DATA 1010
Problem Set 6
Due 19 October 2018 at 11 PM

Problem 1

In this problem, we will justify the idea that the mean of a random variable is the best constant estimator for the random variable, as measured by average squared error.

Suppose that X is a random variable for which $\mathbb{E}[X^2] < \infty$. Show that the minimum of the function $f(a) = \mathbb{E}[(X-a)^2]$ occurs at the value $a = \mathbb{E}[X]$.

Solution

The function f is a quadratic polynomial in a:

$$f(a) = \mathbb{E}[X^2 - 2aX + a^2] = \mathbb{E}[X^2] - 2a\mathbb{E}[X] + a^2.$$

Differentiating gives

$$f'(a) = -2\mathbb{E}[X] + 2a.$$

Setting this expression equal to 0, we get $a = \mathbb{E}[X]$.

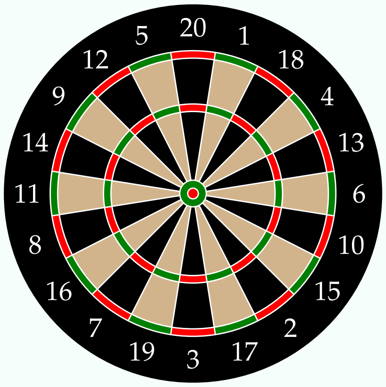
Problem 2

Suppose that the probability density function for the random point where your dart hits the dartboard* $D = \mathbb{R}^2$ is given by

$$f(x,y) = \frac{1}{\pi}e^{-x^2-y^2},$$

where the origin is situated at the dartboard's bull's eye, and where x and y are measured in inches (this function is positive everywhere in $\mathbb{R}^2$, so the "dartboard" includes the disk shown as well as the (infinite) wall it is mounted on—this is realistic insofar as one can indeed hit the wall with a dart throw). Find the probability of scoring triple 20 on your next throw.

Note: the triple 20 region is the smaller of the two thin red strips in the sector labeled "20". The inner and outer radii of this thin strip are 3.85 inches and 4.2 inches, respectively.



Solution

The region in question is described most easily in polar coordinates: it is the set of points whose polar coordinates (r,θ) satisfy $r_i \leq r \leq r_o$ and* $81^\circ \leq \theta \leq 99^\circ$, where $r_i = 3.85$ and $r_o = 4.2$. (Note that the width of each sector is $360^\circ/20 = 18^\circ$, so the angles of the rays bounding the sector labeled 20 are $90^\circ \pm \frac{18^\circ}{2}$)

Therefore, we can obtain the probability of hitting the triple 20 by expressing the density function in polar coordinates and integrating

$$\int_{9\pi/20}^{11\pi/20}\int_{r_i}^{r_o}\frac{1}{\pi}e^{-r^2}r\,dr\,d\theta = \left(\frac{\pi}{10}\right)\left(\frac{1}{\pi}\right)\left(-\tfrac{1}{2}e^{-r_o^2} - \left(-\tfrac{1}{2}e^{-r_i^2}\right)\right).$$

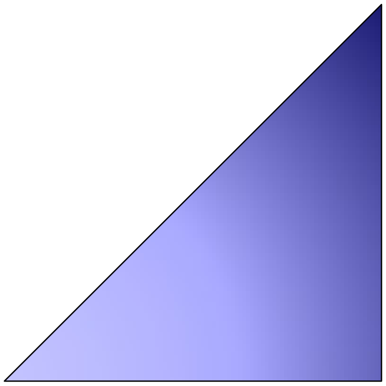
Substituting the given values of r_i and r_o yields a probability of approximately $\boxed{1.717 \times 10^{-8}}$ of scoring 60 on a single throw.

Problem 3

Find the expected distance from the origin to a point (X,Y) selected uniformly at random from the triangle T with vertices at $(0,0)$, $(1,0)$, and $(1,1)$. For fun, make your best estimate of the answer before you do any calculation, and at the end you can comment on whether your estimate ending up being low or high.

Hints: recall that $\mathbb{E}[g(X,Y)] = \iint_{\mathbb{R}^2} g(x,y) f(x,y)\,\mathrm{d}x\,\mathrm{d}y$ if (X,Y) has a joint distribution specified by the probability density function f. Also, you probably want to set the integral up in polar coordinates, and feel free to use a symbolic computation engine (such as Wolfram Alpha) to perform the actual integration.

Solution



Looking at the figure (which shows points shaded according to how far they are from the origin), my guess is 0.85. It looks like lots of points are about distance 1 from the origin, while a smaller area of points (the top right corner) are significantly farther away and a decent chunk of points are closer (bottom left corner).

The word "uniform" means that (X,Y)'s PDF is a constant function. To determine the value of this constant, we use the fact that the PDF integrates to 1. If $f(x,y) = k$ for all $(x,y) \in T$, then

$$\iint_T k\,\mathrm{d}A = k\,\mathrm{area}(T) = 1,$$

which implies that $k = 2$. We are calculating the expected value of $g(X,Y)$, where $g(x,y) = \sqrt{x^2+y^2}$. So the expected value of $g(X,Y)$ is given by $\iint_T g\,2\,\mathrm{d}A$. Because the integrand is conveniently expressed in polar coordinates, we represent T in polar coordinates as the set of points where θ is between 0 and $\pi/4$ and where r is between 0 and $\sec\theta$. Then we have

$$\iint_T g\,2\,\mathrm{d}A = \int_0^{\pi/4}\int_0^{\sec\theta} r\,2\,r\,\mathrm{d}r\,\mathrm{d}\theta = \frac{2}{3}\int_0^{\pi/4}\sec^3\theta\,\mathrm{d}\theta.$$

We can use integration by parts (or a symbolic computation engine) to determine that

$$\int \sec^3\theta\,\mathrm{d}\theta = \frac{1}{2}\left(\sec\theta\tan\theta + \ln|\sec\theta+\tan\theta|\right)$$

. Substitution yields an expected value of

$$\boxed{\frac{1}{3}\left(\sqrt{2}+\ln\left(\sqrt{2}+1\right)\right) \approx 0.765}.$$

We can check our result with Monte Carlo:

Julia

```julia
function randpoint()
    X,Y = rand(2)
    if X > Y
        [X,Y]
    else
        randpoint()
    end
end
mean(norm(randpoint()) for i=1:10^6) # returns approximately 0.765.
```

Problem 4

Suppose that X and Y are random variables whose joint distribution is given by the density $f(x,y) = \frac{3}{2}(x^2+y^2)$ on the unit square $[0,1]^2$.

(a) Find the probability density function of the distribution of X.

(b) Find the probability of the event that $X \geq \frac{1}{2}$ and $Y \geq \frac{1}{2}$.

Solution

(a) The PDF of X is obtained by integrating the joint density over y. We get

$$f_X(x) = \int_0^1 \frac{3}{2}(x^2+y^2)\mathrm{d}y = \frac{3}{2}x^2 + \frac{1}{2}.$$

(b) The probability of the event that $(X,Y) \in [1/2,1] \times [1/2,1]$ is the integral of the joint density over that region:

$$\int_{1/2}^1 \int_{1/2}^1 \frac{3}{2}(x^2+y^2)\,\mathrm{d}x\,\mathrm{d}y = \frac{7}{16}.$$

Problem 5

Consider two events A and B. Show that the indicator random variables $\mathbf{1}_A$ and $\mathbf{1}_B$ are positively correlated if $\mathbb{P}(A \mid B) > \mathbb{P}(A)$ and are negatively correlated if $\mathbb{P}(A \mid B) < \mathbb{P}(A)$.

Solution

We have

$$\operatorname{Cov}(\mathbf{1}_A, \mathbf{1}_B) = \mathbb{E}[\mathbf{1}_A\mathbf{1}_B] - \mathbb{E}[\mathbf{1}_A]\mathbb{E}[\mathbf{1}_B] = \mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B),$$

since $\mathbf{1}_A\mathbf{1}_B = \mathbf{1}_{A\cap B}$, and the expectation of the indicator of an event is the probability of that event. Using the conditional expectation formula, we have

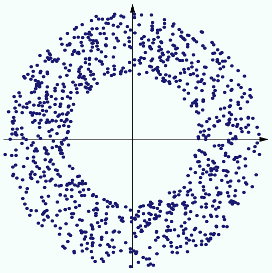
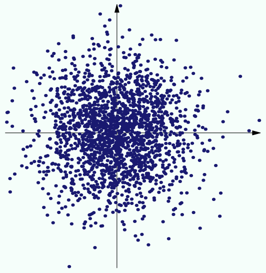
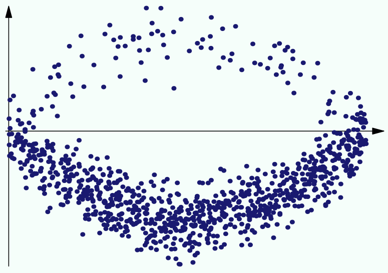
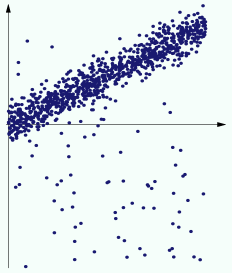
$$\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B) = \mathbb{P}(B)(\mathbb{P}(A \mid B) - \mathbb{P}(A)).$$

So the covariance (and hence also the correlation) is positive if $\mathbb{P}(A \mid B) - \mathbb{P}(A) > 0$, and it's negative if $\mathbb{P}(A \mid B) - \mathbb{P}(A) < 0$.

Problem 6

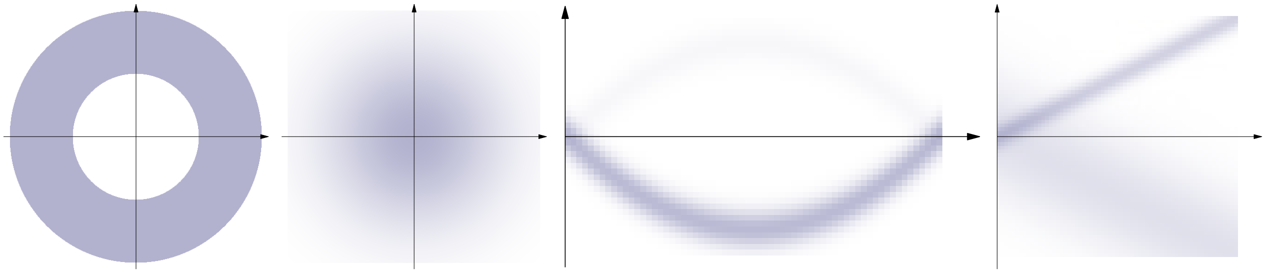
Suppose that many independent samples are drawn from the joint distribution of two random variables X and Y, and the results are as shown in the first figure below. Sketch, by shading, your best guess of the density function of the distribution of (X,Y). Also sketch a graph of the function $x \mapsto \mathbb{E}[Y \mid X = x]$.

Repeat the exercise for each of the remaining figures.

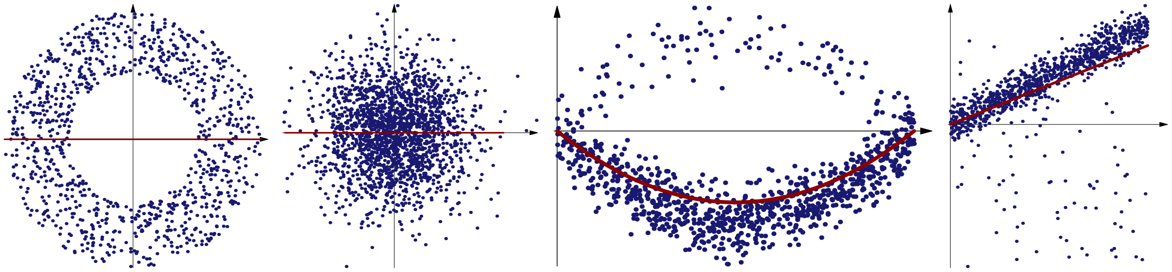





Solution

We expect that the density is large where there are lots of points, and the density is small where there are fewer points. We get something like the following:



The conditional expectation given $X = x$ is obtained by finding the center of mass of the joint density along each vertical line. We get (and these are exact, by the way):



Problem 7

Suppose that X and Y have joint PDF $f(x, y) = \frac{3}{2}y$ on the upper unit disk (that is, the set of points which have positive y-coordinate and are less than one unit from the origin).

(a) Verify that f is indeed a probability density function.

(b) Find the density of the distribution of X.

(c) Find the conditional density of Y given $X = x$.

(d) Find $\mathbb{E}[Y \mid X]$.

Solution

(a) We have $\int_0^\pi \int_0^1 \frac{3}{2} r \sin\theta (r \, \mathrm{d}r \, \mathrm{d}\theta) = 1$, so f is indeed a probability density function.

(b) The distribution of X is obtained by integrating out y:

$$\int_{\mathbb{R}} \frac{3}{2} y \mathbf{1}_{0 \leq y \leq \sqrt{1-x^2}} \, \mathrm{d}y = \int_0^{\sqrt{1-x^2}} \frac{3}{2} y \, \mathrm{d}y = \frac{3}{4}(1 - x^2).$$

(c) The conditional density of Y given $X = x$ is the joint density divided by X's marginal density at x:

$$f_{Y \mid X=x}(y) = \frac{f_{X,Y}(x, y)}{\frac{3}{4}(1 - x^2)} = \frac{\frac{3}{2} y \mathbf{1}_{0 \leq y \leq \sqrt{1-x^2}}}{\frac{3}{4}(1 - x^2)} = \frac{2y \mathbf{1}_{0 \leq y \leq \sqrt{1-x^2}}}{1 - x^2}.$$

(d) The conditional expectation of Y given X is obtained by integrating y times the conditional density of Y given $X = x$ and then substituting X for x:

$$\int_{\mathbb{R}} y \frac{2y \mathbf{1}_{0 \leq y \leq \sqrt{1-x^2}}}{1 - x^2} \, \mathrm{d}y = \int_0^{\sqrt{1-x^2}} \frac{2y^2}{1 - x^2} \, \mathrm{d}y = \frac{2(1 - x^2)^{3/2}}{3(1 - x^2)} = \frac{2}{3}\sqrt{1 - x^2}.$$

Problem 8

You're tasked with loading a pile of stones into your truck. No one has looked carefully at the pile yet, but based on past experience you decide to model the number of stones as a random variable N with $\mathbb{E}[N] = 50$. The stone weights are also random variables (independent of N), and each one has a mean of 10 pounds. Let X be the total weight of the stones in the pile.

(a) For any positive integer $n \geq 1$, find $\mathbb{E}[X \mid N = n]$. Use the result to find $\mathbb{E}[X \mid N]$ in terms of the random variable N.

(b) Use the fact that $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid N]]$ to find $\mathbb{E}[X]$.

Solution

(a) We have

$$\mathbb{E}[X \mid N = n] = \mathbb{E}[X_1 + \cdots + X_N \mid N = n] = \mathbb{E}[X_1 + \cdots + X_n \mid N = n] = \mathbb{E}[X_1 + \cdots + X_n],$$

since the X_i's are independent of $N = n$. By linearity of expectation, this expression is equal to $n\mathbb{E}[X_1] = 10n$. Therefore, $\mathbb{E}[X \mid N] = 10N$.

(b) We have

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid N]] = \mathbb{E}[10N] = 10\mathbb{E}[N] = 10 \cdot 50 = \boxed{500}.$$