

STAT110

Introduction to Probability

Textbook Notes

Summer, 2021

*“Probability theory is nothing but common sense
reduced to calculation”*

– Pierre-Simon Laplace

Contents

PROFESSOR:
Joe Blitzstein

SCRIBE:
Matthew Nazari*

1	Probability and Counting	2
1.1	Naive Probability and Counting	2
1.2	General Probability	3
2	Conditional Probability	5
2.1	Conditioning	5
2.2	Conditional Probability as Probabilities	6
2.3	Independence	7
3	Random Variables	8
3.1	Discrete and Continuous	8
3.2	Discrete Distributions	9
3.3	Continuous Distributions	10
3.4	Functions and Independence of Random Variables	11
4	Expectation	14
4.1	Expectation	14
4.2	The Fundamental Bridge	16
4.3	Variance	17
5	Joint Distributions	19
5.1	Joint, Marginal, and Conditional Distributions	19
5.2	Covariance and Correlation	24
5.3	Multinomial Distribution	24
5.4	Multivariate Normal Distribution	25
6	Transformations	26
6.1	Change of Variables	26
7	Conditional Expectation	28
7.1	Given an Event	28
7.2	Given an R.V.	28

*matthewnazari@college (email)

1 Probability and Counting

1.1 Naive Probability and Counting

Definition 1.1 (sample space Ω). A sample space Ω is the set of all possible outcomes to an experiment.

def. sample space Ω

Definition 1.2 (event A). An event $A \subseteq \Omega$ is a subset of the sample space Ω . We say A occurred if the actual outcome of the experiment is in A .

def. event A

Definition 1.3 (naive probability). If the sample space of an experiment is finite and all outcomes are equally likely, then the probability an event A is

def. naive probability

$$P_{\text{naive}}(A) = \frac{|A|}{|\Omega|}.$$

Example 1.1 Consider a randomly shuffled deck of n cards labelled 1 through n where n is an even number. What is the probability of drawing a card labelled with an odd number? Since all n cards can be drawn, the sample space of this experiment is the set of all cards:

$$\Omega = \{1, 2, \dots, n\}.$$

Since the sample space is finite, $|\Omega| = n$, and each card has an equal probability of being drawn¹, the naive definition of probability applies. The event $A_{\text{odd}} \subseteq \Omega$ that we are interested in is the set of all outcomes where the card labelled a drawn is odd:

$$A_{\text{odd}} = \{a : a \text{ is odd and } a \leq n\} = \{1, 3, \dots, n-1\}.$$

Since the naive definition of probability applies,

$$P_{\text{naive}}(A_{\text{odd}}) = \frac{|A_{\text{odd}}|}{|\Omega|} = \frac{\frac{1}{2}n}{n} = \frac{1}{2}.$$

Proposition 1.1 (multiplication rule). Consider an experiment consisting of two sub-experiments A and B . If there are a possible outcomes for A and for each of these outcomes there are b possible outcomes for B , then the compound experiment has ab outcomes.

Proposition 1.2 (sampling with replacement). Consider an experiment where there are n objects and k choices made from them with replacement (choosing an object does not preclude it from being chosen again). Then there are n^k outcomes.

Proposition 1.3 (sampling without replacement). Consider an experiment where there are n objects and k choices made from them without replacement (choosing an object precludes it from being chosen again). Then there are $n(n-1)\dots(n-k+1) = \frac{n!}{(n-k)!}$ possible outcomes for $k \leq n$ and 0 possible outcomes otherwise.

Definition 1.4 (binomial coefficient). Consider two nonnegative integers k and n . For $k \leq n$,

def. binomial coefficient

$$\binom{n}{k} = \frac{n(n-1)\dots(n-k+1)}{k!} = \frac{n!}{(n-k)!k!}.$$

For $k > n$, $\binom{n}{k} = 0$. The number $\binom{n}{k}$ is the number of subsets of size k for a set of size n .

¹This was assumed because of symmetry and basic intuition.

Example 1.2 A hand is 5 cards dealt from a standard, randomly shuffled 52 card deck. In poker, a hand is a full house if any 2 of the cards are of the same rank and the other 3 cards are of another rank. What is the probability of a full house? The naive probability applies since there are $\binom{52}{5}$ possible outcomes and they are all equally likely. Consider two ways to calculate the number of possible hands that are full houses by applying the multiplication rule (there could be potentially many more ways apply this rule). One way may be to choose one rank, choose 3 cards from this rank, choose one other rank, then choose 2 cards from this rank:

$$13 \binom{4}{3} 12 \binom{4}{2}.$$

Another way is to choose 2 ranks, choose one to be the rank of the three cards, choose 3 cards of that rank, then choose 2 cards of the other rank:

$$\binom{13}{2} 2 \binom{4}{3} \binom{4}{2}.$$

Both of these expressions are equal. Therefore,

$$P(\text{full house}) = \frac{13 \binom{4}{3} 12 \binom{4}{2}}{\binom{52}{5}} = \frac{\binom{13}{2} 2 \binom{4}{3} \binom{4}{2}}{\binom{52}{5}} \approx 0.00144.$$

Example 1.3 In a club of n members there are $n(n-1)(n-2) = \frac{n!}{(n-3)!}$ ways to pick a president, then vice president, then secretary. If we are picking 3 members before assigning positions, then order does not matter. In this case, since we overcounted by a factor of exactly $3!$, there are $\frac{n!}{(n-3)!3!} = \binom{n}{3}$ possible outcomes.

1.2 General Probability

Definition 1.5 (general probability). A probability space $(\Omega, \mathcal{F}, \mathbb{P})$ consists of a sample space Ω , a set of events \mathcal{F} , and a probability measure $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$.² The set \mathcal{F} is a σ -algebra of Ω , meaning \mathcal{F} satisfies the following axioms:

def. general probability

- $\emptyset \in \mathcal{F}$ and $\Omega \in \mathcal{F}$.
- If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.
- If $A_1, A_2, \dots \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

The probability measure \mathbb{P} also satisfies certain axioms:

- $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = 1$.
- If A_1, A_2, \dots are disjoint events, then $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

Note. Unless otherwise stated, we can assume \mathcal{F} is the set of all subsets of Ω and denote probability spaces as simply a sample space and probability function, (Ω, P) .

Proposition 1.4 (basic properties of probability). For any event A and B , the following properties are true:

- $P(A) = 1 - P(A^c)$.

²The probability measure \mathbb{P} is almost always for simplicity written as P .

- $P(A \cup B) = P(A) + P(B) - P(AB)$.
- If $A \subseteq B$, then $P(A) \leq P(B)$.
- $(A \cap B)^c = A^c \cup B^c$ and $(A \cup B)^c = A^c \cap B^c$.³

Proposition 1.5 (inclusion-exclusion principle). For any set of events A_1, \dots, A_n ,

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_i P(A_i) - \sum_{i < j} P(A_i \cap A_j) + \sum_{i < j < k} P(A_i \cap A_j \cap A_k) - \dots + (-1)^{n+1} P(A_1 \cap \dots \cap A_n).$$

³These properties are commonly known as De Morgan's laws.

2 Conditional Probability

2.1 Conditioning

Definition 2.1 (conditional probability). The conditional probability of an event A given another event B , denoted as $P(A|B)$, is

def. conditional probability

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

Note. Conditional probability answers the question, “if the outcome to an experiment ω is in B , then what is the probability that ω is also in A ?” In essence, we update the probability that A occurs given our new information that B occurs.

Definition 2.2 (prior and posterior probability). $P(A)$ is the prior probability of A and $P(A|B)$ is the posterior probability of A after conditioning on B .

def. prior and posterior probability

Proposition 2.1 By rearranging the definition of conditional probability, we get $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$. Generalizing this, we have⁴

$$P(A_1, \dots, A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \dots P(A_n|A_1, \dots, A_{n-1}).$$

Proposition 2.2 (Bayes' Rule).

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

Note. It might seem counter intuitive to use $P(B|A)$ to solve for $P(A|B)$. Often times in statistics, however, $P(A|B)$ is easier to find than $P(B|A)$ (or vice versa).

Proposition 2.3 (Law of Total Probability, LOTP). Let A_1, \dots, A_n be a partition of the sample space⁵ Ω such that $P(A_i) \neq 0$ for all i . Then

$$P(B) = \sum_{i=1}^n P(B|A_i)P(A_i).$$

Note. How we partition the sample space in order to calculate $P(B)$ is crucial. Some partitions could require us to calculate n difficult probabilities, whereas others might be easier.

Example 2.1 There is a fair coin and a coin that lands on heads $3/4$ of the time. You conduct an experiment where you pick up a coin and flip it three times. The coin lands on heads 3 times. What is the probability that the fair coin was picked up? Let A be the event that the coin lands on heads 3 times and F be the event the fair coin was picked. Using Bayes' Rule,

$$P(F|A) = \frac{P(A|F)P(F)}{P(A)}.$$

By assumption, $P(F) = 1/2$. To calculate $P(A)$ requires LOTP,

$$P(A) = P(A|F)P(F) + P(A|F^c)P(F^c) = (1/2)^3(1/2) + (3/4)^3(1/2).$$

⁴“ A, B ” denotes “ $A \cap B$ ” and is read as “ A and B ”.

⁵ A_1, \dots, A_n are disjoint and $\bigcup_{i=1}^n A_i = \Omega$.

Hence, we have

$$P(F|A) = \frac{P(A|F)P(F)}{P(A)} = \frac{(1/2)^3(1/2)}{(1/2)^3(1/2) + (3/4)^3(1/2)} \approx 0.23.$$

Example 2.2 A family has two children. What is the probability that both are girls given the eldest is a girl? Intuitively, we get

$$P(\text{both girls} \mid \text{eldest is a girl}) = \frac{P(\text{both girls, eldest is a girl})}{P(\text{eldest is a girl})} = \frac{1/4}{1/2} = 1/2.$$

What is the probability that both are girls given at least one of them is a girl? Before, conditioning on the event that the eldest girl is a girl eliminates two outcomes from the sample space $\Omega = \{BB, BG, GB, GG\}$. However, “at least one” does not refer to a specific child and thus only one outcome is knocked out of the sample space. We get

$$P(\text{both girls} \mid \text{at least one girl}) = \frac{P(\text{both girls, at least one girl})}{P(\text{at least one girl})} = \frac{1/4}{3/4} = 1/3.$$

What is the probability that both are girls given at least one of them is a girl who was born in winter? We have

$$P(\text{both girls} \mid \text{at least one winter girl}) = \frac{P(\text{both girls, at least one winter girl})}{P(\text{at least one winter girl})}.$$

Notice that

$$P(\text{both girls, at least one winter girl}) = P(\text{both girls, at least one winter child}).$$

Since these are independent events, we have

$$P(\text{both girls, at least one winter child}) = (1/4)(1 - P(\text{both not winter-born})) = (1/4)(1 - (3/4)^2)$$

Therefore,

$$\frac{P(\text{both girls, at least one winter girl})}{P(\text{at least one winter girl})} = \frac{(1/4)(1 - (3/4)^2)}{1 - (7/8)^2} = 7/15.$$

Notice that $7/15 \approx 1/2$. By conditioning on the birth season, we narrow down on a specific child is a child since it is unlikely both children have such a specific characteristic.

2.2 Conditional Probability as Probabilities

Proposition 2.4 Given a probability space (Ω, P) , conditioning on E creates a new probability space $(\Omega \cap E, P')$ where P' , which we have been denoting as $P(\cdot|E)$, abides by the axioms in Definition 1.2.

Note. A way to imagine the new probability space (Ω', P') is to take a subset of the original sample space, namely $\Omega' = \Omega \cap E \subseteq \Omega$, then scaling the probability measure to sum to 1.

Proposition 2.5 (Bayes' Rule with conditioning).

$$P(A|B, E) = \frac{P(A|E)P(B|A, E)}{P(B|E)}.$$

Proposition 2.6 (LOTP with conditioning). Let A_1, \dots, A_n be a partition of the sample

space Ω such that $P(A_i \cap E) \neq 0$ for all i . Then

$$P(B|E) = \sum_{i=1}^n P(B|A_i, E)P(A_i|E).$$

2.3 Independence

Definition 2.3 (independence of two events). If two events A and B are independent, then $P(A \cap B) = P(A)P(B)$.

def. independence of two events

Note. Notice that $P(A \cap B) = P(A)P(B)$ implies $P(A) = P(A|B)$ and $P(B) = P(B|A)$, meaning that knowing B occurs does not update our confidence that A occurs. Also, two disjoint events are usually never independent. If two events A and B are disjoint and independent, then $P(A \cap B) = P(\emptyset) = 0$. Therefore, disjoint events A and B are only independent if and only if $P(A) = 0$ or $P(B) = 0$.

Proposition 2.7 If A and B are independent, then A and B^c are independent, A^c and B are independent, and A^c and B^c are independent.

Definition 2.4 (independence of multiple events). If the events in some set \mathcal{S} are independent, then the events in every subset of \mathcal{S} are independent.

def. independence of multiple events

Note. Pairwise independence is not enough for three events A, B, C to be independent. Independence is a symmetric relationship but not transitive, meaning if A is independent of B and B is independent of C , then A is not necessarily independent of C . Take, for example, the case where $P(A), P(B), P(C) \neq 0$ and A and C are disjoint.

Definition 2.5 (conditional independence). If two events A and B are conditionally independent given E , then $P(A \cap B|E) = P(A|E)P(B|E)$.

def. conditional independence

3 Random Variables

3.1 Discrete and Continuous

Definition 3.1 (random variable, r.v.). Consider a probability space (Ω, P) . A random variable is a function $X : \Omega \rightarrow \mathbb{R}$ that maps outcomes to real numbers.

def. random variable, r.v.

Definition 3.2 (cumulative distribution function, CDF). The cumulative distribution function of an r.v. X is the function $F_X(x) = P(X \leq x)$.⁶ The CDF F_X satisfies the following properties:

def. cumulative distribution function, CDF

- Increasing: if $x_1 \leq x_2$, then $F(x_1) \leq F(x_2)$.
- Right continuous: for any a , $F(a) = \lim_{x \rightarrow a^+} F(x)$.
- Convergence to 0 and 1: $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow \infty} F(x) = 1$

Note. (notation warning). $P(X > x)$ is defined as $P(\{\omega : X(\omega) \leq x\}) = \sum_{x_i \leq x} p_X(x_i)$, i.e., the probability of the event an outcome occurs which maps to x or lower. Likewise, $P(X = x)$ is defined as $P(\{\omega : X(\omega) = x\})$.

Definition 3.3 (discrete r.v.). An r.v. X is discrete if there is a finite or countably infinite number of all values that it maps to with a nonzero probability (e.g., multiples of $1/2$ in the range $[0, 10]$ and integers greater than 0).

def. discrete r.v.

Definition 3.4 (support of a discrete r.v.). The support of a discrete r.v. X is the set S_X of values that X maps to with a nonzero probability:

$$S_X = \{x : P(X = x) > 0\}.$$

def. support of a discrete r.v.

Definition 3.5 (probability mass function, PMF). The probability mass function of a discrete r.v. X is the function $p_X : \mathbb{R} \rightarrow [0, 1]$ given by $p_X(x) = P(X = x)$. The PMF p_X of X must satisfy the following criteria:

def. probability mass function, PMF

- Nonnegative: if $x \in S_X$, then $p_X(x) > 0$. Otherwise, $p_X(x) = 0$.
- Sums to 1: $\sum_{x \in S_X} p_X(x) = 1$.

Definition 3.6 (continuous random variable). An r.v. X is continuous if its CDF is differentiable with possibly the exception of a few points where the CDF is continuous but not differentiable (e.g., the endpoints).

def. continuous random variable

Definition 3.7 (probability density function, PDF). The probability density function of a continuous r.v. X is the derivative f of the CDF F_X of X . The PDF f of X must satisfy the following criteria:

def. probability density function, PDF

- Nonnegative: $f(x) \geq 0$.
- Integrates to 1: $\int_{-\infty}^{\infty} f(x) dx = 1$.

⁶When there is no risk for ambiguity, F_X is denoted simply by F or some other letter.

def. support of a continuous r.v.

Definition 3.8 (support of a continuous r.v.). The support of a continuous r.v. X is the set S_X of values that its PDF f maps to with a nonzero probability density:

$$S_X = \{x : f(x) > 0\}.$$

Proposition 3.1 (PDF to CDF). Let X be a continuous r.v. with PDF f . Then the CDF F_X of X is

$$F_X(x) = \int_{-\infty}^x f(t) dt.$$

Proposition 3.2 (PDF and CDF to probability). Let X be a continuous r.v. with PDF f and CDF F_X . Then

$$P(a \leq x \leq b) = F_X(b) - F_X(a) = \int_a^b f(t) dt.$$

Note. The probability that a continuous r.v. X takes on a particular value is 0, i.e., $P(X = x) = 0$ for all x . The PDF of X can even be greater than 1 at some values.

3.2 Discrete Distributions

Definition 3.9 (Bernoulli distribution). If $X \sim \text{Bern}(p)$,⁷ then $P(X = 1) = p$ and $P(X = 0) = 1 - p$. Any event A has a Bernoulli variable naturally associated with it, where X maps all outcomes in A to 1 and all outcomes in A^c to 0.

def. Bernoulli distribution

Note. There is not one Bernoulli distribution, but a family of Bernoulli distributions indexed by p . If $X \sim \text{Bern}(p)$, it is incorrect to say “ X is Bernoulli”. Instead, we say “ X is Bernoulli with parameter p ” or “ X has the Bernoulli distribution with parameter p ”. This applies to all distributions.

Definition 3.10 (indicator random variable). If an r.v. equals 1 if an event A occurs and 0 if A does not occur, then it is an indicator random variable of A denoted as I_A or $I(A) \sim \text{Bern}(p)$ where $p = P(A)$.

def. indicator random variable

Definition 3.11 (Bernoulli trial). If the only possible outcomes to an experiment are “success” and “failure”, then the experiment is a Bernoulli trial.

def. Bernoulli trial

Definition 3.12 (binomial distribution). If $X \sim \text{Bin}(n, p)$, then

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

def. binomial distribution

for $0 \leq k \leq n$ and 0 otherwise. The binomial distribution describes the number successes in n independent Bernoulli trials of success probability p .

Definition 3.13 (hypergeometric distribution). If $X \sim \text{HGeom}(w, b, n)$, then

$$P(X = k) = \frac{\binom{w}{k} \binom{b}{n-k}}{\binom{w+b}{n}}$$

def. hypergeometric distribution

for $0 \leq k \leq w$ and 0 otherwise. The hypergeometric distribution describes the number of white marbles drawn after drawing n marbles from a jar of w white marbles and b

⁷The “ \sim ” symbol is read as “is distributed as”.

black marbles without replacement.

Definition 3.14 (discrete uniform distribution). If $X \sim \text{DUnif}(C)$ where C is an finite, nonempty set of numbers, then

def. discrete uniform distribution

$$P(X = k) = \frac{1}{|C|}$$

for $k \in C$ and 0 otherwise. The discrete uniform distribution describes the numbers in C as being equally likely.

Definition 3.15 (geometric distribution). If $X \sim \text{Geom}(p)$, then $P(X = k) = q^k p$ for $0 \leq k < \infty$ and 0 otherwise.⁸ The geometric distribution describes the number of failed independent Bernoulli trials until a success.

def. geometric distribution

Definition 3.16 (first success distribution). If $X \sim \text{FS}(p)$, then $P(X = k) = q^{k-1} p$ for $1 \leq k < \infty$ and 0 otherwise. The first success distribution describes the number of total independent Bernoulli trials until a success. If $X \sim \text{FS}(p)$, then $X + 1 \sim \text{Geom}(p)$.

def. first success distribution

Definition 3.17 (negative binomial distribution). If $X \sim \text{NBin}(r, p)$, then

def. negative binomial distribution

$$P(X = k) = \binom{k+r-1}{r-1} p^r q^k$$

for $0 \leq k < \infty$ and 0 otherwise. The negative binomial distribution describes the number of failed independent Bernoulli trials until the r th success.

Definition 3.18 (poisson distribution). If $X \sim \text{Pois}(\lambda)$, where $\lambda > 0$, then

def. poisson distribution

$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

for $0 \leq k < \infty$ and 0 otherwise.

3.3 Continuous Distributions

Definition 3.19 (Uniform distribution). If $U \sim \text{Unif}(a, b)$, then U has the PDF f and CDF F given by

def. Uniform distribution

$$f(x) = \begin{cases} \frac{1}{a-b} & \text{if } a < x < b, \\ 0 & \text{otherwise,} \end{cases} \quad F(x) = \begin{cases} 0 & \text{if } x \leq a, \\ \frac{x-a}{a-b} & \text{if } a < x < b, \\ 1 & \text{if } x \geq b. \end{cases}$$

Definition 3.20 (Logistic distribution). If $X \sim \text{Logistic}$, then X has the PDF f and CDF F given by

def. Logistic distribution

$$f(x) = \frac{e^x}{(1 + e^x)^2}, \quad x \in \mathbb{R}, \quad F(x) = \frac{e^x}{1 + e^x}, \quad x \in \mathbb{R}.$$

Definition 3.21 (Rayleigh distribution). If $X \sim \text{Rayleigh}$, then X has the PDF f and CDF F given by

def. Rayleigh distribution

$$f(x) = x e^{-x^2/2}, \quad x > 0, \quad F(x) = 1 - e^{-x^2/2}, \quad x > 0.$$

⁸Unless specified otherwise, q always denotes $1 - p$.

def. standard Normal distribution

Definition 3.22 (standard Normal distribution). If $Z \sim \mathcal{N}(0, 1)$, then Z has the PDF φ and CDF Φ given by

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad z \in \mathbb{R}, \quad \Phi(z) = \int_{-\infty}^z \varphi(t) dt, \quad z \in \mathbb{R}.$$

Definition 3.23 (Normal distribution). If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $X = \sigma Z + \mu$ has the PDF f and CDF F given by

$$f(x) = \varphi\left(\frac{x - \mu}{\sigma}\right) \frac{1}{\sigma}, \quad x \in \mathbb{R}, \quad F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right), \quad x \in \mathbb{R}.$$

def. Normal distribution

Note. It is mathematically impossible to find a closed form of Φ , so it is left in integral form.

Definition 3.24 (Exponential distribution). If $X \sim \text{Expo}(\lambda)$, then X has the PDF f and CDF F given by

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0, \quad F(x) = 1 - e^{-\lambda x}, \quad x > 0.$$

def. Exponential distribution

3.4 Functions and Independence of Random Variables

Definition 3.25 (function of an r.v.). Consider a probability space (Ω, P) , a discrete r.v. X , and a function $g : \mathbb{R} \rightarrow \mathbb{R}$. Then $Y = g(X)$ is an r.v. given by $Y = g(X(\omega))$ for all $\omega \in \Omega$. If g is injective, then $P(Y = g(x)) = P(X = x)$. Otherwise,

def. function of an r.v.

$$P(Y = y) = \sum_{x \text{ s.t. } g(x)=y} P(X = x).$$

Definition 3.26 (function of two r.v.s). Consider a probability space (Ω, P) , two discrete r.v.s X and Y , and a function $g : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$. Then $Z = g(X, Y)$ is an r.v. given by $Z = g(X(\omega), Y(\omega))$ for all $\omega \in \Omega$. If g is injective, then $P(Z = g(x, y)) = P(X = x, Y = y)$. Otherwise,

def. function of two r.v.s

$$P(Z = z) = \sum_{x, y \text{ s.t. } g(x, y)=z} P(X = x, Y = y).$$

Proposition 3.3 If $X \sim \text{Bin}(n, p)$, then $n - X \sim \text{Bin}(n, q)$.

Proposition 3.4 (sum of binomial r.v.s with same parameter p). If $X \sim \text{Bin}(n, p)$ and $Y \sim \text{Bin}(m, p)$, then $X + Y \sim \text{Bin}(n + m, p)$.

Example 3.1 A particle moves n steps on a number line starting from position 0. All steps are independent and equally probable to be either 1 unit left or right.

- Let X be the number of steps to the right after n steps. What is the distribution of X ? Consider each step as a Bernoulli trial where a success is a move to the right. Since X describes the number of successful steps to the right out of n trials, $X \sim \text{Bin}(n, 1/2)$.
- Let Y be the position of the particle after n steps. What is the distribution of Y ? Notice that since the particle must moves X steps to the right and $n - X$ steps to the left, its final position is $X - (n - X) = 2X - n$. We have expressed Y as an

injection of X , namely $Y = 2X - n$. We can solve for

$$P(Y = k) = P(2X - n = k) = P(X = (k + n)/2) = \binom{n}{\frac{k+n}{2}} \left(\frac{1}{2}\right)^n.$$

- Let D be the distance of the particle after n steps from the origin. What is the distribution of D ? Note that $D = |Y|$. By Definition 3.4, we can solve for the distribution of D despite D not being an injection of Y :

$$P(D = k) = P(Y = k) + P(Y = -k) = 2 \binom{n}{\frac{k+n}{2}} \left(\frac{1}{2}\right)^n.$$

Definition 3.27 (independence of two r.v.s). Two r.v.s X and Y are independent if for all x in the support of X and y in the support of Y ,

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y).$$

def. independence of two r.v.s

Intuitively, X and Y are independent if they do not provide any information about the other.

Definition 3.28 (independence of many r.v.s). Many r.v.s X_1, \dots, X_n are independent if for all x_1, \dots, x_n ,

$$P(X_1 \leq x_1, \dots, X_n \leq x_n) = P(X_1 \leq x_1) \dots P(X_n \leq x_n).$$

def. independence of many r.v.s

Note. Independence of n r.v.s implies independence of every subset of these r.v.s. However, this implication is not true backwards: for example, pairwise independence between all $\binom{n}{2}$ pairs of r.v.s is not enough to imply independence of all n r.v.s.

Proposition 3.5 If X and Y are independent r.v.s, then any function of X is independent of any function of Y .

Definition 3.29 (independent and identically distributed, i.i.d). If r.v.s are i.i.d., then they are independent and have the same CDF.

def. independent and identically distributed, i.i.d

Note. Whether two r.v.s are independent has nothing to do with whether they are identically distributed.

- Two r.v.s can be independent and identically distributed. Let X be the value of a die roll and Y be the value of a second, independent die roll. X and Y are i.i.d.
- Two r.v.s can be independent and not identically distributed. Let X be the value of a die roll and Y be the blackjack value of a card drawn from a standard, well-shuffled deck. Both X and Y provide no information about the other, and they obviously have different distributions.
- Two r.v.s can be dependent and identically distributed. Let X be the number of heads in n consecutive, independent coin tosses and Y be the number of tails. Both X and Y are distributed as $\text{Bin}(n, 1/2)$, but they are obviously dependent.
- Two r.v.s can be dependent and not identically distributed. Let X be an indicator variable for if two coin tosses both land on heads and Y be the number of heads that land.

Proposition 3.6 (conditioning on a Uniform r.v.). Let $U \sim \text{Unif}(a, b)$ and let $(c, d) \subseteq (a, b)$. Then the conditional distribution of U given $U \in (c, d)$ is $\text{Unif}(c, d)$.

Definition 3.30 (location-scale transformation). Let X be an r.v. and $Y = \sigma X + \mu$ where $\sigma, \mu > 0$. Then Y is a location-scale transformation of X where μ controls how the location is changed and σ controls how the scale is changed.

def. location-scale transformation

Proposition 3.7 (Universality of the Uniform). Let F be a CDF which is continuous and strictly increasing on the support of the distribution (i.e., the inverse of the CDF, or quantile function, $F^{-1} : (0, 1) \rightarrow \mathbb{R}$ exists). Then the following is true:

1. If $U \sim \text{Unif}(0, 1)$, then $X = F^{-1}(U)$ is an r.v. with the CDF F .
2. If U has the CDF F , then $F(X) \sim \text{Unif}(0, 1)$.

4 Expectation

4.1 Expectation

Definition 4.1 (expected value of a discrete r.v.). Let X be an discrete r.v. with support S_X . Then the expected value of X is

def. expected value of a discrete r.v.

$$E[X] = \sum_{x \in S_X} xP(X = x).$$

Definition 4.2 (expected value of a continuous r.v.). Let X be an continuous r.v. with PDF f . Then the expected value of X is

def. expected value of a continuous r.v.

$$E[X] = \int_{-\infty}^{\infty} xf(x) dx.$$

Note. The expected value of a discrete r.v. X is the average of all possible values of X weighted by their probability. Sometimes this sum does not converge, but this is never explained.

Example 4.1 Let X be the result of rolling a fair 6-sided die. By definition,

$$E[X] = \sum_{x \in S_X} xP(X = x) = \frac{1}{6}(1) + \frac{1}{6}(2) + \frac{1}{6}(3) + \frac{1}{6}(4) + \frac{1}{6}(5) + \frac{1}{6}(6) = 3.5.$$

Notice that $E[X] \notin S_X$. This is alright in the same way saying “the average household in Cambridge has 1.8 children” is reasonable despite it being impossible for a household to actually have 1.8 children.

Proposition 4.1 (expectation of i.d. r.v.s). If X and Y are i.d., then $E[X] = E[Y]$.

Proposition 4.2 (linearity of expectation). If X and Y are r.v.s, independent or not, then

$$\begin{aligned} E[X + Y] &= E[X] + E[Y], \\ E[cX] &= c E[X] \text{ for all } c \in \mathbb{R}. \end{aligned}$$

Proposition 4.3 (monotonicity of expectation). If X and Y are r.v.s and $X \leq Y$ with a probability of 1, then $E[X] \leq E[Y]$.

Example 4.2 (binomial expectation). Let $X \sim \text{Bin}(n, p)$. There are two ways to calculate $E[X]$. By the definition of expectation, and many algebraic steps,

$$E[X] = \sum_{k=0}^n kP(X = k) = k \binom{n}{k} p^k q^{n-k} = \dots = np.$$

More simply, recall that X can be expressed as the sum of n independent Bernoulli variables distributed as $\text{Bern}(p)$. By the expectation of i.d. r.v.s and linearity of expectation,

$$E[X] = E[I_1 + \dots + I_n] = n E[I_j] = n(1p + 0q) = np.$$

Example 4.3 (hypergeometric expectation). Let $X \sim \text{HGeom}(w, b, n)$. Recall that $X = I_1, \dots, I_n$ where I_1, \dots, I_n are dependent r.v.s distributed as $\text{Bern}(w/(w + b))$. These

r.v.s are not independent, but nevertheless the linearity of expectation applies:

$$E[X] = E[I_1 + \dots + I_n] = n E[I_j] = nw/(w + b).$$

Example 4.4 (geometric expectation). Let $X \sim \text{Geom}(p)$. Then

$$E[X] = \sum_{k=0}^{\infty} kq^k p = pq \sum_{k=0}^{\infty} kq^{k-1} = pq \frac{1}{(1-q)^2} = \frac{q}{p}.$$

Example 4.5 (first success expectation). Let $X \sim \text{FS}(p)$. Recall that $X = Y + 1$ where $Y \sim \text{Geom}(p)$. Then

$$E[X] = E[Y + 1] = \frac{q}{p} + 1 = \frac{1}{p}.$$

Example 4.6 (negative binomial expectation). Let $X \sim \text{NBin}(r, p)$. Recall that $X = X_1 + \dots + X_r$ where X_1, \dots, X_r are independent r.v.s distributed as $\text{Geom}(p)$. Then

$$E[X] = E[X_1 + \dots + X_r] = r E[X_j] = r \frac{q}{p}.$$

Note. Note that if $X = Y + 1$, like in Example 4.1, then $E[X] = E[Y] + 1$ since X is a linear function $g(x) = x + 1$ of Y , so $E[g(X)] = g(E[X])$. However, if g is not linear, then $E[g(X)]$ can be very different from $g(E[X])$. One way to find $E[g(X)]$ is to first find the distribution of $g(X)$, but $E[g(X)]$ can be calculated directly from the distribution of X using the law of the unconscious statistician.

Definition 4.3 (law of the unconscious statistician, LOTUS). Let X be a discrete r.v. with support S_X and let $g : \mathbb{R} \rightarrow \mathbb{R}$ be a function. Then

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

def. law of the unconscious statistician, LOTUS

Let X be a continuous r.v. with PDF f and let $g : \mathbb{R} \rightarrow \mathbb{R}$ be a function. Then

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

Example 4.7 (unbiased estimator). Let $X \sim \text{Pois}(\lambda)$ where λ is unknown but constant. We try to estimate $\theta = e^{-3\lambda}$ with an estimator $g(X)$. An unbiased estimator $g(X)$ is one such that $E[g(X)] - \theta = 0$ (i.e., $g(X)$ is expected to equal the estimator). We can prove $g(X) = (-2)^X$ is an unbiased estimator since

$$\begin{aligned} E(-2)^X - \theta &= \sum_{k=0}^{\infty} (-2)^k P(X = k) - e^{-2\lambda} \\ &= \sum_{k=0}^{\infty} (-2)^k \frac{\lambda^k}{k!} e^{-\lambda} - e^{-2\lambda} \\ &= e^{-\lambda} e^{-2\lambda} - e^{-3\lambda} = 0. \end{aligned}$$

However, $g(X) = (-2)^X$ is a silly estimator because $0 < \theta < 1$ (since $\lambda > 0$). A better estimator for θ is $h(X)$ where $h(X) = 0$ if $g(X) < 0$, $h(X) = 1$ if $g(X) > 1$, and $h(X) = g(X)$ otherwise.

4.2 The Fundamental Bridge

Proposition 4.4 (indicator r.v. properties). Let A and B be any events. Then

1. $(I_A)^k = I_A$ for any $k \in \mathbb{R}$,
2. $I_{A^c} = 1 - I_A$,
3. $I_{A \cap B} = I_A I_B$,
4. $I_{A \cup B} = I_A + I_B - I_A I_B$.

Proposition 4.5 (the fundamental bridge between probability and expectation). For any event A ,

$$P(A) = E[I_A].$$

Example 4.8 (Boole's inequality). We already know that $P(\bigcup_{i=1}^n A_i) \leq \sum_{i=1}^n P(A_i)$, but we can prove this with the fundamental bridge. It is evident that

$$I\left(\bigcup_{i=1}^n A_i\right) \leq \sum_{i=1}^n I(A_i).$$

By the monotonicity of expectation and the fundamental bridge,

$$\begin{aligned} E\left[I\left(\bigcup_{i=1}^n A_i\right)\right] &\leq E\left[\sum_{i=1}^n I(A_i)\right], \\ P\left(\bigcup_{i=1}^n A_i\right) &\leq \sum_{i=1}^n P(A_i). \end{aligned}$$

Example 4.9 (matching). Consider a shuffled deck of n cards labeled 1 through n . A card is a match if its position in the deck corresponds to its label. Let X be the number of matches in the deck. What is $E[X]$? There is no convenient distribution we know for X , but we can write $X = I_1 + \dots + I_n$ where

$$I_j = \begin{cases} 1 & \text{if the } j\text{th card is a match,} \\ 0 & \text{otherwise.} \end{cases}$$

Notice that I_j is just the indicator variable for the event A_j that the j th card is a match. By the fundamental bridge, $E[I_j] = P(A_j) = \frac{1}{n}$. Therefore, $E[X] = n E[I_j] = 1$.

Example 4.10 (Putnam problem). Consider a random permutation of numbers 1 through n . A local maximum is a position whose number is greater than the numbers surrounding it. What is the average number of local maxima in a random permutation of numbers 1 through n ? We are interested in $\sum_{j=1}^n I_j$ where I_j is an indicator variable for position j being a local maximum. We know $E[I_1], E[I_n] = 1/2$ since an end position is a local maximum if its single neighbor is less than it. By symmetry, this event has a probability of $1/2$. Similarly, we know $E_j = 1/3$ for $1 < j < n$ since position j is a local maximum if its two neighbors are less than it. By symmetry, this has a probability of $1/3$. Thus,

$$\sum_{j=1}^n I_j = 2 \cdot \frac{1}{2} + (n-2) \cdot \frac{1}{3} = \frac{n+1}{3}.$$

Proposition 4.6 (algebraic form of the binomial). If $X \sim \text{Bin}(n, p)$, then we can write X as the sum of n independent r.v.s distributed as $\text{Bern}(p)$:

$$X = I_1 + \dots + I_n.$$

Proposition 4.7 (algebraic form of the hypergeometric). If $X \sim \text{HGeom}(w, b, n)$, then we can write X as the sum of n dependent r.v.s distributed as $\text{Bern}(w/(w+b))$ since unconditionally, each I_j has the same probability by symmetry:

$$X = I_1 + \dots + I_n.$$

Proposition 4.8 (algebraic form of the negative binomial). If $X \sim \text{NBin}(r, p)$, then we can write X as the sum of r independent r.v.s distributed as $\text{Geom}(p)$:

$$X = X_1 + \dots + X_n.$$

4.3 Variance

Definition 4.4 (variance). The variance of an r.v. X is $\text{Var}(X) = E[X - EX]^2 = E[X^2] - (EX)^2$. The variance is the average squared difference between an r.v. X and its expectation.⁹

def. variance

Definition 4.5 (standard deviation). The standard deviation of an r.v. X is $\text{SD}(X) = \sqrt{\text{Var}(X)}$. The square root is the average difference in absolute value between an r.v. X and its expectation.

def. standard deviation

Proposition 4.9 (properties of variance). Let X and Y be any r.v.s and c be any scalar. Then

1. $\text{Var}(X + c) = \text{Var}(X)$,
2. $\text{Var}(cX) = c^2 \text{Var}(X)$,
3. $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$ if X and Y are independent,
4. $\text{Var}(X) > 0$ if X is not constant and $\text{Var}(X) = 0$ if X is constant, i.e. $|S_X| = 1$.

Example 4.11 (binomial variance). Let $X \sim \text{Bin}(n, p)$. Recall $X = \sum_{j=1}^n I_j$ where I_1, \dots, I_n are i.i.d $\text{Bern}(p)$. Then $\text{Var}(I_j) = E[I_j^2] - (EI_j)^2 = p - p^2 = p(1 - p)$. Since I_1, \dots, I_n are independent,

$$\text{Var}(X) = \sum_{j=1}^n \text{Var}(I_j) = n \text{Var}(I_j) = np(1 - p).$$

Example 4.12 Let $U \sim \text{Unif}(-1, 1)$. Then $E[U] = 0$ because of symmetry. Calculating $\text{Var}(U)$ requires us to calculate $E[U^2]$ first: by LOTUS,

$$E[U^2] = \int_{-1}^1 u^2 P(U = u) du = \frac{1}{2} \int_{-1}^1 u^2 du = \frac{1}{3}.$$

Therefore,

$$\text{Var}(U) = E[U^2] - (EU)^2 = E[U^2] = \frac{1}{3}.$$

⁹With parentheses, $\text{Var}(X) = E[(X - EX)^2]$. This is unwieldy, so $E[X]$ is often simplified to EX .

Notice that U^2 is not uniform. Let $F(t)$ be the CDF of U^2 . Then $F(t) = 0$ for $t < -1$ and $F(t) = 1$ for $t > 1$. If $-1 \leq t \leq 1$,

$$F(t) = P(U^2 \leq t) = P(-\sqrt{t} \leq U \leq \sqrt{t}) = \sqrt{t}.$$

The last equality is because the probability U is in an interval is proportional to that interval's length.

5 Joint Distributions

We usually care about the relationship between multiple r.v.s in the same experiment. Examining the distribution of an r.v. gives a complete story about it individually, but it explains nothing about its relationship to other r.v.s. Are two Bern(1/2) variables, for example, independent? Or are they indicators of complementary events? Examining the joint distribution of these r.v.s answers these questions.

5.1 Joint, Marginal, and Conditional Distributions

Definition 5.1 (joint CDF). The joint CDF of r.v.s X and Y is the function $F_{X,Y} : \mathbb{R} \times \mathbb{R} \rightarrow [0, 1]$ given by

def. joint CDF

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y).$$

Much like in the univariate case, the joint CDF of discrete r.v.s is unweildy because of its jumps and flat regions. We usually, therefore, work with the joint PMF.

Definition 5.2 (joint PMF). The joint PMF of discrete r.v.s X and Y is the function $p_{X,Y} : \mathbb{R} \times \mathbb{R} \rightarrow [0, 1]$ given by

def. joint PMF

$$p_{X,Y}(x, y) = P(X = x, Y = y)$$

and must satisfy the following criteria:

- Nonnegative: $P(X = x, Y = y) \geq 0$ for all x and y .
- Sums to 1: $\sum_x \sum_y P(X = x, Y = y) = 1$.

The joint CDF and joint PMF of multiple discrete r.v.s is defined analogously. We can get back to the unconditional or marginal distribution of X by adding all possible values of Y . This operation is called marginalizing out Y .

Definition 5.3 (marginal PMF). Let X and Y be discrete r.v.s. The marginal PMF of X is

def. marginal PMF

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

We can obtain the marginal CDF of X by a limit, but this is unweildy:

$$F_X(x) = P(X \leq x) = \lim_{y \rightarrow \infty} P(X \leq x, Y \leq y) = \lim_{y \rightarrow \infty} F_{X,Y}(x, y).$$

Suppose we observe a value of X and want to update our distribution of Y . Instead of marginalizing out X , we can obtain a conditional PMF by fixing y and renormalizing.

Definition 5.4 (conditional PMF). Let X and Y be discrete r.v.s. The conditional PMF $P(Y = y|X = x)$ is given by

def. conditional PMF

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$$

for all x such that $P(X = x) > 0$ and 0 otherwise.

We can relate conditional distributions using Bayes' rule.

$$P(Y = y|X = x) = \frac{P(X = x|Y = y)P(Y = y)}{P(X = x)}.$$

Using LOTP, we have another way of getting the marginal PMF.

$$P(X = x) = \sum_{y \in S_Y} P(X = x|Y = y)P(Y = y).$$

Definition 5.5 (independence of r.v.s). If r.v.s X and Y are independent then

$$F_{X,Y}(x, y) = F_X(x)F_Y(y).$$

def. independence of r.v.s

If X and Y are discrete, this definition is equivalent to the conditions that for all x and y such that $P(X = x) > 0$,

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

$$P(Y = y|X = x) = P(Y = y).$$

Independence means that all conditional PMFs are the same as the marginal PMF.

Example 5.1 Consider a simple discrete joint distribution of two Bernoulli r.v.s X and Y . The joint distribution is fully specified by the four values $P(X = 1, Y = 1)$, $P(X = 1, Y = 0)$, $P(X = 0, Y = 1)$, and $P(X = 0, Y = 0)$ as represented in the table below. To get the marginal probability $P(X = 1)$, we by definition just add $P(X = 1, Y = 0)$ and $P(X = 1, Y = 1)$ to get $25/100$. Likewise, we get that the marginal probability $P(Y = 1)$ is $8/100$. The marginal distribution of X is $\text{Bern}(0.25)$ and the marginal distribution of Y is $\text{Bern}(0.08)$. From the joint distribution, we can tell X and Y are not independent since $P(X = 1, Y = 1) = 5/100 \neq 2/100 = P(X = 1)P(Y = 1)$.

	$Y = 1$	$Y = 0$	Total
$X = 1$	$\frac{5}{100}$	$\frac{20}{100}$	$\frac{25}{100}$
$X = 0$	$\frac{3}{100}$	$\frac{72}{100}$	$\frac{75}{100}$
Total	$\frac{8}{100}$	$\frac{92}{100}$	$\frac{100}{100}$

Example 5.2 Suppose on some day, a chicken lays a random number of eggs. Each egg independently either hatches with a probability p or does not hatch with a probability $q = 1 - p$. Suppose X is the number of eggs that hatch and Y the number that does not hatch, and the total number of eggs $N = X + Y$ is distributed as $\text{Pois}(\lambda)$. What is the joint PMF of x and Y ? By the law of total probability,

$$P(X = i, Y = j) = \sum_{n=0}^{\infty} P(X = i, Y = j|N = n)P(N = n).$$

Notice, though, that $P(X = i, Y = j|N = n) = 0$ unless $n = i + j$. Thus,

$$P(X = i, Y = j) = P(X = i, Y = j|N = i + j)P(N = i + j).$$

Conditional on $N = i + j$, $X = i$ and $Y = j$ are the same exact event, so

$$P(X = i, Y = j) = P(X = i|N = i + j)P(N = i + j).$$

By the story of the binomial distribution, $X|N = n \sim \text{Bin}(n, p)$ since X describe the number of successful hatches out of n eggs given $N = n$. And since it is given that

$N \sim \text{Pois}(\lambda)$, we get

$$\begin{aligned} P(X = i, Y = j) &= P(X = i | N = i + j) P(N = i + j) \\ &= \binom{i+j}{i} p^i q^j \cdot \frac{e^{-\lambda} \lambda^{i+j}}{(i+j)!} \\ &= \frac{e^{-\lambda p} (\lambda p)^i}{i!} \cdot \frac{e^{-\lambda q} (\lambda q)^j}{j!}. \end{aligned}$$

Since the joint PMF factors into the product of the $\text{Pois}(\lambda p)$ PMF as a function of i and the $\text{Pois}(\lambda q)$ PMF as a function of j , this tells us two elegant facts: (1) $X \sim \text{Pois}(\lambda p)$ and $Y \sim \text{Pois}(\lambda q)$, and (2) X and Y are independent since their joint PMF is the product of their marginal PMFs. It makes sense, albeit counterintuitively, that X and Y are unconditionally independent since N is random. Conditionally, of course, X and Y are very dependent.

Proposition 5.1 *If $X \sim \text{Pois}(\lambda p)$, $Y \sim \text{Pois}(\lambda q)$, and X and Y are independent, then $N = X + Y \sim \text{Pois}(\lambda)$ and $X | N = n \sim \text{Bin}(n, p)$.*

Proposition 5.2 *If $N \sim \text{Pois}(\lambda)$ and $X | N = n \sim \text{Bin}(n, p)$, then $X \sim \text{Pois}(\lambda p)$, $Y = N - X \sim \text{Pois}(\lambda q)$, and X and Y are independent.*

We can analogously consider continuous joint distributions. In order for X and Y to have a continuous joint distribution, we require the joint CDF $F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$ be differentiable with respect to x and y .

Definition 5.6 (joint PDF). The joint PDF of continuous r.v.s X and Y is the derivative of their joint CDF with respect to x and y ,

def. joint PDF

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y),$$

and satisfies the following properties:

- Nonnegative: $f_{X,Y}(x, y) \geq 0$ for all x and y .
- Integrates to 1: $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx dy = 1$.

Note that $P(X = x, Y = y) = 0$ for any point (x, y) . For a general area $A \subseteq \mathbb{R}^2$, we get

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

Definition 5.7 (marginal PDF). Let X and Y be continuous r.v.s. The marginal PDF of X is

def. marginal PDF

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy.$$

The joint CDF and joint PDF of multiple continuous r.v.s is defined analogously. If we have the joint PDF of X, Y, Z, W but want the joint PDF of X, Y , we marginalize out Z and W :

$$f_{X,Y}(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y,Z,W}(x, y, z, w) dz dw.$$

Definition 5.8 (conditional PDF). Let X and Y be continuous r.v.s. The conditional

def. conditional PDF

PDF $f_{Y|X}(y|x)$ is given by

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

for all x such that $f_X(x) > 0$ and 0 for $f_X(x) = 0$

Results like Bayes' rule and LOTP work in the continuous case as well:

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

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

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)f_Y(y) dy.$$

If X and Y are continuous, the definition of independence between X and Y is equivalent to the conditions that for all x and y such that $P(X = x) > 0$,

$$f_{X,Y}(x,y) = f_X(x)f_Y(y),$$

$$f_{Y|X}(y|x) = f_Y(y).$$

Example 5.3 (uniform distribution over unit square). Let (X, Y) be a random point in the unit square $\{(x, y) : x, y \in [0, 1]\}$, meaning

$$f_{X,Y}(x,y) = \begin{cases} 1 & \text{if } x, y \in [0, 1], \\ 0 & \text{otherwise.} \end{cases}$$

Here are a few observations: X and Y are $\text{Unif}(0, 1)$ marginally. We verify this by computing

$$f_X(x) = \int_0^1 f_{X,Y}(x,y) dy = \int_0^1 1 dy = 1$$

and similarly for f_Y . Furthermore, X and Y are independent since the joint PDF factors into the product of the marginal PDFs. Geometrically, independence means every horizontal cross section looks like every other horizontal cross section (and same for vertically).

Example 5.4 (uniform distribution over unit circle). Let (X, Y) be a random point in the unit circle $\{(x, y) : x^2 + y^2 \leq 1\}$, meaning

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{\pi} & \text{if } x^2 + y^2 \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Here are a few observations: X and Y are not $\text{Unif}(0, 1)$ marginally unlike the previous example. We verify this by computing

$$f_X(x) = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} f_{X,Y}(x,y) dy = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi} dy = \frac{1}{\pi}(2\sqrt{1-x^2}), \quad -1 \leq x \leq 1$$

and similarly for f_Y . The PDF is denser at 0 than ± 1 , which means X and Y are more likely to take on values near 0. We found the bounds of integration by analyzing what range of values Y can take for $X = x$. Furthermore, X and Y are dependent. Logically, observing that $Y = 0.9$ changes the values that X can take on. Realizing that $f_{X,Y}(0.9, 0.9) \neq f_X(0.9)f_Y(0.9)$ (since $(0.9, 0.9) \notin S_{X,Y}$ while $0.9 \in S_X$ and $0.9 \in S_Y$) is enough to make X and Y dependent.

Example 5.5 (Cauchy PDF). Let X and Y be i.i.d. $N(0, 1)$, and let $T = X/Y$. (Since

$P(Y = 0) = 0$, we can define T arbitrarily when $Y = 0$). Find the PDF of T . We can find an expression for the CDF of T then differentiate:

$$F_T(t) = P(T \leq t) = P\left(\frac{X}{Y} \leq t\right) = P\left(\frac{X}{|Y|} \leq t\right) = P(X \leq t|Y|).$$

We calculate $P(X \leq t|Y|)$ by integrating the joint PDF of X and Y where $X \leq t|Y|$ holds. We know the joint PDF is just the product of the marginal PDFs, by independence:

$$\begin{aligned} F_T(t) &= P(X \leq t|Y|) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{t|y|} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dx dy \\ &= \sqrt{\frac{2}{\pi}} \int_0^{\infty} e^{-y^2/2} \Phi(ty) dy. \end{aligned}$$

We differentiate this to get

$$f_T(t) = F'_T(t) = \frac{1}{\pi(1+t^2)}.$$

Proposition 5.3 (2D LOTUS). Let $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ be any function. If X and Y are discrete, then

$$E[g(X, Y)] = \sum_x \sum_y g(x, y) P(X = x, Y = y).$$

If X and Y are continuous, then

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y) dx dy.$$

Example 5.6 (distance between two points on line). Consider $X, Y \stackrel{\text{i.i.d.}}{\sim} \text{Unif}(0, 1)$. The average distance between X and Y is

$$\begin{aligned} E[|X - Y|] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |x - y| f_{X,Y}(x, y) dx dy \\ &= \int_0^1 \int_0^1 |x - y| dx dy \\ &= \int_0^1 \int_0^y (y - x) dx dy + \int_0^1 \int_y^1 (x - y) dx dy \\ &= 2 \int_0^1 \int_y^1 (x - y) dx dy = \frac{1}{3}. \end{aligned}$$

If we wanted the standard deviation of $|X - Y|$, we would need to calculate $E[|X - Y|^2]$. This can be done by solving the double integral

$$\int_0^1 \int_0^1 (x - y)^2 dx dy$$

or by writing

$$E[|X - Y|^2] = E[X^2] + E[Y^2] - 2E[XY] = 2(E[X^2] - E[X^2]) = 2 \text{Var}(X) = \frac{1}{6}$$

since $E[XY] = E[X^2] = E[X]E[X]$.

5.2 Covariance and Correlation

Definition 5.9 (covariance). The covariance between r.v.s X and Y is $\text{Cov}(X, Y) = E[(X - E X)(Y - E Y)]$ and measures the tendency of X and Y to go up or down together. Simplifying this using linearity, we get

def. covariance

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

Proposition 5.4 (properties of covariance). Let X and Y be any r.v.s. Then

1. $\text{Cov}(X, X) = \text{Var}(X)$,
2. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$,
3. $\text{Cov}(X, c) = 0$ for any constant c ,
4. $\text{Cov}(aX, Y) = a \text{Cov}(X, Y)$ for any constant a ,
5. $\text{Cov}(X + Y, Z + W) = \text{Cov}(X, Z) + \text{Cov}(X, W) + \text{Cov}(Y, Z) + \text{Cov}(Y, W)$,
6. $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y)$.

Definition 5.10 (correlation). The correlation between r.v.s X and Y is

def. correlation

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

(Correlation is undefined in the cases $\text{Var}(X) = 0$ or $\text{Var}(Y) = 0$.)

Definition 5.11 (uncorrelated r.v.s). If $\text{Cov}(X, Y) = 0$, then $\text{Corr}(X, Y) = 0$ and X and Y are uncorrelated.

def. uncorrelated r.v.s

Theorem 5.1 (independence and correlation). If X and Y are independent, then they are uncorrelated.

Proposition 5.5 (correlation bounds). For any r.v.s X and Y , $-1 \leq \text{Corr}(X, Y) \leq 1$.

5.3 Multinomial Distribution

Definition 5.12 If $\mathbf{X} \sim \text{Mult}_k(n, \mathbf{p})$, then

def.

$$P(X_1 = n_1, \dots, X_k = n_k) = \frac{n!}{n_1! \dots n_k!} \cdot p_1^{n_1} \dots p_k^{n_k}$$

for any n_1, \dots, n_k satisfying $n_1 + \dots + n_k = n$ and 0 otherwise. The multinomial distribution describes how many of n objects are independently placed into each of k categories where the p_k is the probability an object is placed into category j .

Proposition 5.6 (multinomial marginals). The marginal distributions of a multinomial variable are binomial. Namely, if $\mathbf{X} \sim \text{Mult}_k(n, \mathbf{p})$, then $X_j \sim \text{Bin}(n, p_j)$.

Proposition 5.7 (multinomial lumping). If $\mathbf{X} \sim \text{Mult}_k(n, \mathbf{p})$, then for any distinct i and j , $X_i + X_j \sim \text{Bin}(n, p_i + p_j)$.

Proposition 5.8 (multinomial conditioning). If $\mathbf{X} \sim \text{Mult}_k(n, \mathbf{p})$, then

$$(X_2, \dots, X_k) | X_1 = n_1 \sim \text{Mult}_{k-1}(n - n_1, (p'_2, \dots, p'_k))$$

where $p'_j = p_j / (p_2 + \dots + p_k)$.

Example 5.7 (superhero Statwoman). Statwoman will have $N \sim \text{Pois}(\lambda)$ battles next year. Each battle is independently with one of three villains. The probability a battle is with Villain j is p_j . Find the joint distribution of X_1, X_2, X_3 , where X_j is the number of battles Statwoman will have with Villain j . By LOTP,

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \sum_{n=0}^{\infty} P(X_1 = x_1, X_2 = x_2, X_3 = x_3 | N = n) P(N = n).$$

Notice there is only one nonzero term in this sum where $n = x_1 + x_2 + x_3$. Also notice that for a given n , $(X_1, X_2, X_3) \sim \text{Mult}_3(n, (p_1, p_2, p_3))$. Hence,

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2, X_3 = x_3) &= P(X_1 = x_1, X_2 = x_2, X_3 = x_3 | N = n) P(N = n), \\ &= \frac{n!}{x_1! x_2! x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3} \cdot e^{-\lambda} \frac{\lambda^n}{n!}, \\ &= \frac{e^{-\lambda p_1} (\lambda p_1)^{x_1}}{x_1!} \cdot \frac{e^{-\lambda p_2} (\lambda p_2)^{x_2}}{x_2!} \cdot \frac{e^{-\lambda p_3} (\lambda p_3)^{x_3}}{x_3!}, \end{aligned}$$

for all $x_1, x_2, x_3 \geq 0$ and 0 otherwise. Therefore, X_1, X_2, X_3 are independent with $X_j \sim \text{Pois}(\lambda p_j)$.

5.4 Multivariate Normal Distribution

Definition 5.13 (multivariate normal distribution, MVN distribution). If a k -dimensional random vector \mathbf{X} has a multivariate normal distribution, then every linear combination of its features has a normal distribution. That is, $t_1 X_1 + \dots + t_k X_k$ has a normal distribution for any constants t_1, \dots, t_k .

def. multivariate normal distribution, MVN distribution

Proposition 5.9 (MVN marginals). The marginal distributions of a MVN variable are normal. Namely, if \mathbf{X} is MVM, then X_j is normal.

Proposition 5.10 (subvectors of MVN variable). If (X_1, X_2, X_3) is MVN, then (X_1, X_2) is MVN.

Proposition 5.11 (concatenation of MVN variables). If (X_1, \dots, X_n) and (Y_1, \dots, Y_m) are MVN, then $(X_1, \dots, X_n, Y_1, \dots, Y_m)$ is MVM.

6 Transformations

6.1 Change of Variables

Theorem 6.1 (change of variables in one dimension). Let X be a continuous r.v. and $Y = g(X)$, where g is differentiable and strictly increasing or decreasing. Then $f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right|$, where $x = g^{-1}(y)$, and $S_Y = \{g(x) : x \in S_X\}$.

Example 6.1 (log-normal PDF). Let $X \in \mathcal{N}(0, 1)$ and $Y = e^X$. Since $g(x) = e^x$ is strictly increasing, we can use the change of variables formula to find the PDF of Y :

$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right| = \varphi(x) \frac{1}{e^x} = \varphi(\log y) \frac{1}{y}, \quad y > 0.$$

Notice we solve for $\left| \frac{dx}{dy} \right|$ by calculating $\left| \frac{dy}{dx} \right|^{-1}$, which, as in this case, is simpler. Also notice we write everything in terms of y and specify S_Y .

Example 6.2 (chi-square pdf). Let $X \in \mathcal{N}(0, 1)$ and $Y = X^2$. We cannot use the change of variables formula since $g(x) = x^2$ is not injective on S_X , so we start from the CDF. We inspect the graph of $y = x^2$ and realize with symmetry

$$F_Y(y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y}) = 2\Phi(\sqrt{y}) - 1.$$

Therefore,

$$f_Y(y) = \frac{\partial}{\partial y} 2\Phi(\sqrt{y}) - 1 = 2\varphi(\sqrt{y}) \cdot \frac{1}{2} y^{-1/2} = \varphi(\sqrt{y}) y^{-1/2}, \quad y > 0.$$

Example 6.3 (lighthouse). A lighthouse on a shore is shining a light toward the ocean at a random angle $U \sim \text{Unif}(-\pi/2, \pi/2)$ in radians. Let X be the point that the light hits on a parallel line 1 mile out from the shore. Namely, $X = \tan(U)$. Since $g(u) = \tan(u)$ is differentiable and injective over S_U , we can use the change of variables formula:

$$f_X(x) = f_U(u) \frac{du}{dx} = \frac{1}{\pi} \cdot \frac{1}{1+x^2}.$$

Theorem 6.2 (change of variables in all dimensions). Let $\mathbf{X} \in \mathbb{R}^n$ be a continuous random vector and $g : A_0 \rightarrow B_0$ be an invertible function where $A_0, B_0 \subseteq \mathbb{R}^n$ are open, $S_X \subseteq A_0$, and $S_Y \subseteq B_0$. Then if $Y = g(X)$,

$$f_Y(\mathbf{y}) = f_X(g^{-1}(\mathbf{y})) \cdot \left| \frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right|, \quad \mathbf{y} \in B_0$$

where $\left| \frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right|$ is the absolute value of the determinant of the Jacobian matrix

$$\frac{\partial \mathbf{x}}{\partial \mathbf{y}} = \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{pmatrix}.$$

Example 6.4 (Box-Muller). Let $U \sim \text{Unif}(0, 2\pi)$, and let $T \sim \text{Expo}(1)$ be independent of U . Let $X = \sqrt{2T} \cos U$ and $Y = \sqrt{2T} \sin U$. Then what is the PDF of X and Y ? Viewing

(X, Y) as a point in the plane,

$$\sqrt{X^2 + Y^2} = \sqrt{2T(\cos^2 U + \sin^2 U)} = \sqrt{2T}$$

is the from the origin and U is the angle. Hence, $(\sqrt{2T}, U)$ expresses (X, Y) in polar coordinates. Note that the transformation is invertible. The Jacobian matrix

$$\frac{\partial(x, y)}{\partial(u, t)} = \begin{pmatrix} -\sqrt{2t} \sin u & \frac{1}{\sqrt{2t}} \cos u \\ \sqrt{2t} \cos u & \frac{1}{\sqrt{2t}} \sin u \end{pmatrix}$$

exists, has continuous entries, and has absolute determinant $|\sin^2 u - \cos^2 u| = 1$. Letting $x = \sqrt{2t} \cos u$, $y = \sqrt{2t} \sin u$ to mirror the transformation from (U, T) to (X, Y) , we have

$$\begin{aligned} f_{X,Y}(x, y) &= f_{U,T}(u, t) \cdot \left| \frac{\partial(u, t)}{\partial(x, y)} \right| \\ &= \frac{1}{2\pi} e^{-t} \cdot 1 \\ &= \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)} \\ &= \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \cdot \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \end{aligned}$$

for all $x, y \in \mathbb{R}$. Since $f_{X,Y}$ factors into a function of x and a function of y , X and Y are independent. Marginally, both X and Y are i.i.d. $\mathcal{N}(0, 1)$ r.v.s.

7 Conditional Expectation

7.1 Given an Event

Consider the following example. If the average American lives to be 80 years old and Fred is 30, then does Fred on average have 50 years left to live? No. The average lifespan is 80 years old, but this is average includes the deaths of people younger than 30 years of age. Given that Fred is already 30, he has on average more than 50 years to live! To put this mathematically, letting T be Fred's lifespan,

$$E[T] < E[T | T \geq 30].$$

Definition 7.1 (conditional expectation (on an event)). Let A be an event with a positive probability. Then

def. conditional expectation (on an event)

$$E[Y | A] = \sum_y yP(y | A) \quad \text{if } Y \text{ is a discrete r.v., or}$$

$$E[Y | A] = \int_{-\infty}^{\infty} y f_{Y|A}(y | A) \quad \text{if } Y \text{ is a continuous r.v..}$$

By the fundamental bridge, since probabilities are expectations, LOTP is a special case of the law of total expectation (which itself is a special case of Adam's Law).

Definition 7.2 (law of total expectation). If A_1, \dots, A_n is a partition of the sample space, then

def. law of total expectation

$$E[Y] = \sum_{i=1}^n E[Y | A_i]P(A_i).$$

7.2 Given an R.V.

Since $X = x$ is an event, intuitively

$$E[Y | X = x] = \sum_y yP(Y = y | X = x) \quad \text{if } Y \text{ is a discrete r.v., or}$$

$$E[Y | X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y | X = x) \quad \text{if } Y \text{ is a continuous r.v..}$$

Notice that if X is unknown, then $E[Y | X]$ is a random variable.

Definition 7.3 (conditional expectation (given an r.v.)). The conditional expectation of Y given X , denoted by $E[Y | X]$, is a random variable $g(x) = E[Y | X = x]$.

def. conditional expectation (given an r.v.)

Theorem 7.1 (Adam's Law). For any r.v.s X and Y ,

$$E[E[Y | X]] = E[Y].$$