Correlation/Cov+Stat

SDS 321

1. (4 points) Consider the following joint pmf test:

	$X = -1$	$X = 0$	$X = 1$
$Y = -1$	0.05	0.15	0.05
$Y = 0$	0.15	0.20	0.05
$Y = 1$	0.10	0.05	0.20

(a) (2 points) Find the covariance of X and Y ($Cov(X, Y)$).

$E[X]$:

$$E[X] = -1 \cdot 0.3 + 0 \cdot 0.4 + 1 \cdot 0.3 = 0$$

$E[Y]$:

$$E[Y] = -1 \cdot 0.25 + 0 \cdot 0.4 + 1 \cdot 0.35 = 0.1$$

$E[XY]$:

$$E[XY] = 1 \cdot 0.05 - 1 \cdot 0.10 - 1 \cdot 0.05 + 1 \cdot 0.20 = 0.10$$

$Cov(X, Y)$:

$$Cov(X, Y) = E[XY] - E[X]E[Y] = 0.10 - 0 \cdot 0.10 = 0.10$$

(b) (2 points) Find the correlation of X and Y ($Cor(X, Y)$).

$E[X^2]$:

$$E[X^2] = 1 \cdot 0.3 + 0 \cdot 0.4 + 1 \cdot 0.3 = 0.6$$

$Var(X)$:

$$Var(X) = E[X^2] - E[X]^2 = 0.6 - 0^2 = 0.6$$

$E[Y^2]$:

$$E[Y^2] = 1 \cdot 0.25 + 0 \cdot 0.4 + 1 \cdot 0.35 = 0.6$$

$$Var(Y) = E[Y^2] - E[Y]^2 = 0.6 - 0.1^2 = 0.59$$

$Cor(X,Y)$:

$$Cor(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{0.10}{\sqrt{0.6}\sqrt{0.59}} \approx 0.17$$

2. (4 points) Consider the following joint pdf:

$$f_{X,Y}(x,y) = \begin{cases} \frac{10}{9}x^2 y, & \text{if } 0 \le x \le 1, 0 \le y \le 3x \\ 0, & \text{otherwise} \end{cases}$$

(a) (2 points) Find the covariance of X and Y ($Cov(X,Y)$).

$E[X]$:

$$\begin{aligned} E[X] &= \int_0^1 x \int_0^{3x} \tfrac{10}{9}x^2 y \, dy \, dx \\ &= \int_0^1 x^3 \left(\tfrac{5}{9}y^2\right)\Big|_0^{3x} dx \\ &= \int_0^1 5x^5 dx \\ &= \tfrac{5}{6}x^6\Big|_0^1 = \tfrac{5}{6} \approx 0.83 \end{aligned}$$

$E[Y]$:

$$\begin{aligned} E[Y] &= \int_0^3 y \int_{\frac{y}{3}}^1 \tfrac{10}{9}x^2 y \, dx \, dy \\ &= \int_0^3 y^2 \left(\tfrac{10}{27}x^3\right)\Big|_{\frac{y}{3}}^1 dy \\ &= \int_0^3 y^2 \left(\frac{10}{27} - \frac{10}{729}y^3\right) dy \\ &= \int_0^3 \tfrac{10}{27}y^2 - \frac{10}{729}y^5 dy \\ &= \tfrac{10}{81}y^3 - \tfrac{10}{4374}y^6\Big|_0^3 \\ &= \tfrac{10}{3} - \tfrac{5}{3} = \tfrac{5}{3} \approx 1.67 \end{aligned}$$

$E[XY]$:

$$\begin{aligned} E[XY] &= \int_0^1 \int_0^{3x} xy\tfrac{10}{9}x^2y\ dy\ dx \\ &= \int_0^1 \int_0^{3x} \tfrac{10}{9}x^3y^2\ dy\ dx \\ &= \int_0^1 \tfrac{10}{27}x^3y^3\Big|_0^{3x}\ dx \\ &= \int_0^1 10x^6 dx \\ &= \tfrac{10}{7}x^7\Big|_0^1 = \tfrac{10}{7} \approx 1.43 \end{aligned}$$

Then $Cov(X, Y)$ is:

$$Cov(X, Y) = E[XY] - E[X]E[Y] = \tfrac{10}{7} - \tfrac{5}{6} \cdot \tfrac{5}{3} = \tfrac{5}{126} \approx 0.04$$

(b) (2 points) Find the correlation of X and Y ($Cor(X, Y)$).

$E[X^2]$:

$$\begin{aligned} E[X^2] &= \int_0^1 x^2 \int_0^{3x} \tfrac{10}{9}x^2y\ dy\ dx \\ &= \int_0^1 x^4\left(\tfrac{5}{9}y^2\right)\Big|_0^{3x} dx \\ &= \int_0^1 5x^6 dx \\ &= \tfrac{5}{7}x^7\Big|_0^1 = \tfrac{5}{7} \approx 0.71 \end{aligned}$$

$Var(X)$:

$$Var(X) = E[X^2] - E[X]^2 = \tfrac{5}{7} - (\tfrac{5}{6})^2 = \tfrac{5}{252} \approx 0.02$$

$E[Y^2]$:

$$\begin{aligned}
E[Y^2] &= \int_0^3 y^2 \int_{\frac{y}{3}}^1 \tfrac{10}{9} x^2 y \, dx \, dy \\
&= \int_0^3 y^3 \left(\tfrac{10}{27} x^3\right)\Big|_{\frac{y}{3}}^1 dy \\
&= \int_0^3 y^3 \left(\frac{10}{27} - \frac{10}{729} y^3\right) dy \\
&= \int_0^3 \tfrac{10}{27} y^3 - \frac{10}{729} y^6 dy \\
&= \tfrac{10}{108} y^4 - \tfrac{10}{5103} y^7 \Big|_0^3 \\
&= \tfrac{15}{2} - \tfrac{30}{7} = \tfrac{45}{14} \approx 3.21
\end{aligned}$$

$Var(Y)$:

$$Var(Y) = E[Y^2] - E[Y]^2 = \tfrac{45}{14} - (\tfrac{5}{3})^2 = \tfrac{55}{126} \approx 0.44$$

Then $Cor(X, Y)$:

$$Cor(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{\frac{5}{126}}{\sqrt{\frac{5}{252}}\sqrt{\frac{45}{14}}} \approx 0.157$$

3. (5 points) Let $E[X] = 1.5$, $E[X^2] = 3$,$E[Y] = 2$,$E[Y^2] = 5$, and $E[XY] = 3.5$. Find the following:

   (a) (1 point) $Cov(X, Y)$

$$Cov(X, Y) = E[XY] - E[X]E[Y] = 3.5 - 1.5 \cdot 2 = 0.5$$

   (b) (1 point) $Var(X - Y)$

$$\begin{aligned}
Var(X - Y) &= Var(X) + Var(Y) - 2Cov(X, Y) \\
&= E[X^2] - E[X]^2 + E[Y^2] - E[Y]^2 - 2Cov(X, Y) \\
&= 3 - 1.5^2 + 5 - 2^2 - 2 \cdot 0.5 = 0.75
\end{aligned}$$

(c) (1 point) $Cov(2X+1, -Y+2)$

$$Cov(2X+1, -Y+2) = -2Cov(X,Y) = -2 \cdot 0.5 = -1$$

(d) (2 points) $Cor(2X+1, -Y+2)$

$$\begin{aligned} Cor(2X+1, -Y+2) &= \frac{Cov(2X+1, -Y+2)}{\sqrt{Var(2X+1)}\sqrt{Var(-Y+2)}} \\ &= \frac{-2Cov(X,Y)}{\sqrt{4Var(X)}\sqrt{Var(Y)}} \\ &= \frac{-2Cov(X,Y)}{\sqrt{4(E[X^2]-E[X]^2)}\sqrt{E[Y^2]-E[Y]^2}} \\ &= \frac{-2 \cdot 0.5}{\sqrt{4(3-1.5^2)}\sqrt{5-2^2}} \\ &= \frac{-1}{\sqrt{3}\sqrt{1}} = -\frac{\sqrt{3}}{3} \approx -0.58 \end{aligned}$$

4. (3 points) Let $Y = 2x+1$ and X have the following pdf:

$$f_X(x) = \begin{cases} 2x, & \text{for } 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

(a) (2 points) If X and Y are independent, find the joint pdf of X and Y ($f_{X,Y}(x,y)$).

First find $f_Y(y)$:

$$f_Y(y) = \frac{1}{2}2\left(\frac{y-1}{2}\right) = \frac{y-1}{2}$$

Then:

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) = \begin{cases} x(y-1) & \text{for } 0 \le x \le 1, 1 \le y \le 3 \\ 0 & \text{otherwise} \end{cases}$$

(b) (1 point) What is the covariance of X and Y?

Since X and Y are independent, the covariance is 0.

5. (4 points) Let X be the number of heads in 10 coin tosses, and Y be the probability that the coin lands heads. Let:

- $X \mid Y \sim Binomial(10, y)$
- $Y \sim Continuos\ Uniform(0.25, 0.65)$

Calculate the following:

(a) (1 point) $E[X]$

$$E[X] = E[E[X \mid Y]] = E[10Y] = 10\left(\frac{0.25 + 0.65}{2}\right) = 4.5$$

(b) (3 points) $Var(X)$

We have:

$$Var(Y) = \frac{(0.65 - 0.25)^2}{12} = \frac{1}{75}$$

$$E[Y] = 0.45$$

$$E[Y^2] = Var(Y) + E[Y]^2 = \frac{1}{75} + 0.45^2 = \frac{259}{1200}$$

Then:

$$\begin{aligned}
Var(X) &= E[Var(X \mid Y)] + Var(E[X \mid Y]) \\
&= E[10Y(1 - Y)] + Var(10Y) \\
&= 10E[Y - Y^2] + 100Var(Y) \\
&= 10\left(E[Y] - E[Y^2]\right) + 100Var(Y) \\
&= 10\left(0.45 - \frac{259}{1200}\right) + 100\left(\frac{1}{75}\right) = \frac{147}{40} = 3.675
\end{aligned}$$

4. The probability of a royal flush in a poker hand is $p = 1/649740$. Below, let X denote the number of royal flushes in n hands.

Hint: In both questions you can write your answer in terms of p – you need not plug in the actual value; but you may if you prefer.

4a. [2pt] Find the probability of no royal flash in $n = 100$ hands. That is, find $p(X = 0)$.

> **Solution**: *$X \sim Bin(n, p)$, with large n and small p. Use the Poisson approx with a $Poi(\lambda)$ distribution, using $\lambda = np$. Thus $p(X = 0) = p(0) \approx e{-}\lambda\frac{\lambda^0}{0!} = e^{-np} = 0.9998$.*

4b. [2pt] How large must n be to render the probability of having no royal flush in n hands smaller than $1/e$? That is, how large must n be for $p(X = 0) < 1/e$?

> **Solution**: *$p(X = 0) \approx e^{-np} < e^{-1} \Rightarrow n > 1/p = 649740$.*

5. In $n = 10,000$ independent tosses of a coin, the coin landed on heads 5126 times. Let X denote the number of heads in n tosses, and let p denote the probability of head.

5a. [2pt] Assuming that the coin is a fair coin, that is $p = 0.5$, find the expected value $\mu = E(X)$ and standard deviation $\sigma = \text{SD}(X)$ of X

> **Solution**: *X is a binomial r.v., $X \sim Bin(n, p)$. Therefore $E(X) = np = 5,000$ and $Var(X) = n\,p(1 - p) = 2500$ and therefore $SD(X) = \sqrt{2500} = 50$.*

5b. [2pt] Find $p(X > 5126)$ if the coin is fair.

> **Solution**: *Note that $np(1-p) > 10$. We can use the normal approximation, using a normal distribution $N(\mu, \sigma)$ with $\mu = np = 5000$ and $\sigma = 50$. We find $p(X > 5126) = p\left(\frac{X-\mu}{\sigma} > \frac{5126.5-5000}{50}\right) \approx p(Z > 2.53) = 0.0057$ for a standard normal r.v. $Z \sim N(0, 1)$. Note the continuity correction in 5126.5.*

5c.[1pt] Is it reasonable to assume that the coin is fair? Explain in at most 20 words. No need for more computations.

> **Solution**: *No, if the coin were fair, a number of heads as extreme or more as the observed 5126 heads would be unlikely beyond reasonable doubt. We conclude the coin must be biased.*

4. Let X_i be independent and identically distibuted random variables, with $E(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 = 30^2$.

Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ denote the sample average over n observations.

4a.[2pt] Find $p(\overline{X}_n - \mu < 1)$ for $n = 900$.

Solution*:*

$$p(\overline{X}_n - \mu < 1) = p\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} < \frac{1}{30/30}\right) \approx p(Z < 1) = 0.84$$

4b.[2pt] Find the minimum n so that

$$p(\mu - 1 < \overline{X}_n < \mu + 1) \geq 0.95.$$

That is, the minimum n to estimate μ with precision ± 1 with probability 0.95.

Solution*:*

$$p(\mu - 1 < \overline{X}_n < \mu + 1) = p\left(\frac{-1}{30/\sqrt{n}} < \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} < \frac{1}{30/\sqrt{n}}\right) \approx p\left(\frac{-1}{30/\sqrt{n}} < Z < \frac{1}{30/\sqrt{n}}\right) \geq .95$$

or $p(Z < \frac{1}{30/\sqrt{n}}) \geq .975$. *We find* $p(Z < 1.96) \geq .975$*, and therefore*

$$\frac{1}{30/\sqrt{n}} \geq 1.96 \Rightarrow n \geq (30 \cdot 1.96)^2 = 3457.4.$$

Alternatively, using Chebyshev's inequality is okay (though you get a much larger number).