Class 7 in-class problems, 18.05, Spring 2022

Concept questions

Concept question 1. Independence I

Roll two dice: X = value on first, Y = value on second

$X\backslash Y$	1	2	3	4	5	6	$p(x_i)$
1	1/36	1/36	1/36	1/36	1/36	1/36	1/6
2	1/36	1/36	1/36	1/36	1/36	1/36	1/6
3	1/36	1/36	1/36	1/36	1/36	1/36	1/6
4	1/36	1/36	1/36	1/36	1/36	1/36	1/6
5	1/36	1/36	1/36	1/36	1/36	1/36	1/6
6	1/36	1/36	1/36	1/36	1/36	1/36	1/6
$p(y_j)$	1/6	1/6	1/6	1/6	1/6	1/6	1

Are X and Y independent? *1. Yes* *2. No*

Solution: Yes. Every cell probability is the product of the marginal probabilities.

Concept question 2. Independence II

Roll two dice: X = value on first, T = sum

$X\backslash T$	2	3	4	5	6	7	8	9	10	11	12	$p(x_i)$
1	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0	0	1/6
2	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0	1/6
3	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	1/6
4	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	1/6
5	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	1/6
6	0	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	1/6
$p(y_j)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36	1

Are X and Y independent? *1. Yes* *2. No*

Solution: No. The cells with probability zero are clearly not the product of the marginal probabilities.

Concept question 3. Independence III

Which of the following joint pdfs are the variables independent? (Each of the ranges is a rectangle chosen so that $\int\int f(x,y)\,dx\,dy = 1$.)

(i) $f(x,y) = 4x^2y^3$.

(ii) $f(x,y) = \frac{1}{2}(x^3y + xy^3)$.

(iii) $f(x,y) = 6e^{-3x-2y}$

(a) *i* **(b)** *ii* **(c)** *iii* **(d)** *i, ii*

(e) *i, iii* **(f)** *ii, iii* **(g)** *i, ii, iii* **(h)** *None*

(i) Independent. The variables can be separated: the marginal densities are $f_X(x) = ax^2$ and $f_Y(y) = by^3$ for some constants a and b with $ab = 4$.

(ii) Not independent. X and Y are not independent because there is no way to factor $f(x, y)$ into a product $f_X(x)f_Y(y)$.

(iii) Independent. The variables can be separated: the marginal densities are $f_X(x) = ae^{-3x}$ and $f_Y(y) = be^{-2y}$ for some constants a and b with $ab = 6$.

Board questions

Problem 1. Joint distributions

Suppose X and Y are random variables and

- *(X, Y) takes values in $[0, 1] \times [0, 1]$.*
- *the pdf is $f(x, y) = x + y$.*

(a) *Show $f(x, y)$ is a valid pdf.*

(b) *Visualize the event A = '$X > 0.3$ and $Y > 0.5$'. Find its probability.*

(c) *Find the cdf $F(x, y)$.*

(d) *Use the cdf $F(x, y)$ to find the marginal cdf $F_X(x)$ and $P(X < 0.5)$.*

(e) *Find the marginal pdf $f_X(x)$. Use this to find $P(X < 0.5)$.*

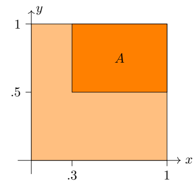
(f) *(New scenario) From the following table compute $F(3.5, 4)$.*

$X\backslash Y$	1	2	3	4	5	6
1	1/36	1/36	1/36	1/36	1/36	1/36
2	1/36	1/36	1/36	1/36	1/36	1/36
3	1/36	1/36	1/36	1/36	1/36	1/36
4	1/36	1/36	1/36	1/36	1/36	1/36
5	1/36	1/36	1/36	1/36	1/36	1/36
6	1/36	1/36	1/36	1/36	1/36	1/36

Solution: **(a)** Validity: Clearly $f(x, y)$ is positive. Next we must show that total probability = 1:

$$\int_0^1 \int_0^1 x + y \, dx \, dy = \int_0^1 \left[\frac{1}{2}x^2 + xy\right]_0^1 dy = \int_0^1 \frac{1}{2} + y \, dy = 1.$$

(b) Here's the visualization



The pdf is not constant so we must compute an integral

$$P(A) = \int_{0.5}^{1} \int_{0.3}^{1} x + y \, dx \, dy = \boxed{0.49}.$$

Make sure you are able to do this integral. Ask if you have any questions.

(c) $F(x,y) = \int_0^y \int_0^x u + v \, du \, dv = \boxed{\frac{x^2 y}{2} + \frac{xy^2}{2}}.$

(d) To find the marginal cdf $F_X(x)$ we simply take y to be the top of the y-range and evalute F: $F_X(x) = F(x,1) = \boxed{\frac{x^2}{2} + \frac{x}{2}.}$ So $\boxed{P(X < 0.5) = 3/8.}$

(e) $f_X(x) = F'_X(x) = x + \frac{1}{2}$.

Or, $f_X(x) = \int_0^1 x + y \, dy = \left[xy + \frac{y^2}{2} \right]_0^1 = \boxed{x + \frac{1}{2}}$. So,

$$P(X < 0.5) = \int_0^{0.5} f_X(x) \, dx = \int_0^{0.5} x + \frac{1}{2} \, dx = \left[\frac{1}{2}x^2 + \frac{1}{2}x \right]_0^{0.5} = \boxed{\frac{3}{8}}.$$

(f) $F(3.5, 4) = P(X \le 3.5, Y \le 4)$.

$X \backslash Y$	1	2	3	4	5	6
1	1/36	1/36	1/36	1/36	1/36	1/36
2	1/36	1/36	1/36	1/36	1/36	1/36
3	1/36	1/36	1/36	1/36	1/36	1/36
4	1/36	1/36	1/36	1/36	1/36	1/36
5	1/36	1/36	1/36	1/36	1/36	1/36
6	1/36	1/36	1/36	1/36	1/36	1/36

Add the probability in the shaded squares: $F(3.5, 4) = 12/36 = 1/3$.

Problem 2. Covariance and correlation

Flip a fair coin 11 times. (The tosses are all independent.)

Let X = number of heads in the first 6 flips

Let Y = number of heads on the last 6 flips.

Compute $\text{Cov}(X, Y)$ and Cor(X, Y).

Solution: Use the properties of covariance.

X_i = the number of heads on the i^{th} flip. (So $X_i \sim \text{Bernoulli}(0.5)$.)

$$X = X_1 + X_2 + ... + X_6 \quad \text{and} \quad Y = X_6 + X_7 + ... + X_{11}.$$

We know $\text{Var}(X_i) = 1/4$. Therefore, using Property 2 (linearity) of covariance

$$\begin{aligned}
\text{Cov}(X, Y) &= \text{Cov}(X_1 + X_2 + ... + X_6,\ X_6 + X_7 + ... + X_{11}) \\
&= \text{Cov}(X_1, X_6) + \text{Cov}(X_1, X_7) + ... + \text{Cov}(X_1, X_{11}) \\
&\quad + \text{Cov}(X_2, X_6) + ... + \text{Cov}(X_2, X_{11}) \\
&\quad + \text{Cov}(X_3, X_6) + ... + \text{Cov}(X_3, X_{11}) \\
&\quad + \text{Cov}(X_4, X_6) + ... + \text{Cov}(X_4, X_{11}) \\
&\quad + \text{Cov}(X_5, X_6) + ... + \text{Cov}(X_5, X_{11}) \\
&\quad + \text{Cov}(X_6, X_6) + ... + \text{Cov}(X_6, X_{11})
\end{aligned}$$

Since the different tosses are independent we know

$$\text{Cov}(X_1, X_6) = 0,\ \text{Cov}(X_1, X_7) = 0,\ \text{Cov}(X_1, X_8) = 0, \text{ etc.}$$

Looking at the expression for $\text{Cov}(X, Y)$ there is only one non-zero term

$$\text{Cov}(X, Y) = \text{Cov}(X_6, X_6) = \text{Var}(X_6) = \boxed{\frac{1}{4}}.$$

For correlation we need σ_X and σ_Y. Since each is the sum of 6 independent Bernoulli(0.5) variables we have $\text{Var}(X) = \text{Var}(Y) = 6/4$. So, $\sigma_X = \sigma_Y = \sqrt{3/2}$.

Thus $\text{Cor}(X, Y) = \dfrac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \dfrac{1/4}{3/2} = 1/6$.

Problem 3. Even more tosses

Toss a fair coin $2n + 1$ times. Let X be the number of heads on the first $n + 1$ tosses and Y the number on the last $n + 1$ tosses.

Compute $\text{Cov}(X, Y)$ and Cor(X, Y).

Solution: As usual let X_i = the number of heads on the i^{th} flip, i.e. 0 or 1. Then

$$X = \sum_{1}^{n+1} X_i, \qquad Y = \sum_{n+1}^{2n+1} X_i$$

X is the sum of $n + 1$ independent Bernoulli(1/2) random variables, so

$$\mu_X = E[X] = \frac{n+1}{2}, \quad \text{and} \quad \text{Var}(X) = \frac{n+1}{4}.$$

Likewise, $\mu_Y = E[Y] = \dfrac{n+1}{2}$, and $\text{Var}(Y) = \dfrac{n+1}{4}$.

Now,

$$\text{Cov}(X, Y) = \text{Cov}\left(\sum_{1}^{n+1} X_i \sum_{n+1}^{2n+1} X_j\right) = \sum_{i=1}^{n+1} \sum_{j=n+1}^{2n+1} \text{Cov}(X_i X_j).$$

Because the X_i are independent the only non-zero term in the above sum is $\text{Cov}(X_{n+1}X_{n+1}) = \text{Var}(X_{n+1}) = \frac{1}{4}$
Therefore,

$$\text{Cov}(X, Y) = \frac{1}{4}.$$

We get the correlation by dividing by the standard deviations.

$$\text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{1/4}{(n+1)/4} = \frac{1}{n+1}.$$

This makes sense: as n increases the correlation should decrease since the contribution of the one flip they have in common becomes less important.

Extra

Discussion: Real-life correlations

- *Over time, amount of ice cream consumption is correlated with number of pool drownings.*
- *In 1685 (and today) being a student is the most dangerous profession. That is, the average age of those who die is less than any other profession.*
- *In 90% of bar fights ending in a death the person who started the fight died.*
- *Hormone replacement therapy (HRT) is correlated with a lower rate of coronary heart disease (CHD).*

Discussion

- Ice cream does not cause drownings. Both are correlated with summer weather.
- In a study in 1685 of the ages and professions of deceased men, it was found that the profession with the lowest average age of death was "student." But, being a student does not cause you to die at an early age. Being a student means you *are* young. This is what makes the average of those that die so low.
- A study of fights in bars in which someone was killed found that, in 90% of the cases, the person who started the fight was the one who died.

  Of course, it's the person who survived telling the story.
- In a widely studied example, numerous epidemiological studies showed that women who were taking combined hormone replacement therapy (HRT) also had a lower-than-average incidence of coronary heart disease (CHD), leading doctors to propose that HRT was protective against CHD. But randomized controlled trials showed that HRT caused a small but statistically significant increase in risk of CHD. Re-analysis of the data from the epidemiological studies showed that women undertaking HRT were more likely to be from higher socio-economic groups (ABC1), with better-than-average diet and exercise regimens. The use of HRT and decreased incidence of coronary heart disease were coincident effects of a common cause (i.e. the benefits associated with a higher socioeconomic status), rather than cause and effect, as had been supposed.

Edward Tufte: "Empirically observed covariation is a necessary but not sufficient condition for causality."

Extra problem 1: Hospitals, binomial, CLT etc.
Here's one more problem. We won't do this in class.

- *A certain town is served by two hospitals.*
- *Larger hospital: about 45 babies born each day.*
- *Smaller hospital about 15 babies born each day.*
- *For a period of 1 year, each hospital recorded the days on which more than 60% of the babies born were boys.*

(a) *Which hospital do you think recorded more such days?*

(i) The larger hospital. *(ii) The smaller hospital.*

(iii) About the same (that is, within 5% of each other).

(b) *Assume exactly 45 and 15 babies are born at the hospitals each day. Let L_i (resp., S_i) be the Bernoulli random variable which takes the value 1 if more than 60% of the babies born in the larger (resp., smaller) hospital on the i^{th} day were boys. Determine the distribution of L_i and of S_i.*

(c) *Let L (resp., S) be the number of days on which more than 60% of the babies born in the larger (resp., smaller) hospital were boys. What type of distribution do L and S have? Compute the expected value and variance in each case.*

(d) *Via the CLT, approximate the 0.84 quantile of L (resp., S). Would you like to revise your answer to part (a)?*

(e) *What is the correlation of L and S? What is the joint pmf of L and S? Visualize the region corresponding to the event $L > S$. Express $P(L > S)$ as a double sum.*

Solution: (a) When this question was asked in a study, the number of undergraduates who chose each option was 21, 21, and 55, respectively. This shows a lack of intuition for the relevance of sample size on deviation from the true mean (i.e., variance).

(b) The random variable X_L, giving the number of boys born in the larger hospital on day i, is governed by a $\text{Bin}(45, 0.5)$ distribution. So L_i has a $\text{Ber}(p_L)$ distribution with

$$p_L = P(X_{:} > 27) = \sum_{k=28}^{45} \binom{45}{k} 0.5^{45} \approx 0.068.$$

Similarly, the random variable X_S, giving the number of boys born in the smaller hospital on day i, is governed by a $\text{Bin}(15, 0.5)$ distribution. So S_i has a $\text{Ber}(p_S)$ distribution with

$$p_S = P(X_S > 9) = \sum_{k=10}^{15} \binom{15}{k} 0.5^{15} \approx 0.151.$$

We see that p_S is indeed greater than p_L, consistent with (ii).

(c) Note that $L = \sum_{i=1}^{365} L_i$ and $S = \sum_{i=1}^{365} S_i$. So L has a Bin$(365, p_L)$ distribution and S has a Bin$(365, p_S)$ distribution. Thus

$$\begin{aligned} E[L] &= 365p_L \approx 25 \\ E[S] &= 365p_S \approx 55 \\ \text{Var}(L) &= 365p_L(1-p_L) \approx 23 \\ \text{Var}(S) &= 365p_S(1-p_S) \approx 47 \end{aligned}$$

(d) By the CLT, the 0.84 quantile is approximately the mean + one sd in each case:

For L, $q_{0.84} \approx 25 + \sqrt{23}$.

For S, $q_{0.84} \approx 55 + \sqrt{47}$.

(e) Since L and S are independent, their correlation is 0 and theirjoint distribution is determined by multiplying their individual distributions. Both L and S are binomial with $n = 365$ and p_L and p_S computed above. Thus

$$P(L = i \text{ and } S = j) = p(i,j) = \binom{365}{i} p_L^i (1-p_L)^{365-i} \binom{365}{j} p_S^j (1-p_S)^{365-j}$$

Thus

$$P(L > S) = \sum_{i=0}^{364} \sum_{j=i+1}^{365} p(i,j) \approx 0.0000916$$

We used the R code below to do the computations.

```
pL = 1 - pbinom(0.6*45, 45, 0.5)
pS = 1 - pbinom(0.6*15, 15, 0.5)
print(pL)
print(pS)

pLGreaterS = 0
for(i in 0:365) {
  for(j in 0:(i-1)) {
    pLGreaterS = pLGreaterS + dbinom(i,365,pL)*dbinom(j,365,pS)
   }
}
print(pLGreaterS)
```

Extra problem 2: Correlation

(a) Flip a coin 3 times. Use a joint pmf table to compute the covariance and correlation between the number of heads on the first 2 and the number of heads on the last 2 flips.

(b) Flip a coin 5 times. Use properties of covariance to compute the covariance and correlation between the number of heads on the first 3 and last 3 flips.

Solution: **(a)** Let X = the number of heads on the first 2 flips and Y the number in the last 2. Considering all 8 possibe tosses: HHH, HHT etc we get the following joint pmf for X and Y

Y/X	0	1	2	
0	1/8	1/8	0	1/4
1	1/8	1/4	1/8	1/2
2	0	1/8	1/8	1/4
	1/4	1/2	1/4	1

Using the table we find

$$E[XY] = \frac{1}{4} + 2\frac{1}{8} + 2\frac{1}{8} + 4\frac{1}{8} = \frac{5}{4}.$$

We know $E[X] = 1 = E[Y]$ so

$$\text{Cov}(X,Y) = E[XY] - E[X]E[Y] = \frac{5}{4} - 1 = \frac{1}{4}.$$

Since X is the sum of 2 independent Bernoulli(0.5) we have $\sigma_X = \sqrt{2/4}$

$$\text{Cor}(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{1/4}{(2)/4} = \frac{1}{2}.$$

(b) As usual let X_i = the number of heads on the i^{th} flip, i.e. 0 or 1.

Let $X = X_1 + X_2 + X_3$ the sum of the first 3 flips and $Y = X_3 + X_4 + X_5$ the sum of the last 3. Using the algebraic properties of covariance we have

$$\begin{aligned}
\text{Cov}(X,Y) &= \text{Cov}(X_1 + X_2 + X_3, X_3 + X_4 + X_5) \\
&= \text{Cov}(X_1, X_3) + \text{Cov}(X_1, X_4) + \text{Cov}(X_1, X_5) \\
&+ \text{Cov}(X_2, X_3) + \text{Cov}(X_2, X_4) + \text{Cov}(X_2, X_5) \\
&+ \text{Cov}(X_3, X_3) + \text{Cov}(X_3, X_4) + \text{Cov}(X_3, X_5)
\end{aligned}$$

Because the X_i are independent the only non-zero term in the above sum is $\text{Cov}(X_3 X_3) = \text{Var}(X_3) = \dfrac{1}{4}$.
Therefore, $\text{Cov}(X,Y) = \frac{1}{4}$.

We get the correlation by dividing by the standard deviations. Since X is the sum of 3 independent Bernoulli(0.5) we have $\sigma_X = \sqrt{3/4}$

$$\text{Cor}(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{1/4}{(3)/4} = \frac{1}{3}.$$