Dr. Katherine Davies

STATS 3D03

Course Notes

Recorded, adapted and illustrated by Zachary Lucier

McMaster University

Fall 2022

Last updated September 16, 2022

Contents

1 Review

We begin our discussion of mathematical statistics with a review of concepts from previous courses. One of these key concepts is that of probability.

1.1 Probability

Recall that a **sample space** Ω is the set of all possible outcomes of an experiment. Subsets of Ω are called **events** and the collection of all events is denoted by $\mathcal{F}$.

Definition 1.1 (probability set function)**.** Let Ω be a sample space and let $\mathcal{F}$ be the collection of all events. Let $P : \mathcal{F} \to \mathbb{R}$ be a real-valued function. Then P is a **probability set function** (also referred to as **probability measure**, **probability distribution** or simply **probability**) if it satisfies the following three conditions:

1. $0 \leq P(A) \leq 1$, for all $A \in \mathcal{F}$.
2. $P(\Omega) = 1$ and $P(\emptyset)$.
3. If $\{A_n\}$ is a sequence of events in $\mathcal{F}$ and $A_m \cap A_n = \emptyset$ for all $m \neq n$, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

In order to formalize quantities which depend on random events, we reintroduce the concept of a random variable and its support.

Definition 1.2 (random variable)**.** Let Ω be a sample space. A **random variable** is a function from Ω into the real numbers. The **support** (also called **space** or **range**) of X is the set of real numbers $\mathcal{S} = \{x : x = X(\omega), \omega \in \Omega\}$.

In cases where $\mathcal{S}$ is a countable set, we say that X is a **discrete random variable**. The set $\mathcal{S}$ may also be an interval of real numbers, in which case we say that X is a **continuous random variable**.

Given a random variable X, its support $\mathcal{S}$ becomes the sample space of interest. Besides inducing the sample space $\mathcal{S}$, X also induces a probability which we call the **distribution** of X.

The probability distribution of a discrete random variable is described completely in terms of its probability mass function and its support.

Definition 1.3 (pmf)**.** Let X be a discrete random variable with support $\mathcal{S}$. The **probability mass function** (pmf) of X p_X is given by

$$p_X(x) = P(X = x), \text{ for } x \in \mathcal{S}.$$

Similarly, the probability distribution of a continuous random variable is described completely in terms of its probability density function and its support.

Definition 1.4 (pdf)**.** Let X be a continuous random variable with support $\mathcal{S}$. The **probability density function** (pdf) of X is a function f_X that satisfies

$$P(X \leq x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t$$

for all $x \in \mathcal{S}$.

The pmf of a discrete random variable and the pdf of a continuous random variable are quite different entities. The cumulative distribution function, though, uniquely determines the probability distribution of a random variable.

Definition 1.5 (cdf)**.** Let X be a random variable. Then its **cumulative distribution function** (cdf) is defined by $F_X(x)$, where

$$F_X(x) = P(X \leq x).$$

1.2 Expectation

One of the most important measures associated with random variables is that of expectation.

Definition 1.6 (expectation)**.** Let X be a random variable with support $\mathcal{S}$.
If X is a *continuous* random variable with pdf $f(x)$ and

$$\int_{-\infty}^{\infty} |x| f(x)\,\mathrm{d}x$$

is finite, then the **expectation** of X, denoted $E(X)$ is defined as

$$E(X) = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x.$$

If X is a *discrete* random variable with pmf $p(x)$ and

$$\sum_{x\in\mathcal{S}} |x| p(x)$$

is finite, then the **expectation** of X is defined as

$$E(X) = \sum_{x\in\mathcal{S}} x p(x).$$

Sometimes the expectation $E(X)$ is called the **expected value** of X or the **mean** of X. When the mean designation is used, we often denote the expected value by μ.

Theorem 1.1 (Law of the unconscious statistician)**.** Let X be a random variable with support $\mathcal{S}_X$ and let $Y = g(X)$ for some real-valued function g.

(a) Suppose X is discrete with pmf $p_X(x)$. If

$$\sum_{x\in\mathcal{S}_X} |g(x)| p_X(x)$$

is finite, then the expectation of Y exists and is given by

$$E(Y) = \sum_{x\in\mathcal{S}_X} g(x) p_X(x).$$

(b) Suppose X is continuous with pdf $f_X(x)$. If

$$\int_{-\infty}^{\infty} |g(x)| f_X(x)\,\mathrm{d}x$$

is finite, then the expectation of Y exists and is given by

$$E(Y) = \int_{-\infty}^{\infty} g(x) f_X(x)\,\mathrm{d}x.$$

An important application of the above theorem shows that expectation is *linear*. That is, $E(aX+$

$b) = aE(X) + b$. It is a useful exercise to show that this is the case. Furthermore, this property can be generalized for $a_1, \dots, a_k$ real numbers and $g_1, \dots, g_k$ real-valued functions.

$$E(a_1 g_1(X) + \cdots + a_k g_k(X)) = a_1 E(g_1(X)) + \cdots + a_k E(g_k(X))$$

1.3 Moments

Expectation allows us to define a countably infinite number of measures associated with random variables, called moments.

Definition 1.7 (moment)**.** Suppose X is a random variable and m is a positive integer. The m**th moment** of X is defined to be $E(X^m)$, provided this expectation exists.

As such, the first moment of a random variable is simply its **mean** μ. It is often useful to think about moments about the mean $E((X-\mu)^m)$. We call these **central moments**.

The second central moment should be familiar to you as the **variance** σ^2. It has the following equivalent formulation which is computationally useful.

$$\mathrm{Var}(X) = E(X^2) - E(X)^2$$

This can be found using the linearity of expectation.

We call the third central moment the **skewness** and call the fourth central moment the **kurtosis**.

Definition 1.8 (mgf)**.** Let X be a random variable such that for some $h > 0$, the expectation of e^{tX} exists for $-h < t < h$. The **moment generating function** (mgf) of X is defined to be the function $M_X(t) = E(e^{tX})$ for $-h < t < h$.

Clearly, $M_X(0) = 1$ for any random variable. Not every random variable has a mgf. For example, the mgf of the Cauchy Distribution with pdf $f(x) = \frac{1}{\pi(1+x^2)}$ is not defined. It can be shown that if the mgf of a random variable exists, then all of its moments exist.

Theorem 1.2. Let X and Y be random variables with mgfs M_X and M_Y, respectively, existing in open intervals about 0. Then $F_X(z) = F_Y(z)$ for all $z \in \mathbb{R}$ if and only if $M_X(t) = M_Y(t)$ in an open interval about 0.

Theorem 1.3. Let X be a random variable with mgf M_X, and let $a, b \in \mathbb{R}$ be fixed. Then the mgf of $Y = aX + b$ also exists and is given by

$$M_Y(t) = e^{bt} M_X(at).$$

Theorem 1.4. Suppose X and Y are independent random variables with mgfs M_X and M_Y. Let $a, b \in \mathbb{R}$ be fixed and define $Z = aX + bY$. Then the mgf of Z exists in an open interval about 0 and is given by

$$M_Z(t) = M_X(at) M_Y(bt).$$

Theorem 1.5. Suppose X is a random variable with mgf M_X and let $M_X^{(m)}(t) = \frac{\mathrm{d}^m}{\mathrm{d}t^m} M_X(t)$. Then the mth moment of X is given by

$$E(X^m) = M_X^{(m)}(0).$$

The above theorem should make clear why we call mgfs as such. The proof is reliant on the Taylor expansion of e^{tX}. We use the linearity of expectation, which will be stated formally in a later section.

$$\begin{aligned} M_X(t) &= E(e^{tX}) \\ &= E\left(\sum_{n=0}^{\infty} \frac{t^n}{n!} X^n \right) \\ &= \sum_{n=0}^{\infty} \frac{t^n}{n!} E(X^n) \end{aligned}$$

1.4 Distributions

We reintroduce some special distributions, starting with those of the discrete kind.

Definition 1.9 (binomial random variable)**.** Assume a sequence of n Bernouilli trials each with probability of success p and let X be the number of successes. Then X is a

binomial random variable with pmf

$$p_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{for } x = 0, 1, \dots, n \\ 0 & \text{otherwise} \end{cases}$$

We write $X \sim b(n,p)$.

If $X \sim b(n,p)$, X has support $\{0,1,\dots,n\}$, mean $\mu = np$, variance $\sigma^2 = np(1-p)$ and mgf $M_X(t) = (1-p+pe^t)^n$.

Definition 1.10 (negative binomial random variable)**.** Assume a sequence of Bernouilli trials each with probability of success p is performed until the rth success occurs. Let Y be the number of trials required. Then Y is a **negative binomial random variable** with pmf

$$p_Y(y) = \begin{cases} \binom{y+r-1}{y-1} p^r (1-p)^y & \text{for } y = 0, 1, \dots \\ 0 & \text{otherwise} \end{cases}$$

We write $Y \sim nb(r,p)$.

If $Y \sim nb(r,p)$, Y has support $\mathbb{Z}_{\geq 0}$, mean $\mu = \frac{pr}{1-p}$, variance $\sigma^2 = \frac{pr}{(1-p)^2}$ and mgf $M_Y(t) = \left(\frac{1-p}{1-pe^t}\right)$ with $t < -\ln p$.

Taking $r = 1$, we obtain the geometric distribution.

Definition 1.11 (Poisson random variable)**.** A discrete random variable X is a **Poisson random variable** if its pmf has the form

$$p_X(x) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!} & \text{for } x = 0, 1, \dots \\ 0 & \text{otherwise} \end{cases}$$

where $\lambda \in \mathbb{R}_{\geq 0}$. We write $X \sim \text{Pois}(\lambda)$.

If $X \sim \text{Pois}(\lambda)$, X has support $\mathbb{Z}_{\geq 0}$, mean $\mu = \lambda$, variance $\sigma^2 = \lambda$ and mgf $M_X(t) = \exp(\lambda(e^t - 1))$.

We now recall some continuous distributions.

Definition 1.12 (uniform random variable)**.** A continuous random variable X is said to be a **uniform random variable** if it has pdf

$$p_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a,b]) \\ 0 & \text{otherwise} \end{cases}$$

where $a, b \in \mathbb{R}$ are fixed. We write $X \sim U(a,b)$.

If $X \sim U(a,b)$, X has support $[a,b]$, mean $\mu = \frac{a+b}{2}$, variance $\sigma^2 = \frac{(b-a)^2}{12}$ and mgf $M_X(t) = \begin{cases} \frac{e^{tb}-e^{ta}}{t(b-a)} & \text{for } t \neq 0 \\ 1 & \text{for } t = 0 \end{cases}$.

Definition 1.13 (normal random variable)**.** A continuous random variable X is said to be a **normal random variable** with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ if its pdf has the form

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

We write $X \sim N(\mu, \sigma^2)$.

If $X \sim N(\mu, \sigma^2)$, X has support $\mathbb{R}$, mean μ, variance σ^2 and mgf $M_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}$.

The derivation of this mgf is left as an exercise.

Theorem 1.6. Let $X_1, \dots, X_n$ be IID random variables such that $X_i \sim N(\mu_i, \sigma_i^2)$ for each $i = 1, \dots, n$. Let $Y = \sum_{i=1}^n a_i X_i$ for some set of real constants $\{a_1, \dots, a_n\}$. Then Y is also normally distributed with

$$E(Y) = \sum_{i=1}^n a_i \mu_i \quad \text{and} \quad \text{Var}(Y) = \sum_{i=1}^n a_i^2 \sigma_i^2.$$

Definition 1.14 (gamma random variable)**.** A continuous random variable X is said to

be a **gamma random variable** with parameters $\alpha, \beta > 0$ if its pdf has the form

$$f_X(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & \text{for } x \in \mathbb{R}_{>0} \\ 0 & \text{otherwise} \end{cases}.$$

We write $X \sim \Gamma(\alpha, \beta)$.

Taking $\alpha = 1$ yields the exponential distribution.

If $X \sim \Gamma(\alpha, \beta)$, X has support $\mathbb{R}_{>0}$, mean $\mu = \frac{\alpha}{\beta}$, variance $\sigma^2 = \frac{\alpha}{\beta^2}$ and mgf $M_X(t) = \left(1 - \frac{t}{\beta}\right)^{-\alpha}$ for $t < \beta$.

Definition 1.15 (beta random variable)**.** A continuous random variable X is said to be a **beta random variable** with parameters $\alpha, \beta > 0$ if its pdf has the form

$$f_X(x) = \begin{cases} \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)} & \text{for } x \in (0,1) \\ 0 & \text{otherwise} \end{cases}$$

where $\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. We write $X \sim \mathrm{Beta}(\alpha, \beta)$.

If $X \sim \mathrm{Beta}(\alpha, \beta)$, X has support $(0,1)$, mean $\mu = \frac{\alpha}{\alpha+\beta}$, variance $\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ and mgf $M_X(t) = 1 + \sum_{k=1}^{\infty} \left(\prod_{r=0}^{k-1} \frac{\alpha+r}{\alpha+\beta+r}\right) \frac{t^k}{k!}$.

2 Multivariate distributions

We will often want to deal with more than one variable based on the same random experiment.

Definition 2.1 (random vector)**.** Consider a random experiment with sample space Ω. Let $X_1, X_2 : \Omega \to \mathbb{R}$ be random variables. We say that (X_1, X_2) is a **random vector**. The support of (X_1, X_2) is the set of ordered pairs $\chi = \{(x_1, x_2) : X_1(\omega) = x_1, X_2(\omega) = x_2, \omega \in \Omega\}$.

2.1 Joint distributions

Of particular interest are events of the form $(X_1 \leq x_1) \cap (X_2 \leq x_2)$ for $(x_1, x_2) \in \mathbb{R}^2$.

Definition 2.2 (joint cdf)**.** Suppose that (X_1, X_2) is a random vector. The **joint cdf** of (X_1, X_2), is the function F_{X_1,X_2} defined as

$$F_{X_1,X_2}(x_1, x_2) = P(X_1 \leq x_1, X_2 \leq x_2).$$

From here we can extend the univariate case for the probability over intervals to rectangular subsets of $\mathbb{R}^2$.

Theorem 2.1 (Rectangular probability formula)**.** Suppose that the random vector (X_1, X_2) has joint cdf F_{X_1,X_2} and let $a_1, b_1, a_2, b_2 \in \mathbb{R}$ be such that $a_1 < b_1$ and $a_2 < b_2$. Then

$$P\left((X_1, X_2) \in [a_1, b_1] \times [a_2, b_2]\right) = F_{X_1,X_2}(b_1, b_2) - F_{X_1,X_2}(a_1, b_2) - F_{X_1,X_2}(b_1, a_2) + F_{X_1,X_2}(a_1, a_2).$$

Recall that, in general, random variables can be of the discrete type or of the continuous type. We extend this idea to random vectors.

A random vector (X_1, X_2) is said to be **discrete** if its support χ is countable. In this case both X_1 and X_2 are discrete random variables. It thus makes sense to define pmfs for discrete random vectors.

Definition 2.3 (joint pmf)**.** Let (X_1, X_2) be a discrete random vector. Then the **joint pmf** of (X_1, X_2), is the function p_{X_1,X_2} given by

$$p_{X_1,X_2}(x_1,x_2) = P(X_1 = x_1, X_2 = x_2).$$

As in the univariate case, this joint pmf satisfies the following properties.

1. $0 \leq p_{X_1,X_2}(x_1,x_2) \leq 1$ for all $(x_1,x_2) \in \chi$.
2. $\sum_{(x_1,x_2)\in\chi} p_{X_1,X_2}(x_1,x_2) = 1$.

Example 2.1. Suppose two dice are rolled. Let X denote the number of dots facing up on the first die and Y the number of dots on the second die. Also, let $W = \max(X,Y)$. We would like to find the joint pmf of the random vector (X,W) and the probability that $X = W$.

Solution. Let χ denote the support of (X,W). It is clear that $\chi = \{1,\dots,6\} \times \{1,\dots,6\}$.

The joint pmf $p_{X,W}(x,w)$ of (X,W) can be summarized by the following table.

$p_{X,W}(x,w)$	$x=1$	$x=2$	$x=3$	$x=4$	$x=5$	$x=6$
$w=1$	$\frac{1}{36}$	0	0	0	0	0
$w=2$	$\frac{1}{36}$	$\frac{2}{36}$	0	0	0	0
$w=3$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{3}{36}$	0	0	0
$w=4$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{4}{36}$	0	0
$w=5$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{5}{36}$	0
$w=6$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{6}{36}$

It is clearly the case that $0 \leq p_{X,W}(x,w) \leq 1$ and $\sum_{(x,w)\in\chi} p_{X,W}(x,w) = 1$.

We can now find the probability that $X = W$. That is, that the first die has the larger number of dots. We sum along the diagonal of the above table.

$$\begin{aligned}\sum_{\substack{(x,w)\in\chi \\ x=w}} p_{X,W}(x,w) &= \frac{1}{36} + \frac{2}{36} + \cdots + \frac{6}{36} \\ &= \frac{7}{12}\end{aligned}$$

If the joint cdf F_{X_1,X_2} of a random vector (X_1,X_2) is continuous, then we say that (X_1,X_2) is **continuous**. Similarly to a joint pmf, we can also define a joint pdf.

Definition 2.4 (joint pdf)**.** Let (X_1,X_2) be a continuous random vector. Then the **joint pdf** of (X_1,X_2) is the function f_{X_1,X_2} satisfying

$$F_{X_1,X_2}(x_1,x_2) = \int_{-\infty}^{x_2}\int_{-\infty}^{x_1} f_{X_1,X_2}(u,v)\,\mathrm{d}u\,\mathrm{d}v$$

for all $(x_1,x_2)\in\mathbb{R}^2$.

As in the univariate case, the joint pdf satisfies the following properties.

1. $f_{X_1,X_2}(x_1,x_2)\geq 0$ for all $(x_1,x_2)\in\mathbb{R}^2$.
2. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X_1,X_2}(u,v)\,\mathrm{d}u\,\mathrm{d}v = 1$.

Often times, we will want to obtain the distributions of the random variables X_1 and X_2 from the joint distribution of (X_1,X_2).

Given the above setup, we may obtain the **marginal cdf** of X_1 from the following equivalent formulations.

$$\begin{aligned} F_{X_1}(x_1) &= P(X_1\leq x_1)\\ &= P(X_1\leq x_1, -\infty < X_2 < \infty)\\ &= \lim_{x_2\to\infty} P(X_1\leq x_1, X_2\leq x_2)\\ &= \lim_{x_2\to\infty} F_{X_1,X_2}(x_1,x_2)\end{aligned}$$

We may also find **marginal pmfs** and **marginal pdfs**.

If (X_1,X_2) is a discrete random vector, then

$$\begin{aligned} p_{X_1}(x_1) &= P(X_1 = x_1)\\ &= \sum_{x_2} p_{X_1,X_2}(x_1,x_2)\end{aligned}$$

If (X_1,X_2) is a continuous random vector, then

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1,x_2)\,\mathrm{d}x_2$$

Example 2.2. We revisit the setup of Example 2.1. We would like to find the marginal pmf of W.

Solution. We apply the definition.

w	$p_W(w)$
1	$\frac{1}{36}$
2	$\frac{3}{36}$
3	$\frac{5}{36}$
4	$\frac{7}{36}$
5	$\frac{9}{36}$
6	$\frac{11}{36}$

We may write

$$p_W(w) = \begin{cases} \frac{2w-1}{36} & \text{if } w = 1, \dots, 6 \\ 0 & \text{otherwise} \end{cases}.$$

Example 2.3. Consider the joint pdf of (X, Y) to be

$$f_{X,Y}(x,y) = \begin{cases} cxy^2 & \text{for } 0 \le x \le 2, 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

We would like to find the value of c such that $f_{X,Y}$ is a valid pdf. We would then like to find the marginal pdfs.

Solution. For $f_{X,Y}$ to be valid, we require it to be non-negative and it must integrate to 1. The second property will be used to determine c.

$$\begin{aligned}
\int_0^1 \int_0^2 cxy^2 \, dx \, dy &= 1 \\
c \int_0^1 y^2 \int_0^2 x \, dx \, dy &= 1 \\
c\left(\frac{1}{3}\right)(2) &= 1 \\
\frac{2}{3}c &= 1 \\
c &= \frac{3}{2}
\end{aligned}$$

Such a c makes $f_{X,Y}$ a valid pdf.

We first find the marginal pdf of X.

$$\begin{aligned}
f_X(x) &= \int_0^1 \frac{3}{2}xy^2 \, dy \\
&= \frac{3}{2}x \int_0^1 y^2 \, dy \\
&= \frac{x}{2}
\end{aligned}$$

So $f_X(x) = \begin{cases} \frac{x}{2} & \text{for } 0 \le x \le 2 \\ 0 & \text{otherwise} \end{cases}$.

$$\begin{aligned}
f_Y(y) &= \int_0^2 \frac{3}{2}xy^2 \, dx \\
&= \frac{3}{2}y^2 \int_0^1 x \, dy \\
&= \frac{3y^2}{4}
\end{aligned}$$

So $f_Y(x) = \begin{cases} \frac{3y^2}{4} & \text{for } 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$.

2.2 Expectation

Expectations in the multivariate case are easily extended from the univariate case to random vectors.

Theorem 2.2 (Law of the unconscious statistician (multivariate)). Let (X_1, X_2) be a random vector and let $Y = g(X_1, X_2)$ for some real-valued function g. Then Y is a random variable and we have the following.

(a) Suppose (X_1, X_2) is discrete with pmf $p_{X_1,X_2}(x_1, x_2)$. If

$$\sum_{x_1} \sum_{x_2} |g(x_1, x_2)| p_{X_1,X_2}(x_1, x_2)$$

is finite, then the expectation of Y exists and is given by

$$E(Y) = \sum_{x_1} \sum_{x_2} g(x_1, x_2) p_{X_1,X_2}(x_1, x_2).$$

(b) Suppose (X_1, X_2) is continuous with pdf $f_{X_1,X_2}(x_1, x_2)$. If

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |g(x_1, x_2)| f_{X_1,X_2}(x_1, x_2) \, dx_1 \, dx_2$$

is finite, then the expectation of Y exists and is given by

$$E(Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x_1, x_2) f_{X_1,X_2}(x_1, x_2) \, dx_1 \, dx_2 .$$

We often abbreviate this theorem LOTUS.

We can now show that expectation is a linear operator.

Theorem 2.3 (Linearity of expectation). Let (X_1, X_2) be a random vector and let $Y_1 = g_1(X_1, X_2)$ and $Y_2 = g_2(X_1, X_2)$ be random variables for some real-valued functions g_1 and g_2. Suppose that both $E(Y_1)$ and $E(Y_2)$ exist. Then for any $k_1, k_2 \in \mathbb{R}$,

$$E(k_1 Y_1 + k_2 Y_2) = k_1 E(Y_1) + k_2 E(Y_2).$$

Proof. We prove for the discrete case.

We show absolute convergence using the triangle inequality.

$$\sum_{x_1}\sum_{x_2}|k_1g_1(x_1,x_2)+k_2g_2(x_1,x_2)|p_{X_1,X_2}(x_1,x_2) \le |k_1|\sum_{x_1}\sum_{x_2}|g_1(x_1,x_2)|p_{X_1,X_2}(x_1,x_2) + |k_2|\sum_{x_1}\sum_{x_2}|g_2(x_1,x_2)|p_{X_1,X_2}(x_1,x_2)$$

But $E(Y_1)$ and $E(Y_2)$ exist, so the above is finite and $E(k_1Y_1+k_2Y_2)$ exists.

$$\begin{aligned}E(k_1Y_1+k_2Y_2) &= \sum_{x_1}\sum_{x_2}(k_1g_1(x_1,x_2)+k_2g_2(x_1,x_2))p_{X_1,X_2}(x_1,x_2)\\ &= k_1\sum_{x_1}\sum_{x_2}g_1(x_1,x_2)p_{X_1,X_2}(x_1,x_2)+k_2\sum_{x_1}\sum_{x_2}g_2(x_1,x_2)p_{X_1,X_2}(x_1,x_2)\\ &= k_1E(Y_1)+k_2E(Y_2)\end{aligned}$$

■

Example 2.4. Consider the random vector (X,Y) with joint pmf $p_{X,Y}(x,y)$ defined as follows.

$p_{X,Y}(x,y)$	$y=1$	$y=2$	$y=3$
$x=0$	$\frac{1}{18}$	$\frac{1}{9}$	$\frac{1}{18}$
$x=1$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{18}$
$x=2$	$\frac{2}{9}$	$\frac{1}{18}$	$\frac{2}{9}$

Find $E(XY)$.

Solution. Let $g(X,Y)=XY$. We use Theorem 2.2 (LOTUS).

$$\begin{aligned}E(X,Y) &= \sum_x\sum_y xyp_{X,Y}(xy)\\ &= (0)(1)\frac{1}{18}+(0)(2)\frac{1}{9}+(0)(3)\frac{1}{18}\\ &\quad+(1)(1)\frac{1}{9}+(1)(2)\frac{1}{9}+(1)(3)\frac{1}{18}\\ &\quad+(2)(1)\frac{2}{9}+(2)(2)\frac{1}{18}+(2)(3)\frac{9}{9}\\ &= \frac{45}{18}\end{aligned}$$

Example 2.5. Recall the setup of Example 2.3. Find $E(XY)$.

Solution.

$$\begin{aligned} E(XY) &= \int_0^1 \int_0^2 (xy)\frac{3}{2}xy^2 \,\mathrm{d}x\,\mathrm{d}y \\ &= \frac{3}{2}\int_0^1 y^3 \int_0^2 x^2 \,\mathrm{d}x\,\mathrm{d}y \\ &= \frac{3}{2}\int_0^1 y^3 \frac{8}{3}\,\mathrm{d}y \\ &= 4\int_0^1 y^3\,\mathrm{d}y \\ &= 4\cdot\frac{1}{4} \\ &= 1 \end{aligned}$$

2.3 Joint measures

For an event B in the support of the discrete random vector (X,Y) with pmf $p_{X,Y}$,

$$P((X,Y)\in B) = \sum_{(x,y)\in B} p_{X,Y}(x,y).$$

For an event A in the support of the continuous random vector (X,Y) with pdf $f_{X,Y}$,

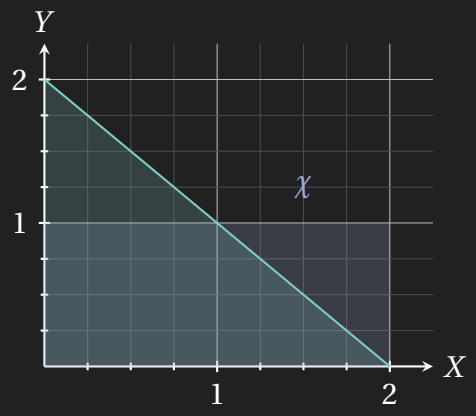
$$P((X,Y)\in A) = \iint_A f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y\,.$$

Example 2.6. Recall the setup of Example 2.3. Find $P(X+Y\le 2)$.

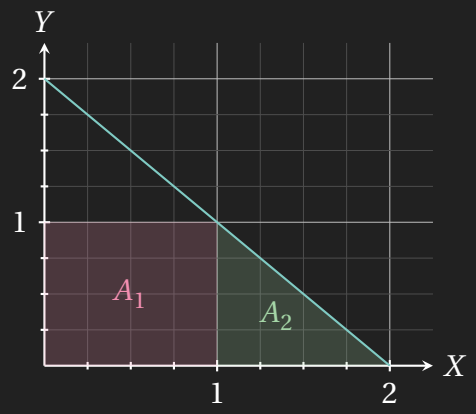
Solution. Let A denote the event $X+Y\le 2$. Then

$$P((X,Y)\in A) = \int\int_A \frac{3}{2}xy^2\,\mathrm{d}y\,\mathrm{d}x\,.$$

The support of (X,Y) is $\chi = \{(x,y) : 0\le x\le 2, 0\le y\le 1\}$ and the set of points under and on the line $X+Y=2$ is $\{(x,y)\in\mathbb{R}^2 : y\le 2-x\}$. As such, A is the intersection of these sets and, notably, $A=\{(x,y)\in\chi : y\le 2-x\}$. We illustrate these sets in the following plot.



We can write A as a disjoint union $A = A_1 \cup A_2$, as shown in the following plot.



We see that $A_1 = \{(x,y) : 0 \leq x \leq y \leq 1\}$ and $A_2 = \{(x,y) : 1 \leq x \leq 2, 0 \leq y \leq 2 - x\}$. So we can compute the required probability as follows.

$$P((X,Y) \in A) = \iint_A f_{X,Y}(x,y)\,dy\,dx = \iint_{A_1} f_{X,Y}(x,y)\,dy\,dx + \iint_{A_2} f_{X,Y}(x,y)\,dy\,dx$$

We solve each integral.

$$\begin{aligned}\iint_{A_1} f_{X,Y}(x,y)\,dy\,dx &= \int_0^1 \int_0^1 \frac{3}{2}xy^2\,dy\,dx \\ &= \frac{3}{2}\int_0^1 x\int_0^1 y^2\,dy\,dx \\ &= \frac{3}{2}\int_0^1 x\frac{1}{3}\,dx \\ &= \frac{1}{2}\int_0^1 x\,dx \\ &= \frac{1}{2}\cdot\frac{1}{2} \\ &= \frac{1}{4}\end{aligned}$$

$$\begin{aligned}\iint_{A_2} f_{X,Y}(x,y)\,dy\,dx &= \int_1^2 \int_0^{2-x} \frac{3}{2}xy^2\,dy\,dx \\ &= \frac{3}{2}\int_1^2 x\int_0^{2-x} y^2\,dy\,dx \\ &= \frac{3}{2}\int_1^2 x\frac{(2-x)^3}{3}\,dy\,dx \\ &= \frac{1}{2}\int_1^2 x(2-x)^3\,dy\,dx \\ &= \frac{1}{2}\int_1^2 -x^4+6x^3-12x^2+8x\,dx \\ &= \frac{1}{2}\cdot\frac{3}{10} \\ &= \frac{3}{20}\end{aligned}$$

Adding the two integrals, $P(X+Y\le 2) = \frac{1}{4}+\frac{3}{20} = \frac{8}{20}$.

Definition 2.5 (covariance)**.** Suppose (X,Y) is a random vector and that $E(X)$ and $E(Y)$ exist. The **covariance** of (X,Y), denoted $\mathrm{Cov}(X,Y)$, is defined by the expectation

$$\mathrm{Cov}(X,Y) = E((X-E(X))(Y-E(Y))).$$

We may obtain a computationally favourable formulation of covariance using the linearity of expectation.

$$\begin{aligned}\operatorname{Cov}(X,Y) &= E((X-E(X))(Y-E(Y)))\\ &= E(XY - E(Y)X - E(X)Y + E(X)E(Y))\\ &= E(XY) - E(Y)E(X) - E(X)E(Y) + E(X)E(Y)\\ &= E(XY) - E(X)E(Y)\end{aligned}$$

We may call $E(XY)$ the **product moment** of X and Y.

When we say *computationally favourable*, we mean that this formulation is useful for computing covariance by hand. It is actually the case that this formulation is numerically unstable and should thus be avoided in computer programs.

Observe that if $E(XY) > E(X)E(Y)$, then $\operatorname{Cov}(X,Y) > 0$. In this case, we say that X and Y are **positively associated**.

Now observe that if $E(XY) < E(X)E(Y)$, then $\operatorname{Cov}(X,Y) < 0$. In this case, we say that X and Y are **negatively associated**.

In the case that $E(XY) = E(X)E(Y)$, then $\operatorname{Cov}(X,Y) = 0$. In this case, we say that X and Y are **not associated**.

We may ask: "Can covariance tell us how closely two random variables are related?" The answer to this question is no. The measure of covariance is *not* invariant to scale.

Suppose we would like to compute $\operatorname{Cov}(aX, bY)$ for $a, b \in \mathbb{R}$. By the linearity of expectation, we find the following.

$$\begin{aligned}\operatorname{Cov}(aX,bY) &= E((aX)(bY)) - E(aX)E(bY)\\ &= ab(E(XY) - E(X)E(Y))\\ &= ab\operatorname{Cov}(X,Y)\end{aligned}$$

As such, we would like a measure of association between random variables that remains invariant under scaling. This measure is correlation.

Definition 2.6 (correlation coefficient)**.** Suppose that (X,Y) is a random vector with covariance $\operatorname{Cov}(X,Y)$. Then the **correlation coefficient** between X and Y, denoted $\operatorname{Corr}(X,Y)$, is defined by

$$\operatorname{Corr}(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}}.$$

We may also use $\rho_{X,Y}$ or simply ρ to denote the correlation coefficient.

Theorem 2.4. For all jointly distributed random variables (X, Y) where $\text{Corr}(X, Y)$ exists, $\text{Corr}(aX + b, cY + d) = \text{sgn}(ac)\,\text{Corr}(X, Y)$ for any scalars $a, b, c, d \in \mathbb{R}$.

Proof. Recall the definition of the correlation coefficient.

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}$$

We compute $\text{Cov}(aX + b, cY + d)$ using the linearity of expectation.

$$\begin{aligned}
\text{Cov}(aX + b, cY + d) &= E((aX + b)(cY + d)) - E(aX + b)E(cY + d) \\
&= E(acXY + adX + bcY + bd) - E(aX + b)E(cY + d) \\
&= acE(XY) + adE(X) + bcE(Y) + bd - (aE(X) + b)(cE(Y) + d) \\
&= acE(XY) + adE(X) + bcE(Y) + bd - acE(X)E(Y) - adE(X) - bcE(Y) - bd \\
&= acE(XY) - acE(X)E(Y) \\
&= ac(E(XY) - E(X)E(Y)) \\
&= ac\,\text{Cov}(X, Y)
\end{aligned}$$

We have that $\text{Var}(aX + b) = a^2\text{Var}(X)$ and $\text{Var}(cY + d) = c^2\text{Var}(Y)$. We substitute these into the definition of the correlation coefficient.

$$\begin{aligned}
\text{Corr}(aX + b, cY + d) &= \frac{\text{Cov}(aX + b, cY + d)}{\sqrt{\text{Var}(aX + b)\,\text{Var}(cY + d)}} \\
&= \frac{ac\,\text{Cov}(X, Y)}{\sqrt{a^2\text{Var}(X)c^2\text{Var}(Y)}} \\
&= \frac{ac\,\text{Cov}(X, Y)}{|ac|\sqrt{\text{Var}(X)\,\text{Var}(Y)}} \\
&= \text{sgn}(ac)\frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}} \\
&= \text{sgn}(ac)\,\text{Corr}(X, Y)
\end{aligned}$$

■

Theorem 2.5. Let (X, Y) be a random vector. Then $\text{Corr}(X, Y) = 0$ if and only if $\text{Cov}(X, Y) = 0$.

Theorem 2.6. For all jointly distributed random variables (X, Y) where $\text{Corr}(X, Y)$ exists, $-1 \leq \text{Corr}(X, Y) \leq 1$.

Proof. We prove the discrete case. That is, when (X, Y) is a discrete random vector.

Recall the Cauchy-Schwarz inequality:

$$\left|\sum_{i=1}^{n} a_i b_i\right| \leq \left(\sum_{i=1}^{n} {a_i}^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^{n} {b_i}^2\right)^{\frac{1}{2}}$$

Consider the following.

$$|\text{Cov}(X, Y)| = \left|\sum (x - E(X))(y - E(Y)) p_{X,Y}(xy)\right|$$

Let $a_i = (x - E(X))\sqrt{p_{X,Y}(xy)}$ and $b_i = (y - E(Y))\sqrt{p_{X,Y}(xy)}$.

$$= \left|\sum a_i b_i\right|$$

We use Cauchy-Schwarz.

$$\begin{aligned} &\leq \left(\sum a_i^2\right)^{\frac{1}{2}} \left(\sum b_i^2\right)^{\frac{1}{2}} \\ &= \left(\sum (x - E(X))^2 p_{X,Y}(xy)\right)^{\frac{1}{2}} \left(\sum (y - E(Y))^2 p_{X,Y}(xy)\right)^{\frac{1}{2}} \\ &= \sqrt{\text{Var}(X)\,\text{Var}(Y)} \end{aligned}$$

So we have that $|\text{Cov}(X, Y)| \leq \sqrt{\text{Var}(X)\,\text{Var}(Y)}$. That is,

$$\left|\frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}\right| = |\text{Corr}(X, Y)| \leq 1$$

■

Example 2.7. Consider the set up of Example 2.1. Find $\text{Cov}(X,W)$ and $\text{Corr}(X,W)$. *Solution.* Recall that $\text{Cov}(X,W) = E(XW) - E(X)E(W)$. Recall that the marginal pmf of W from Example 2.2 is

$$p_W(w) = \begin{cases} \frac{2w-1}{36} & \text{if } w = 1,\dots,6 \\ 0 & \text{otherwise} \end{cases}.$$

We also have that the marginal pmf of X is

$$p_X(x) = \begin{cases} \frac{1}{6} & \text{if } x = 1,\dots,6 \\ 0 & \text{otherwise} \end{cases}.$$

We compute $E(X)$ and $E(W)$.

$$\begin{aligned} E(X) &= \sum_x x p_X(x) \\ &= \sum_{x=1}^{6} x \frac{1}{6} \\ &= \frac{1}{6} + \frac{2}{6} + \cdots + \frac{6}{6} \\ &= \frac{21}{6} \\ &= \frac{7}{2} \end{aligned}$$

$$\begin{aligned} E(W) &= \sum_w w p_W(w) \\ &= \sum_{w=1}^{6} w \frac{2w-1}{36} \\ &= (1)\frac{1}{36} + (2)\frac{3}{36} + \cdots + (6)\frac{11}{36} \\ &= \frac{161}{36} \end{aligned}$$

We find the product moment of X and W using the joint pmf $p_{X,W}$ defined in Example 2.1.

$$\begin{aligned} E(XW) &= \sum_{(x,w)} \sum xwp_{X,W}(x,w) \\ &= \sum_{x=1}^{6} x \sum_{w=1}^{6} wp_{X,W}(x,w) \\ &= \frac{154}{9} \end{aligned}$$

We now see the following.

$$\begin{aligned} \text{Cov}(X,W) &= E(XW) - E(X)E(W) \\ &= \frac{154}{9} - \frac{7}{2} \cdot \frac{161}{36} \\ &= \frac{105}{72} \end{aligned}$$

Recall that $\text{Corr}(X,W) = \frac{\text{Cov}(X,W)}{\sqrt{\text{Var}(X)\text{Var}(W)}}$.

We compute $\text{Var}(X)$ and $\text{Var}(W)$.

$$\begin{aligned} \text{Var}(X) &= E(X^2) - E(X)^2 \\ &= \sum_x x^2 p_X(x) - E(X)^2 \\ &= \sum_{x=1}^{6} \frac{x^2}{6} - E(X)^2 \\ &= \frac{1^2}{6} + \frac{2^2}{6} + \cdots + \frac{6^2}{6} - \left(\frac{7}{2}\right)^2 \\ &= \frac{105}{36} \\ &= \frac{35}{12} \end{aligned}$$

$$\begin{aligned}\mathrm{Var}(W) &= E(W^2) - E(W)^2 \\ &= \sum_w w^2 p_W(w) - E(W)^2 \\ &= \sum_{w=1}^{6} w^2 \frac{2w-1}{36} - E(V)^2 \\ &= 1^2 \cdot \frac{1}{36} + 2^2 \cdot \frac{3}{36} + \cdots + 6^2 \cdot \frac{11}{36} - \left(\frac{161}{36}\right)^2 \\ &= \frac{2555}{1296}\end{aligned}$$

Then we have the following.

$$\begin{aligned}\mathrm{Corr}(X,W) &= \frac{\mathrm{Cov}(X,W)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(W)}} \\ &= \frac{\frac{105}{72}}{\sqrt{\frac{35}{12} \cdot \frac{2555}{1296}}} \\ &\approx 0.6082\end{aligned}$$

Example 2.8. Consider the setup of Example 2.3. Find $\mathrm{Cov}(X,Y)$ and $\mathrm{Corr}(X,Y)$.

Solution. Recall that $E(XY) = 1$ from Example 2.5. We compute $E(X)$ and $E(Y)$.

To be completed

2.4 Conditional distributions

Definition 2.7 (conditional pmf)**.** Let X_1 and X_2 be discrete random variables with joint pmf p_{X_1,X_2}. Let p_{X_1} and p_{X_2} denote the marginal pmfs of X_1 and X_2, respectively. Suppose x_1 is such that $p_{X_1}(x_1) > 0$. The **conditional pmf** of X_2 given that $X_1 = x_1$ is defined by

$$p_{X_2|X_1}(x_2|x_1) = \frac{p_{X_1,X_2}(x_1,x_2)}{p_{X_1}(x_1)}$$

where x_2 is in the support of X_2.

Definition 2.8 (conditional pdf)**.** Let X_1 and X_2 be continuous random variables with joint pdf f_{X_1,X_2}. Let f_{X_1} and f_{X_2} denote the marginal pdfs of X_1 and X_2, respectively. Suppose x_1 is such that $f_{X_1}(x_1) > 0$. The **conditional pdf** of X_2 given that $X_1 = x_1$ is defined by

$$f_{X_2|X_1}(x_2|x_1) = \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_1}(x_1)}$$

where x_2 is in the support of X_2.