Chapter 5 Part 1
STAT 5700: Probability

Prof. Katie Fitzgerald, PhD

Fall 2025

Contents

5.1 Introduction / Motivation

Often we're interested in multiple random variables and how they behave together. Some examples:

- a biologist studying animal survival might be interested in the intersection of two events: the litter contains n animals, and y animals survive
- height & weight of an individual is a pair of events relevant to height-weight measurements used in many health contexts

Importantly for statisticians, when you take a random sample of size n, you end up with a sequence of random variables $Y_1, Y_2, \ldots Y_n$. For example, this could represent the blood pressure of n individuals, or n different measurements for a single person. A specific set of outcomes, or sample measurements, can be expressed in terms of the intersection of the n events $(Y_1 = y_1, Y_2 = y_2, \ldots Y_n = y_n)$. Knowing the probability of this intersection is essential in making inferences about the population from which the sample is drawn.

In chapters 3 & 4, we worked with **univariate** distributions that described the behavior of a single random variable. In Chapter 5, we will consider **bivariate** and **multivariate** distributions; that is, the joint behavior of 2 or more random variables.

5.2 Bivariate and Multivariate Probability Distributions

(discrete version of a pdf)

Joint probability mass functions (pmfs)

Suppose that you randomly select an email sent to your account and record whether or not the email is spam, and whether it is sent as HTML or plain text. Let $S = 1$ if the email is spam and $S = 0$ otherwise, and let $H = 1$ if the email is HTML and $H = 0$ otherwise. The proportion of the population of emails falling into each category of S and H is recorded in the table below.

	H = 0	H = 1
S = 0	0.25	0.65
S = 1	0.06	0.04

joint pmf

Exercise:

What is the probability that a randomly selected email is HTML & spam?

$$P(H = 1, S = 1) = 0.04$$

What is $P(S = 0, H = 0)$?

$$0.25$$

What is the probability that an email is spam?

$$.06 + .04 = .10$$

The table above gives the **joint pmf** of the random variables S and H, where the cells of the table are the probabilities $p(s, h) = P(S = s, H = h)$.

$p(y)$ for discrete prob. dist.

Defintion 5.1/Theorem 5.1 Let Y_1 and Y_2 be two discrete random variables. The probability that $Y_1 = y_1$ and $Y_2 = y_2$ is denoted by

$$p(y_1, y_2) = P(Y_1 = y_1, Y_2 = y_2)$$

, $-\infty < y_1 < \infty, -\infty < y_2 < \infty$. The function $p(y_1, y_2)$ is called the **joint probability mass function (joint pmf)** of Y_1 and Y_2 and has the following properties:

1. $0 \leq p(y_1, y_2) \leq 1$.
2. $\sum_{(y_1,y_2) \in S} p(y_1, y_2) = 1$, where the sum is over all pairs of values (y_1, y_2) in the support S (i.e. that have assigned non-zero probabilities)

Note, you may also see notation that uses two random variables X and Y rather than Y_1 and Y_2.

Exercise (cont'd):
Confirm that the joint pmf above for S and H is a valid pmf.

$$.25 + .65 + .06 + .04 = 1 \checkmark$$

Definition 5.2: For any random variables Y_1 and Y_2, the joint cumulative distribution function* (cdf) is given by

$$F(y_1, y_2) = P(Y_1 \leq y_1, Y_2 \leq y_2)$$

*Note, the book refers to this simply as the joint distribution function.

$\hookrightarrow$ cdf   density function $\longrightarrow$ pdf
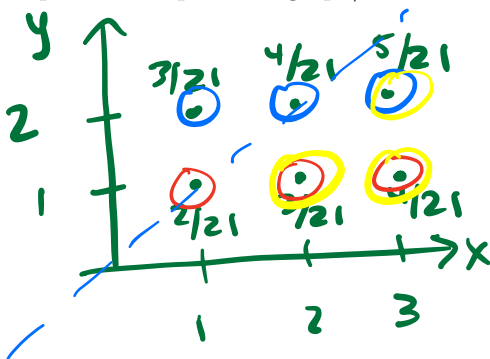
For discrete random variables,

$$F(y_1, y_2) = \sum_{t_1 \leq y_1} \sum_{t_2 \leq y_2} p(t_1, t_2)$$

3

Often we will deal with random variables X and Y whose joint pmf can be represented in functional form:

Example:
Let the joint pmf of X and Y be defined by: $p(x, y) = \frac{x+y}{21}, \quad x = 1, 2, 3, \quad y = 1, 2$

Represent the pmf as a graph/table

$$\frac{3+4+5+2+3+4}{21} = \frac{21}{21} \checkmark$$

$$\sum_{x=1}^{3} \sum_{y=1}^{2} \frac{x+y}{21}$$

$$\sum_{x=1}^{3} \left(\frac{x+1}{21} + \frac{x+2}{21} \right)$$

Confirm this is a valid pmf

$$= \sum_{x=1}^{3} \frac{2x+3}{21} = \frac{2(1)+3 + 2(2)+3 + 2(3)+3}{21} = \frac{21}{21} \checkmark$$

Find $P(Y = 1)$ and $P(Y = 2)$

$$P(Y=1) = \frac{2+3+4}{21} = \frac{9}{21}$$

$$P(Y=2) = \frac{3+4+5}{21} = \frac{12}{21}$$

What is $P(X > Y)$?

$$\{(2,1), (3,1)(3,2)\}$$

$$\frac{3 + 4 + 5}{21} = \boxed{\frac{12}{21}}$$

Find $F(2, 2)$

$$P(X \le 2, Y \le 2)$$

$$\{(1,1)(1,2)(2,1)(2,2)\}$$

$$\frac{2 + 3 + 3 + 4}{21} = \boxed{\frac{12}{21}}$$

→ density → continuous

Joint pdfs

The ideas of bivariate distributions extend from discrete to continuous by replacing sums with integrals.

> Two random variables are said to be jointly continuous if their joint distribution function $F(y_1, y_2)$ is continuous in both arguments.
>
> $$F(y_1, y_2) = \int_{-\infty}^{y_1} \int_{-\infty}^{y_2} f(t_1, t_2) dt_2 dt_1$$
>
> The function $f(y_1, y_2)$ is called the **joint probability density function (joint pdf)** for two jointly continuous random variables and has properties:
>
> 1. $f(y_1, y_2) \geq 0$, for all y_1, y_2
> 2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(y_1, y_2) dy_1 dy_2 = 1$.

- Probabilities are found as volumes under the joint pdf over regions of the support

Reminder about double integrals

Evaluating double integrals amounts to evaluating the inner integral first, then the outer integral:

$$\int \int f(x, y) dx dy = \int \left(\int f(x, y) dx \right) dy$$

- We can usually interchange the order of integration, but we must be careful about the appropriate bounds
- If the bounds are a function of the variables (e.g. $0 < x < y < 1$), then the inner bounds should include the variable, and the outer bounds should be constants (not dependent on the variables).

$$\int_0^1 \int_0^y f(x, y) dx dy = \int_0^1 \int_x^1 f(x, y) dy dx$$

- Draw a picture to determine the appropriate bounds of integration!

Example Suppose that a radioactive particle is randomly located in a square with sides of unit length. That is, if two regions within the unit square and of equal area are considered, the particle is equally likely to be in either region. Let X and Y denote the coordinates of the particle's location. A reasonable model for the relative frequency histogram for X and Y is the bivariate analog of the univariate uniform density function.

$$f(x,y) = \begin{cases} 1, & 0 \le x \le 1, 0 \le y \le 1 \\ 0, & elsewhere \end{cases}$$

a. Sketch the probability density surface.
b. Find $F(0.2, 0.4)$
c. Find $P(0.1 \le X \le 0.3, 0 \le Y \le 0.5)$
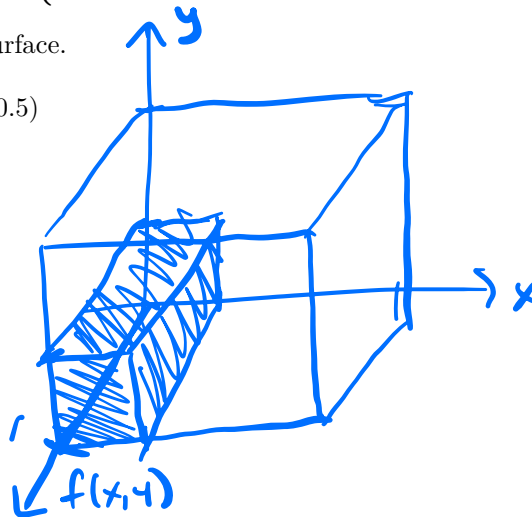
$$\rightarrow P(X \le 0.2, Y \le 0.4)$$

$$= \int_0^{0.2} \int_0^{0.4} 1 \, dy \, dx$$

$$= \int_0^{0.2} y \Big|_0^{0.4} \, dx$$

$$= \int_0^{0.2} 0.4 \, dy$$

$$= 0.4x \Big|_0^{0.2}$$

$$= 0.4(0.2) = \boxed{0.08}$$

$$\int_{0.1}^{0.3} \int_0^{0.5} 1 \, dy \, dx \qquad OR \qquad \int_0^{0.5} \int_{0.5}^{0.3} 1 \, dx \, dy$$

$$\boxed{\underline{0.1}}$$

Example Gasoline is stocked in a bulk tank once at the beginning of each week and then sold to invidual customers. Let X denote the proportion of the capacity of the bulk tank that is ~~available~~ ~~after the tank is~~ stocked at the beginning of the week. Becaues of the limited supplies, X varies from week to week. Let Y denote the prportion of the capacity of the bulk tank that is sold during the week. Both are proportions, so are restricted between 0 and 1, and further, the amount sold y cannot exceed the amount available x. Suppose that the joint density function for X and Y is given by:

$$f(x,y) = \begin{cases} 3x, & 0 \le y \le x \le 1 \\ 0, & elsewhere \end{cases}$$

$X = $ stocked
$Y = $ sold

Find the probability that less than one-half of the tank will be stocked and more than one-quarter of the tank will be sold.

$$P\left(X < \tfrac{1}{2}, Y > \tfrac{1}{4}\right)$$
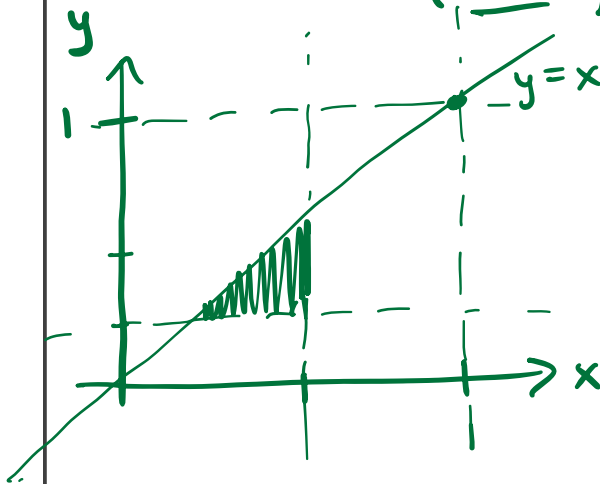
$$\int_{1/4}^{1/2}\int_{1/4}^{x} 3x \, dy \, dx$$

$$\int_{1/4}^{1/2}\int_{y}^{1/2} 3x \, dx \, dy$$

$$\int_{1/4}^{1/2} \left(\frac{3x^2}{2} \Big|_{y}^{1/2} \right) dy$$

$$\int_{1/4}^{1/2} \left(\frac{3\left(\frac{1}{4}\right) - 3y^2}{2} \right) dy$$

$$= \int_{1/4}^{1/2} \left(\frac{3}{8} - \frac{3y^2}{2} \right) dy$$

$$\left(\frac{3}{8}y - \frac{3y^3}{7 \cdot 2} \Big|_{1/4}^{1/2} \right.$$

$$\frac{3}{8}\left(\frac{1}{2} - \frac{1}{4}\right) - \frac{\left(\frac{1}{2}\right)^3 - \left(\frac{1}{4}\right)^3}{2}$$

$$\frac{3}{32} - \frac{1/8 - 1/64}{2}$$

$$\frac{3}{32} - \frac{7}{64(2)}$$

$$\boxed{\frac{5}{128}}$$

5.3 Marginal and Conditional distributions

Definition 5.4a Let X and Y have the joint probability mass function $f(x, y)$ with support S. The probability mass function of X alone, which is called the **marginal probability mass function** of X, is defined by

$$p_X(x) = \sum_y p(x, y) = P(X = x), \qquad x \in S_x,$$

where the summation is taken over all possible y values for each given x in the support of x (S_x). That is, the summation is over all (x, y) in S with a given x value. Similarly, the **marginal probability mass function** of Y, is defined by

$$p_Y(y) = \sum_x p(x, y) = P(Y = y), \qquad y \in S_y,$$

where the summation is taken over all possible x values for each given y in the support of y (S_y).

Spam & HTML example continued:
Find the marginal distributions of S and H. Then confirm that they are valid pmfs.

$$P_s(s) = \sum_n p(s, n)$$

	H = 0	H = 1	
S = 0	0.25	0.65	0.9
S = 1	0.06	0.04	0.1
	.31	.69	1

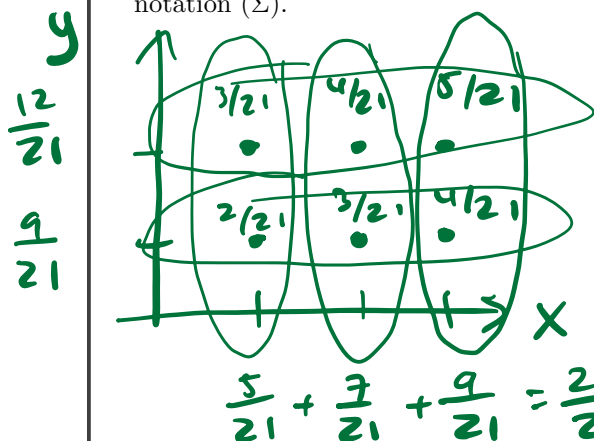| S	$P_s(s)$
0	0.9 ✓
1	0.1

H	$P_H(n)$
0	0.31 ✓
1	0.69

Example (cont'd):
Returning to the example from 5.2 with

$$p(x, y) = \frac{x + y}{21}, \qquad x = 1, 2, 3, \qquad y = 1, 2$$

Find the marginal distributions $p_X(x)$ and $p_Y(y)$, once using the table and once using the summation notation (Σ).

$$P_X(x) = \sum_y p(x, y) = \sum_{y=1}^{2} \frac{x+y}{21}$$

$$= \frac{x+1}{21} + \frac{x+2}{21} = \frac{2x+3}{21}$$

$$P_X(1) = \frac{2(1)+3}{21} = \frac{5}{21}$$

$$P_Y(y) = \sum_{x=1}^{3} \frac{x+y}{21} = \frac{1+y}{21} + \frac{2+y}{21} + \frac{3+y}{21}$$

$$= \frac{6+3y}{21} = \frac{2+y}{7}$$

$$\frac{3}{21} \quad \frac{4}{21} \quad \frac{5}{21}$$

$$\frac{2}{21} \quad \frac{3}{21} \quad \frac{4}{21}$$

$$\frac{5}{21} + \frac{7}{21} + \frac{9}{21} = \frac{21}{21} ✓$$

$$\frac{12}{21} \qquad \frac{9}{21}$$

> **Definition 5.4b**: Let X and Y be jointly continuous random variables with joint pdfs $f(x, y)$. Then the **marginal probability density functions (marginal pdfs)** are given by:
>
> $$f_x(x) = \int_{-\infty}^{\infty} f(x, y)\, dy, \quad x \in S_x,$$
>
> $$f_y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx, \quad y \in S_y$$

- Marginal pdf's are obtained by "integrating out" (or summing over) the other variable

> **Example (cont'd)** Return to the example from 5.2 and find the marginal distributions of X and Y. Confirm that they are valid pdfs.
>
> $$f(x, y) = \begin{cases} 3x, & 0 \le y \le x \le 1 \\ 0, & elsewhere \end{cases} \qquad 3xy^0$$
>
> $$f_x(x) = \int_0^x 3x\, dy = 3xy\Big|_0^x = 3x^2 \qquad 0 \le x \le 1$$
>
> $$f_y(y) = \int_y^1 3x\, dx = \frac{3}{2}x^2\Big|_y^1 = \frac{3}{2} - \frac{3}{2}y^2 \qquad 0 \le y \le 1$$
>
> $$f_y(y) = \int_0^1 \left(\frac{3}{2} - \frac{3}{2}y^2\right) dy = \left(\frac{3}{2}y - \frac{1}{2}y^3\right)\Big|_0^1 = \frac{3}{2} - \frac{1}{2} = \frac{2}{2} = 1 \checkmark$$

Conditional Distributions

Consider a population of students, each of whom have achievement test scores Y and family income levels W. Suppose that we know the joint distribution of Y, W in this population. In the absence of further information, the probability that a student randomly selected from this population has a test score higher than 1000 can be evaluated from the distribution function:

$$P(Y > 1000) = \sum_{k > 1000} P(Y = k)$$

Now suppose that we randomly select a student and learn that their family income level is \$80K per year. What is the probability that this student has a test score higher than 1000? To answer this question, we need conditional probability. Specifically, we need hold family income constant at \$80K, that is, to look at the distribution of test scores among the subset of the population of students with family incomes of exactly \$80K.

$$P(Y > 1000 | W = \$80k)$$

Conditional Probability Mass Functions

Definition: The **conditional probability mass function** of X given that $Y = y$, is defined by

$$p(x|y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{p(x,y)}{p_Y(y)}, \quad p_Y(y) > 0$$

Similarly the **conditional probability mass function** of Y given that $X = x$, is defined by

$$p(y|x) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{p(x,y)}{p_X(x)}, \quad p_X(x) > 0$$

Note this comes from the definition of conditional probability. For $A = \{X = x\}$ and $B = \{Y = y\}$, $p(x,y)$ gives $P(A \cap B)$ and

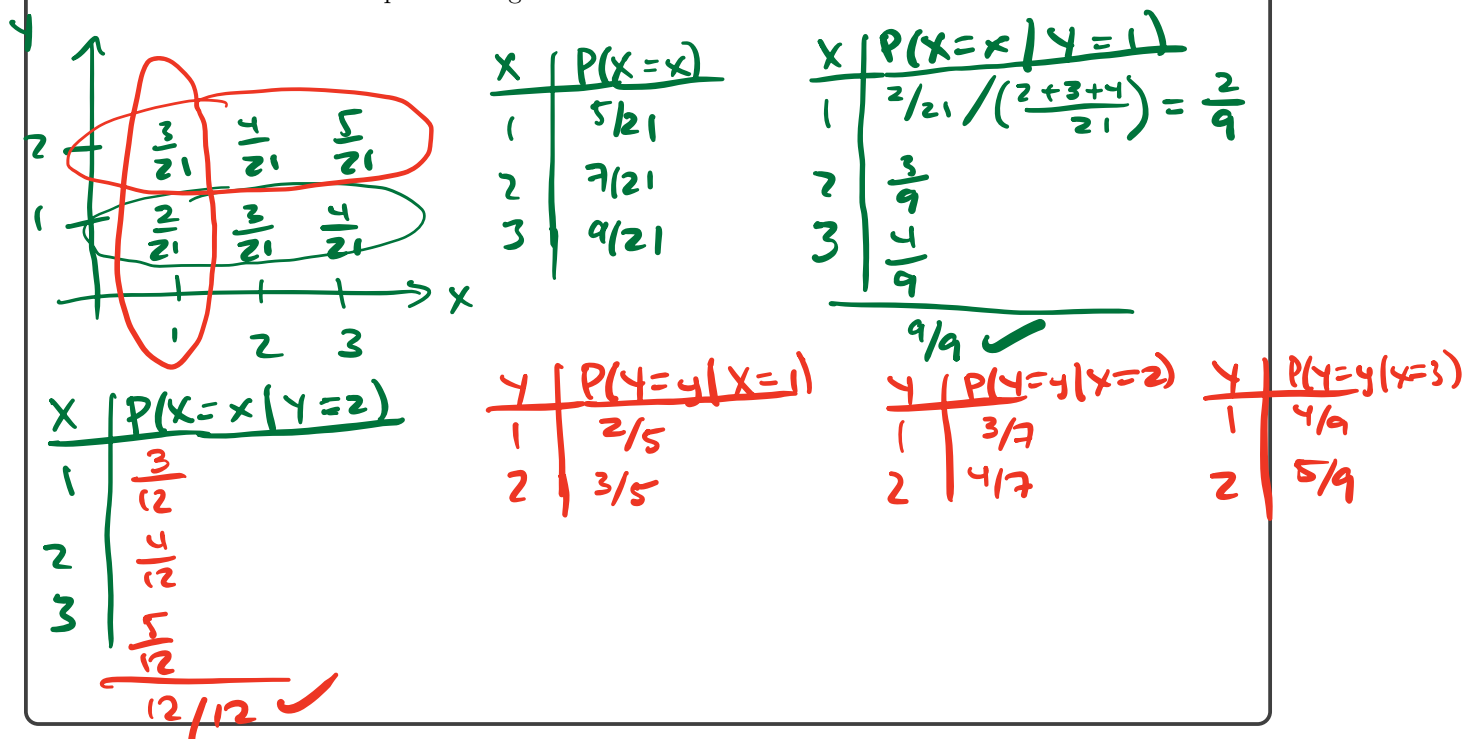$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{p(x,y)}{p_X(x)}$$

Example:
Let

$$f(x,y) = \frac{x+y}{21}, \quad x = 1, 2, 3, \quad y = 1, 2.$$

We previously found marginal pmf's:

$$f_x(x) = \frac{2x+3}{21}, \quad x = 1, 2, 3, \quad \text{and} \quad f_y(y) = \frac{y+2}{7}, \quad y = 1, 2.$$

1. Produce a table for the joint/marginal pmfs
2. Find the conditional pmfs of X given $Y = y$ and add a table for this
3. Find the conditional pmfs of Y given $X = x$ and add a table for this

Example: Let X and Y have a uniform distribution on the set of points with integer coordinates in $S = \{(x, y) : 0 \le x \le 7, \quad x \le y \le x + 2\}$. That is, $f(x, y) = 1/24, \quad (x, y) \in S$, and both x and y are integers. Find:

1. $f_X(x)$ and $f_Y(y)$
2. $p(y|x)$

Definition 5.6: If X and Y are jointly continuous random variables, the **conditional cumulative distribution function** of X given that $Y = y$, is defined by

$$F(x|y) = P(X \le x|Y = y)$$

Definition 5.7: Let X and Y are jointly continuous random variables with joint pdf $f(x, y)$ and marginal pdfs $f_X(x)$ and $f_Y(y)$ respectively. For any y such that $f_Y(y) > 0$, the conditional density of X given $Y = y$ is given by:

$$f(x|y) = \frac{f(x, y)}{f_Y(y)}$$

and, for any x such that $f_X(x) > 0$, the conditional density of Y given $X = x$ is given by:

$$f(y|x) = \frac{f(x, y)}{f_X(x)}$$

Example: A soft drink machine has a rnaomd amount Y in supply at the beginning of a given day and dispenses a random amount X during the day (with measurements in gallons). It is not resupplied during the day, and hence $X \leq Y$. It has been observed that X and Y have a joint pdf given by:

$$f(x,y) = \begin{cases} 1/2, & 0 \leq x \leq y \leq 2 \\ 0, & elsewhere \end{cases}$$

X = sold
Y = supply

That is, the points (x,y) are uniformly distributed over the triangle with the given boundaries.

- Find the conditional pdf of X given $Y = y$
- Evaluate the probability that less than $1/2$ gallon will be sold, given that teh machine contains 1.5 gallons at the start of the day

$$f(x|y) = \frac{f(x,y)}{f_y(y)}$$

$$f_y(y) = \int_0^y f(x,y)\, dx = \int_0^y \frac{1}{2}\, dx = \frac{1}{2}x\Big|_0^y = \frac{1}{2}y \qquad 0 \leq y \leq 2$$

$$f(x|y) = \frac{\frac{1}{2}}{\frac{1}{2}y} = \frac{1}{y} \qquad 0 \leq x \leq y \leq 2$$

$$P(x < 1/2 \mid Y = 1.5) = \int_0^{1/2} \frac{1}{1.5}\, dx = \frac{2}{3}x\Big|_0^{1/2} = \frac{2}{3}\frac{1}{2} = \boxed{\frac{1}{3}}$$

Exam2

5.4 Independent Random Variables

Definition 5.8 Let X have cdf $F_X(x)$, Y have cdf $F_Y(y)$, and X and Y have joint cdf $F(x,y)$. Then X and Y are indpendent are said to be **independent** if any only if

$$F(x,y) = F_X(x)F_Y(y)$$

for every pair of real numbers (x,y). If X and Y are not independent, they are said to be *dependent.*

almost always use pdfs instead of cdg to check indep.

Theorem 5.4 The **discrete** random variables with joint pmf $p(x, y)$ are **independent** if and only if, for every pair of real numbers,

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

or, equivalently,

$$p(x, y) = p_X(x)p_Y(y).$$

The **continuous** random variables with joint pdf $f(x, y)$ are **independent** if and only if, for every pair of real numbers

$$f(x, y) = f_X(x)f_Y(y).$$

Example (cont'd):
Continuing the same example as before with

$$f(x, y) = \frac{x + y}{21}, \quad x = 1, 2, 3, \quad y = 1, 2, \quad f_X(x) = \frac{2x + 3}{21}, \quad f_Y(y) = \frac{2 + y}{7}$$

Are X and Y independent?

$$f_x(x) f_y(y) = \left(\frac{2x+3}{21}\right)\left(\frac{2+y}{7}\right) = \frac{4x + 2xy + 6 + 3y}{21(7)} \neq \frac{x+y}{21}$$

NOT indep.

Example Let

$$f(x, y) = \begin{cases} 6xy^2, & 0 \le x \le 1, 0 \le y \le 1 \\ 0, & elsewhere \end{cases}$$

Show that X and Y are independent.

$$f_x(x) = \int_0^1 6xy^2 \, dy = \frac{6}{3} x y^3 \Big|_0^1 = 2x \qquad 0 < x < 1$$

$$f_y(y) = \int_0^1 6xy^2 \, dx = \frac{6}{2} x^2 y^2 \Big|_0^1 = 3y^2 \qquad 0 < y < 1$$

$$f_x(x) f_y(y) = 2x \cdot 3y^2 = 6xy^2 \checkmark \qquad \text{indep.}$$

Theorem 5.5: Let X and Y have a joint pdf $f(x, y)$ that is positive if and only if $a \le x \le b$ and $c \le y \le d$; and $f(x, y) = 0$ otherwise. Then X and Y are independent random variables if and only if

$$f(x, y) = g(x)h(y),$$

where $g(x)$ is a function of x alone and $h(y)$ is a function of y alone.

Notice, this means we don't actually have to derive the marginal pdfs of x and y. Indeed, $g(x)$ and $h(y)$ need not be marginal densities (though they will be constant multiples of the marginal pdfs).

In the previous example with $f(x, y) = 6xy^2, 0 \le x \le 1, 0 \le y \le 1$, we have x and y bounded between two sets of constants (their bounds do not depend on each other), and we can write $f(x, y)$ as the product of $g(x) = 6x$ and $h(y) = y^2$. Thus, this is enough to say they are independent.

Example: Let

$$f(x,y) = \begin{cases} 2x, & 0 \le x \le 1, 0 \le y \le 1 \\ 0, & elsewhere \end{cases}$$

(annotations above: a over $2x$, b over $0 \le x \le 1$, c, d over $0 \le y \le 1$)

Are X and Y independent?

$2x$

$g(x) = 2x \quad h(y) = 1y^0$

$x,y \quad$ indep $\quad$ by Thm 5.5

5.5 Expected Value of a Function of Random Variables

Let $g(Y_1, Y_2, \ldots, Y_k)$ be a function of **discrete** random variables $Y_1, Y_2, \ldots, Y_k$ that have joint pmf $p(y_1, y_2, \ldots, y_k)$. Then

$$E[g(Y_1, Y_2, \ldots, Y_k)] = \sum_{\text{all } y_k} \cdots \sum_{\text{all } y_2} \sum_{\text{all } y_2} g(y_1, y_2, \ldots, y_k) p(y_1, y_2, \ldots, y_k).$$

(annotation: pmf)

If $Y_1, Y_2, \ldots, Y_k$ are **continuous** random variables with joint pdf $f(y_1, y_2, \ldots, y_k)$, then

$$E[g(Y_1, Y_2, \ldots, Y_k)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(y_1, y_2, \ldots, y_k) f(y_1, y_2, \ldots, y_k) dy_1 dy_2 \ldots dy_k$$

(annotation: pdf)

Exercise (cont'd): Continuing the same example as before with

$$f(x,y) = \frac{x+y}{21}, \quad x = 1, 2, 3, \quad y = 1, 2$$

Find $E(XY^2)$ using the joint distribution.

$$E(XY^2) = \sum_{y=1}^{2} \sum_{x=1}^{3} xy^2 \left(\frac{x+y}{21}\right) = \sum_{y=1}^{2} \sum_{x=1}^{3} \frac{x^2 y^2 + xy^3}{21}$$

$$= \sum_{y=1}^{2} \frac{1^2 y^2 + 1y^3 + 2^2 y^2 + 2y^3 + 3^2 y^2 + 3y^3}{21}$$

$$= \sum_{y=1}^{2} \frac{14y^2 + 6y^3}{21} = \frac{14 + 6 + 14(2) + 6(8)}{21} = \boxed{\frac{124}{21}}$$

Example: Let $f(x,y) = 2$, $\quad 0 < x < y < 1$

1. Find $P(0 < X < 1/2, 0 < Y < 1/2)$ *Hint: draw a picture to determine the bounds of integration*
2. Find the marginal pdfs of X and Y.
3. Find the mean of Y and the variance of Y

$y = x$
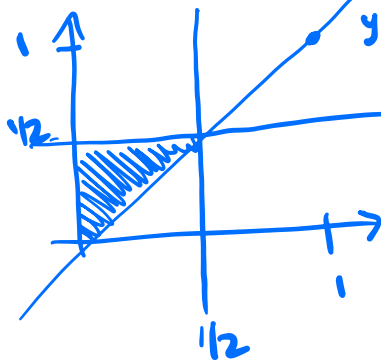
$P(0 < x < 1/2, \ 0 < y < 1/2)$

$= \int_0^{1/2} \int_0^y 2 \, dx \, dy \qquad$ OR $\quad \int_0^{1/2} \int_x^{1/2} 2 \, dy \, dx$

$= \int_0^{1/2} \left(2x \big|_0^y \right) dy$

$= \int_0^{1/2} 2y \, dy$

$= y^2 \big|_0^{1/2} = \boxed{\dfrac{1}{4}}$

② $f_x(x) = \int_x^1 2 \, dy = 2y \big|_x^1 = 2 - 2x = \underline{2(1-x)} \quad 0 < x < 1$

$f_y(y) = \int_0^y 2 \, dx = 2x \big|_0^y = \underline{2y} \qquad 0 < y < 1$

③ 1st: $E(y) = \int y \, f_y(y) \, dy = \int_0^1 y \, 2y \, dy = \int_0^1 2y^2 \, dy = \frac{2}{3} y^3 \big|_0^1$
$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad = \frac{2}{3}$

2nd: $E(y) = \int_0^1 \int_0^y y \, 2 \, dx \, dy$

$\qquad E(y) = \iint y \, f(x,y) \, dx \, dy \quad$ (by def on pg 14)

$\qquad \rightarrow \int_0^1 \left(2xy \big|_{x=0}^y \right) dy = \int_0^1 2y^2 \, dy = \frac{2}{3} y^3 \big|_0^1 = \frac{2}{3}$

take home: can find $E(y)$ from the marginal
$\qquad\qquad\qquad\qquad\qquad$ or the joint

$V(y) = E(y^2) - \left(E(y) \right)^2$

$E(y^2) = \int y^2 \, f_y(y) \, dy$

OR $E(y^2) = \iint y^2 \, f(x,y) \, dx \, dy$

$C = 10X + 0.5X^2$

$E(C) = 10 E(x) + 0.5 E(x^2)$

$E(x^2) \neq \left(E(x) \right)^2$

15

5.6: Special theorems

> **Theorem 5.6** Let c be a constant. Then $E(c) = c$

> **Theorem 5.7** Let $g(X,Y)$ be a function of the random variables X and Y and let c be a constant. Then $E(cg(X,Y)) = cE(g(X,Y))$

> **Theorem 5.8** Let $g_1(X,Y)$, $g_2(X,Y), \ldots g_k(X,Y)$ be functions of the random variables X and Y. Then $E[g_1(X,Y) + g_2(X,Y) + \cdots + g_k(X,Y)] = E(g_1(X,Y)) + E(g_2(X,Y)) + \cdots + E(g_k(X,Y))$

> **Theorem 5.9** Let X and Y be independent random variables and $g(X)$ and $h(Y)$ be functions of X and Y only, respectively. Then
> $$E(g(X)h(Y)) = \cancel{E(g(X)h(Y))} \quad E(g(x))E(h(y))$$
> provided the expectations exist.

new $\;\; E(XY^2) = E(X)E(Y^2)$ iff X, Y indep.

5.7 The Correlation Coefficient

Covariance & Correlation

We now introduce the name of another special expectation:

> If $g(X,Y) = (X - \mu_X)(Y - \mu_Y)$, for two random variables X and Y with means μ_X and μ_Y respectively, then
> $$\sigma_{XY} \equiv Cov(X,Y) \equiv E(g(X,Y)) = E[(X - \mu_X)(Y - \mu_Y)]$$
> is called the **covariance** of X and Y. If the standard deviations σ_X and σ_Y are non-zero, then
> $$\rho_{XY} \equiv Cor(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$
> is called the **correlation** (or **correlation coefficient**) of X and Y.

The **covariance** of two random variables is a measure of the linear dependence between the variables. Higher covariance indicates that one variable can be more strongly predicted by a linear function of the other variable.
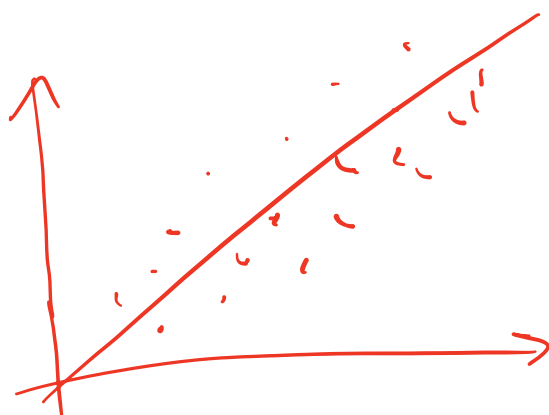
Correlation is a standardized version of the covariance.

Guess the correlation applet: https://www.geogebra.org/m/KE6JfuF9

Graphically understanding Covariance and Correlation

$(x - \mu_x)(y - \mu_y)$

$(-)(+)$ **−** $\quad$ $(+)(+)$ **+** $\quad$ ⎯ μ_y
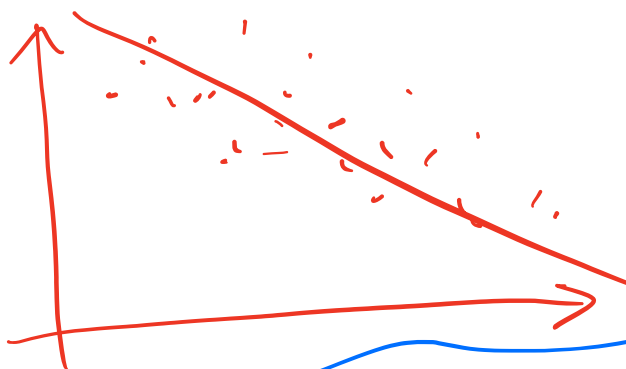
$(-)(-)$ **+** $\quad$ $(+)(-)$ **−**

μ_x
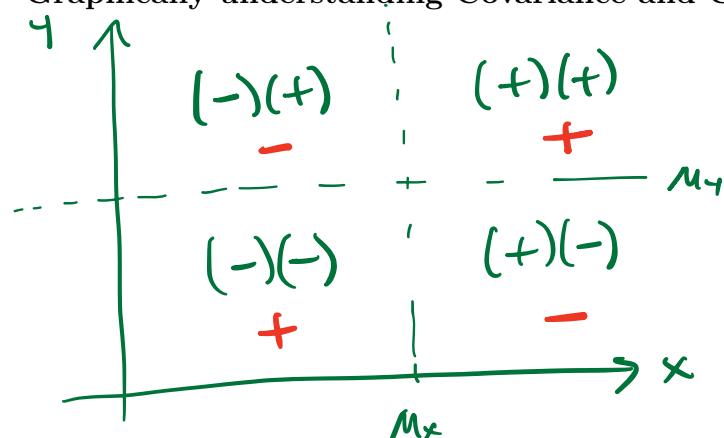
$Cov(X,Y) > 0$

as $X \uparrow Y \uparrow$
OR as $X \downarrow Y \downarrow$
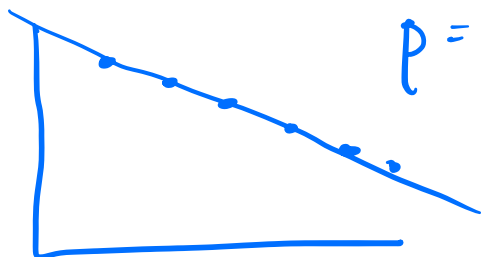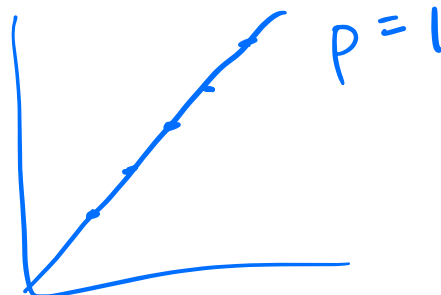move in same direction

$Cov(X,Y) < 0$
as $X \uparrow Y \downarrow$
OR as $X \downarrow Y \uparrow$
move in opposite directions

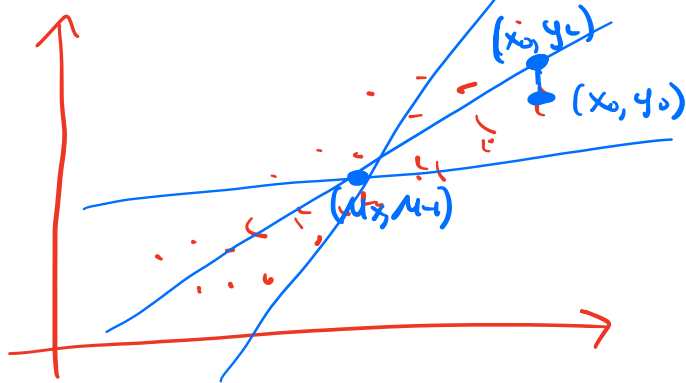(after next proof)

$\rho = -1$

$\rho = 1$

Relationship to regression ("best fit") line

criteria of "best" — must pass thru (μ_x, μ_y). but which one??

want small residuals, on average

$y_c - \mu_y = m(x_o - \mu_x)$ want to find slope m that minimizes

$(y_c - y_o)^2$ on $\underline{average}$ over all y_i

$\underset{E(\ )}{}$

minimize $\quad E\left((y_c - y_o)^2\right) = E\left[(\mu_y + m(X - \mu_x) - Y)^2\right]$

$$= E\left[(m(X - \mu_x) - (Y - \mu_y))^2\right]$$

$$= E\left[m^2(x - \mu_x)^2 - 2m(x - \mu_x)(y - \mu_y) + (y - \mu_y)^2\right]$$

$$= m^2 E\left((x - \mu_x)^2\right) - 2m E\left((x - \mu_x)(y - \mu_y)\right) + E\left((y - \mu_y)^2\right)$$

minimize $g(m) \quad = m^2 V(x) - 2m \, Cov(x,y) + V(y)$

$$g'(m) = 2m V(x) - 2 Cov(x,y) + 0 = 0$$

$$2m V(x) = 2 Cov(x,y)$$

$$m = \frac{2 Cov(x,y)}{2 V(x)} = \frac{\rho_{xy} \sigma_x \sigma_y}{\sigma_x^2} = \boxed{\frac{\rho_{xy} \sigma_y}{\sigma_x}}$$

notice $g(m) = E\left((y_c - y_o)^2\right) \geq 0$ b/c of sq. term

$$g\left(\frac{\rho_{xy}\sigma_y}{\sigma_x}\right) = \left(\frac{\rho_{xy}\sigma_y}{\sigma_x}\right)^2 \sigma_x^2 - 2\left(\frac{\rho_{xy}\sigma_y}{\sigma_x}\right)\rho_{xy}\sigma_x\sigma_y + \sigma_y^2 \geq 0$$

$$= \rho_{xy}^2 \sigma_y^2 - 2\rho_{xy}^2 \sigma_y^2 + \sigma_y^2 \geq 0$$

$$= \sigma_y^2 - \rho_{xy}^2 \sigma_y^2 \geq 0$$

$$\sigma_y^2 (1 - \rho_{xy}^2) \geq 0$$

$$1 \geq \rho_{xy}^2 \implies -1 \leq \rho_{xy} \leq 1$$

correlation restricted bet.
0 and 1

Some important properties of Covariance:

> **Shortcut formula:** $Cov(X,Y) = E(XY) - \mu_X \mu_Y$

> $Cov(X,X) = Var(X)$

> The variance of a sum of two random variables is equal to the sum of their variances and covariances:
> $$V(X+Y) = V(X) + V(Y) + 2Cov(X,Y)$$

> If two random variables are independent, then their covariance is zero. However, zero covariance does NOT imply independence.

Example: Roll a fair four sided die twice. Let X equal the outcome of the first roll and Y the sum of the two rolls. Determine the covariance and correlation coefficient.

d_1/d_2	1	2	3	4
1	(1,2)	(1,3)	(1,4)	(1,5)
2	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,4)	(3,5)	(3,6)	(3,7)
4	(4,5)	(4,6)	(4,7)	(4,8)

$Y^2 P(Y=y)$	Y	$P(Y=y)$	$Y \cdot P(Y=y)$
4/16	2	1/16	2/16
18/16	3	2/16	6/16
16.3/16	4	3/16	12/16
25.4/16	5	4/16	20/16
36.3/16	6	3/16	18/16
49.2/16	7	2/16	14/16
64/16	8	1/16	8/16

$440/16 = E(y^2)$

$80/16 = 5 = E(Y)$

$V(Y) = 27.5 - 5^2 = 2.5$

$Cov(X,Y) = E(XY) - E(X)E(Y)$

$= 13.75 - \frac{5}{2}(5)$

$= 1.25$

$Cor(x,y) = \frac{1.25}{\sqrt{1.25}\sqrt{2.5}} = .707$

$X \sim U(1,4)$

$E(X) = \frac{1+4}{2} = \frac{5}{2}$

$V(X) = \frac{(b-a+1)^2 - 1}{12}$

$= \frac{15}{12} = 1.25$

Need $E(X), E(Y), E(XY)$
$V(X), V(Y)$

$E(XY) = \sum\sum_{(x,y)} xy \, P(x,y)$

$= \frac{1}{16}\left(2 + 3 + 4 + 5 + 6 + 8 + 10 + 12 + 12 + 15 + 18 + 21 + 20 + 24 + 28 + 32\right)$

$- \frac{220}{16} = 13.75$

$M = \frac{Cor(x,y)\sqrt{V(Y)}}{\sqrt{V(X)}}$

$= \frac{1.25}{\sqrt{1.25}\sqrt{2.5}}\frac{\sqrt{2.5}}{\sqrt{1.25}} = 1$

19
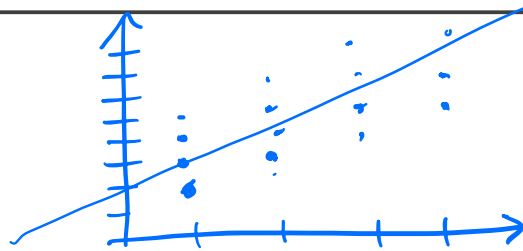
Chapter 5 Group Work

Problem 1

Suppose that you randomly select a student from a large population of 4th grade students and record the student's sex and number of siblings. Let $M = 1$ if the selected student is male and $M = 0$ if the selected student is female. Let S record the student's number of siblings. The proportion of the population of students falling into each category of M and S is recorded in the table below.

# of siblings:	0	1	2	3	4
Female	.10	.18	.12	.07	.04
Male	.12	.18	.14	.03	.02

    a. Verify that the above table is a valid joint pmf.
    b. Find the probability that the randomly selected student will be male with two siblings
    c. Find $P(M = 0, S \leq 2)$
    d. Find the probability that the randomly selected student will be an only child
    e. Find $F(1, 1)$

Problem 2

Of nine executives in a business firm, four are married, three have never married, and two are divorced. Three of the executives are to be selected for promotion. Let Y_1 denote the number of married executives and Y_2 denote the number of never married executives among the three selected for promotion. Assuming that the three are randomly selected from the nine available, find the joint pmf of Y_1 and Y_2.

Problem 3

Let Y_1 and Y_2 have the joint pdf

$$f(y_1, y_2) = \begin{cases} e^{-(y_1 + y_2)}, & y_1 > 0, y_2 > 0 \\ 0, & elsewhere \end{cases}$$

    a. What is $P(Y_1 < 1, Y_2 > 5)$?
    b. What is $P(Y_1 + Y_2 < 3)$?

Problem 4

Return to the scenario in Problem 1.

    a. Find the marginal distribution of M
    b. Find the marginal distribution of S
    c. Find $P(S \leq 2|M = 1)$
    d. Find $P(M = 1|S \leq 2)$

Problem 5

Let $f(x, y) = \frac{4}{3}(1 - xy), \quad 0 \leq x \leq 1, \quad 0 \leq y \leq 1.$

    a. Find the marginal pdfs of X and Y.
    b. Find $P(X \leq Y/2)$

Problem 6

Let the joint pmf of X and Y be

$$f(x, y) = \frac{xy^2}{30}, \quad x = 1, 2, 3, \quad y = 1, 2$$

a. What are the marginal distributions of X and Y?
b. Are X and Y independent?
c. What is $P(X > 1)$?
d. What is $E(X)$?

Problem 7

Let X and Y be two continuous random variables with joint pdf $f(x, y) = \frac{3}{16}xy^2$, $0 \le x \le 2$, $0 \le y \le 2$. Are the two random variables independent?

Problem 8

Let X and Y have the joint pdf $f(x, y) = x + y, 0 \le x \le 1, 0 \le y \le 1$. Find the marginal pdfs $f_X(x)$ and $f_Y(y)$ and show that X and Y are dependent.

Problem 9

Show that $Cov(X, Y) = E(XY) - \mu_X\mu_Y$.

    Hint: start with the definition $Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$

Problem 10

Show that $E(XY) = \mu_X\mu_Y + \rho_{XY}\sigma_X\sigma_Y$

Problem 11

Show that if two random variables X and Y are independent, then their covariance is 0.

    Hint: start with the shortcut formula $Cov(X, Y) = E(XY) - \mu_X\mu_Y$ and re-write $E(XY)$ as a double sum.

Problem 12

Return to the scenario in Problem 1.

a. Are M and S independent?
b. Find $E(M)$ and $V(M)$
c. Find $E(S)$ and $V(S)$
d. Find $Cov(M, S)$
e. Suppose that instead of just one student, you sample 12 students. Let $S_1, \ldots S_{12}$ represent the number of siblings reported by each student (you can assume that these random variables are independent). Let $\overline{S} = \sum_{i=1}^{12} \frac{S_i}{12}$ be the average number of siblings in the sample. Find $E(\overline{S})$ and $V(\overline{S})$