Mathematics/Statistics Bootcamp

Part IV: Probability

Steven Winter Christine Shen

Department of Statistical Science
Duke University

MSS Orientation, August 2022

Outline

Probability

Basic Probability

Axioms:

1. For any event A, $\mathbb{P}(A) \in [0, 1]$;
2. $\mathbb{P}(\Omega) = 1$, where Ω is the sample space.
3. If $A_1, A_2, \ldots$ are disjoint events, then

$$\mathbb{P}\left(\bigcup_i A_i\right) = \sum_{i=1} \mathbb{P}(A_i).$$

Useful consequences (and good exercises):

1. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
2. $\mathbb{P}(\emptyset) = 0$.
3. If $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.
4. For any events $A_1, A_2, \ldots$, we have the bound $\mathbb{P}(\cup_i A_i) \leq \sum_i \mathbb{P}(A_i)$.
5. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Independence

We say events A and B are **independent** (denoted $A \perp B$) if:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

A collection of events $A_1, \ldots, A_n$ is (mutually) **independent** if for *any* sub-collection $A_{i_1}, \ldots, A_{i_K}$:

$$\mathbb{P}\left(\bigcap_{j=1}^{K} A_{i_j}\right) = \prod_{j=1}^{K} \mathbb{P}(A_{i_j}).$$

Dice Example Revisited

Consider tossing two dice. The sample space is

$$\begin{aligned}\Omega &= \{(1,1),(1,2),\dots(1,6),(2,1),\dots,(2,6),\dots,(6,6)\}\\ &= \{(i,j) \mid i,j \in \{1,...,6\}\}.\end{aligned}$$

Define

$$\begin{aligned}A &= \{(i,j) \mid i = 4\}\\ B &= \{(i,j) \mid j \in \{1,6\}\}\\ C &= \{(i,j) \mid i + j = 7\}.\end{aligned}$$

where always $i,j \in \{1,...,6\}$. Are A, B, C independent?

Solution

By counting:

$$\mathbb{P}(A) = \frac{6}{36} = \frac{1}{6}; \quad \mathbb{P}(B) = \frac{12}{36} = \frac{1}{3}; \quad \mathbb{P}(C) = \frac{6}{36} = \frac{1}{6}.$$

Check $A \perp B$:

$$\mathbb{P}(A \cap B) = \mathbb{P}(i = 4 \text{ and } j \in \{1, 6\}) = \frac{2}{36} = \frac{1}{18}$$
$$\mathbb{P}(A)\mathbb{P}(B) = \frac{1}{6} \times \frac{1}{3} = \frac{1}{18}$$

Check $A \perp C$:

$$\mathbb{P}(A \cap C) = \mathbb{P}(i = 4 \text{ and } i + j = 7) = \frac{1}{36}$$
$$\mathbb{P}(A)\mathbb{P}(C) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

Solution

Check $B \perp C$:

$$\mathbb{P}(B \cap C) = \mathbb{P}(j \in \{1, 6\} \text{ and } i + j = 7) = \frac{2}{36} = \frac{1}{18}$$
$$\mathbb{P}(B)\mathbb{P}(C) = \frac{1}{3} \times \frac{1}{6} = \frac{1}{18}$$

Check mutual independence:

$$\mathbb{P}(A \cap B \cap C) = 0$$
$$\mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C) = \frac{1}{6} \times \frac{1}{3} \times \frac{1}{6} = \frac{1}{108} \neq 0$$

Not independent! Remember to check every sub-collection of events.

Discussion

Let A, B and C be events.

1. If $A \perp A$, what do we know about A?
2. If $A \perp B$, is $A \perp B^c$?
3. If $A \perp B$, and $B \perp C$, is $A \perp C$?

Solutions

1. If $A \perp A$, $\mathbb{P}[A] = \mathbb{P}[A \cap A] = \mathbb{P}[A]^2$. Therefore $\mathbb{P}[A] = 0$, or 1.

2. If $A \perp B$, then

$$\begin{aligned}
\mathbb{P}[A \cap B^c] &= \mathbb{P}[A] - \mathbb{P}[A \cap B] \\
&= \mathbb{P}[A] - \mathbb{P}[A]\mathbb{P}[B] \\
&= \mathbb{P}[A](1 - \mathbb{P}[B]) \\
&= \mathbb{P}[A]\mathbb{P}[B^c].
\end{aligned}$$

   Therefore A is independent of B^c.

Solutions

3. Not necessarily. Let $U \sim \mathrm{Unif}[0,1]$,

$$A = [U \leq 1/2], \quad B = [U \leq 1/4] \cup [1/2 < U \leq 3/4]$$
$$C = [U \leq 1/8] \cup [5/8 < U \leq 1].$$

Then

$$\mathbb{P}[A] = \mathbb{P}[B] = \mathbb{P}[C] = 1/2$$
$$\mathbb{P}[A \cap B] = 1/4 = \mathbb{P}[A]\mathbb{P}[B]$$
$$\mathbb{P}[B \cap C] = 1/4 = \mathbb{P}[B]\mathbb{P}[C].$$

But

$$\mathbb{P}[A \cap C] = 1/8 \neq \mathbb{P}[A]\mathbb{P}[C].$$

Conditional Probability

Let A, B, C be events with $\mathbb{P}(B) > 0$. The **conditional probability** of event A given B is

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

We say A and B are **conditionally independent** given C if $\mathbb{P}(C) > 0$, and

$$\mathbb{P}(A \cap B \mid C) = \mathbb{P}(A \mid C)\mathbb{P}(B \mid C).$$

Usually write $A \perp B \mid C$. Will see examples using random variables.

Exercise: prove $A \perp B \mid C$ if and only if

$$\mathbb{P}(A \mid B, C) = \mathbb{P}(A \mid C).$$

Solution

Notice that

$$\mathbb{P}(A \cap B \mid C) = \mathbb{P}(A \mid C)\mathbb{P}(B \mid C)$$

$$\text{iff } \frac{\mathbb{P}(A \cap B \cap C)}{\mathbb{P}(C)} = \frac{\mathbb{P}(A \cap C)}{\mathbb{P}(C)} \frac{\mathbb{P}(B \cap C)}{\mathbb{P}(C)}$$

$$\text{iff } \mathbb{P}(A \cap B \cap C) = \frac{\mathbb{P}(A \cap C)\mathbb{P}(B \cap C)}{\mathbb{P}(C)}$$

$$\text{iff } \frac{\mathbb{P}(A \cap B \cap C)}{\mathbb{P}(B \cap C)} = \frac{\mathbb{P}(A \cap C)}{\mathbb{P}(C)}$$

$$\text{iff } \mathbb{P}(A \mid B, C) = \mathbb{P}(A \mid C).$$

Law of Total Probability and Bayes Rule

A countable collection of events $\{A_1, A_2, \ldots\}$ is a **partition** if $A_i \cap A_j = \emptyset$ for $i \neq j$, and $\cup_j A_j = \Omega$.

Law of Total Probability: for any event B and partition $\{A_j\}$,

$$\mathbb{P}(B) = \sum_j \mathbb{P}(B \mid A_j)\mathbb{P}(A_j).$$

Bayes Rule: for any events A, B with $\mathbb{P}(B) > 0$

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

Bayes Rule Example

Assume we know the following about a specific disease, D:

- the probability of being sick (having the disease) is 0.01,
- the probability of testing positive if sick is 0.95,
- the probability of testing negative if healthy is 0.95.

What is the probability of being sick if the test is positive?

Solution

First find the probability of a positive test using the law of total probability:

$$\begin{aligned}\mathbb{P}(+) &= \mathbb{P}(+|D)P(D) + \mathbb{P}(+|ND)P(ND) \\ &= 0.95 \times 0.01 + (1 - 0.95) \times (1 - 0.01) = 0.059.\end{aligned}$$

Now apply Bayes Rule:

$$\mathbb{P}(D|+) = \frac{\mathbb{P}(+|D)\mathbb{P}(D)}{\mathbb{P}(+)} = \frac{0.95 \times 0.01}{0.059} \approx 0.161$$

Relatively small probability of having the disease given a positive test.

Exercises

1. Consider all length 3 strings constructable from $\{a, b, c\}$:

$$\Omega = \{aaa, bbb, ccc, abc, bca, cba, acb, bac, cab\}.$$

   Assign each string probability $\frac{1}{9}$. For $i = 1, 2, 3$, define A_i as:

$$A_i = \{i^{th}\text{place in the triple is occupied by a}\}.$$

   Are the A_i independent? Prove/disprove.

2. Fix $r, b, c \in \mathbb{N}_+$. An urn starts with r red balls and b blue balls. A ball is drawn uniformly at random and its color is recorded. The ball is then *added back* to the urn along with c balls of the same color. This process is iterated.
   a) How many balls are in the urn right before the nth draw?
   b) Show the probability that the second draw is red is $r/(r + b)$.
   c) Find the probability that the first draw was blue, given that the second draw was red.

Solutions

1. Pairwise independence is satisfied:

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1 \cap A_3) = \mathbb{P}(A_2 \cap A_3) = \frac{1}{9}.$$

   But the joint event:

$$\mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(\{aaa\}) = \frac{1}{9} \neq \mathbb{P}(A_1)P(A_2)\mathbb{P}(A_3).$$

   Hence, the events are **not** mutually independent

Solutions

2. a) At each stage we add c balls; at stage n we have $r+b+(n-1)c$ balls.

   b) Let R_n be the event the nth ball drawn is red; similarly for B_n. By the law of total probability,

$$\begin{aligned} P[R_2] &= P[R_2|R_1]P[R_1] + P[R_2|B_1]P[B_1] \\ &= \left(\frac{r+c}{r+b+c}\right)\left(\frac{r}{r+b}\right) + \left(\frac{r}{r+b+c}\right)\left(\frac{b}{r+b}\right) \\ &= \frac{r}{r+b}. \end{aligned}$$

2. c) By Bayes' rule,

$$P[B_1|R_2] = \frac{P[R_2|B_1]P[B_1]}{P[R_2]} = \frac{\left(\frac{r}{r+b+c}\right)\left(\frac{b}{r+b}\right)}{\left(\frac{r}{r+b}\right)} = \frac{b}{r+b+c}.$$

Multivariate Distributions

Distribution Functions for Multivariate Random Variables

We will cover:

- Joint Distribution
- Marginal Distribution
- Conditional Distribution

Joint Distribution

Joint PDF: A function $f(x_1, \ldots, x_n)$ from $\mathbb{R}^n \to \mathbb{R}$ is called a joint PDF of the random vector $\mathbf{X} = (X_1, \ldots, X_n)$ if for every $A \subset \mathbb{R}^n$,

$$\mathbb{P}(\mathbf{X} \in A) = \int_A f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \mathrm{d}(x_1, \ldots, x_n).$$

Joint PMF: Let R_{X_i} denote the range of discrete variable X_i, $R_{\mathbf{X}} = R_{X_1} \times \cdots \times R_{X_n}$. Let

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$$

be the joint PMF of $\mathbf{X} = (X_1, \ldots, X_n)$. Then for every $A \subset \mathbb{R}^n$,

$$\mathbb{P}(\mathbf{X} \in A) = \sum_{(x_1,\ldots,x_n) \in (A \cap R_{\mathbf{X}})} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n).$$

Marginal Distribution

Given the joint PDF/ PMF, we can find the marginal PDF/ PMF:

Marginal PDF:

$$f_{X_1}(x_1) = \int_{X_2,\ldots,X_n} f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)\mathrm{d}(x_2\ldots x_n).$$

Marginal PMF:

$$f_{X_1}(x_1) = \sum_{(x_2,\ldots,x_n)\in(R_{X_2}\times\cdots\times R_{X_n})} f_{X_1,\ldots,X_n}(x_1,\ldots,x_n).$$

Joint Distribution - Exercise

1. Assume that X and Y have the joint PDF:

$$f_{X,Y}(x,y) = 4xy, \quad 0 < x < 1 \quad 0 < y < 1.$$

   Find $\mathbb{P}(Y < X)$.

2. Random variables X and Y are jointly normal with mean $(\mu_x, \mu_y)^T$ and covariance matrix

$$\begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}.$$

   Find $\mathbb{P}(Y < X)$. Think about what happens if $\mu_x \to \infty$? What about limiting cases of other parameters?
   Hint:
   - What's the distribution of $Y - X$?
   - $\mathbb{V}[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y] + 2Cov(X, Y)$.

Joint Distribution - Exercise Solution

1. We can set up the double integral required for this probability as follows:

$$
\begin{aligned}
\mathbb{P}(Y < X) &= \int_0^1 \int_0^x 4xy \mathrm{d}y \mathrm{d}x \\
&= \int_0^1 [4x\frac{y^2}{2}]\Big|_0^x \mathrm{d}x \\
&= \int_0^1 2x^3 \mathrm{d}x = \frac{1}{2}.
\end{aligned}
$$

Joint Distribution - Exercise Solution

2. We can solve this via a similar approach as the last question, i.e., first identify the joint PDF of X and Y, then set up a double integral.
However, an easier way is to notice that X and Y are jointly normal. Therefore $Y - X$ follows univariate normal with:

$$\mathbb{E}[Y - X] = \mathbb{E}[Y] - \mathbb{E}[X] = \mu_y - \mu_x$$
$$\mathbb{V}[Y - X] = \sigma_y^2 + \sigma_x^2 - 2\rho\sigma_x\sigma_y.$$

Therefore

$$\begin{aligned}\mathbb{P}(Y < X) &= \mathbb{P}(Y - X < 0)\\ &= \Phi\left(\frac{\mu_x - \mu_y}{\sqrt{\sigma_y^2 + \sigma_x^2 - 2\rho\sigma_x\sigma_y}}\right),\end{aligned}$$

where $\Phi(\cdot)$ denotes the CDF for standard normal.

Conditional Distribution

Let X, Y be random variables with joint PDF/ PMF $f_{X,Y}(x,y)$. The **conditional PDF/ PMF** of X given $Y = y$ is:

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

For discrete random variables, this is intuitive:

$$f_{X|Y}(x \mid y) = \mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

How to understand it for continuous random variables? What's $\mathbb{P}(Y = y)$ for a continuous random variable Y?

Conditional Distribution

For continuous random variables X and Y, $\mathbb{P}(Y = y) = 0$, and

$$f_{X|Y}(x \mid y) \neq \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}.$$

Instead, $f_{X|Y}(x \mid y)$ is the function such that

$$\int_B f_{X|Y}(x \mid y) dx = \mathbb{P}(X \in B \mid Y = y), \quad \text{and}$$

$$\mathbb{P}(X \in B \mid Y = y) = \lim_{\epsilon \to 0} \mathbb{P}(X \in B \mid Y \in (y, y + \epsilon)).$$

Conditional Distribution - Exercise

1. Assume that (X, Y) is a continuous random vector with joint pdf given by:

$$f_{X,Y}(x, y) = e^{-y}, \quad 0 < x < y < \infty.$$

   Find the marginal distribution of X, and the conditional distribution $Y|X$.

2. Let $Y \sim N(\mu, \sigma^2)$ with known μ and σ^2. Find the PDF for $Y \mid Y \geq c$, for some $c \in \mathbb{R}$.

   Bonus: Generalize this to a standard multi-variate normal, $\mathbf{Z} \sim N_n(\mathbf{0}, \mathbf{I})$, by finding the PDF for $\mathbf{Z} \mid \mathbf{Z} \in \mathbb{R}^n_+$. What happens in high dimensions (when $n \to \infty$)?

Conditional Distribution - Exercise Solution

1. We start by finding the marginal distribution of X:

$$f_X(x) = \int_x^{\infty} e^{-y} \mathrm{d}y = e^{-x}$$
$$X \sim \text{Exponential}(1).$$

Now by the definition of conditional distributions given earlier:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{e^{-y}}{e^{-x}} \mathbb{I}(x < y).$$

Conditional Distribution - Exercise Solution

2. We start by finding the CDF of $Y \mid Y \geq c$ using conditional probability.

$$\mathbb{P}(Y \leq y \mid Y \geq c) = \frac{\mathbb{P}(Y \leq y \cap Y \geq c)}{\mathbb{P}(Y \geq c)}$$
$$= \begin{cases} \frac{\mathbb{P}(c \leq Y \leq y)}{\mathbb{P}(Y \geq c)} = \frac{\Phi((y-\mu)/\sigma) - \Phi((c-\mu)/\sigma)}{1 - \Phi((c-\mu)/\sigma)} & \text{if } c \leq y \\ 0 & \text{otherwise.} \end{cases}$$

Therefore the PDF can be derived by taking derivative w.r.t. y

$$f_{Y|Y \geq c}(y) = \frac{\mathrm{d}\mathbb{P}(Y \leq y \mid Y \geq c)}{\mathrm{d}y}$$
$$= \begin{cases} \frac{\phi((y-\mu)/\sigma)}{1 - \Phi((c-\mu)/\sigma)} & \text{if } c \leq y \\ 0 & \text{otherwise.} \end{cases}$$

Here $\Phi(\cdot)$ and $\phi(\cdot)$ denote standard normal CDF and PDF.

Conditional Distribution - Exercise Solution

Note that here we are conditioning on an event, or equivalently on an indicator function. Let $X = \mathbb{1}[Y \geq c]$,

$$Y \mid Y \geq c \stackrel{d}{=} Y \mid X = 1.$$

In the multi-variate case, the PDF is

$$\begin{aligned} f_{\mathbf{Z}|\mathbf{Z}\in\mathbb{R}_+^n}(\mathbf{z}) &= \prod_{i=1}^{n} f_{z_i|z_i\geq 0}(z_i) = \prod_{i=1}^{n} \frac{\phi(z_i)}{1-\Phi(0)} \mathbb{1}[z_i \geq 0] \\ &= \frac{1}{\Phi(0)^n} \left(\prod_{i=1}^{n} \phi(z_i) \right) \mathbb{1}[\mathbf{Z} \in \mathbb{R}_+^n]. \end{aligned}$$

As $n \to \infty$, the relative volume of its support vanishes. Creates computational difficulties!

Conditional Independence

Let A, B and C be events. Recall that A and B are said to be **conditionally independent** given C if and only if $\mathbb{P}(C) > 0$, and

$$\mathbb{P}(A \cap B \mid C) = \mathbb{P}(A \mid C)\mathbb{P}(B \mid C).$$

Usually written as $A \perp B \mid C$.

Conditional Independence

Similarly, random variables X and Y are **conditionally independent** given random variable Z if and only if

$$f_{X,Y|Z=z}(x, y) = f_{X|Z=z}(x)f_{Y|Z=z}(y),$$

where $f_{\cdot|Z}(\cdot)$ is the conditional PDF/ PMF given Z.

Usually we denote as $X \perp Y \mid Z$.

Conditional Independence - Example

Suppose we have three discrete random variables Y_1, Y_2, Y_3 that we believe are "independent and identically distributed (i.i.d.)". Does our knowledge about the value of one inform about another? That is:

$$\mathbb{P}(Y_1 = y_1 \mid Y_2 = y_2, Y_3 = y_3) = \mathbb{P}(Y_1 = y_1)?$$

What if Y_1, Y_2, Y_3 are conditionally independent given discrete random variable Θ?

Conditional Independence - Example Solution

If Y_1, Y_2, Y_3 are only conditionally independent, the following equation

$$\mathbb{P}(Y_1 = y_1 \mid Y_2 = y_2, Y_3 = y_3) = \mathbb{P}(Y_1 = y_1)$$

no longer holds. Instead, we have

$$\mathbb{P}(Y_1 = y_1 \mid \Theta = \theta, Y_2 = y_2, Y_3 = y_3) = \mathbb{P}(Y_1 = y_1 \mid \Theta = \theta).$$

Or alternatively,

$$\begin{aligned}\mathbb{P}(Y_1, Y_2, Y_3 \mid \Theta = \theta) =& \mathbb{P}(Y_1 \mid \Theta = \theta) \\ & \mathbb{P}(Y_2 \mid \Theta = \theta)\mathbb{P}(Y_3 \mid \Theta = \theta).\end{aligned}$$

Moments

Expectation of Random Variables

Let X be an integrable random variable,[1] $f_X(x)$ be its PDF/PMF, and $g : \mathbb{R} \to \mathbb{R}$ be any real function. The expectation of $g(X)$ is:

- if X is continuous,

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) \mathrm{d}x.$$

- if X is discrete, let $\mathcal{X}$ denote its range,

$$\mathbb{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x) f_X(x) = \sum_{x \in \mathcal{X}} g(x) \mathbb{P}(X = x).$$

Setting $g(X) = X$ gives $\mathbb{E}[X]$, the expectation of X.

[1] i.e., expectation of X exists. Counter-example: expectation of a Cauchy random variable is undefined.

Variance and Covariance of Random Variables

Let X, Y be square integrable random variables.[2] Variance of X is defined as

$$\begin{aligned}\mathbb{V}[X] &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2.\end{aligned}$$

Covariance between X and Y is defined as

$$\begin{aligned}Cov(X, Y) &= \mathbb{E}[X - \mathbb{E}(X)]\mathbb{E}[Y - \mathbb{E}(Y)] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].\end{aligned}$$

[2] i.e., both expectation and variance exist

Expectation and Variance - Exercise

$X \sim \mathrm{Poisson}(\lambda)$. Show that $\mathbb{E}[X] = \lambda$.

Expectation and Variance - Exercise Solution

We need to compute:

$$\mathbb{E}[X] = \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!}$$
$$= \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!}.$$

Recall the following results from Taylor series expansion:

$$e^y = \sum_{i=0}^{\infty} \frac{y^i}{i!}.$$

Therefore we have:

$$\mathbb{E}[X] = \lambda e^{\lambda} e^{-\lambda} = \lambda.$$

Properties of Expectation

Let

- X, Y be integrable random variables
- $a \in \mathbb{R}$ be a scalar constant
- f and $g : \mathbb{R} \to \mathbb{R}$ be functions such that $f(X)$ and $g(X)$ are integrable

Basic properties of Expectation:

1. Linearity
   - $\mathbb{E}[aX] = a\mathbb{E}[X]$
   - $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
2. Monotonicity
   - $f \leq g \implies \mathbb{E}[f(X)] \leq \mathbb{E}[g(X)]$, or equivalently,
   - $X \leq Y$ with probability 1 $\implies \mathbb{E}[X] \leq \mathbb{E}[Y]$

Jensen's Inequality

Convex function

- A function $\psi : \mathcal{X} \to \mathbb{R}$ is convex iff for all $t \in [0, 1]$, $x_1, x_2 \in \mathcal{X}$,

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2).$$

  It is strictly convex if for any $x_1 \neq x_2$, the inequality is strict.
- Any twice differentiable function ψ is convex iff its second derivative is non-negative. It is strictly convex if its second derivative is positive.

By **Jensen's inequality**, for any integrable random variable X, and convex function ψ,

$$\psi(\mathbb{E}[X]) \leq \mathbb{E}[\psi(X)].$$

Inequality is strict if ψ is strictly convex and X is non-degenerate.

Jensen's Inequality - Optional Example

Let $\|X\|_p = \mathbb{E}[X^p]^{1/p}$ denote the L_p norm of a random variable X.

For $0 < p < q < \infty$, let X be a random variable such that X^q is integrable. Use Jensen's inequality to show

$$\|X\|_p \leq \|X\|_q.$$

Jensen's Inequality - Optional Example Solution

Because $0 < p < q < \infty$, notice that $q/p > 1$, and hence

$$\psi(x) = x^{q/p}$$

is a convex function. Therefore applying Jensen's inequality on $|X|^p$,

$$\psi(\mathbb{E}[|X|^p]) = (\mathbb{E}[|X|^p])^{q/p} \leq \mathbb{E}[\psi(|X|^p)] = \mathbb{E}[|X|^{pq/p}].$$

That is,

$$(\|X\|_p)^q \leq (\|X\|_q)^q$$
$$\|X\|_p \leq \|X\|_q.$$

Cauchy-Schwartz and Hölder's Inequalities

Cauchy-Schwartz inequality

For any square integrable random variables X and Y,

$$\mathbb{E}[XY] \leq \mathbb{E}[|XY|] \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}.$$

Cauchy-Schwartz is a special case of **Hölder's inequality**

For $r \geq 1$, p, $q > 1$ with $1/p + 1/q = 1/r$,

$$\|XY\|_r \leq \|X\|_p \|Y\|_q.$$

Expectation - Example

1. Let $\mathbf{A}$ be an $n \times n$ random matrix, show

$$\mathbb{E}[\mathrm{Tr}(\mathbf{A})] = \mathrm{Tr}(\mathbb{E}[\mathbf{A}]).$$

Expectation - Example

1. Let $\mathbf{A}$ be an $n \times n$ random matrix, show

$$\mathbb{E}[\mathrm{Tr}(\mathbf{A})] = \mathrm{Tr}(\mathbb{E}[\mathbf{A}]).$$

Proof:

$$\begin{aligned}
\mathbb{E}[\mathrm{Tr}(\mathbf{A})] &= \mathbb{E}\left[\sum_{i=1}^{n} a_{ii}\right] = \sum_{i=1}^{n} \mathbb{E}[a_{ii}] \\
&= \mathrm{Tr}\left(\begin{pmatrix} \mathbb{E}[a_{11}] & \dots & \mathbb{E}[a_{1n}] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[a_{n1}] & \cdots & \mathbb{E}[a_{nn}] \end{pmatrix}\right) \\
&= \mathrm{Tr}(\mathbb{E}[\mathbf{A}]).
\end{aligned}$$

Expectation - Example Cont.

2. Consider a random vector $\mathbf{Y} \in \mathbb{R}^n$ with $\mathbb{E}[\mathbf{Y}] = \boldsymbol{\mu}$, and $\mathbb{V}[\mathbf{Y}] = \boldsymbol{\Sigma}$. Show that for any fixed matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$,

$$\mathbb{E}[\mathbf{Y}^T \mathbf{A} \mathbf{Y}] = \boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\mu} + \text{Tr}(\mathbf{A} \boldsymbol{\Sigma}).$$

Expectation - Example Cont.

2. Consider a random vector $\mathbf{Y} \in \mathbb{R}^n$ with $\mathbb{E}[\mathbf{Y}] = \boldsymbol{\mu}$, and $\mathbb{V}[\mathbf{Y}] = \boldsymbol{\Sigma}$. Show that for any fixed matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$,

$$\mathbb{E}[\mathbf{Y}^T \mathbf{A} \mathbf{Y}] = \boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\mu} + \mathrm{Tr}(\mathbf{A}\boldsymbol{\Sigma}).$$

Proof: Notice

$$\begin{aligned} \mathbf{Y}^T \mathbf{A} \mathbf{Y} =& [\boldsymbol{\mu} + (\mathbf{Y} - \boldsymbol{\mu})]^T \mathbf{A} [\boldsymbol{\mu} + (\mathbf{Y} - \boldsymbol{\mu})] \\ =& \boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\mu} + (\mathbf{Y} - \boldsymbol{\mu})^T \mathbf{A} \boldsymbol{\mu} + \boldsymbol{\mu}^T \mathbf{A} (\mathbf{Y} - \boldsymbol{\mu}) \\ & + (\mathbf{Y} - \boldsymbol{\mu})^T \mathbf{A} (\mathbf{Y} - \boldsymbol{\mu}). \end{aligned}$$

Taking expectation on both sides, the first term on the RHS is a constant, the middle two terms become zero. For the last term, we can apply the trace trick.

Expectation - Example Cont.

2. Notice that $(\mathbf{Y}-\mu)^T\mathbf{A}(\mathbf{Y}-\mu)$ is a scalar, therefore

$$\begin{aligned}
&\mathbb{E}[(\mathbf{Y}-\mu)^T\mathbf{A}(\mathbf{Y}-\mu)] \\
=&\mathbb{E}[\mathrm{Tr}[(\mathbf{Y}-\mu)^T\mathbf{A}(\mathbf{Y}-\mu)]] \\
=&\mathbb{E}[\mathrm{Tr}[\mathbf{A}(\mathbf{Y}-\mu)(\mathbf{Y}-\mu)^T]] \\
=&\mathrm{Tr}[\mathbb{E}[\mathbf{A}(\mathbf{Y}-\mu)(\mathbf{Y}-\mu)^T]] \\
=&\mathrm{Tr}[\mathbf{A}\mathbb{E}[(\mathbf{Y}-\mu)(\mathbf{Y}-\mu)^T]] \\
=&\mathrm{Tr}[\mathbf{A}\mathbf{\Sigma}].
\end{aligned}$$

Together with previous results, we have

$$\mathbb{E}[\mathbf{Y}^T\mathbf{A}\mathbf{Y}] = \mu^T\mathbf{A}\mu + \mathrm{Tr}(\mathbf{A}\mathbf{\Sigma}).$$

Properties of Variance

Let X, Y be square integrable random variables, $a, b \in \mathbb{R}$ be scalar constants.

Basic properties of Variance:

1. $\mathbb{V}[X] \geq 0$
2. $\mathbb{V}[X + a] = \mathbb{V}[X]$
3. $\mathbb{V}[aX] = a^2\mathbb{V}[X]$
4. $\mathbb{V}[aX \mp bY] = a^2\mathbb{V}[X] + b^2\mathbb{V}[Y] \mp 2abCov(X, Y)$

Properties of Covariance

Let X, Y, W, V be square integrable random variables, $a, b, c, d \in \mathbb{R}$ be scalar constants.

Basic properties of Covariance:

1. $Cov(X, a) = 0$
2. $Cov(X, X) = \mathbb{V}[X]$
3. $Cov(X, Y) = Cov(Y, X)$
4. Bilinearity

$$\begin{aligned} Cov(aX + bY, cW + dV) = & acCov(X, W) + adCov(X, V) + \\ & bcCov(Y, W) + bdCov(Y, V) \end{aligned}$$

Expectation, Variance and Covariance - Example

Assume

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} \right).$$

We know that the conditional distribution of $X \mid Y$ is also normal. Find its mean and variance.

Expectation, Variance and Covariance - Example Solution

One option is to follow the standard approach:

1. find the joint PDF of X and Y
2. find the marginal PDF of Y
3. find the conditional PDF of $X \mid Y$ from 1 and 2, *complete the square* to identify the mean and variance parameters.

We will not go through the details, but make sure you understand this and are able to derive it.

Expectation, Variance and Covariance - Example Solution

Here we present another approach:

1. First notice that $X - \sigma_{12}/\sigma_{22} Y$ and Y are uncorrelated, and hence independent.

$$
\begin{aligned}
Cov(X - \frac{\sigma_{12}}{\sigma_{22}} Y, Y) &= Cov(X, Y) - \frac{\sigma_{12}}{\sigma_{22}} Cov(Y, Y) \\
&= \sigma_{12} - \frac{\sigma_{12}}{\sigma_{22}} \sigma_{22} = 0
\end{aligned}
$$

2. Rewrite $X \mid Y$ as

$$
X - \frac{\sigma_{12}}{\sigma_{22}} Y + \frac{\sigma_{12}}{\sigma_{22}} Y \mid Y,
$$

and we have

$$
\begin{aligned}
\mathbb{E}[X \mid Y = y] =& \mathbb{E}[X - \frac{\sigma_{12}}{\sigma_{22}} Y + \frac{\sigma_{12}}{\sigma_{22}} Y \mid Y = y] \\
=& \mathbb{E}[X - \frac{\sigma_{12}}{\sigma_{22}} Y \mid Y = y] + \mathbb{E}[\frac{\sigma_{12}}{\sigma_{22}} Y \mid Y = y] \\
=& \mathbb{E}[X - \frac{\sigma_{12}}{\sigma_{22}} Y] + \frac{\sigma_{12}}{\sigma_{22}} y = \mu_X + \frac{\sigma_{12}}{\sigma_{22}} (y - \mu_Y).
\end{aligned}
$$

Expectation, Variance and Covariance - Example Solution

3. Similarly for variance

$$
\begin{aligned}
&\mathbb{V}[X \mid Y = y] \\
=&\mathbb{V}[X - \frac{\sigma_{12}}{\sigma_{22}} Y + \frac{\sigma_{12}}{\sigma_{22}} Y \mid Y = y] \\
=&\mathbb{V}[X - \frac{\sigma_{12}}{\sigma_{22}} Y \mid Y = y] + \mathbb{V}[\frac{\sigma_{12}}{\sigma_{22}} Y \mid Y = y] \\
=&\mathbb{V}[X - \frac{\sigma_{12}}{\sigma_{22}} Y] \\
=&\mathbb{V}[X] + \mathbb{V}[\frac{\sigma_{12}}{\sigma_{22}} Y] - 2Cov(X, \frac{\sigma_{12}}{\sigma_{22}} Y) \\
=&\sigma_{11} + \frac{\sigma_{12}^2}{\sigma_{22}^2}\sigma_{22} - 2\frac{\sigma_{12}}{\sigma_{22}}\sigma_{12} = \sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}}.
\end{aligned}
$$

Try to apply this to multivariate normal and check your results with this link.

Laws of Total Expectation and Total Variance

Let X, Y be square integrable random variables.

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$$
$$\mathbb{V}[Y] = \mathbb{V}[\mathbb{E}[Y|X]] + \mathbb{E}[\mathbb{V}[Y|X]]$$

Laws of Total Expectation and Total Variance - Example

Consider

$$\begin{aligned} X|N &\sim \mathrm{Binomial}(N, p) \\ N &\sim \mathrm{Negative\ Binomial}\left(\tau, r\right). \end{aligned}$$

Find $\mathbb{E}[X]$ and $\mathbb{V}[X]$.

Hint:

$$\mathbb{E}[N] = \frac{\tau r}{1-\tau}, \quad \mathbb{V}[N] = \frac{\tau r}{(1-\tau)^2}.$$

Laws of Total Expectation and Total Variance - Example Solution

First, we apply the law of total expectation.

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[\mathbb{E}[X|N]] \\ &= \mathbb{E}[Np] \\ &= p\frac{\tau r}{1-\tau}.\end{aligned}$$

Next, we apply the law of total variance.

$$\begin{aligned}\mathbb{V}[X] &= \mathbb{E}[\mathbb{V}[X|N]] + \mathbb{V}[\mathbb{E}[X|N]] \\ &= \mathbb{E}[Np(1-p)] + \mathbb{V}[Np] \\ &= p(1-p)\frac{\tau r}{1-\tau} + p^2\frac{\tau r}{(1-\tau)^2}.\end{aligned}$$

Laws of Total Expectation and Total Variance - Exercise

Consider

$$X|P \sim \text{Binomial}(n, P)$$
$$P \sim \text{Beta}(a, b).$$

Find $\mathbb{E}[X]$ and $\mathbb{V}[X]$.

Hint:

$$\mathbb{E}[P] = \frac{a}{a+b}$$
$$\mathbb{V}[P] = \frac{ab}{(a+b)^2(a+b+1)}.$$

Laws of Total Expectation and Total Variance - Exercise Solution

Again start with the marginal expectation:

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|P]] = \mathbb{E}[nP] = n\mathbb{E}[P] = n\frac{a}{a+b}.$$

Then the marginal variance:

$$\begin{aligned}
\mathbb{V}[X] &= \mathbb{V}[\mathbb{E}[X|P]] + \mathbb{E}[\mathbb{V}[X|P]] \\
&= \mathbb{V}[nP] + \mathbb{E}[nP(1-P)] \\
&= n^2\mathbb{V}[P] + n\mathbb{E}[P - P^2] \\
&= n^2\frac{ab}{(a+b)^2(a+b+1)} + n\frac{a}{a+b} \\
&\quad - n(\frac{ab}{(a+b)^2(a+b+1)}) - n(\frac{a}{a+b})^2 \\
&= n\frac{ab(a+b+n)}{(a+b)^2(a+b+1)}.
\end{aligned}$$

Kernel Trick - Example

Consider $X \sim \mathrm{Exponential}(\lambda)$, with PDF $f_X(x) = \lambda e^{-\lambda x}$.

Moments calculation, e.g., the expectation

$$\mathbb{E}[X] = \int_0^\infty x\lambda e^{-\lambda x}\mathrm{d}x.$$

usually requires integration by parts.

Kernel Trick - Example Cont.

Alternatively, we can use the **kernel trick** to avoid the ~~tedious~~ calculus.

First, notice that the PDF for $X \sim \mathrm{Gamma}(\alpha, \beta)$ is

$$g_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}.$$

Recall the integral from the previous slide:

$$\mathbb{E}[X] = \int_0^\infty \lambda x e^{-\lambda x} \mathrm{d}x.$$

Here the integrand is almost like a Gamma PDF with $\alpha = 2$, $\beta = \lambda$.

Kernel Trick - Example Cont.

The PDF of a random variable integrates to 1. Therefore if we consider $X \sim \mathit{Gamma}(2, \lambda)$, we have

$$\int_0^\infty \frac{\lambda^2}{\Gamma(2)} x e^{-\lambda x} \mathrm{d}x = 1.$$

Therefore

$$\begin{aligned}\mathbb{E}[X] &= \int_0^\infty \lambda x e^{-\lambda x} \mathrm{d}x \\ &= \frac{1}{\lambda/\Gamma(2)} = \frac{1}{\lambda}.\end{aligned}$$

Kernel Trick

The **kernel** of a distribution is the form of the PDF/PMF in which any factors that are not functions of any of the random variable(s) are omitted.

The **kernel trick** utilizes the fact that PDF/PMF integrates/ sums to 1, to help us:

1. solve integration problems (as shown in the last example);
2. identify distributions (see optional exercise in next slide, and also later in Bayesisan inference).

Note that the term *kernel* here is different from the *kernel functions* in machine learning.

Kernel Trick - Exercise

Still let $X \sim \text{Exponential}(\lambda)$, use the kernel trick to find $\mathbb{V}[X]$.

Kernel Trick - Exercise Solution

We first find $\mathbb{E}[X^2]$:

$$\mathbb{E}[X^2] = \frac{\Gamma(3)}{\lambda^2} \int_0^{\infty} \frac{\lambda^2}{\Gamma(3)} x^{3-1} \lambda e^{-\lambda x} \mathrm{d}x = \frac{2}{\lambda^2} \cdot 1 = \frac{2}{\lambda^2}.$$

Therefore the variance is:

$$\begin{aligned} \mathbb{V}[X] &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\ &= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}. \end{aligned}$$

Note: For integration, always try the kernel trick first. This will also be very useful for getting started on Bayesian inference.

Moment Generating Functions

The **moment generating function** (MGF) for a random variable X (if it exists) is defined as:

$$M_x(t) = \mathbb{E}[e^{tX}].$$

Let $\mathcal{X}$ denote the range of X, $f_X(x)$ denote the PDF/ PMF.

- If X is discrete

$$M_x(t) = \sum_{x \in \mathcal{X}} e^{tx} f_X(x).$$

- If X is continuous

$$M_x(t) = \int_{\mathcal{X}} e^{tx} f_X(x) dx.$$

Properties of MGF

Let X, Y be random variables with well defined MGFs.

1. If $M_X(t) = M_Y(t)$, then $X \stackrel{d}{=} Y$, i.e., MGF *uniquely* defines the distribution of a random variable.
   Exercise: anything else you have learned that can uniquely characterize a distribution?
2. To calculate the n^{th} moment of X
$$\mathbb{E}[X^n] = M_X^{(n)}(0).$$
3. If X and Y are independent,
$$\begin{aligned} M_{X+Y}(t) &= \mathbb{E}[e^{t(X+Y)}] \\ &= \mathbb{E}[e^{tX}]\mathbb{E}[e^{tY}] \\ &= M_X(t)M_Y(t). \end{aligned}$$

MGFs are helpful for determining distributions of sums of independent random variables.

MGF - Example

Let $X \sim Gamma(\alpha, \beta)$ (rate parameterization). Find $M_X(t)$.

MGF - Example Solution

Recall the kernel trick!

$$
\begin{aligned}
M_x(t) &= \int_0^\infty e^{tx} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} dx \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{(\alpha-1)} e^{-(\beta-t)x} dx \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha)}{(\beta-t)^\alpha} \int_0^\infty \frac{(\beta-t)^\alpha}{\Gamma(\alpha)} x^{(\alpha-1)} e^{-(\beta-t)x} dx \\
&= \frac{\beta^\alpha}{(\beta-t)^\alpha} \cdot 1 \\
&= \left(\frac{\beta}{\beta-t}\right)^\alpha, \quad \text{for } t < \beta.
\end{aligned}
$$

MGF - Exercise

1. Let $X_1, \dots, X_n \overset{i.i.d.}{\sim}$ Gamma(α, β), $Y = \sum_{i=1}^{n} X_i$.

   Find $M_Y(t)$, and identify the distribution of Y.

2. (Optional) Let $X_1, \dots, X_N \overset{i.i.d.}{\sim}$ Exponential(β), $N \sim$ Poisson(λ), and $Y = \sum_{i=1}^{N} X_i$. Find $M_Y(t)$.

   Hint:
   - Exponential$(\beta) \overset{d}{=}$ Gamma$(1,\ \beta)$.
   - Recall the law of total expectation.

MGF - Exercise Solution

1. As X_i's are independent, applying the property of MGF, we have

$$\begin{aligned} M_Y(t) &= \prod_{i=1}^{n} M_{X_i}(t) \\ &= \left[\left(\frac{\beta}{\beta - t} \right)^{\alpha} \right]^{n} \\ &= \left(\frac{\beta}{\beta - t} \right)^{n\alpha}. \end{aligned}$$

By the uniqueness property of MGF, we know $Y \sim \text{Gamma}(n\alpha, \beta)$.

MGF - Exercise Solution

2. By the law of total expectation, we have

$$M_Y(t) = \mathbb{E}[e^{tY}] = \mathbb{E}[e^{t\sum_{i=1}^{n} X_i}]$$
$$= \mathbb{E}_N[\mathbb{E}(e^{t\sum_{i=1}^{n} X_i} \mid N)].$$

Because Exponential(β) $\overset{d}{=}$ Gamma(1, β), applying results from the last exercise, we know that $Y \mid N \sim \text{Gamma}(N, \beta)$, therefore

$$\mathbb{E}(e^{t\sum_{i=1}^{n} X_i} \mid N) = \left(\frac{\beta}{\beta - t}\right)^N.$$

MGF - Exercise Solution

Now we can expand the terms for the outer expectation

$$
\begin{aligned}
&\mathbb{E}_N[\mathbb{E}(e^{t\sum_{i=1}^{n} X_i} \mid N)] \\
=&\mathbb{E}_N\left[\left(\frac{\beta}{\beta - t}\right)^N\right] \\
=&\sum_{k=0}^{\infty}\left(\frac{\beta}{\beta - t}\right)^k \frac{\lambda^k e^{-\lambda}}{k!} \\
=&\left[\sum_{k=0}^{\infty}\left(\frac{\lambda\beta}{\beta - t}\right)^k \frac{e^{-\frac{\lambda\beta}{\beta - t}}}{k!}\right] e^{\frac{\lambda\beta}{\beta - t} - \lambda}.
\end{aligned}
$$

Recognizing terms in the bracket are sum of probabilities for a Poisson distribution with parameter $\lambda\beta/(\beta - t)$, the equation simplifies to just

$$
\mathbb{E}_N[\mathbb{E}(e^{t\sum_{i=1}^{n} X_i} \mid N)] = e^{\frac{t\lambda}{\beta - t}}.
$$

Acknowledgement

Past contributors:

- Jordan Bryan, PhD student
- Brian Cozzi, MSS alumni
- Michael Valancius, MSS alumni
- Graham Tierney, PhD student
- Becky Tang, PhD student