_1 + Z_2 - Z_3 - Z_4 > 1) = \mathbb{P}\left(\frac{Z_1 + Z_2 - Z_3 - Z_4}{2} > \frac{1}{2}\right) = 1 - \Phi(0.5),$$

since $Z_1 + Z_2 - Z_3 - Z_4 \sim N(0, 4)$, so $\frac{Z_1+Z_2-Z_3-Z_4}{2} \sim N(0,1)$.

(c) $E(3Z_1 + 2Z_2 - Z_3 - 4Z_4 + 10)$

$$= 3E(Z_1) + 2E(Z_2) - E(Z_3) - 4E(Z_4) + 10 = 10$$

(d) $SD(3Z_1 + 2Z_2 - Z_3 - 4Z_4 + 10)$

$$= \sqrt{Var(3Z_1 + 2Z_2 - Z_3 - 4Z_4 + 10)} = \sqrt{Var(3Z_1) + Var(2Z_2) + Var(-Z_3) + Var(-4Z_4)}$$
$$= \sqrt{9 + 4 + 1 + 16} = \sqrt{30}$$

2. Distance Between Normal Points

(a) Show that X has the normal (μ, σ^2) distribution if and only if $X = \sigma Z + \mu$ where Z has the standard normal distribution.

Since any linear transformation of a normal variable is normal, we just have to check the scaling of the parameters. If $Z \sim N(0,1)$, then $E(\sigma Z + \mu) = \sigma E(Z) + \mu = \mu$ and $Var(\sigma Z + \mu) = \sigma^2 Var(Z) = \sigma^2$. The reverse is calculation is similar: if $X \sim N(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma$ has $E(Z) = E(X)/\sigma - \mu/\sigma = 0$ and $Var(Z) = Var(X)/\sigma^2 = 1$.

(b) Consider two points thrown independently on the plane, such that the two coordinates of each point are i.i.d. normal (μ, σ^2) random variables. Find the expectation and variance of the distance between the two points.

If the two points have coordinates (X_1, Y_1) and (X_2, Y_2), then the distance between them is $D = \sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}$. Now $X_1 - X_2 \sim N(0, 2\sigma^2)$, and also $Y_1 - Y_2 \sim N(0, 2\sigma^2)$, independently. Hence, $D = \sqrt{(\sqrt{2}\sigma Z_1)^2 + (\sqrt{2}\sigma Z_2)^2} =$

$\sqrt{2}\sigma\sqrt{Z_1^2+Z_2^2}$ with $Z_1, Z_2 \sim N(0,1)$ i.i.d. We know that $R = \sqrt{Z_1^2+Z_2^2}$ has the (standard) Rayleigh distribution with density $f_R(r) = re^{-r^2/2}$. This has expectation

$$E(R) = \int_0^\infty r f_R(r)dr = \int_0^\infty r \cdot \left(re^{-r^2/2}\right)dr = \left[r(-e^{-r^2/2})\right]_{r=0}^\infty + \int_0^\infty e^{-r^2/2}dr$$
$$= \int_0^\infty e^{-r^2/2}dr = \sqrt{2\pi}\cdot\frac{1}{2}\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}}e^{-r^2/2}dr = \sqrt{2\pi}\cdot\frac{1}{2}\cdot 1 = \sqrt{\frac{\pi}{2}}.$$

Hence, $E(D) = E(\sqrt{2}\sigma R) = \sqrt{2}\sigma\sqrt{\frac{\pi}{2}} = \sigma\sqrt{\pi}$.

For the variance,

$$E(D^2) = E(2\sigma^2(Z_1^2+Z_2^2)) = 2\sigma^2(EZ_1^2+EZ_2^2) = 2\sigma^2(1+1) = 4\sigma^2,$$

thus,

$$Var(D) = ED^2 - (ED)^2 = 4\sigma^2 - \pi\sigma^2 = (4-\pi)\sigma^2.$$

3. The Cauchy Density

Let X be uniform on the interval $(-\pi/2, \pi/2)$, and let $Y = \tan(X)$.

(a) Find the density of Y. This is called the *Cauchy* density.

f_X is constant $\frac{1}{\pi}$ on $(-\pi/2, \pi/2)$, so with $y = \tan(x)$, $-\infty < y < \infty$ we have

$$f_Y(y) = \frac{f_X(x)}{\left|\frac{dy}{dx}\right|} = \frac{1/\pi}{(\tan(x))'} = \frac{1/\pi}{1/\cos^2(x)} = \frac{\cos^2(x)}{\pi} = \frac{\cos^2(\arctan(y))}{\pi}.$$

We want to express $\cos^2(x)$ in terms of $\tan(x)$ to simplify the expression on the right-hand side. For this we use the fact that

$$\tan^2(x) = \frac{\sin^2(x)}{\cos^2(x)} = \frac{1-\cos^2(x)}{\cos^2(x)} = \frac{1}{\cos^2(x)} - 1,$$

hence $\cos^2(x) = \frac{1}{1+\tan^2(x)}$. This implies

$$f_Y(y) = \frac{1}{\pi}\cdot\frac{1}{1+\tan^2(x)} = \frac{1}{\pi(1+y^2)}.$$

(Alternatively, $f_Y(y) = f_X(x)\cdot\left|\frac{dx}{dy}\right| = \frac{1}{\pi}\cdot(\arctan(y))' = \frac{1}{\pi}\cdot\frac{1}{1+y^2}$.)

(b) Show that the distribution of Y is symmetric about 0 but $E(Y)$ is undefined.

$f_Y(y) = f_Y(-y)$ for all y, so the density is symmetric about 0.

$E(Y)$ is undefined, because the function $y \cdot f_Y(y)$ is not integrable on $(-\infty, \infty)$: if we just integrate on the positive half-line and use the change of variables $u = 1+y^2$, $\frac{du}{dy} = 2y$ we have

$$\int_0^\infty y\cdot\frac{1}{\pi}\cdot\frac{1}{1+y^2}dy = \int_1^\infty \frac{1}{2\pi}\cdot\frac{1}{u}du = \infty.$$

4. Spacings

Let U_1, U_2, U_3, U_4 be i.i.d. uniform $(0,1)$ random variables. Let $U_{(1)}, U_{(2)}, U_{(3)}, U_{(4)}$ denote the four variables arranged in increasing order. As a visualization, if you mark the points U_1, U_2, U_3, U_4 on the unit interval, then from left to right the marks will be at $U_{(1)}, U_{(2)}, U_{(3)}, U_{(4)}$.

(a) Find the density of $U_{(1)}$.

The cdf of $U_{(1)}$ is

$$F_{U_{(1)}}(x) = \mathbb{P}(U_{(1)} < x) = 1 - \mathbb{P}(U_{(1)} > x) = 1 - \mathbb{P}(U_1, U_2, U_3, U_4 > x) = 1 - (\mathbb{P}(U_i > x))^4 = 1 - (1-x)^4$$

for $0 \le x \le 1$. Hence, the density of $U_{(1)}$ is

$$f_{U_{(1)}}(x) = \frac{dF_{U_{(1)}}(x)}{dx} = 4(1-x)^3$$

for $0 < x < 1$.

(b) Find the density of $U_{(4)}$.

The cdf of $U_{(4)}$ is

$$F_{U_{(4)}}(y) = \mathbb{P}(U_{(4)} < y) = \mathbb{P}(U_1, U_2, U_3, U_4 < y) = (\mathbb{P}(U_i < y))^4 = y^4$$

for $0 \le y \le 1$. Hence, the density is

$$f_{U_{(4)}}(y) = \frac{dF_{U_{(4)}}(y)}{dy} = 4y^3$$

for $0 < y < 1$.

(c) Find the density of $1 - U_{(4)}$, and compare with the answer to Part **a**.

Applying the change of variables $x = 1 - y$ to the density of $f_{U_{(4)}}$ above we get

$$f_{1-U_{(4)}}(x) = f_{U_{(4)}}(y) = 4y^3 = 4(1-x)^3$$

for $0 < x < 1$, which is the same as $f_{U_{(1)}}$. (This can also be seen by symmetry: if we flip the interval, that won't change the distribution of the points, and $U_{(1)}$ becomes $1 - U_{(4)}$.)

(d) Review the method of Exercise **6c** of Week 2 Exercises and explain why $U_{(2)}$ and $U_{(3)}$ have the joint density given by

$$f(x,y) = \begin{cases} cx(1-y), & 0 < x < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

Here $c > 0$ is a constant. Find it without integration.

For any x and y such that $0 < x < y < 1$,

$$f(x,y)dxdy$$

is the (infinitesimal) probability that $U_{(2)}$ is in a dx neighborhood of x and $U_{(3)}$ is in a dy neighborhood of y. This probability is

$$4! \cdot x \cdot dx \cdot dy \cdot (1-y),$$

coming from the fact that there are 4! ways to order the points, the first has to come before x (prob. x), the second has to fall in the dx neighborhood of x (prob. dx), the third has to fall in the dy neighborhood of y (prob. dy) and the forth has to go between y and 1 (prob. $1-y$). Therefore, $f(x,y) = 24x(1-y)$ for $0 < x < y < 1$.

(e) Find the density of $U_{(3)} - U_{(2)}$ and compare it with the answer to Part **a**.

One way to do it is to compute the cdf of $U_{(3)} - U_{(2)}$ first: with $0 < z < 1$

$$F_{U_{(3)}-U_{(2)}}(z) = \mathbb{P}(U_{(3)} - U_{(2)} < z).$$

This is the integral of the joint density of $U_{(2)}$ and $U_{(3)}$ over the region where $y < x + z$. If we carefully look at the boundaries of this region, the integral is

$$F_{U_{(3)}-U_{(2)}}(z) = \int_0^{1-z} \int_x^{x+z} 24x(1-y)dydx + \int_{1-z}^{1} \int_x^{1} 24x(1-y)dydx.$$

Calculating this yields

$$F_{U_{(3)}-U_{(2)}}(z) = 1 - (1-z)^4,$$

the same as the cdf of $U_{(1)}$ in part (a). Hence, the density is also the same as that of $U_{(1)}$, $4(1-z)^3$ for $0 < z < 1$.

5. A Ratio

Let X and Y have the joint density given by

$$f(x,y) \;=\; \begin{cases} \frac{1}{y}e^{-(x+y^2)/y}, & x > 0, y > 0 \\ 0 & \text{otherwise} \end{cases}$$

Show that X/Y and Y are i.i.d. exponential variables and hence find $E(X)$ and $Var(X)$.

Let's apply the change of variables $(X,Y) \to (U,V)$ with $U = X/Y$ and $V = Y$. If $f_{X,Y}$ is the original and $f_{U,V}$ is the new joint density and $u = x/y$, $v = y$, then $f_{U,V}(u,v) = f_{X,Y}(x,y)/|\det(J(x,y))|$, where $J(x,y)$ is the matrix

$$\begin{bmatrix} \frac{du}{dx} & \frac{du}{dy} \\ \frac{dv}{dx} & \frac{dv}{dy} \end{bmatrix} = \begin{bmatrix} \frac{1}{y} & \frac{-x}{y^2} \\ 0 & 1 \end{bmatrix}.$$

This has determinant $1/y$, hence, for $u, v > 0$

$$f_{U,V}(u,v) = f_{X,Y}(x,y) \cdot y = e^{(-x+y^2)/y} = e^{-x/y-y} = e^{-u-v} = e^{-u}e^{-v},$$

the product of two $Exp(1)$ densities. Therefore, $U = X/Y$ and $V = Y$ are independent $Exp(1)$ variables. Using this fact, we have

$$E(X) = E\left(\frac{X}{Y} \cdot Y\right) = E\left(\frac{X}{Y}\right) \cdot E(Y) = 1 \cdot 1 = 1,$$

and

$$E(X^2) = E\left(\left(\frac{X}{Y}\right)^2 \cdot Y^2\right) = E\left(\frac{X}{Y}\right)^2 \cdot E(Y^2) = 2 \cdot 2 = 4,$$

so

$$Var(X) = E(X^2) - (E(X))^2 = 4 - 1 = 3.$$

6. The Chi-Squared Distributions

Let n be a positive integer. In statistics, the gamma $(n/2, 1/2)$ distribution is known as the *chi-squared distribution with n degrees of freedom*. We will denote that distribution χ^2_n.

(a) Let Z be a standard normal variable. We showed in class that Z^2 has the gamma $(1/2, 1/2)$ distribution. Now let $Z_1, Z_2, \ldots, Z_n$ be i.i.d. standard normal variables,. Explain why $Z_1^2 + Z_2^2 + \cdots + Z_n^2$ has the χ_n^2 distribution.

We showed that in general if $X \sim \Gamma(r, \lambda)$ and $Y \sim \Gamma(s, \lambda)$, then $X + Y \sim \Gamma(r + s, \lambda)$. Using this recursively, $Z_1^2 + Z_2^2 \sim \Gamma(2/2, 1/2)$, $Z_1^2 + Z_2^2 + Z_3^2 \sim \Gamma(3/2, 1/2), \ldots$, so $Z_1^2 + Z_2^2 + \ldots + Z_n^2 \sim \Gamma(n/2, 1/2)$, which is the χ_n^2 distribution.

(b) Find the expectation and variance of the χ_n^2 distribution. Use Exercise **5ef** of Week 4 Exercises.

Exercise 5e from Week 4 shows that the $\Gamma(r, \lambda)$ distribution has expectation r/λ. Thus, the expectation of the $\Gamma(n/2, 1/2)$ distribution is $\frac{n/2}{1/2} = n$.

5f from Week 4 shows that the variance of $\Gamma(r, \lambda)$ is r/λ^2, thus, the variance of $\Gamma(n/2, 1/2)$ is $\frac{n/2}{(1/2)^2} = 2n$.

(c) Sketch the graph of the χ_n^2 distribution for large n, and explain your choice of shape.

For large n, the χ_n^2 distribution is close to the $Normal(n, 2n)$ distribution, since by part (a) we are adding up a large number of i.i.d. variables (the Z_i^2 variables), so the CLT applies with the expectation and variance computed in part (b).

Problems left from last week:

5/4. 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 $-1 \leq r(X, Y) \leq 1$.

[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 $a \neq 0$, 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.

5/5. 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 $X \neq Y$, 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.

5/9. 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.