Joint distributions: common distributions

Chrysafis Vogiatzis

Lecture 13

Learning objectives

After these lectures, we will be able to:

- Find the pmf/pdf of a function of a random variable.
- Recognize multinomial distributions.
- Calculate probabilities (including marginal and conditional ones) for multinomially distributed random variables.
- Recognize bivariate normal distributions.
- Describe and explain bivariate normal distributions and their correlations.
- Calculate probabilities (including marginal and conditional ones) for bivariate normally distributed random variables.

Motivation: Success or failure? More like full success, or somewhat success, or ...

In Lectures 5-6, we introduced a lot of discrete distributions. One of the most fundamental ones is the binomial distribution. Its premise is simple: perform an experiment n times and count the number of successes, assuming the remainders are failures. This works pretty well when we have two outcomes: for example, a patient may have an infection or not, a student may pass a class or not, etc.

What happens when the number of outcomes is higher than 2? What if a patient may have a severe infection, or a moderate infection, or no infection? What if a student can get an A, a B, a C, a D, or fail a class?

Motivation: Normally distributed random variables with correlation

We sometimes are aware that a specific random variable is normally distributed. However, its exact parameters may depend (and in turn may also affect) another normally distributed random variable. For example, consider a Sunday night at HBO. A TV series starting at 10pm may expect a normally distributed share of viewers with a known mean and standard deviation. However, if the TV series showing at 9pm has its grand finale, we may anticipate a higher viewership for the 10pm show, too! This relationship needs to be modeled somehow...

Distribution of a function

We have already discussed what we expect will happen for a function of a random variable. [1] We repeat the definitions here for convenience:

Definition 1 (Expectation of a function of a random variable) *Let $g(X)$ be a function of a random variable X. Then, the expectation of $g(X)$ is denoted by $E[g(X)]$ and is equal to:*

- *for discrete random variable X with sample space S:*

$$E[g(X)] = \sum_{x \in S} g(x) \cdot p(x).$$

- *for continuous random variable X:*

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x) \cdot f(x) dx.$$

It is time we discuss how the function is **distributed**: rather than addressing questions of expectation ("what should I expect the function value to be?"), we will be addressing questions of probability ("what is the probability the function value is...").

Some examples of why this would be useful:

- What is the probability my profits are higher than $2000 today?
  - My profits depend on the number of customers, discrete random variable X.
- What is the probability the circuit overheats?
  - The heat of the circuit is a function of its current, continuous random variable X.
- What is the probability the crop has high yield?
  - The yield of a crop is a function of the location temperature, continuous random variable X.

Formally, let $Y = h(X)$ be a **one-to-one** transformation of a random variable X to a random variable Y. The one-to-one transformation is important: it implies that solving $y = h(x)$ provides us with a unique solution. Assume that the solution is [2]

[1] Recall Lecture 9 and the properties of expectation section.

[2] Recall the definition of inverse functions.

$$x = h^{-1}(y) = u(y).$$

Examples of inverses

- $Y = X^2 \implies x = u(y) = \sqrt{y}$.
- $Y = 2\ln x \implies x = e^{y/2}$.

Definition 2 (Distribution of a function) *Let $Y = h(X)$ be a one-to-one function of random variable X to Y. X is distributed with pmf/pdf $f_X(x)$. Then, the pmf/pdf of random variable $Y = h(X)$ can be found using the* ***chain rule****:*

1. *Discrete X:* $f_Y(y) = f_X(u(y))$.
2. *Continuous X:* $f_Y(y) = f_X(u(y)) \cdot |u'(y)|$,
   where $u'(y)$ is the derivative of function $u(y)$.

Printer speed

A printer has speed that is equal to $h(x) = \frac{\sqrt{x+1}}{x+1}$, where x is the condition of the printer. The condition of the printer is a continuous random variable distributed exponentially with rate $\lambda = 1$. What is the probability the printer is faster than 0.5?

First of all, let's see what we have:

- X is the condition of the printer, a random variable with $f_X(x) = \lambda \cdot e^{-\lambda x}$ and $x \geq 0$.
- Y is the speed of the printer, a random variable which is a function of X and has $Y = h(x) = \frac{\sqrt{x+1}}{x+1}$. By definition $0 \leq y \leq 1$.
- We may solve for $u(y)$:

$$y = \frac{\sqrt{x+1}}{x+1} \implies x = 1 - \frac{1}{y^2} \implies u(y) = \frac{y^2-1}{y^2}.$$

Based on the chain rule: $f_Y(y) = e^{-\frac{y^2-1}{y^2}} \cdot \frac{2}{y^3}$. Finally, we have:

$$P(Y > 0.5) = \int\limits_{0.5}^{1} e^{-\frac{y^2-1}{y^2}} \cdot \frac{2}{y^3} dy = 0.9502.$$

The multinomial distribution

Flashback! Let's review together the **binomial distribution**, one of the first discrete probability distributions we studied together. We had the following setup.

What if we perform n independent trials of the same experiment? Each trial may result in a success (with probability p) or a failure (with probability $q = 1 - p$). Let X be the number of successes we observe: then X is said to be binomially distributed with parameters n and p. Some interesting things about the binomial distribution:

- pmf: $P(X = x) = p(x) = \binom{n}{x} p^x (1-p)^{n-x}$, for $0 \leq x \leq n$.
- expectation: $E[X] = np$.
- variance: $Var[X] = np(1-p)$.

What if we tried to generalize this? We will still perform n independent trials; however now each trial will result in one of k outcomes (instead of just two). Each outcome appears with each own probability p_i. Clearly we need $\sum_{i=1}^{k} p_i = 1$. This is called the **multinomial distribution**. Formally:

Definition 3 (The multinomial distribution) *Let X_i be the number of times that outcome i appears in n independent trials. Each outcome i appears with probability p_i such that $\sum_{i=1}^{k} p_i = 1$. Then, $(X_1, X_2, \ldots, X_k)$ is distributed following a multinomial distribution with parameters n and $p_i, i = 1, \ldots, k$. The joint probability mass function is given by:*

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3, \ldots, X_k = x_k) = \frac{n!}{x_1! x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}.$$

Note that $\sum_{i=1}^{k} x_i = n$.

Grades

According to the grade disparity website at UIUC, a student registering for CS 101 gets an A 73% of the time, a B 17% of the time, a C 6% of the time, a D 2% of the time, and an F the remaining 2% of the time. In a section of the class, there are 20 students. What is the probability that:

a) 10 students get an A and 10 students get a B?

b) everyone gets an A?

c) 12 students get an A, 5 students get a B, 2 students get a C, and 1 student gets a D?

Grades

This is a multinomial distribution with $n = 20$, $p_1 = 0.73$, $p_2 = 0.17$, $p_3 = 0.06$, $p_4 = 0.02$, $p_5 = 0.02$ for A, B, C, D, and F, respectively. Let X_1, X_2, X_3, X_4, X_5 be the number of students getting an A, B, C, D, F. Then, we have:

a) 10 students get an A and 10 students get a B?

$$\begin{aligned} P(X_1 = 10, X_2 = 10, X_3 = 0, X_4 = 0, X_5 = 0) = \\ = \frac{20!}{10!10!0!0!0!} 0.73^{10} 0.17^{10} 0.06^{0} 0.02^{0} 0.02^{0} = \\ = 0.00016. \end{aligned}$$

b) everyone gets an A?

$$\begin{aligned} P(X_1 = 20, X_2 = 0, X_3 = 0, X_4 = 0, X_5 = 0) = \\ = \frac{20!}{20!0!0!0!0!} 0.73^{20} 0.17^{0} 0.06^{0} 0.02^{0} 0.02^{0} = \\ = 0.00185. \end{aligned}$$

c) 12 students get an A, 5 students get a B, 2 students get a C, and 1 student gets a D?

$$\begin{aligned} P(X_1 = 12, X_2 = 5, X_3 = 2, X_4 = 1, X_5 = 0) = \\ = \frac{20!}{12!5!2!1!0!} 0.73^{12} 0.17^{5} 0.06^{2} 0.02^{1} 0.02^{0} = \\ = 0.00495. \end{aligned}$$

The marginal distribution

Let us derive the marginal distribution of $f_{X_1 X_2 \cdots X_k}(x_1, x_2, \ldots, x_k) = P(X_1 = x_1, X_2 = x_2, X_3 = x_3, \ldots, X_k = x_k)$ for the multinomial distribution.

First of all, a little notation. Like we did in Lectures 11 and 12, the marginal distribution of, say, X_i can be written as $f_{X_i}(x_i)$. Remember that X_i is the random variable capturing the number of outcomes i we see in n tries.

Now, let outcome i be a "success"; otherwise, a "failure". Does this ring a bell? Also, remember that outcome i happens with probability p_i; everything else with probability $1 - p_i$. Hence, X_i is the number of successes we see in n independent tries. What is X_i distributed like when we view it like this?

The marginal distribution of X_i is the binomial distribution: i.e., every single one of the X_i is **binomially distributed** with parameters n, p_i.

Grades

For the example discussed earlier, what is the probability that:

a) 10 students get an A?

b) at most 1 students fails?

Also, how many students are expected to get each grade?

The first one is binomially distributed with $n = 20, p = 0.73$. The second one is binomially distributed with $n = 20, p = 0.02$. Overall, we have:

a) $P(X_1 = 10) = \binom{20}{10} 0.73^{10} 0.27^{10} = 0.01635$.

b) $P(X_5 \leq 1) = \binom{20}{0} 0.02^0 0.98^{20} + \binom{20}{1} 0.02^1 0.98^{19} = 0.6676 + 0.2725 = 0.9401$.

To answer the expectation question, if each outcome is binomially distributed with $n = 20$ and p_i, the expectations are:

a) A: 14.6 b) B: 3.4 c) C: 1.2 d) D: 0.4 e) F: 0.4

The conditional distribution

Now, let us consider the conditional distribution of the multinomial distribution. Say that among all outcomes we already know that X_j has happened x_j times. This implies that there is no uncertainty about x_j of the n tries. Let's keep that in mind.

Furthermore, if we were to remove these x_j outcomes, what we are left with is $n - x_j$ tries; however these tries do not have all k outcomes happening, but instead only $k - 1$.

Let us consider this with an example: if we have 20 students taking a class, and we know that 15 students ended up with an A, then the stochastic nature of this distribution only affects the remaining 5 students – after we remove the students whose grade is known to be an A.

Finally, in the remaining outcomes, we know that x_j is missing. However, we also know that if we sum the probabilities of all outcomes we should be getting 1. In the case of the grades from before, after we remove the As, we get a summation of probabilities equal to $p_2 + p_3 + p_4 + p_5 = 0.27 \neq 1$. To fix this issue, we renormalize the remainder of the probabilities. Instead of p_i, we now use $q_i = \frac{p_i}{\sum\limits_{\ell \neq j} p_\ell}$.

Summing up:

The conditional distribution of $X_1, X_2, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k$ given $X_j = x_j$ is the multinomial distribution again but with parameters $n - x_j$, $q_i = \frac{p_i}{\sum\limits_{\ell \neq j} p_\ell}$.

Grades

Back to the example we have been using. We have just been informed that 3 students failed. What is the probability that:

a) 10 students get an A and 7 students get a B?

b) 10 students get an A, 5 students get a B, and 2 students get a C?

Both are multinomial distributions with parameters $n = 20 - 3 = 17$ and $q_1 = \frac{p_1}{p_1+p_2+p_3+p_4} = \frac{0.73}{0.98} = 0.7449$; $q_2 = \frac{p_2}{p_1+p_2+p_3+p_4} = \frac{0.17}{0.98} = 0.1735$; $q_3 = 0.0612$; and $q_4 = 0.0204$. Then, we have:

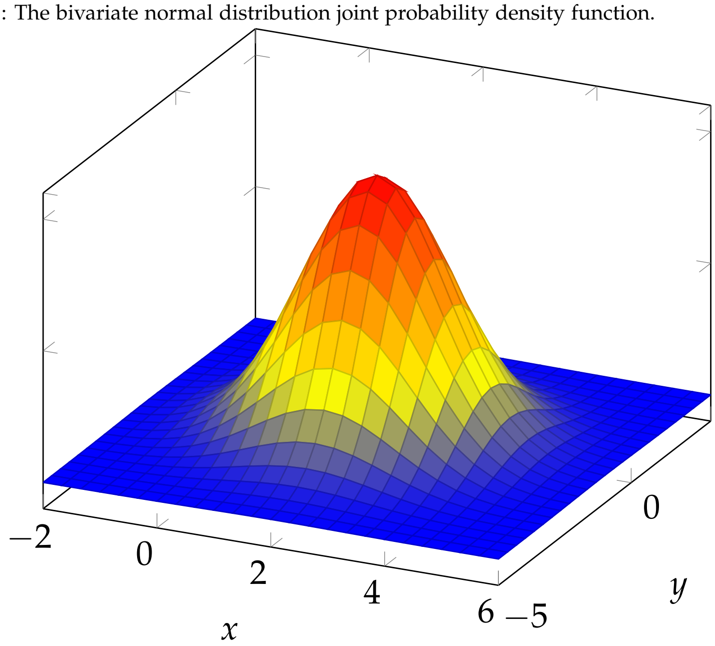
a) $P(X_1 = 10, X_2 = 7, X_3 = 0, X_4 = 0) = \frac{17!}{10!7!0!0!} 0.7449^{10} 0.1735^{7} 0.0612^{0} 0.0204^{0} = 0.0048$.

b) $P(X_1 = 10, X_2 = 5, X_3 = 2, X_4 = 0) = \frac{17!}{10!5!2!0!} 0.7449^{10} 0.1735^{5} 0.0612^{2} 0.0204^{0} = 0.01265$.

The bivariate normal distribution

Similarly to what we did for the binomial and its extension to the multinomial, we will also extend the normal distribution. What if, we have two jointly distributed variables that are *individually* normally distributed with their own means and variances? In essence, what if

Figure 1: The bivariate normal distribution joint probability density function.



we have the three-dimensional pdf portrayed in Figure ???

Formally, we provide the definition that follows:

Definition 4 (Bivariate normal distribution) *Consider two normally distributed random variables X, Y with means μ_X, μ_Y and variances σ_X^2, σ_Y^2. That is, $X \sim \mathcal{N}\left(\mu_X, \sigma_X^2\right)$ and $Y \sim \mathcal{N}\left(\mu_Y, \sigma_Y^2\right)$. We also assume that the two random variables are correlated with correlation ρ_{XY}.*

*Then, two random variables X and Y with the above parameters are **jointly distributed with a bivariate random distribution** if:*

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{XY}^2}} \cdot e^{\frac{-z}{2\left(1-\rho_{XY}^2\right)}},$$

where $z = \frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho_{XY}(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}$

We observe that ρ_{XY} plays an important role. When $\rho_{XY} = 0$, we have the simplified version of the bivariate random distribution as:

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y} \cdot e^{-\left(\frac{(x-\mu_X)^2}{2\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right)}.$$

However, we know that $\rho_{XY} = 0$ implies that X and Y are independent. And, for two independent random variables, we know that their joint pdf is equal to the product of the individual pdfs. Let's see if that is the case here:

Figure 2: The bivariate normal distribution and its contour plot. Here, we have that $\rho_{XY} = 0$.

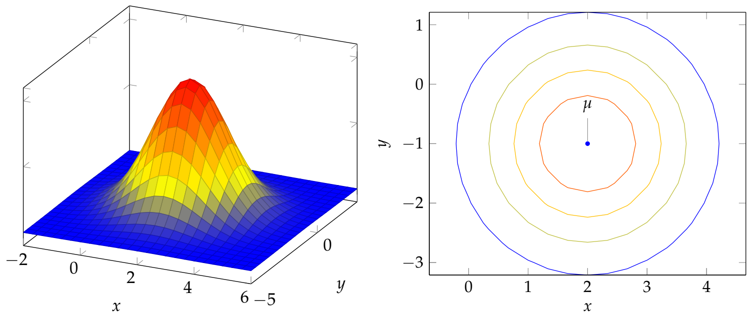
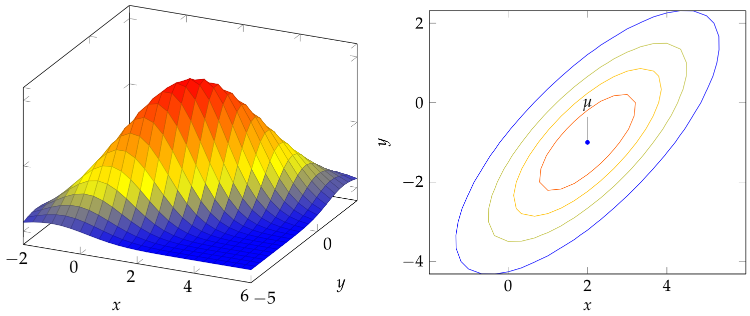


Figure 3: The bivariate normal distribution and its contour plot. Here, we have that $\rho_{XY} > 0$.



$$f_X(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma_X} e^{-\frac{(x-\mu_X)^2}{2\sigma_X^2}}$$

$$f_Y(y) = \frac{1}{\sqrt{2\pi} \cdot \sigma_Y} e^{-\frac{(y-\mu_Y)^2}{2\sigma_Y^2}}$$

$$f_X(x) \cdot f_Y(y) = \frac{1}{\sqrt{2\pi} \cdot \sigma_X} e^{-\frac{(x-\mu_X)^2}{2\sigma_X^2}} \cdot \frac{1}{\sqrt{2\pi} \cdot \sigma_Y} e^{-\frac{(y-\mu_Y)^2}{2\sigma_Y^2}} =$$

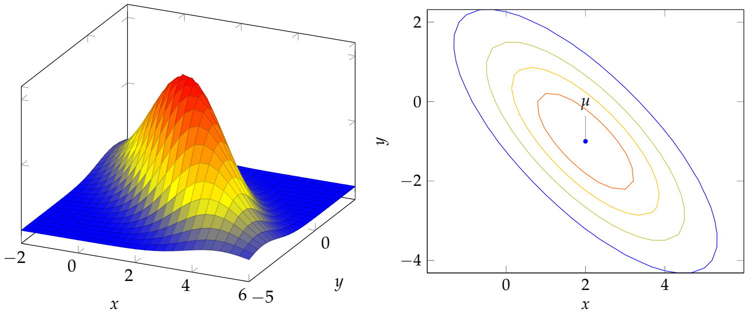
$$= \frac{1}{2\pi\sigma_X\sigma_Y} \cdot e^{-\left(\frac{(x-\mu_X)^2}{2\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right)} = f_{XY}(x, y).$$

This independence is shown in Figure **??**. What happens when $\rho > 0$ or $\rho < 0$? What happens when $\rho = 1$ or $\rho = -1$?

When we have positive correlation, this implies that higher/lower values of X will imply higher/lower values of Y and vice-versa (for Y and X). This is showcased with the pdf and the contour in Figure **??**, which appears to be "positively" skewed.

On the other hand, if $\rho_{XY} < 0$, this means that higher/lower values of X will lead to lower/higher values of Y and vice-versa (for Y and X). This is exactly the opposite. This is again shown visually with the pdf and the contour in Figure **??**, which appears to be "negatively" skewed.

Figure 4: The bivariate normal distribution and its contour plot. Here, we have that $\rho_{XY} = 0$.



What happens when $\rho = 1$ or $\rho = -1$? Let's leave this as food for thought.

The marginal and the conditional distribution

Much like what we did for the multinomial distribution, we may also derive the marginal and conditional distributions for the bivariate normal distribution. More specifically, both the *marginal* and the *conditional* distributions for the bivariate normal distribution are normal distributions themselves!

$$\begin{aligned}
\text{Marginal pdf:}\quad & X \sim \mathcal{N}\left(\mu_X, \sigma_X^2\right) \\
& Y \sim \mathcal{N}\left(\mu_Y, \sigma_Y^2\right) \\
\text{Conditional pdf:}\quad & X|Y = y \sim \mathcal{N}\left(\mu_{X|Y=y}, \sigma^2_{X|Y=y}\right) \\
& \mu_{X|Y=y} = \mu_X + \rho_{XY}\left(\frac{\sigma_X}{\sigma_Y}\right)(y - \mu_Y) \\
& \sigma^2_{X|Y=y} = \sigma_X^2\left(1 - \rho_{XY}^2\right) \\
& Y|X = x \sim \mathcal{N}\left(\mu_{Y|X=x}, \sigma^2_{Y|X=x}\right) \\
& \mu_{Y|X=x} = \mu_Y + \rho_{XY}\left(\frac{\sigma_Y}{\sigma_X}\right)(x - \mu_X) \\
& \sigma^2_{Y|X=x} = \sigma_Y^2\left(1 - \rho_{XY}^2\right)
\end{aligned}$$

For the conditional pdf, the question from earlier comes back. What if X and Y are independent? What if they are perfectly correlated ($\rho_{XY} = 1$ or $\rho_{XY} = -1$)?

We finish this lecture with a big, comprehensive example that combines information from Lectures 12 and 13, as well as Lecture 7. Pay close attention to the derivations and calculations that follow!

Bivariate normal distribution example

A class has two exams, both of which have grades that are normally distributed with $\mu_1 = 80, \mu_2 = 82.5$ and $\sigma_1^2 = 100, \sigma_2^2 = 225$. Finally the two exams are positively correlated with $\rho = 0.6$. What is the probability that:

a) a random student scores over 75 in Exam 2?

b) a random student scores over 75 in Exam 2 given that they scored an 85 in the first exam?

c) the sum of the two exams of a random student is less than or equal to 175?

d) a random student did better on the second exam than the first exam?

Let's get to it. Let X_1 be the grade of the first exam, and X_2 the grade of the second exam. Then:

a) We know that $X_2 \sim \mathcal{N}(82.5, 225)$. Hence:

- $z = \frac{75-82.5}{15} = -0.5$.
- $P(X_2 > 75) = 1 - P(X_2 \leq 75) = 1 - \Phi(z) = \Phi(-z) = \Phi(0.5) = 0.6915$.

b) We also know that $X_2|X_1 \sim \mathcal{N}\left(\mu_{X_2|X_1}, \sigma^2_{X_2|X_1}\right)$. We calculate:

- $\mu_{X_2|X_1=85} = \mu_{X_2} + \rho_{X_1X_2}\left(\frac{\sigma_{X_2}}{\sigma_{X_1}}\right)(85 - \mu_{X_1}) = 82.5 + 0.6 \cdot \frac{15}{10} \cdot 5 = 87$.
- $\sigma^2_{X_2|X_1=85} = \sigma^2_{X_2}\left(1 - \rho^2_{X_1X_2}\right) = 225 \cdot 0.64 = 144$.
- $z = \frac{75-87}{12} = -1$.
- $P(X_2 > 75|X_1 = 85) = 1 - P(X_2 \leq 75|X_1 = 85) = 1 - \Phi(z) = \Phi(-z) = \Phi(1) = 0.8413$.

Hence, knowing that the student did better than average in the first exam changes our perspective for their probability to do well in the second exam, too.

Bivariate normal distribution example

c) The sum of two normally distributed random variables is also normally distributed! Additionally, we have:

$$\begin{aligned} E\left[\sum_{i=1}^{n} a_i X_i\right] &= \sum_{i=1}^{n} a_i E\left[X_i\right] \\ Var\left[\sum_{i=1}^{n} a_i X_i\right] &= \sum_{i=1}^{n} a_i^2 Var\left[X_i\right] + \sum_{i=1}^{n} \sum_{j=1:i\neq j}^{n} Cov\left[X_i, X_j\right] \\ &= \sum_{i=1}^{n} a_i^2 Var\left[X_i\right] + 2 \cdot \sum_{i=1}^{n} \sum_{i<j} Cov\left[X_i, X_j\right] \end{aligned}$$

In our case, we have two random variables, so:

$$\begin{aligned} E\left[X_1 + X_2\right] &= E\left[X_1\right] + E\left[X_2\right] = 162.5 \\ Var\left[X_1 + X_2\right] &= Var\left[X_1\right] + Var\left[X_2\right] + 2Cov\left[X_1, X_2\right] = \\ &= 325 + 2\sigma^2_{X_1 X_2}. \end{aligned}$$

To calculate $\sigma^2_{X_1 X_2}$ we use the definition of correlation (see Lecture 12):

$$\rho_{X_1 X_2} = \frac{\sigma^2_{X_1 X_2}}{\sigma_{X_1}\sigma_{X_2}} \implies 0.6 = \frac{\sigma^2_{X_1 X_2}}{10 \cdot 15} = \sigma^2_{X_1 X_2} = 90.$$

This leads to a final variance of $Var\left[X_1 + X_2\right] = 505$. Finally:

- $X_1 + X_2 \sim \mathcal{N}\left(162.5, 505\right)$.
- $z = \frac{175-162.5}{\sqrt{505}} = 0.56$.
- $P(X_1 + X_2 \leq 175) = \Phi(z) = \Phi(0.56) = 0.7123$.

d) For this question, we want $X_2 > X_1 \implies X_2 - X_1 > 0$. The difference of two normally distributed random variables is—again—normally distributed! Its details:

$$\begin{aligned} E\left[X_2 - X_1\right] &= E\left[X_2\right] - E\left[X_1\right] = 2.5 \\ Var\left[X_2 - X_1\right] &= Var\left[X_1\right] + Var\left[X_2\right] + 2Cov\left[X_1, X_2\right] = 505. \end{aligned}$$

- $X_2 + X_1 \sim \mathcal{N}\left(2.5, 505\right)$.
- $z = \frac{0-2.5}{\sqrt{505}} = -0.11$.
- $P(X_2 > X_1) = P(X_2 - X_1 > 0) = 1 - P(X_2 - X_1 \leq 0) = 1 - \Phi(z) = \Phi(-z) = \Phi(0.11) = 0.5438$.