Notes for probability and statistics

John Kerl

February 4, 2009

Abstract

This is my primary reference for probability and statistics: I include what I feel to be the most important definitions and examples.

Probability content is taken from Dr. Tom Kennedy's splendid lectures for Math 564 (probability) at the University of Arizona in spring of 2007. That is, these are (except for my appendices) simply my hand-written class notes (with some examples omitted for brevity), made legible and searchable.

Statistics content is merged from Dr. Rabi Bhattacharya's 567A (theoretical statistics) course in spring 2008, Casella and Berger's *Statistical Inference*, and my own attempts to create a unified probability/statistics vocabulary and example base. The statistics content is, as of this writing, under construction.

Contents

1 What's the difference?

In probability, we start with a model describing what events we think are going to occur, with what likelihoods. The events may be random, in the sense that we don't know for sure what will happen next, but we do quantify our degree of surprise when various things happen.

The standard example is flipping a fair coin. "Fair" means, technically, that the probability of heads on a given flip is 50%, and the probability of tails on a given flip is 50%. This doesn't mean that every other flip will give a head — after all, three heads in a row is no surprise. Five heads in a row would be more surprising, and when you've seen twenty heads in a row you're sure that something fishy is going on. What the 50% probability of heads does mean is that, as the number of flips increases, we expect the number of heads to approach half the number of flips. Seven heads on ten flips is no surprise; 700,000 heads on 1,000,000 tosses is highly unlikely.

Another example would be flipping an unfair coin, where we know ahead of time that there's a 60% chance of heads on each toss, and a 40% chance of tails.

A third example would be rolling a loaded die, where (for example) the chances of rolling 1, 2, 3, 4, 5, or 6 are 25%, 5%, 20%, 20%, 20%, and 10%, respectively. Given this setup, you'd expect rolling three 1's in a row to be much more likely than rolling three 2's in a row.

As these examples illustrate, the probabilist starts with a probability model (something which assigns various percentage likelihoods of different things happening), then tells us which things are more and less likely to occur.

Key points about probability:

- Rules $\longrightarrow$ data: Given the rules, describe the likelihoods of various events occurring.
- Probability is about prediction — looking forward.
- Probability is mathematics.

The statistician turns this around:

- Rules $\longleftarrow$ data: Given only the data, try to guess what the rules were. That is, some probability model controlled what data came out, and the best we can do is guess — or approximate — what that model was. We might guess wrong; we might refine our guess as we get more data.
- Statistics is about looking backward.
- Statistics is an art. It uses mathematical methods, but it is more than math.
- Once we make our best *statistical* guess about what the probability model is (what the rules are), based on looking *backward*, we can then use that *probability* model to predict the *future*. (This is, in part, why I say that probability doesn't need statistics, but statistics uses probability.)

Here's my favorite example to illustrate. Suppose I give you a list of heads and tails. You, as the statistician, are in the following situation:

- You do not know ahead of time that the coin is fair. Maybe you've been hired to decide whether the coin is fair (or, more generally, whether a gambling house is committing fraud).

- You may not even know ahead of time whether the data come from a coin-flipping experiment at all.

Suppose the data are three heads. Your first guess might be that a fair coin is being flipped, and these data don't contradict that hypothesis. Based on these data, you might hypothesize that the rules governing the experiment are that of a fair coin: your probability model for predicting the future is that heads and tails each occur with 50% likelihood.

If there are ten heads in a row, though, or twenty, then you might start to reject that hypothesis and replace it with the hypothesis that the coin has heads on both sides. Then you'd predict that the next toss will certainly be heads: your new probability model for predicting the future is that heads occur with 100% likelihood, and tails occur with 0% likelihood.

If the data are "heads, tails, heads, tails, heads, tails", then again, your first fair-coin hypothesis seems plausible. If on the other hand you have heads alternating with tails not three pairs but 50 pairs in a row, then you reject that model. It begins to sound like the coin is not being flipped in the air, but rather is being flipped with a spatula. Your new probability model is that if the previous result was tails or heads, then the Next result is heads or tails, respectively, with 100% likelihood.

2 Probability

2.1 Events

2.1.1 Fundamental definitions

Definitions 2.1. When we do an **experiment**, we obtain an **outcome**. The set of all outcomes, conventionally written Ω, is called the **sample space**. (Mathematically, we only require Ω to be a set.)

Definition 2.2. An **event** is, intuitively, any subset of the sample space. Technically, it is any *measurable subset* of the sample space.

Example 2.3. The experiment is rolling a 6-sided die once. The outcome is the number of pips on the top face after the roll. The sample space Ω is $\{1, 2, 3, 4, 5, 6\}$. Example events are "the result of the roll is a 3" and "the result of the roll is odd".

Definition 2.4. Events A and B are **disjoint** if $A \cap B = \emptyset$.

Definition 2.5. A collection $\mathcal{F}$ of subsets of Ω is called a **σ-field** or **event space** if $\emptyset \in \mathcal{F}$, $\mathcal{F}$ is closed under countable unions, and $\mathcal{F}$ is closed under complements. (Note in particular that this means $\Omega \in \mathcal{F}$ and $\mathcal{F}$ is closed under countable intersections.)

Remark 2.6. There is a superficial resemblance with topological spaces: A topological space X has a topology $\mathcal{T}$ which is a collection of "open" subsets that is closed under complements, *finite* (rather than countable) intersection and *arbitrary* (rather than countable) union.

Examples 2.7. The smallest (or *coarsest*) σ-field for any Ω is $\{\emptyset, \Omega\}$; the largest (or *finest*) is 2^Ω, the set of all subsets of Ω. There may be many σ-fields in between. For example, if A is any subset of Ω which isn't $\emptyset$ or Ω, then one can check that the 4-element collection $\{\emptyset, A, A^c, \Omega\}$ is a σ-field.

Definition 2.8. A **probability measure** P is a function from a σ-field $\mathcal{F}$ to $[0, 1]$ such that:

- For all $A \in \mathcal{F}$, $P(A) \geq 0$;
- $P(\Omega) = 1$ and $P(\emptyset) = 0$;
- If $A_1, A_2, \ldots$, is a finite or countable subset of $\mathcal{F}$, with the A_i's all (pairwise) disjoint, then

$$P(\cup_i A_i) = \sum_i P(A_i).$$

  This is the **countable additivity** property of the probability measure P.

Remark 2.9. For uncountable Ω, 2^Ω is a σ-field, but it is impossible to define a probability measure on it; it is "too big". Consult your favorite textbook on Lebesgue measure for the reason why. For finite or countable Ω, on the other hand, 2^Ω is in fact what we think of for $\mathcal{F}$.

Definition 2.10. A **probability space** is a triple $(\Omega, \mathcal{F}, P)$ of a sample space Ω, a σ-field $\mathcal{F}$ on Ω, and a probability measure P on $\mathcal{F}$.

Remark 2.11. Technically, a probability space is nothing more than a measure space Ω with the additional requirement that $P(\Omega) = 1$.

2.1.2 Conditioning and independence

Definition 2.12. Let A, B be two events with $P(B) > 0$. Then we define the **conditional probability** of A **given** B to be

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

Mnemonic: *intersection over given*.

Notation 2.13. We will often write $P(A, B)$ in place of $P(A \cap B)$.

Definition 2.14. Two events A and B are **independent** (or **pairwise independent**) if

$$P(A \cap B) = P(A)P(B).$$

Mnemonic: write down $P(A) = P(A|B) = P(A \cap B)/P(B)$ and clear the denominators.

Remark 2.15. This is not the same as disjoint. If A and B are disjoint, then by countable additivity of P, we have

$$P(A \cup B) = P(A) + P(B).$$

Definition 2.16. Events $A_1, \ldots, A_n$ are **independent** if for all $I \subseteq \{1, 2, \ldots, n\}$,

$$P(\cap_{i \in I} A_i) = \prod_{i \in I} P(A_i).$$

Mnemonic: We just look at all possible factorizations.

Example 2.17. Three events A, B, and C are independent if

$$P(A \cap B) = P(A)P(B), P(A \cap C) = P(A)P(C), P(B \cap C) = P(B)P(C), \quad \text{and} \quad P(A \cap B \cap C) = P(A)P(B)P(C).$$

Theorem 2.18 (Partition theorem, or law of total probablity)**.** *Let $\{B_i\}$ be a countable partition of Ω and let A be an event. Then*

$$P(A) = \sum_i P(A|B_i)P(B_i).$$

Proof. Note that

$$P(A) = \sum_i P(A \cap B_i)$$

since the B_i's partition Ω. □

2.2 Discrete random variables

2.2.1 Definitions

Definition 2.19. A **random variable** X is a function $X : \Omega \to \mathbb{R}$. We say X is a **discrete random variable** if its range, $X(\Omega)$, is finite or countable.

Example 2.20. Roll two 6-sided dice and let X be their sum.

Definition 2.21. Given a random variable X, the **probability mass function** or **PMF** of X, written $f(x)$ or $f_X(x)$, is

$$f_X(x) = P(X = x) = P(X^{-1}(x)).$$

This is the probability that X's value is some specific real number x. Note that $X^{-1}(x)$, the **preimage** of x, is an event and so we can compute its probability using the probability measure P.

Definition 2.22. Let X_1 and X_2 be two random variables on two probability spaces $(\Omega_1, \mathcal{F}_1, P_1)$ and $(\Omega_2, \mathcal{F}_2, P_2)$, respectively. Then X_1 and X_2 are **identically distributed** if $X_1(x) = X_2(x)$ for all $x \in \mathbb{R}$.

Remark 2.23. Identically distributed discrete random variables X and Y have the same PMF.

Remark 2.24. Just as with events, which have a general measure-theoretic definition (definition 2.2), there is also a general measure-theoretic definition for random variables: they are simply measurable functions from a probability space to a measurable space.

2.2.2 Catalog of discrete random variables

Note: **mean** and **variance** are defined in section 2.2.3. They are included here for ready reference.

Bernoulli DRV:

- Parameter $p \in [0, 1]$.
- Range $X = \{0, 1\}$.
- PMF $P(X = 0) = p$, $P(X = 1) = 1 - p$.
- Example: Flip a p-weighted coin once.
- Mean: $1 - p$.
- Variance: $p(1 - p)$.

Binomial DRV:

- Two parameters: $p \in [0, 1]$ and $n \in \mathbb{Z}^+$.
- Range $X = \{0, 1, 2, \ldots, n\}$.
- PMF $P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$.
- Example: Flip a p-weighted coin n times; X is the number of heads.
- Mean: np.

- Variance: $np(1-p)$.

Remark 2.25. There is a trick to see that the sum of probabilities is 1 — recognize the sum of probabilities as the expansion of $(p+(1-p))^n$ using the binomial theorem:

$$1 = 1^n = (p+(1-p))^n = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k}.$$

Poisson DRV:

- Parameter $\lambda > 0$.
- Range $X = \{0, 1, 2, \ldots\}$.
- PMF $P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$.
- Example: Limiting case of binomial random variable with large n, small p, and $\lambda = np \approx 1$.
- Mean: λ.
- Variance: λ.

Geometric DRV:

- Parameter $p \in [0, 1]$.
- Range $X = \{1, 2, 3, \ldots\}$.
- PMF $P(X = k) = p(1-p)^{k-1}$.
- Example: Flip a p-weighted coin until you get heads; X is the number of flips it takes. (Note: some authors count the number of tails *before* the head.)
- Mean: $1/p$.
- Variance: $(1-p)/p^2$.

Negative binomial DRV:

- Two parameters $p \in [0, 1]$ and $n = 1, 2, 3, \ldots$.
- Range $X = \{n, n+1, n+2, \ldots\}$.
- PMF $P(X = k) = \binom{k-1}{n-1} p^n (1-p)^{k-n}$.
- Example: Flip a p-weighted coin until you get n heads; X is the number of flips it takes.
- Note: Deriving the $\binom{k-1}{n-1}$ factor is a non-trivial counting problem; it is deferred until later in the course.
- Mean: n/p.
- Variance: $n(1-p)/p^2$.

2.2.3 Expectations

Definition 2.26. Functions of a discrete random variable: If $X : \Omega \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ then $g(X) : \Omega \to \mathbb{R}$ is another random variable. Let $Y = g(X)$.

To find the PMF of $g(X)$ given PMF of (X): write the latter as $f_X(x) = P(X = x)$. Then

$$P(g(X) = y) = \sum_{x \in g^{-1}(y)} P(X = x) = \sum_{x \in g^{-1}(y)} f_X(x).$$

Definition 2.27. Let X be a discrete random variable with PMF $f_X(x)$. If

$$\sum_x |x| f_X(x) < \infty$$

(i.e. if we have absolute convergence) then we define the **expected value** (also call **expectation** or **mean**) of X to be

$$E[X] = \sum_x x f_X(x) = \sum_x x P(X = x).$$

Mnemonic: This is just the weighted sum of possible X values, weighted by their probabilities.

Theorem 2.28 (Law of the Unconscious Statistician)**.** *Let X be a discrete random variable and $g : \mathbb{R} \to \mathbb{R}$. Let $Y = g(X)$. If*

$$\sum_x |g(x)| \, f_X(x) < \infty$$

then

$$E[Y] = \sum_x g(x) f_X(x).$$

Definition 2.29. The **variance** of X, written $\sigma^2(X)$ or $\mathrm{Var}(X)$, is

$$\sigma^2(X) = E[(X - E[X])^2].$$

Definition 2.30. The square root of the variance is the **standard deviation**, written $\sigma(X)$ or σ_X.

Proposition 2.31. *Let $\mu = E[X]$. Then*

$$\sigma^2(X) = E[X^2] - E[X]^2.$$

Proof. FOIL, and use linearity of expectation. □

Proposition 2.32. $\mathrm{Var}(cX) = c^2 \mathrm{Var}(X)$.

Proof.

$$\mathrm{Var}(cX) = E[c^2X^2] - E[cX]^2 = c^2E[X^2] - c^2E[X]^2 = c^2\left(E[X^2] - E[X]^2\right) = c^2\mathrm{Var}(X).$$

□

Theorem 2.33. *Let X be a discrete random variable; let $a, b \in \mathbb{R}$. Then*

(i) $E[aX + b] = aE[X] + b$.

(ii) If $P(X = b) = 1$ then $E[X] = b$.

(iii) If $a \leq X \leq b$ then $a \leq E[X] \leq b$.

(iv) If $g, h : \mathbb{R} \to \mathbb{R}$ and $g(X), h(X)$ have finite means, then $E[g(X) + h(X)] = E[g(X)] + E[h(X)]$.

Definition 2.34. Let X be a discrete random variable and let B be an event. The **conditional PMF** of X given B is

$$f(x|B) = P(X = x|B).$$

The **conditional expectation** of X given B is

$$E[X|B] = \sum_x x f(x|B)$$

provided (as usual) that the sum converges absolutely.

Theorem 2.35 (Partition theorem, or law of total expectation)**.** *Let $\{B_i\}$ be a countable partition of Ω and let X be a random variable. Then*

$$E[X] = \sum_i E[X|B_i]P(B_i).$$

2.3 Multiple discrete random variables

2.3.1 Definitions

Definition 2.36. We define the **joint PMF** or **joint density** of X and Y to be

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

Proposition 2.37. *For $A \subseteq \mathbb{R}^2$ we have*

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

Corollary 2.38. *Let $g(x,y) : \mathbb{R}^2 \to \mathbb{R}$. Let X and Y be discrete random variables and let $Z = g(X,Y)$. Then*

$$f_Z(z) = P(Z = z) = \sum_{z\in g^{-1}(x,y)} f_{X,Y}(x,y).$$

We can use joint densities to recover **marginal** densities:

Corollary 2.39. *Let X and Y be discrete random variables with joint density $f_{X,Y}$. Then*

$$f_X(x) = \sum_y f_{X,Y}(x,y) \qquad \text{and} \qquad f_Y(y) = \sum_x f_{X,Y}(x,y).$$

2.3.2 Expectations

Theorem 2.40 (Law of the Unconscious Statistician)**.** *Let X and Y be discrete random variables and let $g(x,y) : \mathbb{R}^2 \to \mathbb{R}$. Let $Z = g(X,Y)$. Then*

$$E[Z] = \sum_{x,y} g(x,y) f_{X,Y}(x,y).$$

Proof. As in the single-variable case (theorem 2.40). □

Corollary 2.41. *Let X and Y be discrete random variables, and let $a, b \in \mathbb{R}$. Then*

$$E[aX + bY] = aE[X] + bE[Y].$$

Proof. Use $g(x,y) = ax + by$, and use the theorem twice. □

2.3.3 Independence

Recall definition 2.14 of independent events: A and B are independent if $P(A \cap B) = P(A)P(B)$. We use this to define independence of discrete random variables.

Definition 2.42. Two discrete random variables X and Y are **independent** if, for all x and y,

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

Using PMF notation, we say X and Y are independent if

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

for all x and y, i.e. if the joint PMF **factors**.

Notation 2.43. We often abbreviate *independent and identically distributed* (definitions 2.42 and 2.22) as **IID**.

Question: given only the joint density of X and Y, can we tell if X and Y are independent? From corollary 2.39, we can recover the PMFs of X and Y:

$$f_X(x) = \sum_y f_{X,Y}(x,y) \qquad \text{and} \qquad f_Y(y) = \sum_x f_{X,Y}(x,y).$$

Then we can multiply them back together and see if we get the joint density back.

Point:

- If X and Y are independent, then we can go from marginals to joint PMFs by multiplying.
- We can always go from joint PMFs to marginals by summing as above.

Theorem 2.44. *If X and Y are independent discrete random variables, then*

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

Theorem 2.45. *If X and Y are independent discrete random variables and $g, h : \mathbb{R} \to \mathbb{R}$ then $g(X)$ and $h(Y)$ are independent discrete random variables.*

Corollary 2.46. *In particular,*

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$$

provided $g(X)$ and $h(Y)$ have finite mean.

Theorem 2.47. *If X and Y are independent discrete random variables then*

$$\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y).$$

Proof. Use definition 2.29 and theorem 2.44. □

Definition 2.48. We say that X and Y are **uncorrelated** if

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

Remark 2.49. Independent implies uncorrelated but not vice versa.

Remark 2.50. Theorem 2.47 holds for uncorrelated discrete random variables.

2.3.4 Sample mean

Definition 2.51. Let $X_1, \ldots, X_n$ be independent and identically distributed. (Think of multiple trials of the same experiment.) Since each X_i has the same expectation, we call their common mean the **population mean** and denote it by μ. Likewise, we call their common variance the **population variance** and denote it by σ^2. Let

$$\overline{X} = \frac{\sum_{i=1}^n X_i}{n}.$$

This is a new random variable, called the **sample mean** of $X_1, \ldots, X_n$.

By linearity of expectation,

$$E[\overline{X}] = \frac{1}{n}\sum_{i=1}^{n} E[X_j] = \mu.$$

Recall from proposition that the variance scales as $\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$. Also, since the X_i's are independent, their variances add. Thus

$$\mathrm{Var}(\overline{X}) = \mathrm{Var}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{n}{n^2}\sigma^2 = \frac{\sigma^2}{n}.$$

2.3.5 Moment-generating functions and characteristic functions

Definitions 2.52. Let X be a discrete random variable. Then the **moment-generating function** or **MGF** of X is

$$M_X(t) = E[e^{tX}].$$

By the Law of the Unconscious Statistician (theorem 2.28), this is

$$M_X(t) = \sum_x e^{tx} f_X(x).$$

Likewise, the **characteristic function** of X is

$$\beta_X(t) = E[e^{itX}]$$

which is

$$\beta_X(t) = \sum_x e^{itx} f_X(x).$$

Remark 2.53. These functions are just computational tricks. There is no intrinsic meaning in these functions.

Proposition 2.54. *Let $M_X(t)$ be the moment-generating function for a discrete random variable X. Then*

$$E[X^k] = M_X^{(k)}(0) = \frac{d^k}{dt^k} M_X(t)\bigg|_{t=0}.$$

Proposition 2.55. *If X and Y are identically distributed, they have the same PMF (remark 2.23) and thus they also have the same MGF.*

Proposition 2.56. *Let X and Y be independent discrete random variables and let $Z = X + Y$. Then*

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

2.3.6 Sums of discrete random variables

Moment-generating functions are perhaps a better approach than the following.

Let X and Y be discrete random variables and let $Z = X + Y$. We can find the PMF of Z by

$$f_Z(z) = P(Z = z) = P(X + Y = z) = \sum_{x+y=z} f_{X,Y}(x, y) = \sum_x \sum_{y=z-x} f_{X,Y}(x, y) = \sum_x f_{X,Y}(x, z - x).$$

Now further suppose that X and Y are independent. Then

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

so

$$f_Z(z) = \sum_x f_X(x) f_Y(z-x).$$

This is the **convolution** of f_X and f_Y. Note that, as always in convolutions, we are summing over all the ways in which x and y can add up to z.

2.4 Continuous random variables

2.4.1 Definitions

A continuous random variable, mimicking definition 2.19, is a function from Ω to $\mathbb{R}$ with the property that for all $x \in \mathbb{R}$, $P(X = x) = 0$. Thus, the PMF which we used for discrete random variables is not useful. Instead we first define another function, namely, $P(X \leq x)$.

Definition 2.57. The **cumulative distribution function** or **CDF** for a random variable (whether discrete or continuous) is

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

Theorem 2.58. *The CDF $F_X(x)$ for a random variable satisfies the following properties:*

- $0 \leq F_X(x) \leq 1$, *and* F_X *is non-decreasing.*
- $\lim_{x\to-\infty} F_X(x) = 0$ *and* $\lim_{x\to+\infty} F_X(x) = 1$.
- $F_X(x)$ *is right-continuous (in the sense from introductory calculus).*

Definition 2.59. A random variable X is a **continuous random variable** if there exists a function $f_X(x)$, called the **probability density function** of **PDF**, such that the CDF $F_X(x)$ is given by $\int_{-\infty}^{+\infty} f_X(x)\,dx$.

Remark 2.60. Note that the integral is done using Lebesgue measure (or, for the purposes of this course, Riemann integration). If we allow counting measure and require countable range, then we can subsume discrete random variables into this definition. However, that is outside the scope of this course.

Remark 2.61. Some random variables are neither discrete nor continuous; these appear for example in dynamical systems.

Remark 2.62. The PDF and CDF are related as follows (making use of the second fundamental theorem of calculus):

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x} f_X(t)\,dt \\ f_X(x) &= \frac{d}{dx} F_X(x). \end{aligned}$$

Remark 2.63. The PDF of X is the probability that X lies in some interval. E.g.

$$P(a \leq X \leq b) = \int_a^b f_X(x)\,dx.$$

Thus PDFs are non-negative and have integral 1.

Remark 2.64. We can now neatly define some terminology from statistics. Namely:

- Let X be a random variable from $\mathbb{R}$ to $\mathbb{R}$ with CDF $F_X(x)$ and $f_X(x)$.
- The **mean** of X is the expectation value $E[X]$ as defined below.
- The **median** of X is $F_X^{-1}(0.5)$.
- A **mode** of X is a local maximum of $f_X(x)$. If the PDF has two local maxima, we say that X is **bimodal**. If the PDF has a single maximum, we call it *the* mode of X.

2.4.2 Catalog of continuous random variables

Note: **mean** and **variance** are defined in section 2.4.6. They are included here for ready reference. I thought about including graphs of the PDFs and CDFs, but instead I will refer you to the excellent Wikipedia article on *Probability distribution*, and the pages linking from there.

Uniform CRV:

- Parameters $a < b$.
- Range $[a, b]$.
- PDF
$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, \text{elsewhere.} \end{cases}$$
- CDF
$$F_X(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \le x < b \\ 0, & b \le x \end{cases}$$
- Mean: $(a+b)/2$.
- Variance: $(b-a)^2/12$.

Exponential CRV:

- Parameter $\lambda > 0$.
- Range $X = \{0, \infty\}$.
- PDF
$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \ge 0 \\ 0, & x < 0. \end{cases}$$
- CDF
$$F_X(x) = \begin{cases} 1 - e^{-\lambda x}, & x \ge 0 \\ 0, & x < 0. \end{cases}$$
- Mean: $1/\lambda$.
- Variance: $1/\lambda^2$.

Cauchy CRV:

- No parameters.
- Range $X = \{-\infty, \infty\}$.
- PDF
$$f_X(x) = \frac{1}{\pi(1+x^2)}.$$

- CDF
$$F_X(x) = \frac{\tan^{-1}(x)}{\pi} + \frac{1}{2}.$$
- Mean: infinite.
- Variance: infinite.

Normal CRV (see section 2.4.3):

- Parameters $\mu \in \mathbb{R}$, $\sigma > 0$.
- Note: with $\mu = 0$ and $\sigma = 1$ we have the **standard normal distribution**. One says it has *zero mean and unit variance*.
- Range $X = \{-\infty, \infty\}$.
- PDF
$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right].$$
- CDF: No closed-form expression. Related to $\text{erf}(x)$ but not quite the same: The standard normal CDF is $\frac{1}{2}(1 + \text{erf}(x/\sqrt{2}))$. Note that some computing systems may have $\text{erf}(x)$ but not `normalcdf`, or vice versa. Thus this conversion formula comes in handy.
- Mean: μ.
- Variance: σ^2.

Gamma CRV (see section 2.4.4):

- Parameters $\lambda > 0$, $w > 0$.
- Range $x \geq 0$.
- PDF
$$f_X(x) = \frac{\lambda^w}{\Gamma(w)}\, x^{w-1} e^{-\lambda x}, x \geq 0.$$
- CDF TBD.
- Mean: w/λ.
- Variance: w/λ^2.

Beta CRV:

- Parameters $\alpha > 0$, $\beta > 0$.
- Range $x \in [0, 1]$.
- PDF
$$f_X(x) = \frac{x^\alpha (1-x)^\beta}{B(\alpha, \beta)}.$$
- CDF TBD.
- Mean: $\alpha/(\alpha + \beta)$.
- Variance: $\alpha\beta$...

2.4.3 The normal distribution

Definition 2.65. The **normal distribution** has two parameters: real μ and positive σ. A random variable with the normal distribution has PDF

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right].$$

Remark 2.66. The normal distribution has mean μ and variance σ^2.

Definition 2.67. With $\mu = 0$ and $\sigma = 1$ we have the **standard normal distribution**.

2.4.4 The gamma distribution

Definition 2.68. The **gamma function** is defined by

$$\Gamma(\omega) = \int_0^\infty x^{w-1}e^{-x}dx.$$

Remark 2.69. Integration by parts shows that $\Gamma(n) = (n-1)!$. Thus, the gamma function is a generalized factorial function.

Definition 2.70. The **gamma distribution** has two parameters $\lambda, w > 0$. A random variable with the gamma distribution has PDF

$$f(x) = \begin{cases} \frac{\lambda^w}{\Gamma(w)}x^{w-1}e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0. \end{cases}$$

Remark 2.71. With $w = 1$ we have an exponential distribution. Thus, the gamma distribution generalizes the exponential distribution.

Proposition 2.72. *The gamma distribution has mean w/λ.*

2.4.5 Functions of a single random variable

Given a continuous random variable X and $g(x) : \mathbb{R} \to \mathbb{R}$, we have a new continuous random variable $Y = g(X)$. Often one wants to find the PDF of Y given the PDF of X. The **method** is to first find the CDF of Y, then differentiate to find the PDF.

Example 2.73. Let X be uniform on $[0, 1]$ and let $Y = X^2$. The CDF of Y is

$$P(Y \leq y) = P(X^2 \leq y) = P(X \leq \sqrt{y}) = \int_0^{\sqrt{y}} 1\,dx = \sqrt{y}$$

i.e.

$$F_Y(y) = \begin{cases} 0, & y < 0 \\ \sqrt{y}, & 0 \leq y < 1 \\ 1, & 1 \leq y. \end{cases}$$

Then the PDF is

$$f_Y(y) = \begin{cases} 0, & y < 0 \\ \frac{1}{2\sqrt{y}}, & 0 \leq y < 1 \\ 0, & 1 \leq y. \end{cases}$$

The critical step in the method is going from $P(X^2 \leq y)$ to $P(X \leq \sqrt{y})$.

2.4.6 Expectations

Definition 2.74. Let X be a continuous random variable with PDF $f_X(x)$. If

$$\int_{-\infty}^{+\infty} |x| f_X(x)\, dx < \infty$$

(i.e. if we have absolute convergence) then we define the **expected value** (also call **expectation** or **mean**) of X to be

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

Mnemonic: As in the discrete case, this is just the weighted sum of possible X values, weighted by their probabilities.

Definition 2.75. Let X be a continuous random variable. The **variance** of X is

$$E[(X-\mu)^2]$$

where $\mu = E[X]$. By the corollary to the Law of the Unconscious Statistician (below),

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

Theorem 2.76 (Law of the Unconscious Statistician)**.** *Let X be a continuous random variable and $g : \mathbb{R} \to \mathbb{R}$. Let $Y = g(X)$. If*

$$\int_{-\infty}^{+\infty} |g(x)|\, f_X(x)\, dx < \infty$$

then

$$E[Y] = \int_{-\infty}^{+\infty} g(x)\, f_X(x)\, dx < \infty$$

Corollary 2.77. *If $g, h : \mathbb{R} \to \mathbb{R}$ and $a, b \in \mathbb{R}$ then*

$$E[ag(X) + bh(X)] = aE[g(X)] + bE[h(X)].$$

2.5 Multiple continuous random variables

2.5.1 Definitions

Definition 2.78. Given two continuous random variables X and Y, we define their **joint CDF** to be

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

Definition 2.79. Two random variables X and Y are **jointly continuous** if there exists a function $f_{X,Y}(x,y)$ (their **joint PDF**) such that

$$F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u,v)\, dv\, du.$$

The **joint PDF** is thought of as

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_{x=a}^{x=b} \int_{y=c}^{y=d} f_{X,Y}(x,y)\, dy\, dx$$

and more generally for $A \subseteq \mathbb{R}^2$

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

The joint CDF and joint PDF are related by

$$\begin{aligned} f_{X,Y}(x,y) &= \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y) \\ F_{X,Y}(x,y) &= \int_{-\infty}^{x} \int_{-\infty}^{y} f(u,v)\, dv\, du. \end{aligned}$$

Given a joint CDF $F_{X,Y}(x,y)$ we may recover the **marginal** CDFs $F_X(x)$ and $F_Y(y)$ by

$$F_X(x) = \lim_{y \to +\infty} F_{X,Y}(x,y)$$

and similarly for $F_Y(y)$.

How do we recover the **marginal** PDFs given the joint PDFs? To derive the formula, differentiate the CDF:

$$\begin{aligned} f_X(x) &= \frac{d}{dx} F_X(x) = \frac{d}{dx} F_{X,Y}(x, +\infty) \\ &= \frac{d}{dx} \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{x} f_{X,Y}(u,v)\, du \right] dv \\ &= \frac{d}{dx} \int_{-\infty}^{x} \left[\int_{-\infty}^{+\infty} f_{X,Y}(u,v)\, dv \right] du \\ &= \int_{-\infty}^{+\infty} f_{X,Y}(x,v)\, dv \end{aligned}$$

i.e. the formula is

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

That is, we **integrate away** one variable.

2.5.2 Independence

Definition 2.80. Two continuous random variables X and Y are **independent** iff their CDFs factor:

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

Remark 2.81. For discrete random variables, we defined independence (definition 2.42) in terms of the factoring of the PMFs; for continuous random variables, we define independence in terms of the factoring of the CDFs. Factorization of PDFs does hold (as shown in the following theorem), but it is a consequence rather than a definition.

Theorem 2.82. *Two continuous random variables X and Y are **independent** iff their PDFs factor:*

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

Proof.

$$\begin{aligned} f_{X,Y}(x,y) &= \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y) \\ &= \frac{\partial^2}{\partial x \partial y} \left[F_X(x)F_Y(y)\right] \\ &= \frac{\partial}{\partial x}\left[F_X(x)\right] \frac{\partial}{\partial y}\left[F_Y(y)\right] \\ &= f_X(x)f_Y(y). \end{aligned}$$

□

2.5.3 Expectations

Theorem 2.83 (Law of the Unconscious Statistician)**.** *Let X and Y be continuous random variables and $g(x,y) : \mathbb{R}^2 \to \mathbb{R}$. Let $Z = g(X,Y)$. If*

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |g(x,y)|\, f_{X,Y}(x,y)\, dx\, dy < \infty$$

then

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

Corollary 2.84. *Let X and Y be continuous random variables and let $a, b \in \mathbb{R}$. Then*

$$E[aX + bY] = aE[X] + bE[Y].$$

Theorem 2.85. *Let X and Y be independent continuous random variables and let $g, h : \mathbb{R} \to \mathbb{R}$. Then*

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)].$$

Corollary 2.86. *Let X and Y be independent continuous random variables. Then*

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

Corollary 2.87. *Let X and Y be independent continuous random variables. Then*

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

Proof. Computation. □

Remark 2.88. Expectations always add. They multiply only when X and Y are independent. Variances add only when X and Y are independent.

Theorem 2.89. *Let X and Y be independent random variables and let $g, h : \mathbb{R} \to \mathbb{R}$. Then $g(X)$ and $h(Y)$ are independent.*

2.5.4 The IID paradigm: S_n and $\overline{X}_n$

Notation: Let X_n be a sequence of IID random variables, with common mean μ and common variance σ^2. We write

$$S_n = \sum_{i=1}^{n} X_n$$

and

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_n.$$

The latter is called the **sample mean** of the X_n's.

Mean and variance of S_n: We already know

$$E[S_n] = E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i] = n\mu$$

and (using independence of the trials to split up the variance)

$$\mathrm{Var}(S_n) = \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = n\sigma^2.$$

Mean and variance of $\overline{X}_n$: Likewise,

$$E[\overline{X}_n] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n} n\mu = \mu$$

and (using independence of the trials to split up the variance)

$$\mathrm{Var}(\overline{X}_n) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

2.5.5 Functions of multiple random variables

Let X, Y be continuous random variables and let $g(x, y) : \mathbb{R}^2 \to \mathbb{R}$. Let $Z = g(X, Y)$. How do we find the PDF of Z? As in the univariate case (section 2.4.5), the **method** is to first find the CDF of Z, then differentiate.

[Need to type up the example from 3-21.]

[Need to type up convolution notes for independent X, Y from 3-23.]

$$f_Z(z) = \int f_X(x) f_Y(z - x) dx.$$

2.5.6 Moment-generating functions and characteristic functions

Definition 2.90. Much as in the discrete case (definitions 2.52),

$$M_X(t) = E[e^{tX}] = \int_{-\infty}^{+\infty} e^{tx} f_X(x) dx$$

and

$$\beta_X(t) = E[e^{itX}] = \int_{-\infty}^{+\infty} e^{itx} f_X(x) dx.$$

Proposition 2.91. *We have:*

(i) $M_X^{(k)}(0) = E[X^k]$.

(ii) If $Y = aX + b$, *then* $M_Y(t) = e^{bt} M_X(at)$.

(iii) If X *and* Y *are independent, then* $M_{X+Y}(t) = M_X(t) M_Y(t)$.

Example 2.92. The standard normal random variable Z has (tedious computations omitted here) moment-generating function

$$M_Z(t) = \exp(t^2/2).$$

The general normal random variable $X = \mu + \sigma Z$ has moment-generating function

$$M_X(t) = \exp(\mu t + \sigma^2 t^2/2).$$

Proposition 2.93. *Let* $X_1, \ldots, X_n$ *be independent normal random variables with means* μ_i *and variances* σ_i^2. *Then* $Y = \sum X_i$ *has normal distribution with mean* $\sum \mu_i$ *and variance* $\sum \sigma_i^2$.

2.5.7 Change of variables

Let X and Y be continuous random variables. If $T : \mathbb{R}^2 \to \mathbb{R}^2$ is invertible, sending $(U, V) = T(X, Y)$, then

$$\int_D f(x, y)\, dx\, dy = \int_{T(D)} f(x(u, v), y(u, v))\ |J(u, v)|\ du\, dv$$

where $J(u, v)$ is the **Jacobian matrix**

$$J(u, v) = \begin{pmatrix} \partial x/\partial u & \partial x/\partial v \\ \partial y/\partial u & \partial y/\partial v \end{pmatrix}.$$

Theorem 2.94. *Let* X *and* Y *be jointly continuous random variables with PDF* $f_{X,Y}(x, y)$. *Let*

$$D = \{(x, y) : f_{X,Y}(x, y) > 0\}$$

i.e. the ***range*** *of* X, Y. *Suppose* $T : D \to S \subseteq \mathbb{R}^2$ *is 1-1 and onto. Define new random variables* U *and* V *by* $(U, V) = T(X, Y)$. *Then*

$$f_{U,V}(u, v) = \begin{cases} f_{X,Y}(x(u, v), y(u, v))\ |J(u, v)| & \text{if} (u, v) \in S \\ 0 & \text{otherwise} \end{cases}$$

where $J(u, v)$ *is as above.*

2.5.8 Conditional density and expectation

Given two continuous random variables X and Y, we want to define $P(X|Y = y)$. Since $Y = y$ is a null event, the usual intersection-over-given notion of conditional probability (see section 2.1.2) will give us zero divided by zero. Somewhat as in l'Hôpital's rule in calculus, we can nonetheless make sense of it.

Definition 2.95. Let X, Y be jointly continuous with PDF $f_{X,Y}(x, y)$. The **conditional density** of X given Y is

$$f_{X|Y}(x|y) = \begin{cases} \frac{f_{X,Y}(x,y)}{f_Y(y)}, & f_Y(y) \neq 0 \\ 0, & f_Y(y) = 0. \end{cases}$$

Remark 2.96. Recall that

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

so we can think of the conditional density of X given Y as being

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{\int_{x=-\infty}^{x=+\infty} f_{X,Y}(x, y)dx}$$

whenever the denominator is non-zero.

Definition 2.97. Let X, Y be jointly continuous with PDF $f_{X,Y}(x, y)$. The **conditional expectation** of X given Y is

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

Definition 2.98. $E[X|Y = y]$ is a function of y; it must depend only on y. Call it $g(y)$. Then $g(Y)$ is a random variable, which we write as

$$E[X|Y].$$

This new random variable has the following properties.

Theorem 2.99. *Let X, Y, Z be random variables, $a, b \in \mathbb{R}$, and $g : \mathbb{R} \to \mathbb{R}$. Then*

- $E[a|Y] = a$.
- $E[aX + bZ|Y] = eE[X|Y] + bE[Z|Y]$. *(Note: this is linearity on the left; linearity on the right emphatically does* not *hold.)*
- *If $X \geq 0$ then $E[X|Y] \geq 0$.*
- *If X, Y are independent then $E[X|Y] = E[X]$. (Mnemonic: Y gives no information about X.)*
- $E[E[X|Y]] = E[X]$. *(This is the partition theorem in disguise. See below.)*
- $E[Xg(Y)|Y] = g(Y)E[X|Y]$. *(Mnemonic: given a specific y, $g(y)$ is constant.)*
- *Special case: $E[g(Y)|Y] = g(Y)$.*

2.5.9 The bivariate normal distribution

If X and Y are independent and standard normal, then their joint PDF is

$$f_{X,Y}(x,y) = \frac{1}{2\pi}\exp\left(-\frac{x^2+y^2}{2}\right). \qquad (*)$$

It is possible for X and Y to *not* be independent, while their marginals *are* still standard normal. In fact, there is a 1-parameter family of such X, Y pairs.

Definition 2.100. Let $-1 < \rho < 1$. The **bivariate normal distribution** with parameter ρ has PDF

$$f_{X,Y}(x,y)\frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left(-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2}\right).$$

Remark 2.101. Note the following:

- It is straightforward but tedious to verify that $\int\int f_{X,Y}(x,y) = 1$.
- With $\rho = 0$ we obtain equation (*) as a special case.
- One may complete the square and use translation invariance of the integral to find that the marginals are in fact univariate standard normals.
- Again completing the square, one finds that $f_{Y|X}(y|x)$ is normal with $\mu = \rho x$ and $\sigma^2 = 1-\rho^2$.

2.5.10 Covariance and correlation

Definition 2.102. Let X and Y be random variables with means μ_X and μ_Y, respectively. The **covariance** of X and Y is

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

Mnemonic 2.103. With $Y = X$ we recover the familiar formula (definition 2.75) for the variance of X: $E[X^2] - E[X]^2$.

Theorem 2.104. *Let X and Y be random variables. Then*

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + 2\mathrm{Cov}(X,Y) + \mathrm{Var}(Y).$$

Remark 2.105. Bill Faris calls this the "most important theorem in probability".

Corollary 2.106. *If* $\mathrm{Cov}(X,Y) = 0$ *then* $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Theorem 2.107. *If X and Y are independent then* $\mathrm{Cov}(X,Y) = 0$. *The converse does not hold.*

Definition 2.108. Let X and Y be random variables with variances σ_X^2 and σ_Y^2, respectively. The **correlation coefficient** of X and Y is

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}.$$

Remark 2.109. The covariance is quadratic in X and Y (with respect to linear rescaling); the correlation coefficient is scale-invariant.

Theorem 2.110. *The correlation coefficient satisfies*

$$-1 \le \rho(X,Y) \le 1.$$

Remark 2.111. Equivalently,

$$\mathrm{Cov}(X,Y)^2 \leq \mathrm{Var}(X)\mathrm{Var}(Y).$$

Proof. It suffices to show $|\rho(X,Y)| \leq 1$, which is equivalent to showing $|\mathrm{Cov}(X,Y)| \leq \sigma_X\sigma_Y$.

This follows abstractly from the Cauchy-Schwarz inequality,

$$|\langle f, g\rangle| \leq \|f\|\,\|g\|,$$

with $f = X - \mu_X$ and $g = Y - \mu_Y$. Namely,

$$\begin{aligned}
\langle f, g\rangle^2 &\leq \|f\|^2\,\|g\|^2 = \langle f, f\rangle\langle g, g\rangle \\
\left(\int_\Omega (X-\mu_X)(Y-\mu_Y)\,dP\right)^2 &\leq \int_\Omega (X-\mu_X)^2\,dP \int_\Omega (Y-\mu_Y)^2\,dP \\
E[(X-\mu_X)(Y-\mu_Y)]^2 &\leq E[(X-\mu_X)^2]\,E[(Y-\mu_Y)^2] \\
\mathrm{Cov}(X,Y)^2 &\leq \mathrm{Var}(X)\mathrm{Var}(Y).
\end{aligned}$$

Faris sketches another route, which I complete here. **Normalize** X and Y as follows. Let μ_X, μ_Y, σ_X, and σ_Y be their means and standard deviations, respectively. Then

$$\frac{X-\mu_X}{\sigma_X} \quad \text{and} \quad \frac{Y-\mu_Y}{\sigma_Y}$$

each have zero mean and unit standard deviation. (In particular this will mean, below, that their second moments are 1.) We can create a new pair of random variables

$$\left(\frac{X-\mu_X}{\sigma_X} \pm \frac{Y-\mu_Y}{\sigma_Y}\right)^2.$$

Since each takes non-negative values, the means are non-negative as well [xxx xref forward to where this is proved ... it seems obvious but actually requires proof]:

$$E\left[\left(\frac{X-\mu_X}{\sigma_X} \pm \frac{Y-\mu_Y}{\sigma_Y}\right)^2\right] \geq 0.$$

FOILing out we have

$$E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)^2 \pm \frac{2(X-\mu_X)(Y-\mu_Y)}{\sigma_X\sigma_Y} + \left(\frac{Y-\mu_Y}{\sigma_Y}\right)^2\right] \geq 0.$$

Using the linearity of expectation and recalling that the normalized variables have second moments equal to 1, we have

$$\begin{aligned}
2 \pm 2E\left[\frac{(X-\mu_X)(Y-\mu_Y)}{\sigma_X\sigma_Y}\right] &\geq 0 \\
-1 \leq E\left[\frac{(X-\mu_X)(Y-\mu_Y)}{\sigma_X\sigma_Y}\right] &\leq 1 \\
-\sigma_X\sigma_Y \leq E\left[(X-\mu_X)(Y-\mu_Y)\right] &\leq \sigma_X\sigma_Y \\
-\sigma_X\sigma_Y \leq \mathrm{Cov}(X,Y) &\leq \sigma_X\sigma_Y \\
|\mathrm{Cov}(X,Y)| &\leq \sigma_X\sigma_Y.
\end{aligned}$$

□

2.6 Laws of averages

Here is a statistics paradigm. Run an experiment n with IID random variables X_n (whether continuous or discrete). The n-tuple $(X_1, \ldots, X_n)$ is called a **sample**. The average of X_1 through X_n is called the **sample mean**, written $\overline{X}_n$; it is also a random variable.

The big question is: what does $\overline{X}_n$ look like as n gets large? For example, roll a 6-sided die (so $\mu = 3.5$) a million times. What is the probability of the event $|\overline{X}_n - 3.5| > 0.01$? One would hope this probability would be small, and would get smaller as n increases.

We have two main theorems here:

- The law of large numbers says that $\overline{X}_n \to \mu$, although we need to define the notion of convergence of a random variable to a real number. There are two flavors of convergence: weak and strong.
- The central limit theorem describes the PDF of $\overline{X}_n$.

2.6.1 The weak law of large numbers

Definition 2.112. Let X_n and X be random variables. We say $X_n \to X$ **in probability** if for all $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left(\{\omega \in \Omega : |X_n(\omega) - X(\omega)| \geq \varepsilon\}\right) = 0.$$

More tersely, we may write

$$P(|X_n - X| \geq \varepsilon) \to 0 \qquad \text{or} \qquad P(|X_n - X| < \varepsilon) \to 1.$$

Theorem 2.113 (Weak law of large numbers)**.** *Let X_n be an IID sequence with common mean μ and finite variance σ^2. Then*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \to \mu$$

in probability.

Here is another notion of convergence.

Definition 2.114. Let X_n and X be random variables. We say $X_n \to X$ **in mean square** if

$$E[(X_n - X)^2] \to 0.$$

Theorem 2.115 (Chebyshev's inequality)**.** *Let X be a random variable with finite variance. Let $a > 0$. Then*

$$P(|X - \mu| \geq a) \leq \frac{E[(X-\mu)^2]}{a^2}$$

Remark 2.116. This means

$$P(|X - \mu| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}.$$

Proof. Use the partition theorem with two partitions on some as-yet-unspecified event A:

$$E[(X-\mu)^2] = E\left[(X-\mu)^2 \,\middle|\, A\right] P(A) \;+\; E\left[(X-\mu)^2 \,\middle|\, A^c\right] P(A^c).$$

Regardless of what A is, the last term is non-negative. So we have

$$E[(X-\mu)^2] \;\geq\; E\left[(X-\mu)^2 \,\middle|\, A\right] P(A).$$

Now let A be the particular event that $|X-\mu| \geq a$. Then we have

$$\begin{aligned} E[(X-\mu)^2] &\geq E\left[(X-\mu)^2 \,\middle|\, |X-\mu| \geq a\right] P(|X-\mu| \geq a) \\ &= E\left[(X-\mu)^2 \,\middle|\, (X-\mu)^2 \geq a^2\right] P(|X-\mu| \geq a) \\ &\geq a^2 P(|X-\mu| \geq a). \end{aligned}$$

Dividing through by a^2 we have

$$P(|X-\mu| \geq a) \leq \frac{E[(X-\mu)^2]}{a^2}$$

as desired. □

Theorem 2.117. *Convergence in mean square implies convergence in probability.*

Proof. Let $\varepsilon > 0$. We need to show $P(|X_n - X| > \varepsilon) \to 0$. By Chebyshev's inequality,

$$P(|X_n - X| > \varepsilon) \geq \frac{1}{\varepsilon^2} E[(X_n - X)^2].$$

□

Remark 2.118. Notes about Chebyshev's inequality:

- It is used in the proof of the weak law, which I am omitting.
- The bounds provided by Chebyshev's inequality are rather loose, but they are certain. The central limit theorem gives tighter bounds, but only probabilistically.

2.6.2 The strong law of large numbers

Here is a third notion of convergence.

Definition 2.119. We say $X_n \to X$ **with probability one** (w.p. 1) or **almost surely** (a.s.) if

$$P\left(\left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 1.$$

More tersely, we may write

$$P(X_n \to X) = 1.$$

Theorem 2.120 (Strong law of large numbers)**.** *Let X_n be an IID sequence with common mean μ. Then $\overline{X}_n \to \mu$ with probability 1.*

Theorem 2.121. *$X_n \to X$ w.p. 1 implies $X_n \to X$ in probability.*

Remark 2.122. This means the strong law is stronger than the weak law.

2.6.3 The central limit theorem

Motivation. Let X_n be an IID sequence and let $S_n = \sum X_n$. We know from section 2.5.4 that $E[S_n] = n\mu$ and $\text{Var}(S_n) = n\sigma^2$. We expect the PDF of S_n to be centered at $n\mu$ with width approximately $\sqrt{n}\sigma$. Likewise, if $\overline{X}_n = \sum X_n/n$ then we know that $E[\overline{X}_n] = \mu$ and $\text{Var}(\overline{X}_n) = \sigma^2/n$. We expect the PDF of S_n to be centered at $n\mu$ with width approximately $\sigma/\sqrt{n}$. For various distributions, one finds empirically that these PDFs look approximately normal if n is large.

Definition 2.123. Let X be a random variable with mean μ and variance σ^2. The **standardization** or **normalization** of X is

$$\frac{X-\mu}{\sigma}.$$

In particular, if we standardize S_n, we get

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

with mean 0 and variance 1:

$$\text{Var}(Z_n) = \frac{1}{\sigma^2 n}\text{Var}(S_n - n\mu) = \frac{1}{\sigma^2 n}\text{Var}(S_n) = \frac{n\sigma^2}{n\sigma^2} = 1.$$

Likewise, if we standardize $\overline{X}_n$, we get

$$\frac{\overline{X}_n - \mu}{\sigma\sqrt{n}}.$$

The **central limit theorem** says that the standardizations of S_n and $\overline{X}_n$ both approach standard normal for large n.

Definition 2.124. Let $\Phi(x)$ be the CDF of the standard normal:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-x^2/2}dx.$$

Theorem 2.125. *Let X_n be an IID sequence with finite mean μ and finite non-zero variance σ^2. Let Z_n be the standardization of S_n as above. Then*

$$P(Z_n \leq x) \to \Phi(x)$$

for all x.

Remark 2.126. This convergence is called **convergence in distribution**.

xxx include some examples here.

2.6.4 Confidence intervals

Here is an application of the Central Limit Theorem to statistics. The **population** is a random variable X. A **sample** of size n is n IID copies of X. The random variable X has a (true) **population mean** μ_X but we do not know what it is; all we have is the **sample mean** $\overline{X}_n = \sum_{k=1}^n X_k/n$, which is an estimate of the population mean. We would like to put some error bars on this estimate.

We quantify this problem using the notion of **confidence intervals**. We look for $\varepsilon > 0$ such that

$$P\left(|\overline{X}_n - \mu_X| \geq \varepsilon\right) = 0.05$$

or, alternatively,

$$P\left(|\overline{X}_n - \mu_X| < \varepsilon\right) = 0.95.$$

(Five percent is a conventional value in statistics.)

Using the CLT, we treat $\overline{X}_n$ as being approximately normal. We standardize it (statisticians call this *taking the z-score*) in the usual way:

$$Z = \frac{\overline{X}_n - \mu_{\overline{X}_n}}{\sigma_{\overline{X}_n}} = \frac{\overline{X}_n - \mu_X}{\sigma_X/\sqrt{n}}.$$

Then

$$\begin{aligned} P(|\overline{X}_n - \mu_X| \geq \varepsilon) &= P\left(|\overline{X}_n - \mu_X| \geq \varepsilon\right) \\ &= P\left(\frac{|\overline{X}_n - \mu_X|}{\sigma_X/\sqrt{n}} \geq \frac{\varepsilon}{\sqrt{n}}\right) \\ &\approx P\left(Z \geq \frac{\varepsilon}{\sqrt{n}}\right) = 0.05. \end{aligned}$$

It's worth memorizing (or you can compute it if you prefer) that the standard normal curve has area 0.95 for Z running from -1.96 to $+1.96$. So, $\varepsilon/(\sigma_X/\sqrt{n})$ should be 1.96. Solving for ε in terms of n gives

$$\varepsilon = \frac{1.96\sigma_X}{\sqrt{n}}.$$

Note that this requires the population standard deviation to be known. (We can compute the sample standard deviation and use that as an estimate of the population standard deviation σ_X, but we've not developed any theory as to the error in *that* estimate.)

Example 2.127. ▷ Let X have the Bernoulli distribution with parameter p — flip a coin with probability p of heads; assign the value 0 to tails and 1 to heads. Recall from section 2.2.2 that X has mean p. (In section 2.2.2 we took tails to be 1; I have changed the convention.) Suppose you flip the coin 1000 times (i.e. $n = 1000$) and obtain 520 heads. Then $\overline{X}_n = 0.520$. Then

$$\varepsilon = \frac{1.96\sqrt{p(1-p)}}{\sqrt{1000}} \approx 0.0619\sqrt{p(1-p)}.$$

For $p = 0.5$, $\varepsilon \approx 0.031$. Thus we are 95% certain that μ_X is within 0.031 on either side of 0.520, i.e. between 0.489 and 0.551. ◁

2.7 Stochastic processes

xxx Examples:

- Sequence of die rolls (independent).
- Sequence of die tips (non-independent but Markov).
- Sequence of coin flips (independent).
- Sum of coin flips (Martingale).

2.7.1 Die tips

2.7.2 Coin flips

$$\Omega = \{1, -1\}^\infty, \qquad X_n = \omega_n \qquad S_n = \sum_{j=1}^{n} X_j.$$

$n = 3$:

k	0	1	2	3
$2k - n$	-3	-1	1	3

$$P(S_n = 2k - n) = \binom{n}{k} p^k (1-p)^{n-k}.$$

2.7.3 Filtrations

Show the filtration tree (refinement of partitions of Ω) for the sum of coin flips.

Have done sigma of finite generating set; show the size is 2^{2^n}.

2.7.4 Markov processes

Homogeneous (die tips).

Non-homogeneous (sum of coin flips).

xref back to the partition theorem.

Key point: PMFs which evolve in time.

Show some matrix products for the sum of coin flips.

2.7.5 Martingales

Sum of coin flips.

sub and super and at, depending on p.

3 Statistics

The key concept here is parameter estimation. My goals are:

- Present a few concepts from parametric statistics, with an example-heavy approach.
- Give unified notation for probability and statistics.
- Work out several parameter-estimation examples concretely.

3.1 Sampling

3.1.1 Finite-population example

Work out an example with finite population $\{x_1, \ldots, x_n\}$ and simple random sample with replacement $\{X_1, \ldots, X_N\}$.

Define μ and σ^2.

Define **unbiased estimator**.

Show that $\overline{X}$ and s^2 are unbiased estimators of μ and σ^2 respectively. Explain the factor of $n-1$ in s^2.

3.1.2 Infinite-population example

3.2 Decision theory

The presentation here follows [**Bha**]. I am simply tabulating and elaborating upon his definitions.

One has the following:

- A **population space** Ω.
- An **observation space** or **sample space** $\mathfrak{X} = \Omega^n$ containing **observations** $\mathbf{X} = (X_1, \ldots, X_n)$ which are n-tuples of IID random variables on Ω.
- A **parameter space** Θ indexing (for $\theta \in \Theta$) a family of probability measures $\{P_\theta\}$ on $(\Omega, \mathcal{F})$. Then for each P_θ one obtains a probability space $(\Omega, \mathcal{F}, P_\theta)$. One can think of these measures P_θ as **conditional probabilities** $P(\mathbf{x} \mid \theta)$.

  Nominally, Θ is $\mathbb{R}^d$ or some subset thereof.
- An **action space** $\mathcal{A}$ indexing (for $\theta \in \Theta$) a family of probability measures P_θ on $(\Omega, \mathcal{F})$.

  Nominally, $\mathcal{A}$ is all or part of Θ. This is best explained by example: Suppose $\theta = (\mu, \sigma^2)$ where each θ is a 2-tuple defining a normal probability distribution on the real line. Then one might want to use one's observation only to estimate μ. Then $\Theta = \mathbb{R} \times \mathbb{R}^+$ whereas $\mathcal{A}$ is merely $\mathbb{R}$.
- The population space, sample space, parameter space, and action space are all measurable spaces with their respective σ-algebras.

- A **loss function** $L : (\Theta, \mathcal{A}) \to \mathbb{R}$. This quantifies the loss incurred when θ (e.g. the population mean μ) is estimated by a (e.g. the sample mean $\overline{X}$). The most common loss function is the **least-squared error**

$$L(\theta, a) = \|\theta - a\|^2.$$

- A **decision rule** $d : \mathfrak{X} \to \mathcal{A}$. For example,

$$a = d(\mathbf{X}) = \overline{X} := \frac{\sum_{i=1}^n X_i}{n}.$$

  [xxx to do: Define admissibility. Sufficient: rabi thm 3.3 and cor 3.1. Make some plots.]

- The **risk function** associated with a given decision rule is then

$$R(\theta, d) := E_\theta \left[L(\theta, d(\mathbf{X})) \right].$$

  The subscript on the E reminds us (for cases when we need reminding) which variable(s) not to integrate out: E_θ of something will be (potentially) a function of θ and so we won't integrate over θ. In particular,

$$R(\theta, d) = E_\theta \left[L(\theta, d(\mathbf{X})) \right] \tag{3.2.1}$$

$$= \int_{\mathfrak{X}} L(\theta, d(\mathbf{x}))\, dP_\theta(\mathbf{x}). \tag{3.2.2}$$

To summarize, the players in a decision problem are:

Ω	$\mathfrak{X}$	
Θ	$\mathcal{A}$	d
$L(\theta, a)$	$R(\theta, d)$	

Example 3.1. ▷ Let the X_i's be IID with mean μ and variance σ^2. If the decision rule is $a = \overline{X}$ with $\theta = \mu$ and least-squared-error loss function, then

$$\begin{aligned} R(\theta, d) &= E_\mu \left[(\mu - \overline{X})^2 \right] \\ &= \mu^2 - 2\mu E[\overline{X}] + E[\overline{X}^2] \\ &= \mu^2 - \frac{2\mu}{n} E\left[\sum_{i=1}^n X_i \right] + \frac{1}{n^2} E\left[\sum_{i=1}^n \sum_{j=1}^n X_i X_j \right] \\ &= \mu^2 - \frac{2\mu}{n} \sum_{i=1}^n E[X_i] + \frac{1}{n^2} \sum_{i=1}^n \sum_{j \neq i} E[X_i X_j] + \frac{1}{n^2} \sum_{i=1}^n E[X_i^2]. \end{aligned}$$

Now, the X_i's are IID so $E[X_i] = \mu$ is the same for all i, and $E[X_i X_j] = E[X_i]E[X_j]$ for $i \neq j$. Then

$$R(\theta, d) = \mu^2 - 2\mu^2 + \frac{n(n-1)}{n^2} \mu^2 + \frac{n}{n^2} E[X_1^2].$$

Recall that the variance was

$$\sigma^2 = E[(X_1 - \mu)^2] = E[X_1^2] - 2\mu E[X_1] + \mu^2 = E[X_1^2] - \mu^2$$

so

$$E[X_1^2] = \sigma^2 + \mu^2.$$

Then

$$R(\theta, d) = -\mu^2 + \frac{(n-1)}{n}\mu^2 + \frac{1}{n}\sigma^2 + \frac{1}{n}\mu^2 = \frac{\sigma^2}{n}.$$

This means that, if we use the sample mean $\overline{X}$ to estimate the population mean μ, our risk increases with larger population variance and decreases with larger sample size. ◁

3.3 Parameter estimation

3.3.1 Maximum-likelihood estimation

log-likelihood example as well.

3.3.2 Method of moments

3.3.3 Bayes estimation

Here $\vartheta \in \Theta$ is thought of as a random variable.

Notation:

Value	Random variable	Space
x	X	$\mathfrak{X}$
θ	ϑ	Θ

One somehow knows (or guesses) a probability distribution τ on Θ. This is called the **prior distribution** or simply **prior**. This encodes what we know about θ values prior to making an observation $\mathbf{X}$. (Below we'll have a posterior distribution which is conditioned on the observation $\mathbf{X}$.) If τ has a density $f_\vartheta(\theta)$ with respect to Lebesgue measure then we write

$$d\tau(\theta) = f_\vartheta(\theta)\, d\theta.$$

We have a loss function $L(\theta, a)$ as defined in section 3.2. This will always be least-squared error unless otherwise noted.

Definition 3.2. The **Bayes risk** of a decision rule d is

$$r(\tau, d) = \int_\Theta R(\theta, d)\, d\tau(\theta).$$

By equation 3.2.1, this is

$$r(\tau, d) = \int_\Theta E_\theta \left[L(\theta, d(\mathbf{X}))\right]\, d\tau(\theta)$$

which by equation 3.2.2 is, in turn,

$$r(\tau, d) = \int_\Theta \left[\int_\mathfrak{X} L(\theta, d(\mathbf{x}))\, dP_\theta(\mathbf{x})\right]\, d\tau(\theta)$$

Definition 3.3. A **Bayes rule** d_0 is a decision rule which minimizes Bayes risk:

$$r(\tau, d_0) = \inf_d r(\tau, d)$$

where the infimum is taken across all decision rules d. Note that a minimizer may not exist.

Recall from section 2.5.8 that if we have two random variables $\mathbf{X}$ and ϑ, then:

- Given the joint density $f_{\mathbf{X},\vartheta}(\mathbf{x}, \theta)$ we can integrate out $\mathbf{x}$ to obtain the marginal density $f_\vartheta(\theta)$:

  $$f_\vartheta(\theta) = \int_{\mathfrak{X}} f_{\mathbf{X},\vartheta}(\mathbf{x}, \theta)\, d\mathbf{x}.$$

  Likewise, we can integrate out θ to obtain the marginal density $f_{\mathbf{X}}(\mathbf{x})$:

  $$f_{\mathbf{X}}(\mathbf{x}) = \int_\Theta f_{\mathbf{X},\vartheta}(\mathbf{x}, \theta)\, d\theta.$$

- We have new random variables $\mathbf{X} \mid \vartheta$ and $\vartheta \mid \mathbf{X}$.
- We can compute their conditional expectations

  $$E[\mathbf{X} \mid \vartheta] \qquad \text{and} \qquad E[\vartheta \mid \mathbf{X}].$$

- We have conditional density which is joint over marginal:

  $$f_{\mathbf{X}|\vartheta}(\mathbf{x} \mid \theta) = \frac{f_{\mathbf{X},\vartheta}(\mathbf{x}, \theta)}{f_\vartheta(\theta)} = \frac{f_{\mathbf{X},\vartheta}(\mathbf{x}, \theta)}{\int_{\mathfrak{X}} f_{\mathbf{X},\vartheta}(\mathbf{x}, \theta)\, d\mathbf{x}}$$

  and likewise

  $$f_{\vartheta|\mathbf{X}}(\mathbf{x} \mid \theta) = \frac{f_{\mathbf{X},\vartheta}(\mathbf{x}, \theta)}{f_{\mathbf{X}}(\mathbf{x})} = \frac{f_{\mathbf{X},\vartheta}(\mathbf{x}, \theta)}{\int_\Theta f_{\mathbf{X},\vartheta}(\mathbf{x}, \theta)\, d\theta}.$$

- Given these two facts, we can solve (just as in Bayes' theorem, theorem B.1) for one conditional density in terms of the other:

  $$f_{\vartheta|\mathbf{X}}(\theta \mid \mathbf{x}) = f_{\mathbf{X}|\vartheta}(\mathbf{x} \mid \theta)\ \frac{f_\vartheta(\theta)}{f_{\mathbf{X}}(\mathbf{x})}.$$

These facts motivate the following definition.

Definition 3.4. The **posterior distribution** of ϑ given $\mathbf{X}$ is

$$f_{\vartheta|\mathbf{X}}(\theta \mid \mathbf{x}) = f_{\mathbf{X}|\vartheta}(\mathbf{x} \mid \theta)\ \frac{f_\vartheta(\theta)}{f_{\mathbf{X}}(\mathbf{x})}$$

Definition 3.5. The **posterior mean** is the expectation of the posterior distribution $f_{\vartheta|\mathbf{X}}(\theta \mid \mathbf{x})$.

The theorem is that the posterior mean

$$d_0(\mathbf{X}) = E[\vartheta \mid \mathbf{X}]$$

is a Bayes estimator of θ, satisfying definition 3.3. Following [**CB**], I currently think this is because of the following (assuming the distributions of $\mathbf{X}$ and ϑ both have densities as above, and using Bayes' theorem):

$$\begin{aligned} r(\tau, d) &= \int_\Theta R(\theta, d)\, d\tau(\theta) \\ &= \int_\Theta E_\theta \left[L(\theta, d(\mathbf{X})) \right]\, d\tau(\theta) \\ &= \int_\Theta \left[\int_{\mathfrak{X}} L(\theta, d(\mathbf{x}))\, dP_\theta(\mathbf{x}) \right]\, d\tau(\theta) \\ &= \int_\Theta \left[\int_{\mathfrak{X}} L(\theta, d(\mathbf{x}))\, dP(\mathbf{x} \mid \theta) \right]\, d\tau(\theta) \\ &= \int_\Theta \left[\int_{\mathfrak{X}} L(\theta, d(\mathbf{x}))\, f_{\mathbf{X}|\vartheta}(\mathbf{x} \mid \theta)\, d\mathbf{x} \right] f_\vartheta(\theta)\, d\theta \\ &= \int_{\mathfrak{X}} \left[\int_\Theta L(\theta, d(\mathbf{x}))\, f_{\vartheta|\mathbf{X}}(\theta \mid \mathbf{x})\, d\theta \right] f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}. \end{aligned}$$

The quantity in square brackets, which is a function of $\mathbf{x}$, is called the **posterior expected loss**. (The conditionals in the integrals are reminiscent of the partition theorem [theorem 2.35].) When computing a Bayes rule, one selects d to minimize the posterior expected loss for each $\mathbf{x}$.

To compute $d_0(\mathbf{X})$, we need to find all three right-hand terms in

$$f_{\vartheta|\mathbf{X}}(\theta \mid \mathbf{x}) = f_{\mathbf{X}|\vartheta}(\mathbf{x} \mid \theta)\, \frac{f_\vartheta(\theta)}{f_{\mathbf{X}}(\mathbf{x})}.$$

These are found as follows:

- $f_\vartheta(\theta)$ is the given prior.
- $f_{\mathbf{X}|\vartheta}(\mathbf{x} \mid \theta)$ is given as the distribution of $\mathbf{X}$ with parameter θ.
- $f_{\mathbf{X}}(\mathbf{x})$ is the marginal of $f_{\mathbf{X}|\vartheta}(\mathbf{x} \mid \theta)$, found by integrating out θ. However, as usual in probability, one may find a trick to avoid doing the integral.

3.3.4 Minimax estimation

def'n. Sufficient: rabi thm 3.6 and 3.7.

A The coin-flipping experiments

This section is an extended worked example, tying together various concepts. We apply the Central Limit Theorem first to repeated tosses of a single coin, then to repeated collections of tosses.

A.1 Single coin flips

The first experiment is tossing a single coin which has probability p of heads. Then $\Omega = \{\mathrm{T}, \mathrm{H}\}$. Let X be the random variable which takes value 0 for tails and 1 for heads. As discussed in section 2.2.2, X has the Bernoulli distribution with parameter p. I will allow p to vary throughout this section, although I will focus on $p = 0.5$ and $p = 0.6$. Recall that X has mean

$$\mu_X = p.$$

(In section 2.2.2 we took 1 for tails which is the opposite convention from the one here.) Its standard deviation is

$$\sigma_X = \sqrt{p(1-p)},$$

which is $\sqrt{0.25} = 0.5$ and $\sqrt{0.24} \approx 0.4899$ for $p = 0.5$ and $p = 0.6$, respectively.

Now flip the coin a large number n of times — say, $n = 1000$ — and count the number of heads. Using the notation of section 2.5.4, the number of heads is S_n. There are two ways to look at this.

- On one hand, from section 2.2.2 we know that S_n is geometric with parameters p and n — if we think of the 1000 tosses as a single experiment. (This is precisely what we will do in the next section.) The PMF of S_n is the one involving binomial coefficients; we would expect

  $$\mu_{S_n} = np$$

  i.e. 500 or 600 and

  $$\sigma_{S_n} = \sqrt{np(1-p)}$$

  which is $\sqrt{250} \approx 15.81$ or $240 \approx 15.49$ for $p = 0.5$ and $p = 0.6$, respectively.

- On the other hand, the Central Limit Theorem (section 2.6.3) says that as n increases, the distribution of S_n begins to look normal. This PDF involves the exponential function, as shown in section 2.5.4. The mean of the sums S_n is

  $$\mu_{S_n} \approx n\mu_X = np$$

  (again 500 or 600). The standard deviation of those sums about the means 500 or 600 is

  $$\sigma_{S_n} \approx \sqrt{n}\sigma_X = \sqrt{np(1-p)}$$

  which are again 15.81 and 15.49, respectively.

Note that the binomial PMF and the normal PDF are *not* the same, even though they produced the same means and standard deviations: the geometric random variable has an integer-valued PMF; $P(499.1 \leq S_n \leq 499.9) = 0$ and likewise S_n can never be anything out of the range from 0 to 1000. The normal PDF, on the other hand, gives $P(499.1 \leq S_n \leq 499.9) \neq 0$ since we are taking area under a curve where the function is non-negative. Since the output of the exponential function is never 0, the normal PDF gives non-zero (although admittedly very, very tiny) probability of S_n being less than 0 or greater than 1000. (See also [**MM**] for some very nice plots.)

Now we can ask about fairness of coins. The probabilistic point of view is to fix p and ask about the probabilities of various values of S_n. If the coin is fair, what are my chances of flipping anywhere between 470 and 530 heads? Using the geometric PMF is a mess — in fact, my calculator can't compute $\binom{1000}{470}$ without overflow. Using the normal approximation, though, is easy. I asked my TI-83 to integrate its `normalpdf` function, with $\mu = 500$ and $\sigma = 15.81$, from 470 to 530 and it told me 0.9422.

How surprised should I be if I toss 580 heads? The standardization, which [**MM**] call the z-**score**, of S_n is (definition 2.123)

$$z = \frac{S_n - \mu_{S_n}}{\sigma_{S_n}}.$$

This counts how many standard deviations away from the mean a given observation is. I have $80/15.81 \approx 5.06$ so this result is more than five standard deviations away from the mean. I would not think the coin is fair. If I re-do that computation with $p = 0.6$, $\mu_{S_n} = 600$, and $\sigma_{S_n} = 15.49$, I get a z-score of -1.29 which is not surprising if the coin has parameter $p = 0.6$.

The point of view used in statistics is to start with the data, and from that try to estimate with various levels of confidence what the parameters are. Given a suspicious coin, what experiments would we have to run to be 95% sure that we've find out what the coin's parameter p is, to within, say, ± 0.01? Continuing the example from section 2.6.4, we ask for $\varepsilon = 0.01$. We had

$$\varepsilon = \frac{1.96\sigma_X}{\sqrt{n}}$$

so, setting $\varepsilon = 0.01$ and solving for n, we have

$$n = \left(\frac{1.96\sigma_X}{0.01}\right)^2 = \left(\frac{1.96\sqrt{p(1-p)}}{0.01}\right)^2 \approx 38416\, p(1-p).$$

Now, $p(1-p)$ has a maximum at $p = 0.5$, for which $n = 9604$, so that many flips would determine p to within ± 0.01 with 95% confidence. Re-doing the arithmetic with 0.001 in place of 0.01 gives $n = 960,400$. Generalizing, we see that each additional decimal place costs 100 times as many runs.

A.2 Batches of coin flips

The second experiment is 1000 tosses of a coin where each coin has probability p of heads. (Or, think of simultaneously tossing 1000 identical such coins.) Then $\#\Omega = 2^{1000} \approx 10^{301}$. Let $Y : \Omega \to \mathbb{R}$ be the random variable which counts the number of heads. This is an example where the random variable is far easier to deal with than the entire sample space (which is huge).

As discussed in section 2.2.2, Y has the binomial distribution with parameters p and $n = 1000$. Recall from the previous section that that Y has mean

$$\mu_Y = 1000p,$$

e.g. 500 or 600. Likewise, its standard deviation is

$$\sigma_Y = \sqrt{1000p(1-p)},$$

which is $\sqrt{250} \approx 15.81$ or $240 \approx 15.49$ for $p = 0.5$ and $p = 0.6$, respectively.

Let $\overline{Y}_N$ be the average of N runs of this experiment — that is, the **sample mean**. The Central Limit Theorem (section 2.6.3) says that as N increases, the distribution of $\overline{Y}_N$ begins to look normal. The mean of the sample means is

$$\mu_{\overline{Y}_N} \approx \mu_Y = 1000p,$$

which is again 500 or 600. The standard deviation of the sample means is

$$\sigma_{\overline{Y}_N} \approx \frac{\sigma_Y}{\sqrt{N}} = \frac{\sqrt{np(1-p)}}{N}.$$

This is $15.81/\sqrt{N}$ or $15.49/\sqrt{N}$, respectively.

Here is the (crucial) interpretation: Given the parameters p and n of Y, there is a true population mean and a true population standard deviation. If $p = 0.5$, then $\sigma_Y \approx 15.81$ and even if we're on the three millionth iteration of the flip-1000-coins experiment, it's still going to be quite likely as ever that we'll get a 514 or a 492 and so on. If we don't know the true p of the coins then, while the true population mean μ_Y and the true population standard deviation σ_Y exist, we don't know what they are. All we have is some suspicious-looking identical coins and our laboratory equipment. As we run and re-run the flip-1000-coins experiment, the following happens:

- The sample mean $\mu_{\overline{Y}_N}$, for increasingly larger N, will approach the population mean μ_Y.
- The variations *of the sample mean* will decrease. We might say the error in our estimate of the population mean is shrinking.
- The population standard deviation appears in the above formulas via its effects on the standard deviation *of the sample mean*.
- Nothing we have done so far has given us a reliable connection between the sample standard deviation and the population standard deviation. We can guess that the sample standard deviation approaches the population standard deviation, but (in this course) we have not developed any information about the error in that computation.

Here is a numerical example. I run (simulated on a computer) the flip-1000-coins experiment 400 times. The first time I get 482, so the sample mean is 482. The second time I get 521, so the sample mean is $(482 + 521)/2$. The third time I get 494, so the sample mean is $(482 + 521 + 494)/3$ and so on.

Here is $p = 0.5$:

N	Y	$\overline{Y}_N$	Sample std. dev.
1	482	482.000	N/A
2	521	501.500	27.577
3	494	499.000	19.975
4	512	502.250	17.557
5	485	498.800	17.050
6	507	500.167	15.613
⋮	⋮	⋮	⋮
395	493	500.258	16.163
396	505	500.270	16.144
397	501	500.272	16.124
398	494	500.256	16.107
399	501	500.258	16.086
400	474	500.192	16.120

Here is $p = 0.6$:

N	Y	$\overline{Y}_N$	Sample std. dev.
1	616	616.000	N/A
2	584	600.000	22.627
3	620	606.667	19.732
4	617	609.250	16.919
5	583	604.000	18.775
6	613	605.500	17.190
⋮	⋮	⋮	⋮
395	608	599.484	17.031
396	615	599.523	17.027
397	609	599.547	17.013
398	607	599.565	16.995
399	604	599.576	16.975
400	611	599.605	16.964

B Bayes' theorem

Bayes' theorem is so important that it merits multiple points of view: algebraic, graphical, and numerical.

B.1 Algebraic approach

Recall from definition 2.12 that if A is an event, and if B is another event with non-zero probability, the **conditional probability** of A given B is

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

Bayes' theorem tells us how to invert this: how to compute the probability of B given A. First, the algebraic treatment.

Theorem B.1 (Bayes' theorem)**.** *Let A and B be events with non-zero probability. Then*

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

Proof. Using the definition (intersection over given), we have

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

Multiplying top and bottom by $P(B)$ (which is OK since $P(B) \neq 0$) we get

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

Now notice that $B \cap A$ is the same as $A \cap B$, so in particular $P(B \cap A)$ is the same as $P(A \cap B)$. Transposing the terms in the denominator gives us

$$\begin{aligned} P(B|A) \quad &= \quad \frac{P(A \cap B)}{P(B)}\ \frac{P(B)}{P(A)} \\ &= \quad P(A|B)\ \frac{P(B)}{P(A)} \end{aligned}$$

as desired. □

B.2 Graphical/numerical approach

The following example is adapted from [**KK**]. Suppose that 0.8% of the general population has a certain disease, and suppose that we have a test for it. Specifically, if a person actually has the disease, the test says so 90% of the time. If a person does not have the disease, the test gives a false diagnosis 7% of the time. When a particular patient tests positive, what is the probability they have the disease?

We can write this symbolically as follows. Let D be the event that the person has the disease and $\overline{D}$ be its complement; let Y (for "yes") be the event that the test *says* the person has the disease, with complement

$\overline{Y}$. Writing the given information in these terms, we have:

$$\begin{aligned} P(D) &= 0.008 \\ P(Y|D) &= 0.90 \\ P(Y|\overline{D}) &= 0.07. \end{aligned}$$

Before computing any conditional probabilities, let's find the probability of Y by itself. Using the partition theorem (theorem 2.18) we know this is

$$\begin{aligned} P(Y) &= P(Y|D)\,P(D) \;+\; P(Y|\overline{D})\,P(\overline{D}) \\ &= 0.90 \cdot 0.008 + 0.07 \cdot 0.992 \\ &= 0.077. \end{aligned}$$

So the non-conditional probabilities are

$$P(\overline{D}) = 0.992 \qquad P(\overline{Y}) = 0.923$$
$$P(D) = 0.008 \qquad P(Y) = 0.077.$$

Looking at D and Y separately, we can think of the population at large as being split into those with and without the disease, and those for whom the test is positive or negative. Suppose in particular that we have a sample of 1000 people who are representative of the general population. Here are some (very rectangular) Venn diagrams:

$\overline{D}$	D	
992	8	

923	$\overline{Y}$
77	Y

Bayes' theorem has to do with how these two partitions intersect to make four groups:

$\overline{D}$	D	
$P(\overline{Y}\|\overline{D}) \cdot 992 = ?$	$P(\overline{Y}\|D) \cdot 8 = ?$	$\overline{Y}$
$P(Y\|\overline{D}) \cdot 992 = ?$	$P(Y\|D) \cdot 8 = ?$	Y

and

$\overline{D}$	D	
$P(\overline{D}\|\overline{Y}) \cdot 923 = ?$	$P(D\|\overline{Y}) \cdot 923 = ?$	$\overline{Y}$
$P(\overline{D}\|Y) \cdot 77 = ?$	$P(D\|Y) \cdot 77 = ?$	Y

We can use the theorem to find the probability our patient has the disease, given the positive test result:

$$\begin{aligned} P(D|Y) &= P(Y|D)\,\frac{P(D)}{P(Y)} \\ &= 0.90 \;\cdot\; \frac{0.008}{0.077} \\ &= 0.094. \end{aligned}$$

That is, there's only a one-in-eleven chance the patient actually has the disease. Kaplan and Kaplan's idea is to look at this surprising result in the context of the other 999 people also tested. I will elaborate on this, working out all the math.

We have two conditional probabilities given: $P(Y|D)$ and $P(Y|\overline{D})$. We found $P(D|Y)$. What about $P(D|\overline{Y})$? We can use the partition theorem (theorem 2.18) again to solve for what we don't know in terms of what

we do know:

$$\begin{aligned} P(D) &= P(D|Y)\,P(Y) \;+\; P(D|\overline{Y})\,P(\overline{Y}) \\ P(D|\overline{Y}) &= \frac{P(D) - P(D|Y)\,P(Y)}{P(\overline{Y})} \\ &= \frac{0.008 - 0.094 \cdot 0.077}{0.923} \\ &= 0.0008. \end{aligned}$$

We now have all four conditional probabilities:

$$\begin{aligned} P(Y|\overline{D}) &= 0.07 & \qquad P(D|\overline{Y}) &= 0.0008 \\ P(Y|D) &= 0.90 & \qquad P(D|Y) &= 0.094. \end{aligned}$$

Now we can fill out the four-square table:

- Since $P(Y|\overline{D}) = 0.07$, seven percent of the 992 disease-free people (70 of them) get false positives; the rest (922) get a correct negative result.
- Since $P(Y|D) = 0.90$, ninety percent of the 8 people with the disease test positive (i.e. all but one of them); one of the 8 gets a false sense of security.
- Since $P(D|\overline{Y}) = 0.0008$, 0.08% of the 923 with negative test results (one person) does in fact have the disease; the other 922 (as we found just above) get a correct negative result.
- Since $P(D|Y) = 0.094$, only 9.4% of the 77 people with positive test results (7 people) have the disease; the other 70 get a scare (and, presumably, a re-test).

So, the sample of 1000 people splits up as follows:

$\overline{D}$	D	
922	1	$\overline{Y}$
70	7	Y

Moreover, we can rank events by likelihood:

(1) Healthy people correctly diagnosed: 92.2%.

(2) False positives: 7%.

(3) People with the disease, correctly diagnosed: 0.7%.

(4) False negatives: 0.1%.

Now it's no surprise our patient got a false positive: this happens 10 times as often as a correct positive diagnosis.

B.3 Asymptotics

The specific example provided some insight, but what happens when we vary the parameters? We had

$$P(D|Y) = \frac{P(Y|D)\,P(D)}{P(Y)} = \frac{P(Y|D)\,P(D)}{P(Y|D)\,P(D) \;+\; P(Y|\overline{D})\,P(\overline{D})}.$$

Let's turn this into a design problem. Let

$$\begin{aligned} p = &= P(D) \\ a = &= P(Y|D) \\ b = &= P(Y|\overline{D}) \\ 1-\varepsilon &= P(D|Y) \\ \varepsilon &= P(\overline{D}|Y). \end{aligned}$$

How would you choose the test-design parameters a and b (i.e. how good would your test have to be) to get ε small? What if the disease is rarer (p smaller)?

Suppose we want a high likelihood of correct detection, i.e. ε small. Then

$$\begin{aligned} P(D|Y) &= \frac{P(Y|D)\,P(D)}{P(Y|D)\,P(D) \;+\; P(Y|\overline{D})\,P(\overline{D})} \\ &= \frac{ap}{ap + b(1-p)} \\ P(\overline{D}|Y) &= \frac{b(1-p)}{ap + b(1-p)} \\ \frac{b(1-p)}{ap + b(1-p)} &< \varepsilon. \end{aligned}$$

Solving for a and b we get

$$\frac{a}{b} > \frac{(1-\varepsilon)(1-p)}{\varepsilon p}.$$

There are two free parameters, a and b, so I'll just consider their ratio. Now, the function

$$\frac{1-x}{x}$$

blows up near zero; for small x, it's approximately $1/x$. For small ε and p we have

$$\frac{b}{a} < \varepsilon p.$$

If, say, we have p and ε both 0.001, then a and b need to differ by a factor of a million. Recall that a and b are both probabilities, and so range between zero and one. To test a one-in-a-million event with 99.99% confidence ($p = 10^{-6}$ and $\varepsilon = 10^{-4}$), b must be less than 10^{-10}.

That tells us how to choose $P(Y|\overline{D})$ to in order to get $P(\overline{D}|Y)$, i.e. the probability of false positives, small. What about false negatives: $P(D|\overline{Y})$? If you do similar algebra to the above you should find that getting less than ε requires

$$1 - a < \frac{\varepsilon}{p}.$$

This is a less strict constraint: $P(Y|D)$ needs to be very close to 1 only when $\varepsilon \ll p$.

B.4 Conclusions

Some points:

- $P(Y|D)$ is information known to the person who creates the test — say, at a pharmaceutical company; $P(D|Y)$ is information relevant to the people who give and receive the test — for example, at the doctor's office. This duality between design and implementation suggests that Bayes' theorem has important consequences in many practical situations.
- The results can be surprising — after all, in the example above the test was 90% accurate, was it not? Bayes' theorem is important to know precisely because it is counterintuitive.
- We can see from the example and the asymptotics above that rare events are hard to test accurately for. If we want certain testing for rare events, the test might be impossible (or overly expensive) to design in practical terms.

C Probability and measure theory

Probability theory is measure theory with a soul.
— Mark Kac.

Modern probability is a special case of measure theory, but this course avoids the latter. Here we draw the connections for the reader with a measure-theoretic background. (Full information may be found in [**FG**], but I like to have a brief, handy reference. See also [**Fol**], [**Rud**], or [**Roy**].)

C.1 Dictionary

Measure theory / analysis	**Probability**
The **sample space** Ω is simply a set.	Same.

Measure theory / analysis	**Probability**
A **σ-field** $\mathcal{F}$ on Ω is a subset of 2^Ω satisfying the axioms of definition 2.5; $\mathcal{F}$ must contain at least $\emptyset$ and Ω, and it must be closed under complements, countable unions, and countable intersections.	Same.

Note that even if Ω is uncountable (for example, $\Omega = \mathbb{R}$), 2^Ω still satisfies the axioms for a σ-field. If Ω is a topological space (e.g. $\mathbb{R}^d$), the standard σ-field is the **Borel σ-field** which is the one generated by all the open sets of Ω.

Measure theory / analysis	**Probability**
The pair $(\Omega, \mathcal{F})$ is called a **measurable space**. This is an unfortunate misnomer since it may not be possible to put a measure on it — in which case we would certainly think of it as "unmeasurable"!	Same.

Measure theory / analysis	**Probability**
Elements of $\mathcal{F}$ are called **measurable sets**. This is also a misnomer because we haven't defined measures yet!	An **event** is nothing more than a measurable set.

Measure theory / analysis	Probability
A **measure** is a function $\mu : \mathcal{F} \to [0, +\infty]$ with the following properties: • $\mu(\emptyset) = 0$. • For all $A \in \mathcal{F}$, $\mu(A) \geq 0$. • If $A_1, A_2, \ldots$, is a finite or countable subset of $\mathcal{F}$, with the A_i's all (pairwise) disjoint, then $P(\cup_i A_i) = \sum_i P(A_i)$. This is the **countable additivity** property of the probability measure μ.	A **probability measure** is a measure with the additional requirement that $\mu(\Omega) = 1$. Thus, we have $\mu : \mathcal{F} \to [0, 1]$.

Note that if Ω is finite or countably infinite, it is possible to define a measure on the biggest σ-field on Ω, namely, $\mathcal{F} = 2^\Omega$. If $\Omega = \mathbb{R}$, then it is not possible to define a measure on $\mathcal{F} = 2^\Omega$. See [**Fol**] for a proof.

Measure theory / analysis	Probability
A **measure space** is a triple $(\Omega, \mathcal{F}, \mu)$ where μ is a measure on $\mathcal{F}$.	A **probability space** is a triple $(\Omega, \mathcal{F}, P)$ where P is a probability measure on $\mathcal{F}$.

Measure theory / analysis	Probability
A **measurable function** is a function f from one measurable space $(\Omega, \mathcal{F})$ to another measurable space $(\Psi, \mathcal{G})$ such that the preimage under f of each measurable set in Ψ is a measurable set in Ω. That is, for all $B \in \mathcal{G}$, $f^{-1}(B) \in \mathcal{F}$.	A **random variable** is a measurable function X from a probability space $(\Omega, \mathcal{F}, P)$ to a measurable space $(\Psi, \mathcal{G})$. For this course, that measurable space has been $\Psi = \mathbb{R}$, with $\mathcal{G}$ being the Borel sets in $\mathbb{R}$.

Measure theory / analysis	Probability
Expectation: $\int_\Omega X(\omega)\, dP(\omega)$.	Expectation: $E[X]$.

Remark C.1. In an undergraduate context, we use the terms *probability density function* and *cumulative distribution function*. In a graduate context, we use the following:

- We have a real-valued random variable X. That is, we have $X : \Omega \to \mathbb{R}$ where $(\Omega, \mathcal{F}, P)$ is a probability space and the measurable space is $(\mathbb{R}, \mathcal{B})$, namely, the reals with the Borel σ-algebra.
- This gives us a probability measure μ_X on $\mathbb{R}$ by $\mu_X(B) = P(X^{-1}(B))$. This is called simply the **distribution** of X.
- We can then measure the particular Borel sets $(-\infty, x]$. This defines a function $F_X(x) = \mu_X((-\infty, x])$. This (the CDF) is called the **distribution function** of X.
- If $d\mu_X$ is absolutely continuous with respect to Lebesgue measure, i.e. $d\mu_X = f_X(x)\, dx$ (where $f_X(x)$ is the familiar PDF), we say $f_X(x)$ is the **density** of X.

Measure theory / analysis	**Probability**
The Laplace and Fourier transforms, respectively, of $f : \mathbb{R} \to \mathbb{R}$ are $$(\mathcal{L}f)(t) = \int_{\mathbb{R}} e^{tx} f(x)\, dt$$ and $$(\mathcal{F}f)(t) = \int_{\mathbb{R}} e^{itx} f(x)\, dt.$$	The moment-generating and characteristic functions, respectively, of a real-valued random variable X are $$M_X(t) = E[e^{tX}] \quad \text{and} \quad \beta_X(t) = E[e^{itX}].$$ Let μ be the distribution of X. We say that the moment-generating and characteristic functions are the **Laplace transform** and **Fourier transform**, respectively, of the measure μ: $$\mathcal{L}\mu = \int_{\mathbb{R}} e^{tx}\, d\mu \quad \text{and} \quad \mathcal{F}\mu = \int_{\mathbb{R}} e^{itx}\, d\mu.$$ If the distribution of X has a density function $f(x)$, i.e. $d\mu = f(x)dx$ where $d\mu$ is absolutely continuous with respect to Lebesgue measure dx, then these are $$M_X(t) = E[e^{tX}] = \int_{\mathbb{R}} e^{tx} f(x)\, dx = (\mathcal{L}f)(t)$$ and $$\beta_X(t) = E[e^{itX}] = \int_{\mathbb{R}} e^{itx} f(x)\, dx = (\mathcal{F}f)(t).$$

xxx 8-29 cty of finite measures.

xxx three kinds of measures: discrete, cts, singular.

xxx 9-5 and 9-24: MCTx2, DCTx2, Fub, Fatou

xxx cvgce: a.s./w.p.1; i.p.; i.m.s.; in dist (10-29 & 10-31) flavors.

xxx 9-14 Borel-Cantelli.

xxx 9-26 σ-field indep.

xxx 9-28 prod meas

xxx 10-19 Kolm 0-1 and tail field?

C.2 Measurability

xxx include the not-less-refined-than pictures.

C.3 Independence and measurability

xxx type up handwritten notes. xxx emph summarizing content from the main body of the paper, connecting it with the (more abstract and more puzzling) measure-theoretic notions. xxx mention calculus of expectations, in the independent and measurable cases. xxx perhaps mention SDEs ....

xxx caveat I'm writing everything in terms of discrete random variables.

xxx recall the following:

$$P(X = x) = P(X^{-1}(x))$$

and

$$P(A, B) = P(A \cap B).$$

The right-hand sides are set-theoretic; the left-hand sides are in more common use.

Events:

- Events A and B are independent if $P(A \cap B) = P(A)P(B)$.
- Definition of conditional probability:

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

- Partition theorem:

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i)\, P(B_i).$$

Random variables:

- Partition theorem:

$$E[X] = \sum_{i=1}^{n} E[X \mid B_i]\, P(B_i).$$

- Independence of a random variable and an event:

$$P((X = x) \cap B) = P(X = x)P(B) \text{ for all } x \in \mathbb{R}.$$

- Independence of two random variables:

$$P\big((X = x) \cap (Y = y)\big) = P(X = x)\, P(Y = y) \text{ for all } x, y \in \mathbb{R}.$$

- Conditional expectation:

$$E[X \mid B] \;=\; \sum_x x\, P((X = x) \mid B) \;=\; \sum_x x\, P(X^{-1}(x) \mid B) \;=\; \frac{\sum_x x\, P(X^{-1}(x) \cap B)}{P(B)}.$$

  Note: *given* means *relative to*.

- If X is independent of B, then $P(B)$ factors out of the numerator:

$$E[X \mid B] \;=\; \frac{\sum_x x\, P(X^{-1}(x) \cap B)}{P(B)} \;=\; \frac{\sum_x x\, P(X^{-1}(x))P(B)}{P(B)} \;=\; \sum_x x\, P(X^{-1}(x)) \;=\; E[X].$$

- If X is B-measurable then $X^{-1}(x) = \emptyset, B, B^c, \Omega$:

$$E[X \mid B] = \frac{\sum_x x\, P(X^{-1}(x) \cap B)}{P(B)} = \frac{X(B)P(B \cap B) + X(B^c)P(B^c \cap B)}{P(B)} = \frac{X(B)P(B)}{P(B)} = X(B);$$
$$E[X \mid B^c] = X(B^c);$$
$$E[X \mid \Omega] = \frac{\sum_x x\, P(X^{-1}(x) \cap \Omega)}{P(\Omega)} = \sum_x x\, P(X^{-1}(x)) = X(B)P(B) + X(B^c)P(B^c) = E[X];$$
$$E[X \mid \emptyset] = 0.$$

  [Claim this is just X, when written on partitions.]

  [Claim: a finite σ-algebra is generated by a partition.]

  [xxx include the not-less-refined-than pictures.]

- Example: Let $\Omega = \{1, 2, 3, 4\}$ with $P(k) = 1/4$ for $k = 1, 2, 3, 4$. Let $X(1) = X(3) = 2$ and $X(2) = X(4) = 3$. Let $B = \{1, 3\}$. Then $E[X] = 2.5$ and $E[X \mid B] = 2$ which is what one would expect.

- Conditional PMF:
$$P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{\sum_x P(X = x, Y = y)}.$$

  xxx 4x4 dot figure here.

Example C.2. ▷

Let
$$\Omega = \{1, 2, 3, 4\}, \qquad \mathcal{F} = 2^{\Omega}, \qquad \text{and} \qquad P(k) = 1/4 \text{ for } k = 1, 2, 3, 4.$$

Let
$$\begin{aligned} X(1) &= 1 \\ X(2) &= 0 \\ X(3) &= 1 \\ X(4) &= 0. \end{aligned}$$

(That is, X is the parity random variable.)

Here is a σ-algebra $\mathcal{G}$ such that X is not $\mathcal{G}$-measurable. Let
$$\mathcal{G} = \sigma(\{1\}) = \Big\{\{\}, \{1\}, \{2, 3, 4\}, \{1, 2, 3, 4\}\Big\}.$$

Then
$$X^{-1}(1) = \{1, 3\} \notin \mathcal{G}.$$

Thus X is not $\mathcal{G}$-measurable. However,
$$X^{-1}(1) = \{1, 3\} \in \mathcal{F}$$

so X is $\mathcal{F}$-measurable.

Next, I want to enumerate the subsets G of Ω such that X is independent of G. We have [xxx xref] X is independent of G if for all $x \in \mathbb{R}$,
$$P((X = k) \cap G) = P(X = k)P(G).$$

For brevity (so that the table below will fit on the page), write

$$A_0 = X^{-1}(0) \qquad \text{and} \qquad A_1 = X^{-1}(1).$$

Ordered by sizes of G's:

G	$A_0 \cap G$	$A_1 \cap G$	$P(G)$	$P(A_0)P(G)$	$P(A_0 \cap G)$	$P(A_1)P(G)$	$P(A_1 \cap G)$	Indep.?
$\{\}$	$\{\}$	$\{\}$	0	0	0	0	0	**yes**
$\{1\}$	$\{\}$	$\{1\}$	1/4	1/8	0	1/8	1/4	no
$\{2\}$	$\{2\}$	$\{\}$	1/4	1/8	1/4	1/8	0	no
$\{3\}$	$\{\}$	$\{3\}$	1/4	1/8	0	1/8	1/4	no
$\{4\}$	$\{4\}$	$\{\}$	1/4	1/8	1/4	1/8	0	no
$\{1,3\}$	$\{\}$	$\{1,3\}$	1/2	1/4	0	1/4	1/2	no
$\{2,4\}$	$\{2,4\}$	$\{\}$	1/2	1/4	1/2	1/4	0	no
$\{1,2\}$	$\{2\}$	$\{1\}$	1/2	1/4	1/4	1/4	1/4	**yes**
$\{1,4\}$	$\{4\}$	$\{1\}$	1/2	1/4	1/4	1/4	1/4	**yes**
$\{2,3\}$	$\{2\}$	$\{3\}$	1/2	1/4	1/4	1/4	1/4	**yes**
$\{3,4\}$	$\{4\}$	$\{3\}$	1/2	1/4	1/4	1/4	1/4	**yes**
$\{1,2,3\}$	$\{2\}$	$\{1,3\}$	3/4	3/8	1/4	3/8	1/2	no
$\{1,2,4\}$	$\{2,4\}$	$\{1\}$	3/4	3/8	1/2	3/8	1/4	no
$\{1,3,4\}$	$\{4\}$	$\{1,3\}$	3/4	3/8	1/4	3/8	1/2	no
$\{2,3,4\}$	$\{2,4\}$	$\{3\}$	3/4	3/8	1/2	3/8	1/4	no
$\{1,2,3,4\}$	$\{2,4\}$	$\{1,3\}$	1	1/2	1/2	1/2	1/2	**yes**

Ordered by independence and dependence:

G	$A_0 \cap G$	$A_1 \cap G$	$P(G)$	$P(A_0)P(G)$	$P(A_0 \cap G)$	$P(A_1)P(G)$	$P(A_1 \cap G)$	Indep.?
$\{\}$	$\{\}$	$\{\}$	0	0	0	0	0	**yes**
$\{1,2\}$	$\{2\}$	$\{1\}$	1/2	1/4	1/4	1/4	1/4	**yes**
$\{1,4\}$	$\{4\}$	$\{1\}$	1/2	1/4	1/4	1/4	1/4	**yes**
$\{2,3\}$	$\{2\}$	$\{3\}$	1/2	1/4	1/4	1/4	1/4	**yes**
$\{3,4\}$	$\{4\}$	$\{3\}$	1/2	1/4	1/4	1/4	1/4	**yes**
$\{1,2,3,4\}$	$\{2,4\}$	$\{1,3\}$	1	1/2	1/2	1/2	1/2	**yes**
$\{1\}$	$\{\}$	$\{1\}$	1/4	1/8	0	1/8	1/4	no
$\{2\}$	$\{2\}$	$\{\}$	1/4	1/8	1/4	1/8	0	no
$\{3\}$	$\{\}$	$\{3\}$	1/4	1/8	0	1/8	1/4	no
$\{4\}$	$\{4\}$	$\{\}$	1/4	1/8	1/4	1/8	0	no
$\{1,3\}$	$\{\}$	$\{1,3\}$	1/2	1/4	0	1/4	1/2	no
$\{2,4\}$	$\{2,4\}$	$\{\}$	1/2	1/4	1/2	1/4	0	no
$\{1,2,3\}$	$\{2\}$	$\{1,3\}$	3/4	3/8	1/4	3/8	1/2	no
$\{1,2,4\}$	$\{2,4\}$	$\{1\}$	3/4	3/8	1/2	3/8	1/4	no
$\{1,3,4\}$	$\{4\}$	$\{1,3\}$	3/4	3/8	1/4	3/8	1/2	no
$\{2,3,4\}$	$\{2,4\}$	$\{3\}$	3/4	3/8	1/2	3/8	1/4	no

$\triangleleft$

D A proof of the inclusion-exclusion formula

Proposition (**Inclusion-exclusion formula**). *Let $A_1, \ldots, A_n$ be events. Then*

$$P(\cup_{i=1}^n A_i) = \sum_{1\le i\le n} P(A_i) - \sum_{1\le i<j\le n} P(A_i\cap A_j) + \sum_{1\le i<j<k\le n} P(A_i\cap A_j\cap A_k) - \ldots + (-1)^{n+1} P(A_1\cap\ldots\cap A_n).$$

Proof. The proof is by strong induction. For the base case $n = 1$, the left-hand side is $P(A)$ and the right-hand side is also $P(A)$.

* * *

A bonus case, $n = 2$, is not necessary but helps to illustrate what's going on. The formula is

$$P(A\cup B) = P(A) + P(B) - P(A\cap B).$$

This is easy to prove using Venn diagrams and finite additivity of P on the disjoint sets $A\setminus B$, $A\cap B$, and $B\setminus A$. I am not including a picture in this note. The point, though, is that if A and B overlap, then $A\cap B$ is counted twice (*overcounted*) in $P(A\cap B)$, so we need to subtract off $P(A\cap B)$ to compensate.

* * *

Now for the induction step. Suppose the inclusion-exclusion formula is true for $1, 2, \ldots, n-1$ (we'll only need 2 and $n-1$), and show it's true for n. Notationally, this is a mess. I'll do the $n = 3$ case since it is easier to understand. This will illuminate how to proceed in the messier general case.

For $n = 3$, we are asked to show that

$$P(A\cup B\cup C) = P(A) + P(B) + P(C) - P(A\cap B) - P(A\cap C) - P(B\cap C) + P(A\cap B\cap C).$$

Since we want to use induction, we can try to isolate C from A and B. We can write

$$\begin{aligned} P((A\cup B)\cup C) = \quad & P(A) + P(B) \quad && +P(C) \\ & -P(A\cap B) && -P(A\cap C) - P(B\cap C) \\ & && +P(A\cap B\cap C). \end{aligned}$$

We have

$$P(A) + P(B) - P(A\cap B) = P(A\cup B)$$

by the induction hypothesis at $n - 1 = 2$. (By isolating terms with C, we have found terms involving one fewer set). For the moment, to make things a little clearer, write $X = A\cup B$. Then we need to show that

$$P(X\cup C) = P(X) + P(C) - P(A\cap C) - P(B\cap C) + P(A\cap B\cap C).$$

Using the induction hypothesis for the two sets X and C, we know that

$$P(X\cup C) = P(X) + P(C) - P(X\cap C).$$

Looking at these last two equations — the first of which we need to prove, and the second of which we already know is true — we see that we'll be done if only we can show that

$$-P(A\cap C) - P(B\cap C) + P(A\cap B\cap C) = -P(X\cap C).$$

Toggling the negative signs, this is

$$P(X \cap C) = P(A \cap C) + P(B \cap C) - P(A \cap B \cap C).$$

I put $X = A \cup B$ only for convenience; I'm done with it now. The statement I need to prove is

$$P((A \cup B) \cap C) = P(A \cap C) + P(B \cap C) - P(A \cap B \cap C).$$

The trick is that

$$A \cap B \cap C = (A \cap C) \cap (B \cap C)$$

and

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C).$$

That is, I distribute the C's. So, the statement I need to prove is

$$P((A \cap C) \cup (B \cap C)) = P(A \cap C) + P(B \cap C) - P((A \cap C) \cap (B \cap C)).$$

Now we again have one fewer set involved: this is the inclusion-exclusion formula for the *two* sets $(A \cap C)$ and $(B \cap C)$. Thus this statement is true by the induction hypothesis. And, that was the last thing we needed to prove.

* * *

Now, guided by the $n = 3$ case, we can confidently wade into the morass of subscripts which is the induction step. We are asked to show that

$$P(\cup_{i=1}^{n} A_i) = \sum_{1 \leq i \leq n} P(A_i) - \sum_{1 \leq i < j \leq n} P(A_i \cap A_j) + \sum_{1 \leq i < j < k \leq n} P(A_i \cap A_j \cap A_k) - \ldots + (-1)^{n+1} P(A_1 \cap \ldots \cap A_n).$$

Since we want to use induction, we can try to isolate A_n from the others. We can write

$$\begin{array}{lll} P((\cup_{i=1}^{n-1} A_i) \cup A_n) = & \sum_{1 \leq i \leq n-1} P(A_i) & + P(A_n) \\ & - \sum_{1 \leq i < j \leq n-1} P(A_i \cap A_j) & - \sum_{1 \leq i \leq n-1} P(A_i \cap A_n) \\ & + \sum_{1 \leq i < j < k \leq n-1} P(A_i \cap A_j \cap A_k) & + \sum_{1 \leq i < j \leq n-1} P(A_i \cap A_j \cap A_n) \\ & - \ldots & - \ldots \\ & + (-1)^n P(A_1 \cap \ldots \cap A_{n-1}) & + (-1)^n \sum_{1 \leq i \leq n-1} P(A_1 \cap \ldots \cap \hat{A}_i \cap \ldots \cap A_{n-1} \cap A_n) \\ & & + (-1)^{n+1} P(A_1 \cap \ldots \cap A_n). \end{array}$$

where the notation $\hat{A}_i$ means omit the A_i from the intersection. We have

$$P(\cup_{i=1}^{n-1} A_i) = \sum_{1 \leq i \leq n-1} P(A_i) - \sum_{1 \leq i < j \leq n-1} P(A_i \cap A_j) + \sum_{1 \leq i < j < k \leq n-1} P(A_i \cap A_j \cap A_k) - \ldots + (-1)^n P(A_1 \cap \ldots \cap A_{n-1})$$

by the induction hypothesis at $n-1$. (By isolating terms with A_n, we have again found terms involving one fewer set). As above, for clarity, temporarily write $X = \cup_{i=1}^{n-1}$. Then we need to show that

$$\begin{aligned} P(X \cup A_n) &= P(X) + P(A_n) \\ &\quad - \sum_{1 \le i \le n-1} P(A_i \cap A_n) \\ &\quad + \sum_{1 \le i < j \le n-1} P(A_i \cap A_j \cap A_n) \\ &\quad - \dots \\ &\quad + (-1)^n \sum_{1 \le i \le n-1} P(A_1 \cap \dots \cap \hat{A}_i \cap \dots \cap A_{n-1} \cap A_n) \\ &\quad + (-1)^{n+1} P(A_1 \cap \dots \cap A_n). \end{aligned}$$

Using the induction hypothesis for the two sets X and A_n, we know that

$$P(X \cup A_n) = P(X) + P(A_n) - P(X \cap A_n).$$

Looking at these last two equations — the first of which we need to prove, and the second of which we already know is true — we see that we'll be done if only we can show that

$$\begin{aligned} &- \sum_{1 \le i \le n-1} P(A_i \cap A_n) \\ &+ \sum_{1 \le i < j \le n-1} P(A_i \cap A_j \cap A_n) \\ &- \dots \\ &+ (-1)^n \sum_{1 \le i \le n-1} P(A_1 \cap \dots \cap \hat{A}_i \cap \dots \cap A_{n-1} \cap A_n) \\ &+ (-1)^{n+1} P(A_1 \cap \dots \cap A_n) \\ = \; &- P(X \cap A_n). \end{aligned}$$

Toggling the negative signs, this is

$$\begin{aligned} P(X \cap A_n) &= \sum_{1 \le i \le n-1} P(A_i \cap A_n) \\ &\quad - \sum_{1 \le i < j \le n-1} P(A_i \cap A_j \cap A_n) \\ &\quad + \dots \\ &\quad - (-1)^n \sum_{1 \le i \le n-1} P(A_1 \cap \dots \cap \hat{A}_i \cap \dots \cap A_{n-1} \cap A_n) \\ &\quad - (-1)^{n+1} P(A_1 \cap \dots \cap A_n). \end{aligned}$$

Note that $-(-1)^n$ is the same as $(-1)^{n-1}$. So we need to show

$$\begin{aligned} P(X \cap A_n) &= \sum_{1 \le i \le n-1} P(A_i \cap A_n) \\ &\quad - \sum_{1 \le i < j \le n-1} P(A_i \cap A_j \cap A_n) \\ &\quad + \dots \\ &\quad + (-1)^{n-1} \sum_{1 \le i \le n-1} P(A_1 \cap \dots \cap \hat{A}_i \cap \dots \cap A_{n-1} \cap A_n) \\ &\quad + (-1)^n P(A_1 \cap \dots \cap A_n). \end{aligned}$$

As before, I don't need to write $X = \cup_{i=1}^{n-1}$ anymore. The statement I need to prove is

$$\begin{aligned} P((\cup_{i=1}^{n-1} A_i) \cap A_n) &= \sum_{1 \le i \le n-1} P(A_i \cap A_n) \\ &\quad - \sum_{1 \le i < j \le n-1} P(A_i \cap A_j \cap A_n) \\ &\quad + \ldots \\ &\quad + (-1)^{n-1} \sum_{1 \le i \le n-1} P(A_1 \cap \ldots \cap \hat{A}_i \cap \ldots \cap A_{n-1} \cap A_n) \\ &\quad + (-1)^n P(A_1 \cap \ldots \cap A_n). \end{aligned}$$

The distribution tricks are

$$A_1 \cap \ldots \cap A_n = (A_1 \cap A_n) \cap \ldots \cap (A_{n-1} \cap A_n)$$

and

$$(A_1 \cup \ldots \cup A_{n-1}) \cap A_n = (A_1 \cap A_n) \cup \ldots \cup (A_{n-1} \cap A_n).$$

So, the statement I need to prove is

$$\begin{aligned} P(\cup_{i=1}^{n-1} (A_i \cap A_n)) &= \sum_{1 \le i \le n-1} P(A_i \cap A_n) \\ &\quad - \sum_{1 \le i < j \le n-1} P((A_i \cap A_n) \cap (A_j \cap A_n)) \\ &\quad + \ldots \\ &\quad + (-1)^{n-1} \sum_{1 \le i \le n-1} P((A_1 \cap A_n) \cap \ldots \cap (\hat{A}_i \cap A_n) \cap \ldots \cap (A_{n-1} \cap A_n)) \\ &\quad + (-1)^n P((A_1 \cap A_n) \cap \ldots \cap (A_{n-1} \cap A_n)). \end{aligned}$$

Now we again have one fewer set involved: this is the inclusion-exclusion formula for the $n-1$ sets $(A_1 \cap A_n)$ through $(A_{n-1} \cap A_n)$. Thus this statement is true by the induction hypothesis. Since that is all that remained to be shown, we are done. □

References

[**Bha**] Bhattacharya, R. *Theoretical Statistics.* Course notes, University of Arizona Math 567A, spring 2008.

[**CB**] Casella, G. and Berger, R.L. *Statistical Inference* (2nd ed.). Duxbury Press, 2001.

[**Fol**] Folland, G.B. *Real Analysis: Modern Techniques and Their Applications* (2nd ed.). Wiley-Interscience, 1999.

[**FG**] Fristedt, B. and Gray, L. *A Modern Approach to Probability Theory.* Birkhauser, 1997.

[**GS**] Grimmett, G. and Stirzaker, D. *Probability and Random Processes*, 3rd ed. Oxford, 2001.

[**Kennedy**] Kennedy, T. Math 564. Course at University of Arizona, spring 2007.

[**KK**] Kaplan, M. and Kaplan, E. *Chances Are . . . : Adventures in Probability.* Penguin, 2007.

[**MM**] Moore, D.S. and McCabe, G.P. *Introduction to the Practice of Statistics.* Freeman and Co., 2005.

[**Roy**] Royden, H.L. *Real Analysis* (2nd ed.). MacMillan, 1968.

[**Rud**] Rudin, W. *Principles of Mathematical Analysis* (3rd ed.). McGraw-Hill, 1976.

Index