Probability and Statistics Basics

Kevin Kircher — Cornell MAE — Spring '14

These notes summarize some basic probability and statistics material. The primary sources are *A Modern Introduction to Probability and Statistics* by Dekking, Kraaikamp, Lopuhaä and Meester, *Introduction to Probability* by Dimitri Bertsekas, and the lectures of Profs. Gennady Samorodnitsky and Mark Psiaki.

Contents

Part I
Probability

1 Outcomes, Events and Probability

Definitions

- A *sample space* Ω is a set of the outcomes of an *experiment*.
- An *event* is a subset of the sample space.
- Two events A and B are *disjoint* if they have no elements (*outcomes*) in common.

Axioms

- **Nonnegativity**: $\mathbf{P}(A) \geq 0$ for all events A
- **Normalization**: $\mathbf{P}(\Omega) = 1$
- **Disjoint Unions**: for all disjoint events A_i, $\mathbf{P}(A_1 \cup A_2 \cup \dots) = \mathbf{P}(A_1) + \mathbf{P}(A_2) + \dots$

Results

- **DeMorgan's Laws**. For any two events A and B,

$$(A \cup B)^c = A^c \cap B^c$$
$$(A \cap B)^c = A^c \cup B^c$$

  Mnemonic: distribute the c and flip the set operator.

- For unions of intersections and intersections of unions,

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

- The probability of a union of (non-disjoint) events is

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

  Intuition: subtract the intersection of A and B to avoid double counting. For three events,

$$\mathbf{P}(A \cup B \cup C) = \mathbf{P}(A) + \mathbf{P}(B) + \mathbf{P}(C) - \mathbf{P}(A \cap B) - \mathbf{P}(A \cap C) - \mathbf{P}(B \cap C) + \mathbf{P}(A \cap B \cap C)$$

- **The Complement Rule**:
$$\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$$

- A *permutation* $P_{n,k}$ is an ordering of k objects out of a pool of n. Such a permutation can be done in
$$P_{n,k} = \frac{n!}{(n-k)!}$$
ways.

- A *combination* $\binom{n}{k}$ (pronounced "n choose k") is a choice of k objects from a pool of n, where order doesn't matter.
$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$
Example: choosing 3 medalists out of a heat of 8 runners is a combination because order doesn't matter. On the other hand, choosing the gold, silver and bronze medalists is a permutation because order matters.

2 Conditional Probability and Independence

Definitions

- The *conditional probability of A given C* (C is called the *conditioning event*), provided $\mathbf{P}(C) > 0$, is

$$\mathbf{P}(A \mid C) = \frac{\mathbf{P}(A \cap C)}{\mathbf{P}(C)}$$

  Note that the Complement Rule works for conditional probabilities. For all events A,

$$\mathbf{P}(A \mid C) + \mathbf{P}(A^c \mid C) = 1$$

  For three events A, B and C,

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

- Events A and B are *independent* if any of the following are true:

$$\mathbf{P}(A \mid B) = \mathbf{P}(A)$$
$$\mathbf{P}(B \mid A) = \mathbf{P}(B)$$
$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\,\mathbf{P}(B)$$

  where A can be replaced with A^c or B with B^c. All twelve of these statements are equivalent.

- Two or more events A_1, A_2, ... , A_m are independent if

$$\mathbf{P}(A_1 \cap A_2 \cap \cdots \cap A_m) = \mathbf{P}(A_1)\,\mathbf{P}(A_2) \ldots \mathbf{P}(A_m)$$

  and if the above equation also holds when any number of events are replaced by their complements., e.g.

$$\mathbf{P}(A_1 \cap A_2^c \cap A_3 \cap \cdots \cap A_m) = \mathbf{P}(A_1)\,\mathbf{P}(A_2^c)\,\mathbf{P}(A_3) \ldots \mathbf{P}(A_m)$$

  In general, establishing the independence of m events requires checking 2^m equations.

  A useful rule: if events $A_1, \ldots, A_n$ are independent, then so are any derived events constructed from disjoint groupings of the A_i.

Results

- **The Multiplication Rule**. For events A and C,

$$\mathbf{P}(A \cap C) = \mathbf{P}(A \mid C) \cdot \mathbf{P}(C)$$

Note that this works even if $\mathbf{P}(C) = 0$. This allows us to break the probability of a complicated intersection up into a sequence of less complicated conditional probabilities. Handy for iterative calculations.

The general form of the Multiplication Rule, for events $A_1, \ldots, A_n$ with positive probability, is

$$\mathbf{P}(\cap_{i=1}^{n} A_i) = \mathbf{P}(A_1)\,\mathbf{P}(A_2 \mid A_1)\,\mathbf{P}(A_3 \mid A_1 \cap A_2) \ldots \mathbf{P}(A_n \mid \cap_{i=1}^{n-1} A_i)$$

- **The Law of Total Probability**. For disjoint events $C_1, C_2, \ldots, C_m$ that partition Ω,

$$\mathbf{P}(A) = \mathbf{P}(A \mid C_1)\,\mathbf{P}(C_1) + \mathbf{P}(A \mid C_2)\,\mathbf{P}(C_2) + \cdots + \mathbf{P}(A \mid C_m)\,\mathbf{P}(C_m)$$

  This allows us to write a probability $\mathbf{P}(A)$ as a weighted sum of conditional probabilities. Useful when the conditional probabilities are known or easy. A special case:

$$\mathbf{P}(B) = \mathbf{P}(B \mid A)\,\mathbf{P}(A) + \mathbf{P}(B \mid A^c)\,\mathbf{P}(A^c)$$

- **Bayes' Rule**. For disjoint events $C_1, C_2, \ldots, C_m$ that partition Ω,

$$\mathbf{P}(C_i \mid A) = \frac{\mathbf{P}(A \mid C_i) \cdot \mathbf{P}(C_i)}{\mathbf{P}(A \mid C_1)\,\mathbf{P}(C_1) + \mathbf{P}(A \mid C_2)\,\mathbf{P}(C_2) + \cdots + \mathbf{P}(A \mid C_m)\,\mathbf{P}(C_m)}$$

  Note that we can also write Bayes' Rule in a simpler form, and use the Law of Total Probability to expand the denominator. This simpler form is

$$\mathbf{P}(C_i \mid A) = \frac{\mathbf{P}(A \mid C_i) \cdot \mathbf{P}(C_i)}{\mathbf{P}(A)}$$

3 Discrete Random Variables

Definitions

- A *discrete random variable* is a function $X : \Omega \to \mathbb{R}$ that takes on a countable (possibly infinite, if $n \to \infty$) number of discrete values $x_1, x_2, \ldots, x_n$.
- The *probability mass function* p_X of a discrete random variable X is the function $p_X : \mathbb{R} \to [0, 1]$, defined by
$$p_X(x_i) = \mathbf{P}(X = x_i).$$
Equivalently, for any set B,
$$\mathbf{P}(X \in B) = \sum_{x_i \in B} p_X(x_i).$$
The *pmf* is non-zero only at the discrete values $x_1, x_2, \ldots$
More precisely, the *pmf* obeys
$$\begin{aligned} p_X(x_i) &> 0 \\ \sum_i p_X(x_i) &= 1 \\ p_X(x) &= 0 \text{ for all } x \neq x_i \end{aligned}$$
- The *cumulative distribution function* F_X of a discrete random variable X is the function $F_X : \mathbb{R} \to [0, 1]$, defined by
$$F_X(x) = \mathbf{P}(X \leq x) \text{ for } x \in \mathbf{R}$$
The *cdf* of a discrete RV is piecewise continuous from the right. For a *pmf* defined as above, F_X obeys
$$\begin{aligned} F_X(x) &= \sum_{x_i \leq x} p_X(x_i) \\ x_1 \leq x_2 &\Rightarrow F_X(x_2) \leq F(x_2) \\ \lim_{x \to +\infty} F_X(x) &= 1 \\ \lim_{x \to -\infty} F_X(x) &= 0 \end{aligned}$$

Common Discrete Distributions

- X has the *Bernoulli distribution* $\text{Ber}(p)$ with parameter $0 \leq p \leq 1$ if its *pmf* is given by
$$p_X(x_i) = \begin{cases} p, & \text{if } x_i = 1, \\ 1 - p, & \text{if } x_i = 0, \\ 0, & \text{otherwise} \end{cases}$$

Expectation: $\mathbf{E}\,X = p$. Variance: $\mathrm{Var}(X) = p(1-p)$.

Bernoulli trials form the basis of all the most important discrete RVs. The Bernoulli distribution models a sequence of independent binary trials (coin flips), with probability p of success in each trial.

- X has the *binomial distribution* $\mathrm{Bin}(n,p)$ with parameters $n = 1, 2, \ldots$ and $0 \le p \le 1$ if its *pmf* is given by

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k} \text{ for } k = 0, 1, \ldots, n.$$

  Expectation: $\mathbf{E}\,X = np$. Variance: $\mathrm{Var}(X) = np(1-p)$.

  The binomial RV counts the number of successes in n Bernoulli trials, with probability p of success in each trial.

  NB. The Bernoulli RV is a special case of the binomial RV: $\mathrm{Bin}(1,p)$ is $\mathrm{Ber}(p)$.

- The *multinomial distribution* $\mathrm{Mult}(n, p_1, \ldots, p_k)$ counts the number of times, out of n independent trials with k types of outcome in every trial, that an outcome of type i is observed, for $i \in \{1, \ldots, k\}$. The i^{th} type of outcome has probability p_i of success.

  Let m_i be the number of times an outcome of type i is observed, so that $\sum_i m_i = n$. Then the multinomial distribution is

$$p_{\mathbf{M}}(\mathbf{m}) = \mathbf{P}(M_1 = m_1, \ldots, M_k = m_k) = \frac{n!}{m_1! \ldots m_k!} p_1^{m_1} \ldots p_k^{m_k}.$$

  Note that this gives the distribution of the *multiplicities* of outcomes of type 1 through k. In Matlab, the RHS is easily computed with mnpdf($\mathbf{m}$, $\mathbf{p}$).

  For $i \in \{1, \ldots, k\}$,

  - Expectation: $\mathbf{E}\,M_i = np_i$
  - Variance: $\mathrm{Var}(M_i) = np_i(1-p_i)$
  - Covariance: for $i \ne j$, $\mathrm{Cov}(M_i, M_j) = -np_ip_j$

  NB. $\mathrm{Bin}(n,p)$ is $\mathrm{Mult}(n, p, 1-p)$.

- X has a *negative binomial distribution* $\mathrm{NB}(n,p)$ with parameters $n = 1, 2, \ldots$ and $0 \le p \le 1$ if its *pmf* is given by

$$p_X(k) = \binom{k-1}{n-1} p^n (1-p)^{k-n} \text{ for } k = n, n+1, \ldots$$

  Expectation: $\mathbf{E}\,X = n/p$. Variance: $\mathrm{Var}(X) = n(1-p)/p^2$.

  The negative binomial RV counts the number of trials until the n^{th} success, with probability p of success in each trial.

- X has the *geometric distribution* Geo(p) with parameter $0 < p \leq 1$ if its *pmf* is given by

$$p_X(k) = p(1-p)^{k-1} \text{ for } k = 0, 1, \ldots, n.$$

Expectation: $\mathbf{E}\, X = 1/p$. Variance: $\mathrm{Var}(X) = (1-p)/p^2$.

NB. The geometric RV is a special case of the negative binomial: NB(1,p) = Geo(p). The geometric RV counts the number of trials until the first success.

The geometric RV has the **memoryless property**.

- X has the *hypergeometric distribution* Hyp(n, N, M) with parameters $n = 1, 2, \ldots,$ $N > 0$, and $M > 0$ if its *pmf* is given by

$$p_X(k) = \frac{\binom{N}{k}\binom{M}{n-k}}{\binom{N+M}{n}}, \quad \max(n-M, 0) \leq\ k \leq \min(n, N)$$

Expectation: $\mathbf{E}\, X = nN/(N+M)$.

Variance: $\mathrm{Var}(X) = \frac{nNM}{(N+M)^2}\left(1 - \frac{n-1}{N+M+1}\right)$.

The hypergeometric RV counts the number of successes – defined by choosing a white ball – in n trials of choosing a ball *without replacement* from an initial population of N white balls and M black balls.

- X has the *Poisson distribution* Poiss(λ) with parameter λ if its *pmf* is given by

$$p_X(k) = \frac{\lambda^k}{k!}e^{-\lambda}, \text{ for } k = 0, 1, 2, \ldots$$

$\mathbf{E}\, X = \mathrm{Var}(X) = \lambda$.

The Poisson RV is the limiting case of the binomial RV as $n \to \infty$ and $p \to 0$, while the product $np \to \lambda > 0$ (infinite trials, infinitesimal probability of success per trial, but a finite product of the two).

An example: in a huge volume of dough ($n \to \infty$), the probability of scooping out any particular raisin is vanishingly small ($p \to 0$), but there's a constant raisin density ($np \to \lambda$).

Reproducing property: if $X_1, \ldots, X_n$ are Poisson RVs with parameters $\lambda_1, \ldots, \lambda_n$, then the sum $Y = X_1 + \cdots + X_n$ is a Poisson RV with parameter $\lambda_1 + \cdots + \lambda_n$.

4 Continuous Random Variables

Definitions

- A random variable X is *continuous* if for some function $f_X : \mathbb{R} \to \mathbb{R}$ with $f_X(x) \geq 0$ for all x and $\int_{-\infty}^{\infty} f_X(x)\mathrm{d}x = 1$, and for real numbers a and b with $a \leq b$,

$$\mathbf{P}(a \leq X \leq b) = \int_a^b f_X(x)\mathrm{d}x.$$

  In particular,

$$F_X(a) = \int_{-\infty}^{a} f_X(x)\mathrm{d}x.$$

  The function f_X is called the *probability density function (pdf)* of X. As in the discrete case, F_X is called the *cdf* of X.

- For continuous RV X and for $0 \leq p \leq 1$, the p^{th} *quantile* or $100p^{\text{th}}$ *percentile* of the distribution of X is the smallest number q_p such that

$$F_X(q_p) = p$$

  The *median* of a distribution is its 50^{th} percentile.

- The *pdf* f_X and *cdf* F_X of a continuous random variable X are related by

$$F_X(b) = \int_{-\infty}^{b} f_X(x)\mathrm{d}x$$

and

$$f_X(x) = \frac{d}{\mathrm{d}x} F_X(x)$$

- *Specifying the distribution* of a RV X means identifying the characteristic that uniquely determines the probabilities associated with X. This can be done by specifying any one of the following:

  1. The *cdf* of X (works for RVs that are discrete, continuous or neither)
  2. The *pmf* (discrete) or *pdf* (continuous) of X
  3. The name of a standard RV
  4. The moment generating function, $\phi_X(t)$
  5. The Laplace transform, $\mathcal{L}_X t$
  6. The characteristic function, $\psi_X(t)$

Common Continuous Distributions

- The *uniform distribution* on $[a, b]$, Uni(a, b), is given by the *pdf* $f_X(x) = 0$ if $x \notin [a, b]$, and

$$f_X(x) = \frac{1}{b-a} \text{ for } a \le x \le b$$
$$F_X(x) = \frac{x-a}{b-a}$$

  Expectation: $\mathbf{E}\, X = (a+b)/2$. Variance: $\text{Var}(X) = (b-a)^2/12$.

  Note that the uniform RV is a special case of the beta RV: Beta(1,1) = Uni(0,1).

- The *beta distribution* on [0,1], Beta(α, β) with parameters $\alpha > 0$, $\beta > 0$, is given by $f_X(x) = 0$ if $x < 0$ or $x > 1$, and

$$f_X(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} \text{ for } 0 < x < 1$$

  where $B(\alpha, \beta)$ is the *Beta function*:

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\mathrm{d}x = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$
$$\text{for } \alpha > 0, \beta > 0$$

  and $\Gamma(\alpha)$ is the *Gamma function*, a generalized factorial:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}\mathrm{d}x \quad \text{for} \quad \alpha > 0$$
$$\Rightarrow \quad \Gamma\Big(\frac{1}{2}\Big) = \sqrt{\pi}, \quad \Gamma(m+1) = m\Gamma(m)$$
$$\text{and } m \in \mathbf{Z} \quad \Rightarrow \quad \Gamma(m+1) = m!$$

  The beta RV on [0,1] with parameters α and β has mean and variance given by

$$\mathbf{E}\, X = \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad \text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

  The beta RV offers flexible shapes of a density on a bounded interval. The standard beta RV is defined on [0,1], but it can be shifted and scaled to other intervals.

  On an interval $[a, b]$, the beta distribution with parameters $\alpha > 0$, $\beta > 0$, is given by $f_X(x) = 0$ if $x < a$ or $x > b$, and

$$f_X(x) = \frac{1}{b-a}\frac{1}{B(\alpha,\beta)}\Big(\frac{x-a}{b-a}\Big)^{\alpha-1}\Big(\frac{b-x}{b-a}\Big)^{\beta-1} \text{ for } a < x < b.$$

- The *Gamma distribution* with shape parameter $\alpha > 0$ and scale parameter $\lambda > 0$ has the density

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

If $X \sim \text{Gamma}(\alpha, \lambda)$, then $\mathbf{E}\, X = \alpha/\lambda$, $\text{Var}(X) = \alpha/\lambda^2$, and

$$\phi_X(t) = \big(\frac{\lambda}{\lambda - t}\big)^{\alpha}.$$

If the shape parameter $\alpha > 0$ is an integer, then the Gamma distribution is also called the *Erlang distribution* with α degrees of freedom and scale λ.

Like the Poisson RV, the Gamma RV is **reproducing**: if $X_1, \ldots, X_n$ are independent Gamma RVs with the same scale parameter λ, and different shape parameters $\alpha_1, \ldots, \alpha_n$, then the sum $Y = X_1 + \cdots + X_n$ is also Gamma distributed with scale λ and shape $\alpha_1 + \cdots + \alpha_n$.

- The *exponential distribution* with parameter λ, $\text{Exp}(\lambda)$, is given by $f_X(x) = 0$ if $x < 0$, and

$$f_X(x) = \lambda e^{-\lambda x} \text{ for } x \geq 0$$
$$F_X(a) = 1 - e^{-\lambda a} \text{ for } a \geq 0$$

Like the geometric distribution, the exponential distribution is **memoryless**. If $X \sim \text{Exp}(\lambda)$, then for any $x, y > 0$,

$$\mathbf{P}(X > x + y \;\big|\; X > x) = \mathbf{P}(X > y) = e^{-\lambda y}.$$

If $X \sim \text{Exp}(\lambda)$, then $\mathbf{E}\, X = 1/\lambda$ and $\text{Var}(X) = 1/\lambda^2$.

NB. $\text{Exp}(\lambda)$ is the Gamma(1, λ) distribution.

Competition between exponential RVs.

Let $X_i \sim \text{Exp}(\lambda_i)$, $i = 1, \ldots, n$. Then $Y = \min\{X_i\}$ has *cdf*

$$F_Y(y) = 1 - e^{-(\lambda_1 + \cdots + \lambda_n)y},$$

and the probability that the winner is X_j is

$$\mathbf{P}(Y = x_j) = \frac{\lambda_j}{\lambda_1 + \cdots + \lambda_n}.$$

- The *Pareto distribution* with parameter $\alpha > 0$, $\text{Par}(\alpha)$, is given by $f_X(x) = 0$ if $x < 1$, and

$$f_X(x) = \frac{\alpha}{x^{\alpha+1}} \text{ for } x \geq 1$$
$$F_X(x) = 1 - \frac{1}{x^{\alpha}} \text{ for } x \geq 1$$

5 The Normal Distribution

This is the most important continuous distribution, by far. The normal distribution is also called *Gaussian*.

Definitions

- The *normal distribution* with parameters μ (expectation) and $\sigma^2 > 0$ (variance), $\mathcal{N}(\mu, \sigma^2)$ is given by

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \text{ for } -\infty < x < \infty$$

and

$$F_X(a) = \int_{-\infty}^{a} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \mathrm{d}x \text{ for } -\infty < x < \infty$$

and

$$\phi_X(t) = e^{\mu t + \sigma^2 t^2/2}, \text{ for } -\infty < t < \infty.$$

- The *standard normal distribution* $\mathcal{N}(0, 1)$ is the normal distribution with zero mean and variance $\sigma^2 = 1$. If $X \sim \mathcal{N}(0, 1)$, then

$$f_X(x) = \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \text{ for } -\infty < x < \infty$$

and

$$F_X(x) = \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \mathrm{d}x \text{ for } -\infty < x < \infty$$

  Note the symmetry of $\phi(x)$ about $x = 0$.

  Table B.1 has right-tail probabilities, a vs. $1 - \Phi(a)$, for $\mathcal{N}(0, 1)$.

- *Normal approximation to the binomial distribution.* Let $X \sim \text{Bin}(n, p)$, with n large and p not too close either to 0 or 1, and let

$$Y = \frac{X - np}{\sqrt{np(1-p)}}.$$

  Then Y has, approximately, the standard normal distribution $\mathcal{N}(0, 1)$.

  When is this approximation valid? Rule of thumb: $np > 5$ and $n(1-p) > 5$.

- *Normal approximation to the Poisson distribution.* Let $X \sim \text{Poiss}(\lambda)$, with λ large. Let

$$Y = \frac{X - \lambda}{\sqrt{\lambda}}.$$

Then Y has, approximately, the standard normal distribution $\mathcal{N}(0,1)$.

When is this approximation valid? Rule of thumb: $\lambda > 5$.

- *Hierarchy of approximations.* Often we have a model with behavior that's well-described by the hypergeometric distribution (drawing balls without replacement). That's hard to work with, so we approximate it with the binomial distribution. In the limit that $n \to \infty$, $p \to 0$, $np \to \lambda > 0$, the binomial distribution simplifies to the Poisson distribution with parameter λ. Finally, we can approximate the Poisson distribution with the Normal distribution, because it's super easy to work with.

  hypergeometric $\Rightarrow$ binomial $\Rightarrow$ Poisson $\Rightarrow$ normal

- *Chi-square distribution with n degrees of freedom.* This RV is closely related to the normal RV. If $X \sim \chi^2_n$, then

$$f_X(x) = \begin{cases} \frac{x^{\frac{n}{2}-1}}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} e^{-\frac{x}{2}}, & x > 0 \\ 0, & x \le 0. \end{cases}$$

  Expectation: $\mathbf{E}\, X = n$. Variance: $\mathrm{Var}(X) = 2n$.

  If $X_1, \ldots, X_n$ are standard normal RVs, and $Y = X_1^2 + \cdots + X_n^2$ then $Y \sim \chi^2_n$.

  NB. The χ^2_n distribution is Gamma$(n/2, 1/2)$, so it has the reproducing property: if $X_1, \ldots, X_n$ are independent $\chi^2_{k_i}$ RVs, and $Y = X_1 + \cdots + X_n$, then $Y \sim \chi^2_k$, where $k = k_1 + \cdots + k_n$.

- *Student t distribution with n degrees of freedom.* This RV is also closely related to the normal RV. If X is a Student t RV with n degrees of freedom, then

$$f_X(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\,\Gamma(\frac{n}{2})} \Big(1 + \frac{x^2}{n}\Big)^{-(n+1)/2}, \quad -\infty < x < \infty.$$

  Expectation: $\mathbf{E}\, X = 0$. Variance: $\mathrm{Var}(X) = n/(n-2)$.

  How does the Student t RV arise?

  If $X \sim \mathcal{N}(0,1)$, $Y \sim \chi^2_n$, and X and Y are independent, then $T = X/\sqrt{Y/n}$ is a Student t RV with n DOF.

- *F distribution with n_1 and n_2 degrees of freedom.* This RV is again closely related to the normal RV. It has a complicated density, defined on $(0, \infty)$.

  How does the F RV arise?

  If $X_1 \sim \chi^2_{n_1}$, $X_2 \sim \chi^2_{n_2}$, and X_1 and X_2 are independent, then $Z = \frac{X_1/n_1}{X_2/n_2}$ is an F RV with n_1 and n_2 DOF.

- The *bivariate normal distribution.* This is the most important multivariable distribution. The bivariate normal distribution of a random vector (X, Y) is fully determined by:

  1. the mean μ_X and variance σ_X^2 of X;
  2. the mean μ_Y and variance σ_Y^2 of Y; and
  3. the correlation coefficient $\rho_{X,Y}$ of X and Y.

  If (X, Y) is jointly normal, then

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{X,Y}^2}} \exp\left\{ -\frac{\frac{(x-\mu_x)^2}{\sigma_X^2} - 2\rho_{X,Y}\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}}{2(1-\rho_{X,Y}^2)} \right\},$$

  defined for all $(x, y) \in \mathbf{R}^2$.

  The *mean vector* and *covariance matrix* of (X, Y) are

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}$$

  and

$$\Sigma = \begin{pmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(X,Y) & \mathrm{Var}(X) \end{pmatrix} = \begin{pmatrix} \sigma_X^2 & \rho_{X,Y}\sigma_X\sigma_Y \\ \rho_{X,Y}\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}.$$

  Since $\mathrm{Var}(X) = \mathrm{Cov}(X, X)$ and $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$, we can write the entries of the covariance more simply as $\big[\Sigma\big]_{ij} = \mathrm{Cov}(X_i, Y_j)$. The covariance matrix is positive semidefinite symmetric.

- More generally, the *multivariate normal distribution* of a random vector $\mathbf{X} = (X_1, \ldots, X_n)^T$ is defined by a vector of expectations $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ and the covariance matrix Σ. If $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, then

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det\Sigma}} \exp\Big\{ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big\}, \quad \mathbf{x} \in \mathbf{R}^n$$

Properties

- If X has an $\mathcal{N}(\mu_X, \sigma_X^2)$ distribution, then for any $r \neq 0$ and any s, $Y = rX + s$ has an $\mathcal{N}(r\mu + s, r^2\sigma^2)$ distribution.

  A handy application of this result: to find the probability $F_X(a)$ of a RV X with a $\mathcal{N}(\mu_X, \sigma_X^2)$ distribution,

  1. transform X to $Z = \frac{X-\mu_X}{\sigma_X}$, which has a $\mathcal{N}(0, 1)$ distribution
  2. note that $F_X(a) = F_Z(\frac{a-\mu_X}{\sigma_X}) = \Phi(\frac{a-\mu_X}{\sigma_X})$

3. use Table B.1 to look up the right-tail probability $1 - \Phi(\frac{a-\mu_X}{\sigma_X})$.

- The mean and median of the normal distribution coincide.
- Normal RVs stay normal under linear transformation: if $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y = aX+b$ for some scalars a, b, then $Y \sim \mathcal{N}(a\mu + b,\ a^2\sigma^2)$.
- A linear combination of independent normal RVs is again normal: if a_i are scalars and $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, $i = 1, \ldots, n$, and if $Y = \sum_{i=1}^n X_i$, then $Y \sim \mathcal{N}(\sum_{i=1}^n a_i\mu_i,\ \sum_{i=1}^n a_i^2\sigma_i^2)$.
- Like Poisson and Gamma RVs, normal RVs are *reproducing*: if $X_1, \ldots, X_n$ are independent normal RVs, with $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, and $Y = X_1 + \cdots + X_n$, then $Y \sim \mathcal{N}(\mu_1 + \cdots + \mu_n,\ \sigma_1^2 + \cdots + \sigma_n^2)$.
- If (X, Y) is jointly normal with means μ_X and μ_Y, variances σ_X^2, σ_Y^2, and correlation $\rho_{X,Y}$, then
  - marginal *pdfs* are normal
    - $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$
    - $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$
  - conditioning preserves normality
    - $(X|Y = y) \sim \mathcal{N}(\frac{\sigma_X}{\sigma_Y}(y - \mu_Y)\rho_{X,Y} + \mu_X,\ (1 - \rho_{X,Y}^2)\sigma_X^2)$
    - $(Y|X = x) \sim \mathcal{N}(\frac{\sigma_Y}{\sigma_X}(x - \mu_X)\rho_{X,Y} + \mu_Y,\ (1 - \rho_{X,Y}^2)\sigma_Y^2)$
  - linear combinations preserve normality
    - $Z = aX + bY \Rightarrow Z \sim \mathcal{N}(a\mu_X + b\mu_Y,\ a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\rho_{X,Y}\sigma_X\sigma_Y)$
  - for jointly normal RVs, uncorrelation implies independence (this is NOT true for general RVs)
    - $\rho_{X,Y} = 0 \Rightarrow f_{X,Y}(x, y) = f_X(x)f_Y(y) \Rightarrow X$ and Y are independent
  - all level sets of $f_{X,Y}(x, y)$ are ellipses (they're circles if $\sigma_X = \sigma_Y$ and $\rho_{X,Y} = 0$).
- If $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_x, P_{xx})$, $\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}_z, P_{zz})$, $\mathbf{cov}(\mathbf{X}, \mathbf{Z}) = P_{xz}$, then
  - marginal normality $\iff$ joint normality

$$\mathbf{Y} = \begin{bmatrix}\mathbf{X}\\ \mathbf{Z}\end{bmatrix} \quad \Rightarrow \quad \mathbf{E}\,\mathbf{Y} = \boldsymbol{\mu}_y = \begin{bmatrix}\boldsymbol{\mu}_x\\ \boldsymbol{\mu}_z\end{bmatrix} \quad \text{and} \quad \mathbf{cov}(\mathbf{Y}) = P_{yy} = \begin{bmatrix}P_{xx} & P_{xz}\\ P_{xz}^T & P_{zz}\end{bmatrix}$$

  - linear combination preserves normality

$$\mathbf{Y} = A\mathbf{X} \quad \Rightarrow \quad \mathbf{Y} \sim \mathcal{N}(A\mu_x, AP_{xx}A^T)$$

    (this is actually true for any random vectors, not necessarily Gaussian)
  - conditioning preserves normality

$$(\mathbf{X}|\mathbf{Z} = \mathbf{z}) \sim \mathcal{N}(\boldsymbol{\mu}_x + P_{xz}P_{zz}^{-1}(\mathbf{z} - \boldsymbol{\mu}_z),\ P_{xx} - P_{xz}P_{zz}^{-1}P_{xz}^T)$$

6 Expectation and Variance

Definitions

- The *expectation* of a discrete random variable X taking values $a_1, a_2, \ldots$ and with *pmf* p_X is the weighted average

$$\mathbf{E}\, X = \sum_i a_i \, \mathbf{P}(X = a_i) = \sum_i a_i p_X(a_i)$$

  For a continuous random variable X with *pdf* f_X, the expectation is

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

  If the integral diverges, e.g. for the Cauchy distribution, then we say the expectation DNE.

- The *variance* of X is the scalar

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

  An equivalent formula that's easier to compute:

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

  The number $\mathbf{E}\, X^2$ is called the *second moment* of X, so the variance is the second moment minus the square of the first moment.

  Why do we care about variance? Because it's a measure of the *dispersion* of a RV. There are other measures of dispersion, such as Maximum Average Dispersion (MAD), but variance is the simplest and most commonly used.

- The *standard deviation* of a RV X is $\sigma = \sqrt{\mathrm{Var}[X]}$.

Results

- **Expectation of a non-negative RV.** If $\mathbf{P}(X < 0) = 0$, then

$$\mathbf{E}\, X = \int_0^{\infty} (1 - F_X(x)) \mathrm{d}x.$$

  This formula is valid for *any* type of nonnegative RV – discrete, continuous or neither. This is useful when the *cdf* is known but the *pdf* or *pmf* is unknown.

  In the special case of a discrete RV taking values $0, 1, 2, \ldots$, this reduces to

$$\mathbf{E}\, X = \sum_{n=0}^{\infty} (1 - F_X(n)).$$

- **Expectation of g(X)**. Let X be a random variable and let $g : \mathbf{R} \to \mathbf{R}$. If X is discrete, with values $a_1, a_2, \ldots$, then

$$\mathbf{E}\, g(X) = \sum_i g(a_i)\ p_X(a_i)$$

  If X is continuous, with *pdf* f_X, then

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

- **Expectation of a product of independent RVs.** If RVs X and Y are independent, then

$$\mathbf{E}\,[XY] = (\mathbf{E}\, X)(\mathbf{E}\, Y).$$

- **Expectation and variance under linear transformation.** For any RV X and scalars r and s,

$$\mathbf{E}\,[rX + s] = r\,\mathbf{E}\,[X] + s$$

and

$$\mathrm{Var}(rX + s) = r^2\mathrm{Var}(X) \quad \Rightarrow \quad \sigma_{rX+s} = |r|\,\sigma_X.$$

- **Expectation and variance of a linear combination of RVs.** If $X_1, \ldots, X_n$ are RVs and $a_1, \ldots, a_n$ are scalars, then

$$\mathbf{E} \sum_{i=1}^{n} a_i X_i = \sum_{i=1}^{n} a_i\, \mathbf{E}\, X_i$$

  and

$$\mathrm{Var}(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i^2 \mathrm{Var}(X_i) + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} a_i a_j \mathrm{Cov}(X_i, X_j).$$

  In the special case of two RVs X and Y, and for scalars r, s, and t, we have

$$\mathbf{E}\,[rX + sY + t] = r\,\mathbf{E}\,[X] + s\,\mathbf{E}\,[Y] + t.$$

  and

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

7 Joint Distributions and Independence

Definitions

- The *joint probability mass function* $p_{X,Y}$ of two discrete RVs X and Y is the function $p_{X,Y} : \mathbf{R}^2 \to [0,1]$, defined by

$$p_{X,Y}(a,b) = \mathbf{P}(X = a, Y = b) \text{ for } -\infty < a, b < \infty$$

  Equivalently, for any set B,

$$\mathbf{P}((X,Y) \in B) = \sum_{(a_i,b_j)\in B} p_{X,Y}(a_i, b_j).$$

  The joint *pmf* $p_{X,Y}$ satisfies

  - **Non-negativity:** $p_{a_i,b_j} \geq 0$ for all a_i and b_j
  - **Normalization:** $\sum_{a_i,b_j} p_{X,Y}(a_i, b_j) = 1$

  Note: the joint *pmf* contains more information than can be obtained from the marginal *pmfs* p_X and p_Y. In fact, sometimes the joint *pmf* can't be be retrieved from the marginal *pmf*s.

- The *joint probability density function* $f_{X,Y}$ of two continuous RVs X and Y is the function $f_{X,Y} : \mathbf{R}^2 \to \mathbf{R}$, defined for all numbers a_1, a_2, b_1, b_2 with $a_1 \leq b_1$ and $a_2 \leq b_2$ by

$$\mathbf{P}(a_1 \leq X \leq b_1, a_2 \leq Y \leq b_2) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y$$

  The joint *pdf* $f_{X,Y}$ satisfies

  - **Non-negativity:** $f_{X,Y} \geq 0$ for all x and y
  - **Normalization:** $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y = 1$

  The joint *pdf* of two RVs is also called the *bivariate probability density.*

- The *joint cumulative distribution function* $F_{X,Y}$ of two RVs X and Y is the function $F : \mathbf{R}^2 \to [0,1]$ defined by

$$F_{X,Y}(a,b) = \mathbf{P}(X \leq a, Y \leq b) \text{ for } -\infty < a, b < \infty$$

  In practice, the joint *cdf* is rarely known. Most practical work is done with the joint *pmf* or *pdf.*

- Two RVs X and Y with joint *cdf* $F_{X,Y}$ are *independent* if

$$F_{X,Y}(a,b) = F_X(a)F_Y(b) \quad \text{for all } a, b$$

or, equivalently,

$$\mathbf{P}(X \leq a, Y \leq b) = \mathbf{P}(X \leq a)\,\mathbf{P}(Y \leq b).$$

For continuous RVs X and Y with joint *pdf* $f_{X,Y}$, the definition of independence implies that

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

For discrete RVs X and Y, the definition of independence implies that

$$\begin{aligned} &\mathbf{P}(X = a, Y = b) = \mathbf{P}(X = a)\,\mathbf{P}(Y = b) \\ \iff \quad &p_{X,Y}(a,b) = p_X(a)p_Y(b) \end{aligned}$$

- **Necessary and sufficient conditions for independence.** RVs X and Y are independent if and only if they satisfy *both* of the following conditions.
  - **Separable joint *pdf*:** $f_{X,Y}(x,y) = g(x)h(y)$ for some functions g and $h : \mathbf{R} \to \mathbf{R}$, and
  - **Rectangular feasible region:** The bounds on x and y over which $f_{X,Y}(x,y)$ is defined form a rectangle, i.e. $a \leq x \leq b$ and $c \leq y \leq d$ for some scalars a, b, c, d.

  So if the bounds on x depend on y or vice versa, then X and Y are dependent.

Results

- **Marginal *cdfs* from joint *cdf*.** The marginal *cdfs* F_X and F_Y are obtained from the joint *cdf* $F_{X,Y}$, for each a and b, by

$$F_X(a) = \mathbf{P}(X \leq a) = F(a,\infty) = \lim_{b\to\infty} F_{X,Y}(a,b)$$

and

$$F_Y(b) = \mathbf{P}(Y \leq b) = F(\infty,b) = \lim_{a\to\infty} F_{X,Y}(a,b)$$

- **Marginal *pdfs* from joint *pdf*.** The marginal *pdfs* f_X and f_Y are obtained from the joint *pdf* $f_{X,Y}$ by integrating out the other variable, i.e.

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

and

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

- **Marginal *pmfs* from joint *pmf*.** The marginal *pmfs* p_X and p_Y are obtained from the joint *pmf* $p_{X,Y}$ by summing out the other variable, i.e.

$$p_X(a_i) = \sum_j p_{X,Y}(a_i, b_j)$$

and

$$p_Y(b_j) = \sum_i p_{X,Y}(a_i, b_j)$$

- **Relating joint *pdf* and joint *cdf*.** The joint *pdf* $f_{X,Y}$ and the joint *cdf* $F_{X,Y}$ are related by

$$F_{X,Y}(a,b) = \int_{-\infty}^{a} \int_{-\infty}^{b} f_{X,Y}(a,b)\mathrm{d}x\mathrm{d}y$$

and

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

- **Propagation of independence.** Let RVs $X_1, X_2, \ldots, X_N$ be independent. For each i, let $h_i : \mathbf{R} \to \mathbf{R}$ be a function, and let $Y_i = h_i(X_i)$. Then $Y_1, Y_2, \ldots, Y_N$ are independent. (Note that the functions h_i do not have to be the same.)

- **Expectation of g(X,Y).** If (X,Y) is a discrete random vector with *pmf* $p_{X,Y}$, then for any function $g : \mathbf{R}^2 \to \mathbf{R}$,

$$\mathbf{E}\, g(X,Y) = \sum_{x_i} \sum_{y_j} g(x_i, y_j) p_{X,Y}(x_i, y_j)$$

  If (X,Y) is a continuous random vector with *pdf* $f_{X,Y}$, then for any function $g : \mathbf{R}^2 \to \mathbf{R}$,

$$\mathbf{E}\, g(X,Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y$$

  This generalizes in the natural way to n RVs.

8 Covariance and Correlation

Definitions

- The *covariance* between RVs X and Y is defined by

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

  Or, equivalently,

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

  If X and Y are *positively correlated*, then $\mathrm{Cov}(X,Y) > 0$, and when $X > \mathbf{E}\,X$, it tends to be true that $Y > \mathbf{E}\,Y$ as well. If, on the other hand, when $X > \mathbf{E}\,X$ we tend to observe that $Y < \mathbf{E}\,Y$, then X and Y are *negatively correlated* and $\mathrm{Cov}(X,Y) < 0$. If $\mathrm{Cov}(X,Y) = 0$, then X and Y are *uncorrelated*.

  If X and Y are independent, then X and Y are uncorrelated. However, *dependent* RVs X and Y can also be uncorrelated. Independence implies uncorrelation, but uncorrelation does not imply independence.

- The *correlation coefficient* $\rho(X,Y)$, for RVs X and Y, is defined to be 0 if $\mathrm{Var}(X) = 0$ or $\mathrm{Var}(Y) = 0$, and otherwise

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

  The correlation coefficient is dimensionless, and is invariant (except for a possible sign change) under linear transformation of X and Y. Because of this, it's used as a standardized version of covariance.

  The correlation coefficient has the following properties.

  - **Linearity.** For RVs X and Y, and for constants r, s, t and u with $r, t \neq 0$,

$$\rho(rX + s, tY + u) = \begin{cases} -\rho(X,Y) \text{ if } rt < 0, \\ \rho(X,Y) \text{ if } rt > 0. \end{cases}$$

  - **Bounds.** For nonconstant RVs X and Y,

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

  - **NB.** Correlation measures the *directional dependence* between X and Y. X and Y are "most correlated" if $X = Y$, in which case $\rho(X,Y) = 1$, or if $X = -Y$, in which case $\rho(X,Y) = -1$.

Results

- **Covariance from joint *pdf* or *pmf*.** If (X, Y) is *discrete* with possible values (x_i, y_j) and joint *pmf* $p_{X,Y}$, then

$$\mathrm{Cov}(X, Y) = \sum_{x_i} \sum_{y_j} (x_i - \mu_X)(y_j - \mu_Y) p_{X,Y}(x_i, y_j)$$

  or, equivalently,

$$\mathrm{Cov}(X, Y) = \sum_{x_i} \sum_{y_j} x_i y_j p_{X,Y}(x_i, y_j) - \mu_X \mu_Y.$$

  If (X, Y) is *continuous* with joint *pdf* $f_{X,Y}$, then

$$\mathrm{Cov}(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f_{X,Y}(x, y) \mathrm{d}x \mathrm{d}y$$

  or, equivalently,

$$\mathrm{Cov}(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_{X,Y}(x, y) \mathrm{d}x \mathrm{d}y - \mu_X \mu_Y.$$

- **Covariance under linear transformation.** For RVs X and Y,

$$\mathrm{Cov}(rX + s, tY + u) = rt\mathrm{Cov}(X, Y).$$

- **Symmetry of covariance.** For RVs X and Y,

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

- **General additivity of covariance.** For RVs $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_k$,

$$\mathrm{Cov}(\sum_{i=1}^{n} X_i, \sum_{j=1}^{k} Y_j) = \sum_{i=1}^{n} \sum_{j=1}^{k} \mathrm{Cov}(X_i, Y_j).$$

  Special case:

$$\mathrm{Cov}(X_1 + X_2, Y) = \mathrm{Cov}(X_1, Y) + \mathrm{Cov}(X_2, Y).$$

9 Random Vectors

The theory for two RVs extends naturally to n RVs. This section lays out some notation and linear algebraic conveniences.

Definitions

- Let $X_1, \ldots, X_n$ RVs. Then $\mathbf{X} = (X_1, \ldots, X_n)^T$ is a **random vector** with realization $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbf{R}^n$.
- ***pmf.*** Let $\mathbf{X}$ be a vector of discrete RVs $X_1, \ldots, X_n$. Then the *pmf* of $\mathbf{X}$ is the joint *pmf* of $X_1, \ldots, X_n$:
$$p_{\mathbf{X}}(\mathbf{x}) = p_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$$
- ***pdf.*** Let $\mathbf{X}$ be a vector of continuous RVs $X_1, \ldots, X_n$. Then the *pdf* of $\mathbf{X}$ is the joint *pdf* of $X_1, \ldots, X_n$:
$$f_{\mathbf{X}}(\mathbf{x}) = f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$$
- **Expectation**. The expectation of $\mathbf{X}$, $\mathbf{E}\,\mathbf{X} \in \mathbf{R}^n$, is just a list of the expectations of $X_1, \ldots, X_n$:
$$\begin{aligned}\mathbf{E}\,\mathbf{X} &= \begin{bmatrix}\mathbf{E}\,X_1 \\ \vdots \\ \mathbf{E}\,X_n\end{bmatrix} = \begin{bmatrix}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} x_1 f_{\mathbf{X}}(\mathbf{x})\mathrm{d}x_1\cdots\mathrm{d}x_n \\ \vdots \\ \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} x_n f_{\mathbf{X}}(\mathbf{x})\mathrm{d}x_1\cdots\mathrm{d}x_n\end{bmatrix} \\ &= \begin{bmatrix}\int_{-\infty}^{\infty} x_1 f_{\mathbf{X}}(\mathbf{x})\mathrm{d}\mathbf{x} \\ \vdots \\ \int_{-\infty}^{\infty} x_n f_{\mathbf{X}}(\mathbf{x})\mathrm{d}\mathbf{x}\end{bmatrix} = \int_{\mathbf{R}^n} \mathbf{x} f_{\mathbf{X}}(\mathbf{x})\mathrm{d}\mathbf{x}\end{aligned}$$
where $\mathrm{d}\mathbf{x} = \mathrm{d}x_1\cdots\mathrm{d}x_n \in \mathbf{R}$. Expectation is defined similarly for discrete random vectors. In the above notation, we use the fact that the integral of a vector $\mathbf{v}$ is a vector of (scalar) integrals of the components of $\mathbf{v}$.
- **Covariance matrix**. The covariance matrix of a random vector $\mathbf{X}$, denoted $\mathbf{cov}(\mathbf{X}) \in \mathbf{R}^{n\times n}$, is
$$\begin{aligned}\mathbf{cov}(\mathbf{X}) &= \mathbf{E}\left[(\mathbf{X} - \mathbf{E}\,\mathbf{X})(\mathbf{X} - \mathbf{E}\,\mathbf{X})^T\right] \\ &= \begin{bmatrix}\mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \ldots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2,) & & \vdots \\ \vdots & & \ddots & \\ \mathrm{Cov}(X_n, X_1) & \ldots & & \mathrm{Var}(X_n)\end{bmatrix}\end{aligned}$$
Since $\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i)$, $\mathbf{cov}(\mathbf{X})$ is symmetric. If $X_1, \ldots, X_n$ are linearly independent, then $\mathbf{cov}(\mathbf{X}) \succ 0$.

Results

- **Expectation of a quadratic form.** If $A = A^T \in \mathbf{R}^n$ and $\mathbf{X}$ is a random n-vector with expectation $\boldsymbol{\mu}_x$, then

$$\mathbf{E}\left[\mathbf{X}^T A \mathbf{X}\right] = \mathrm{tr}[A(\boldsymbol{\mu}_x \boldsymbol{\mu}_x^T + P_{xx})]$$

10 Transformations of Random Variables

The problem we consider in this chapter: given a RV X with a known distribution (*pdf*, *pmf*, or *cdf*), and a transformation $T : \mathbf{R} \to \mathbf{R}$, find the distribution of $Y = T(X)$. In the scalar case, we consider both monotonic and non-monotonic transformations.

In the context of random vectors, the question is slightly different: given a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ with a known joint distribution (joint *pdf*, joint *pmf*, or joint *cdf*), and a transformation $\mathbf{Y} = T(\mathbf{X}) = (Y_1, \ldots, Y_n)$, find the joint distribution of $\mathbf{Y}$. In the vector case, we consider only one-to-one transformations between spaces of the same dimension, i.e. transformations T such that $T : \mathbf{R}^n \to \mathbf{R}^n$ where T^{-1} is well-defined.

We also discuss a few standard transformations $T : \mathbf{R}^2 \to \mathbf{R}$ (sum, difference, product, quotient) for independent RVs.

Definitions

- **Jacobian.** Let $\mathbf{X}$ and $\mathbf{Y}$ be n-dimensional RVs, with $\mathbf{Y} = T(\mathbf{X})$, where T is one-to-one, and hence T^{-1} is well-defined. Then $\mathbf{x} = T^{-1}(\mathbf{y}) \iff \mathbf{x} = \mathbf{h}(\mathbf{y})$ for some function $\mathbf{h}(\mathbf{y}) = T^{-1}(\mathbf{y}) = (h_1(\mathbf{y}), \ldots, h_n(\mathbf{y}))^T$. The Jacobian of T^{-1} is the n by n determinant
$$J_{T^{-1}}(\mathbf{y}) = \begin{vmatrix} \frac{\partial h_1}{\partial y_1} & \cdots & \frac{\partial h_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_n}{\partial y_1} & \cdots & \frac{\partial h_n}{\partial y_n} \end{vmatrix} = \begin{vmatrix} \leftarrow & \boldsymbol{\nabla} h_1(\mathbf{y}) & \rightarrow \\ & \vdots & \\ \leftarrow & \boldsymbol{\nabla} h_n(\mathbf{y}) & \rightarrow \end{vmatrix},$$
where $\boldsymbol{\nabla} h_1(\mathbf{y})$ is the gradient of h_1 with respect to $\mathbf{y}$, viewed as a row vector.

- A twice differentiable function $g : \mathbf{R} \to \mathbf{R}$ is *convex* on an interval I if $g''(x) \geq 0$ for all $x \in I$, and *strictly convex* if $g''(x) > 0$ for all $x \in I$. Graphical interpretation: a line drawn between any two points on the graph of a convex function will always lie above the graph.

Results

- ***cdf* under scalar transformation.** Let $Y = T(X)$, where X has *pmf* $p_X(x)$ or *pdf* $f_X(x)$. Then the *cdf* of Y is, in the discrete case,
$$F_Y(y) = \sum_{\{x_i | T(x_i) \leq y\}} p_X(x_i), \quad -\infty < y < \infty$$
or, in the continuous case,
$$F_Y(y) = \int_{\{x | T(x) \leq y\}} f_X(x) \mathrm{d}x, \quad -\infty < y < \infty$$
NB. These computations are often hard or impossible to do directly!

- ***pmf* under scalar transformation.** Let X have *pmf* $p_X(x)$, and let $Y = T(X)$. Then the *pmf* of Y is

$$p_Y(y_j) = \sum_{\{x_i | T(x_i) = y_j\}} p_X(x_i).$$

  In the special case where T is one-to-one,

$$p_Y(y_j) = p_X(T^{-1}(y_j)).$$

- ***pdf* under scalar transformation.** Let X have *pdf* $f_X(x)$, and let $Y = T(X)$. If T is a general transformation, not necessarily monotonic, and if $T(x) = y$ has roots $T_1^{-1}(y), T_2^{-1}(y), \ldots$, then Y has *pdf*

$$f_Y(y) = \sum_i f_X(T_i^{-1}(y)) |\frac{\mathrm{d}}{\mathrm{d}y} T_i^{-1}(y)|.$$

  In the special case that T is one-to-one (and hence T^{-1} is well-defined), $Y = T(X)$ has *pdf*

$$f_Y(y) = f_X(T^{-1}(y)) |\frac{\mathrm{d}}{\mathrm{d}y} T^{-1}(y)|.$$

- ***pmf* under vector transformation.** Let $\mathbf{X}$ have joint *pmf* $p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = p_{\mathbf{X}}(\mathbf{x})$, and let $\mathbf{Y} = T(\mathbf{X})$ with T one-to-one. Then $\mathbf{Y}$ is also discrete, with *pmf*

$$p_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = p_{\mathbf{Y}}(\mathbf{y}) = \sum_{\{\mathbf{x} | T(\mathbf{x}) = \mathbf{y}\}} p_{\mathbf{X}}(\mathbf{x}).$$

- ***pdf* under vector transformation.** Let $\mathbf{X}$ have joint *pdf* $f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{\mathbf{X}}(\mathbf{x})$, and let $\mathbf{Y} = T(\mathbf{X})$ with T one-to-one. Then $\mathbf{Y}$ *may* also be continuous (no guarantees). If so, its *pdf* is

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(T^{-1}(\mathbf{y})) |J_{T^{-1}}(\mathbf{y})|.$$

- **Standard transformation: sum.** Let X_1 and X_2 be continuous, *independent* RVs with *pdfs* $f_{X_1}(x_1)$ and $f_{X_2}(x_2)$, and let $Z = X_1 + X_2$. Then $f_Z(z)$ is given by the **convolution** of X_1 and X_2:

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X_1}(z - v) f_{X_2}(v) \mathrm{d}v.$$

- **Standard transformation: difference.** Let X_1 and X_2 be continuous, *independent* RVs with *pdfs* $f_{X_1}(x_1)$ and $f_{X_2}(x_2)$, and let $Z = X_1 - X_2$. Then $f_Z(z)$ is given by:

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X_1}(z + v) f_{X_2}(v) \mathrm{d}v.$$

- **Standard transformation: product.** Let X_1 and X_2 be continuous, *independent* RVs with *pdfs* $f_{X_1}(x_1)$ and $f_{X_2}(x_2)$, and let $Z = X_1 X_2$. Then $f_Z(z)$ is given by

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{|v|}\ f_{X_1}(z/v)\ f_{X_2}(v)\mathrm{d}v.$$

- **Standard transformation: quotient.** Let X_1 and X_2 be continuous, *independent* RVs with *pdfs* $f_{X_1}(x_1)$ and $f_{X_2}(x_2)$, and let $Z = X_1/X_2$. Then $f_Z(z)$ is given by

$$f_Z(z) = \int_{-\infty}^{\infty} |v|\ \ f_{X_1}(zv)\ f_{X_2}(v)\mathrm{d}v.$$

- ***cdf* and *pdf* under linear transformation.** Let X be a continuous RV with *cdf* F_X and *pdf* f_X. If we change units to $Y = rX + s$ for real numbers r and s, then

$$F_Y(y) = F_X(\frac{y-s}{r}) \text{ and } f_Y(y) = \frac{1}{r} f_X(\frac{y-s}{r})$$

- **Y = 1/X transformation.** If RV X has *pdf* f_X and $Y = 1/X$, then the *pdf* of Y is

$$f_Y(y) = \frac{1}{y^2} f_X\Big(\frac{1}{y}\Big)$$

  for $y < 0$ and $y > 0$. Any value for $f_Y(0)$ is fine.

- **Jensen's inequality.** Let g be a convex function, and let X be a RV. Then

$$g(\mathbf{E}\, X) \leq \mathbf{E}\, g(X)$$

  If g is strictly convex, then Jensen's inequality is strict.

  The exception to Jensen's inequality is when $\mathrm{Var}(X) = 0$, i.e. X is not random at all.

11 The Law of Large Numbers

Definitions

- A sequence of RVs $X_1, X_2, X_3, \ldots$ is *independent and identically distributed (iid)* if the RVs in the sequence are independent and all have the same distribution.

- The *sample mean* $\bar{X}_n$ of n *iid* RVs is defined by

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}.$$

  If the expectation and variance of the X_i are μ and σ^2, then the expectation and variance of the sample mean are

$$\mathbf{E}\,\bar{X}_n = \mu$$

  and

$$\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}.$$

  Remark: as n increases, the sample mean becomes more and more concentrated around the true mean μ of the X_i. Loosely speaking, this is the Law of Large Numbers:

$$\bar{X}_n \to \mu \quad \text{as} \quad n \to \infty.$$

Results

- **Markov's Inequality.** For any *nonnegative* RV X and any positive scalar a,

$$\mathbf{P}(X > a) \leq \frac{\mathbf{E}\,X}{a}.$$

- **Chebyshev's Inequality.** For a RV Y and any $a > 0$,

$$\mathbf{P}(|Y - \mathbf{E}\,Y| \geq a) \leq \frac{1}{a^2}\mathrm{Var}(Y).$$

  Qualitatively, this says that for any distribution, most of the probability mass is within a few standard deviations of the expectation. This is the so-called **"$\mu\pm$ a few rule."**

- **The Weak Law of Large Numbers.** If $\bar{X}_n$ is the sample mean of n *iid* RVs with expectation μ and variance σ^2, then for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbf{P}(\left|\bar{X}_n - \mu\right| > \varepsilon) = 0.$$

  The law of large numbers, in combination with simulation and histograms, allows us to recover the probability of any particular event, and therefore the entire *pdf*.

NB. The law of large numbers is valid for distributions with **finite expectation.** Not all distributions have finite expectation, however. For example, the expectation of the Cauchy distribution DNE, and the expectation of the Pareto distribution for $\alpha < 1$ is infinite. $\text{Par}(\alpha < 1)$ is called a *heavy-tailed distribution*, because $\bar{X}_n$ grows without bound as n increases. This is due to the occasional VERY LARGE realization of X_i.

- **The Strong Law of Large Numbers.** If $\bar{X}_n$ is the sample mean of n *iid* RVs with expectation μ, then
$$P\big(\lim_{n\to\infty} \bar{X}_n = \mu\big) = 1.$$

12 Moment Generating Function, Laplace Transform, and Characteristic Function

- The **moment generating function** ϕ_X of a RV X is defined by

$$\phi_X(t) = \mathbf{E}\, e^{tX}, \quad -\infty < t < \infty$$

The reason for the name: for every $n = 1, 2, \ldots,$

$$\frac{\mathrm{d}^n}{\mathrm{d}t^n}\phi_X(t)\Big|_{t=0} = \mathbf{E}\,[X^n],$$

the n^{th} moment of X.

- The **Laplace transform** $\mathcal{L}\,[X]$ of a RV X is defined by

$$\mathcal{L}\,[X]\,(t) = \mathbf{E}\, e^{-tX} = \phi_X(-t), \quad t > 0.$$

The Laplace transform (when it is defined) is always finite.

The Laplace transform can be used to generate moments of X:

$$\frac{\mathrm{d}^n}{\mathrm{d}t^n}\mathcal{L}\,[X]\,(t)\Big|_{t=0} = (-1)^n\, \mathbf{E}\,[X^n],$$

- The **characteristic function** ψ_X of a RV X is defined by

$$\psi_X(t) = \mathbf{E}\, e^{itX} = \phi_X(-t), -\infty < t < \infty$$

The characteristic function is used when the moment generating function and Laplace transform are undefined. The characteristic function is always defined.

The characteristic function can be used to generate moments of X:

$$\frac{\mathrm{d}^n}{\mathrm{d}t^n}\psi_X(t)\Big|_{t=0} = i^n \;\; \mathbf{E}\,[X^n],$$

- If RVs X and Y are independent, then

$$\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$$

and

$$\mathcal{L}_{X+Y}(t) = \mathcal{L}_X(t)\mathcal{L}_Y(t)$$

and

$$\psi_{X+Y}(t) = \psi_X(t)\psi_Y(t).$$

13 Conditional Distributions

Definitions

- **Conditional *pmf* and *pdf*.** Let (X, Y) be a discrete random vector. Then the conditional distribution of X given $Y = y_j$ is

$$p_{X|Y}(x_i|y_j) = \frac{p_{X,Y}(x_i, y_j)}{p_Y(y_j)}.$$

  If X and Y are independent, then $p_{X|Y}(x_i, y_j) = p_X(x_i)$ for all x_i.

  If (X,Y) is a continuous random vector, then the conditional *pdf* of X given $Y = y$ is

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

  If X and Y are independent, then $f_{X|Y}(x, y) = f_X(x)$ for all x.

- **Conditional expectation.** Let (X, Y) be a discrete random vector. Then the conditional expectation of X given $Y = y_j$ is

$$\mathbf{E}[X|Y = y_j] = \sum_{x_i} x_i p_{X|Y}(x_i|y_j).$$

  If (X,Y) is a continuous random vector, then the conditional expectation of X given $Y = y$ is

$$\mathbf{E}[X|Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\mathrm{d}x.$$

  NB. The conditional expectation of X given Y is itself a random vector (or a random variable, for a particular realization of $Y = y$).

- **Conditional variance.** Let (X, Y) be a discrete random vector. Then the conditional variance of X given $Y = y_j$ is

$$\mathrm{Var}(X|Y = y_j) = \mathbf{E}[X^2|Y = y_j] - \mathbf{E}[X|Y = y_j]^2$$

  where

$$\mathbf{E}[X^2|Y = y_j] = \sum_{x_i} x_i^2 p_{X|Y}(x_i|y_j).$$

  If (X,Y) is a continuous random vector, then the conditional expectation of X given $Y = y$ is

$$\mathrm{Var}(X|Y = y) = \mathbf{E}[X^2|Y = y] - \mathbf{E}[X|Y = y]^2$$

  where

$$\mathbf{E}[X^2|Y = y] = \int_{-\infty}^{\infty} x^2 f_{X|Y}(x|y)\mathrm{d}x.$$

  NB. The conditional variance of X given Y is itself a random vector.

Results

- **Iterated expectation.** For any random variables X and Y (discrete, continuous or mixed),
$$\mathbf{E}\,Y = \mathbf{E}\,[\mathbf{E}\,[Y|X]].$$
This is a useful tool for calculating the expectation of a RV Y that may be difficult or impossible to calculate directly.

  Unpacking this formula for the case of a discrete conditioning RV:
$$\mathbf{E}\,Y = \sum_{x_i} \mathbf{E}\,[Y|X = x_i] p_X(x_i).$$

  Unpacking this formula for the case of a continuous conditioning RV:
$$\mathbf{E}\,Y = \int_{-\infty}^{\infty} \mathbf{E}\,[Y|X = x] f_X(x)\mathrm{d}x.$$

- **Iterated variance.** For any random variables X and Y (discrete, continuous or mixed),
$$\mathrm{Var}(Y) = \mathbf{E}\,[\mathrm{Var}(Y|X)] + \mathrm{Var}(\mathbf{E}[Y|X]).$$
This is a useful tool for calculating the variance of a RV Y that may be difficult or impossible to calculate directly.

  Unpacking this formula for the case of a discrete conditioning RV:
$$\begin{aligned}\mathbf{E}\,\mathrm{Var}(Y|X) &= \sum_{x_i} \mathrm{Var}(Y|X = x_i) p_X(x_i), \text{ and} \\ \mathrm{Var}(\mathbf{E}\,[Y|X]) &= \sum_{x_i} \mathbf{E}\,[Y|X = x_i]^2 p_X(x_i) - \Big(\sum_{x_i} \mathbf{E}\,[Y|X = x_i] p_X(x_i)\Big)^2.\end{aligned}$$

  Unpacking this formula for the case of a continuous conditioning RV:
$$\begin{aligned}\mathbf{E}\,\mathrm{Var}(Y|X) &= \int_{-\infty}^{\infty} \mathrm{Var}(Y|X = x) f_X(x)\mathrm{d}x, \text{ and} \\ \mathrm{Var}(\mathbf{E}\,[Y|X]) &= \int_{-\infty}^{\infty} x\,\mathbf{E}\,[Y|X = x]^2 f_X(x)\mathrm{d}x - \Big(\int_{-\infty}^{\infty} \mathbf{E}\,[Y|X = x] f_X(x)\mathrm{d}x\Big)^2.\end{aligned}$$

- **Law of Total Probability for random variables.** Let $X_1, \ldots, X_n$ be random variables. Some may be discrete, some continuous. Let A be an event expressed in terms of $X_1, \ldots, X_n$.

If the conditioning event X_i is discrete, then the LTP is

$$\mathbf{P}(A) = \sum_{x_i} p_{X_i}(x_i)\,\mathbf{P}(A|X = x_i) \quad \text{(one conditioning event), and}$$

$$\mathbf{P}(A) = \sum_{x_i}\sum_{x_j} p_{X_i,X_j}(x_i, x_j)\,\mathbf{P}(A|X_i = x_i, X_j = x_j) \quad \text{(two conditioning events).}$$

If the conditioning event X_i is continuous, then the LTP is

$$\mathbf{P}(A) = \int_{-\infty}^{\infty} f_X(x)\,\mathbf{P}(A|X = x)\mathrm{d}x \quad \text{(one conditioning event), and}$$

$$\mathbf{P}(A) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X_i,X_j}(x_i, x_j)\,\mathbf{P}(A|X_i = x_i, X_j = x_j)\mathrm{d}x_i\mathrm{d}x_j \quad \text{(two conditioning events).}$$

14 Order Statistics

Order statistics are functions of a set of n *iid* RVs, $X_1, \ldots, X_n$. They're used to investigate the relative values – the biggest, the smallest, the middle, etc. Let f_X and F_X be the *pdf* and *cdf* of every X_i, for $i \in [1, n]$.

Definitions

- Let the RVs $X_{(1)}, \ldots, X_{(n)}$, be the **order statistics** of the *iid* **sample** $X_1, \ldots, X_n$. Then
$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n-1)} \leq X_{(n)}.$$
The inequalities are strict in the continuous case, since the probability of two continuous RVs being equal is zero.
- The **first ordered statistic**, $X_{(1)}$, is the minimum.
- The **last order statistic**, $X_{(n)}$, is the maximum.

Results

- ***cdf* and *pdf* of the minimum.** The distribution of the first order statistic is
$$F_{X_{(1)}}(x) = 1 - (1 - F_X(x))^n$$
and
$$f_{X_{(1)}}(x) = n(1 - F_X(x))^{n-1} f_X(x).$$
- ***cdf* and *pdf* of the maximum.** The distribution of the last order statistic is
$$F_{X_{(n)}}(x) = F_X(x)^n$$
and
$$f_{X_{(n)}}(x) = nF_X(x)^{n-1} f_X(x).$$
- ***cdf* and *pdf* of a general order statistic.** The distribution of the k^{th} order statistic is
$$F_{X_{(k)}}(x) = \sum_{i=k}^{n} \binom{n}{i} F_X(x)^i (1 - F_X(x))^{n-i},$$
and
$$f_{X_{(k)}}(x) = nf_X(x) \binom{n-1}{k-1} F_X(x)^{k-1} (1 - F_X(x))^{n-k}.$$

- **Joint density of two order statistics.** Let $1 \leq j \leq k \leq n$. Then the joint density of $X_{(j)}$ and $X_{(k)}$ is

$$f_{X_{(j)},X_{(k)}}(x,y) = f_X(x)f_X(y)\frac{n!}{(j-1)!(k-j-1)!(n-k)!} \\ * F_X(x)^{j-1}(F_X(y)-F_X(x))^{k-j-1}(1-F_X(y))^{n-k}, \quad \text{for } x < y.$$

- **Joint density of all order statistics.** Let $1 \leq j \leq k \leq n$. Then the joint density of $X_{(j)}$ and $X_{(k)}$ is

$$f_{X_{(1)},\ldots,X_{(n)}}(x_1,\ldots,x_n) = \begin{cases} n!f_X(x_1)\ldots f_X(x_n), & \text{if } x_1 \leq \cdots \leq x_n. \\ 0, & \text{otherwise.} \end{cases}$$

NB. This looks just like the *unordered* joint *pdf* $f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = f_X(x_1)\ldots f_X(x_n)$, except for two differences. First, we have a factor of $n!$ in the ordered version. Second, the ordered version has a greatly restricted feasible region ($x_1 \leq \cdots \leq x_n$).

15 The Central Limit Theorem

Let $X_1, \ldots, X_n$ be *iid* RVs with common mean $\mathbf{E}\, X_i = \mu$ and common variance $\mathrm{Var}(X_i) = \sigma^2$, for all $i \in [1, n]$.

Definitions

- **Sample mean.** Define the sample mean by

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}.$$

  Recall that the Law of Large Numbers says that

$$\bar{X}_n \to \mu \quad \text{as} \quad n \to \infty.$$

Results

- **Central Limit Theorem.** Let $\sigma^2 = \mathrm{Var}(X_i)$ be finite. Then as $n \to \infty$, the distribution of $\sqrt{n}(\bar{X}_n - \mu)$ converges to the $\mathcal{N}(0, \sigma^2)$ distribution.

- **Corollary 1.**

$$\left|\bar{X}_n - \mu\right| \approx 1/\sqrt{n}.$$

  This gives an order of magnitude estimate of the error between the sample mean and the true mean. A consequence of this corollary: if you want your sample mean to approximate the true mean μ to within one part in 10^p (relative error), then you need $n \approx 10^{2p}/\mu^2$ samples.

- **Corollary 2.**

  For large n,

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \sim \mathcal{N}(0, 1), \text{ (approximately)}$$
$$\Rightarrow \mathbf{P}\left(\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq x\right) \approx \Phi(x).$$

  How large should n be to use this result? Rule of thumb: $n > n^*$, where $n^* \approx 30$.

  NB. Really, the value of n^* depends on the distribution. For distributions that "look like" the standard normal distribution, $n^* \approx 5$ may suffice.

- **Corollary 3.**

  For large n,

$$\left(X_1 + \cdots + X_n\right) \sim \mathcal{N}(n\mu, n\sigma^2).$$

  This is a useful but very loosely formulated result.

- **Infinite variance case.** If $\mathrm{Var}(X) \to \infty$, then the CLT *may* still hold, but

$$\left|\bar{X}_n - \mu\right| \approx \frac{1}{n^{1-1/\alpha}} \text{ for some } 1 < \alpha < 2.$$

The closer α gets to 1, the heavier the tails of the distribution, and the less benefit you get from averaging. How to pick out heavy tails on a graph? Look at the ratio between the extremes and the main body of the data.

For large n, $X_1 + \cdots + X_n$ has approximately the *α-stable Stable distribution.*

16 Stochastic Processes

Definitions

- Consider the sequence of random n-vectors

$$\ldots, \mathbf{X}(-1), \mathbf{X}(0), \mathbf{X}(1), \ldots \tag{1}$$

  where the argument denotes (discrete) time. We call such a sequence a **stochastic process**. In general, each $\mathbf{X}(k)$ has a different density $f_{\mathbf{X}(k)}(\mathbf{x}(k))$.

- Let

$$X^j = \{\ldots, \mathbf{X}(-1), \mathbf{X}(0), \mathbf{X}(1), \ldots, \mathbf{X}(j)\}$$

  be the set of random vectors up to and including time j.

  A stochastic process is **Markovian** if

$$f_{\mathbf{X}(k)|X^j}(\mathbf{x}(k)|X^j) = f_{\mathbf{X}(k)|\mathbf{X}(j)}(\mathbf{x}(k)|\mathbf{x}(j)) \quad \text{for all } j < k$$

  i.e. the current state contains all useful information for the purposes of predicting the future.

- The **autocorrelation** of the stochastic process (1) is

$$R(k,j) = \mathbf{E}\left[\mathbf{X}(k)\mathbf{X}(j)^T\right] \in \mathbf{R}^{n\times n}$$

  NB. $R(k,j) = R(j,k)^T$ in general.

- Let $\mathbf{E}\,\mathbf{X}(k) = \boldsymbol{\mu}_x(k)$ for all k. Then the **autocovariance** of the stochastic process (1) is

$$\begin{aligned} V(k,j) &= \mathbf{E}\left[(\mathbf{X}(k) - \boldsymbol{\mu}_x(k))(\mathbf{X}(j) - \boldsymbol{\mu}_x(j))^T\right] \\ &= R(k,j) - \boldsymbol{\mu}_x(k)\boldsymbol{\mu}_x(j)^T \end{aligned}$$

  If $\boldsymbol{\mu}_x(k) = \mathbf{0}$ for all k, then the autocovariance and autocorrelation of (1) are the same.

- The stochastic process (1) is (wide-sense) **stationary** if its first two moments don't change over time. More precisely, a stationary process satisfies

  1. $\boldsymbol{\mu}_x(k) = \boldsymbol{\mu}_x$ for all k
  2. $R(k,j) = \tilde{R}(l)$ for all k, j, where $l = k - j$, for some function $\tilde{R} : \mathbf{R} \to \mathbf{R}^{n\times n}$

  Interpretation of 2: the autocorrelation of a stationary stochastic process may depend on the time lag between two samples, but not on the particular time index of either of them.

  For a stationary stochastic process, $\tilde{R}(l) = \tilde{R}(-l)^T$.

  NB. There's a stricter version of stationarity that's not typically achievable in practice. It requires time invariance of the *pdf*, rather than just the first two moments.

- The **power spectral density** of a *stationary* stochastic process is the Fourier transform of its autocorrelation,

$$S(\omega) = \mathcal{F}\left[\tilde{R}(\tau)\right] = \int_{-\infty}^{\infty} e^{i\omega\tau}\tilde{R}(\tau)\mathrm{d}\tau$$

  where $\tilde{R}(\tau)$ is the autocorrelation between samples separated by time τ.

- The stochastic process (1) is called **white noise** if its autocovariance is zero for any two different sample times:

$$V(k,j) = \mathbf{E}\left[(\mathbf{X}(k) - \boldsymbol{\mu}_x(k))(\mathbf{X}(j) - \boldsymbol{\mu}_x(j))^T\right] = Q(k)\delta_{jk} \text{ for all } j,k$$

  where $Q(k)$ is the covariance matrix of $\mathbf{X}(k)$ and δ_{kj} is the Kronecker delta. If the white noise is zero mean, then the whiteness condition simplifies to

$$V(k,j) = \tilde{R}(k-j) = \mathbf{E}\left[\mathbf{X}(k)\mathbf{X}(j)^T\right] = Q(k)\delta_{jk} \text{ for all } j,k$$

  NB. White noise is stationary iff $\mathbf{E}\,\mathbf{X}(j) = \mathbf{E}\,\mathbf{X}(k)$ for all j,k and $Q(k) = Q$ for all k. The power spectral density of white noise is constant across all frequencies. The Matlab commands *xcorr.m* and *fft.m* are useful for analyzing the cross-correlation and spectrum of a random sequence. *xcorr.m* can compute the autocorrelation of a sequence, and *fft.m* can compute the Fourier transform of the autocorrelation, a.k.a. the power spectral density. If the transform is flat over all frequencies, then the sequence is probably white noise.

- A stationary stochastic process with mean $\mathbf{E}\,\mathbf{X}(k) = \boldsymbol{\mu}_x$ for all k is **ergodic** if expectations agree with simple time averages of the realizations:

$$\lim_{N\to\infty} \frac{1}{2N}\sum_{k=-N}^{N} \mathbf{x}(k) = \boldsymbol{\mu}_x$$

- Suppose the stochastic process $\ldots, \mathbf{V}(-1), \mathbf{V}(0), \mathbf{V}(1), \ldots$ is stationary white noise, i.e.

  1. $\mathbf{E}\,\mathbf{V}(k) = \mathbf{0}$ for all k
  2. $\mathbf{E}\left[\mathbf{V}(k)\mathbf{V}(j)\right] = Q\delta_{jk}$ for all j,k

  Define the random vectors $\mathbf{X}(k)$ as follows:

  1. $\mathbf{X}(0) = \mathbf{0}$ with probability 1
  2. $\mathbf{X}(k+1) = \mathbf{X}(k) + \mathbf{V}(k)$ for $k \in \{0, 1, 2, \ldots\}$

Then the stochastic process $\mathbf{X}(0), \mathbf{X}(1), \ldots$ is called a **random walk**. One interpretation: a drunk man wandering around. At each time step he picks a random direction and wanders in that direction until the next step.

Some properties of random walks:

- each $\mathbf{X}(k)$ can be written as a sum of the past white noise:

$$\mathbf{X}(k) = \sum_{j=0}^{k-1} \mathbf{V}(j) \text{ for all } k \geq 1$$

- every $\mathbf{X}(k)$ is zero mean

$$\mathbf{E}\,\mathbf{X}(k) = \mathbf{0} \text{ for all } k \geq 0$$

- the covariance matrix of $\mathbf{X}(k)$ is

$$\mathbf{E}\,[\mathbf{X}(k)\mathbf{X}(k)^T] = kQ$$

  so a random walk is non-stationary. In fact, the uncertainty in $\mathbf{X}(k)$ grows with time – "walk long enough and you could end up anywhere."

- Let $\ldots, \mathbf{V}(-1), \mathbf{V}(0), \mathbf{V}(1), \ldots$ be Gaussian white noise, and let $\mathbf{X}(0)$ also be Gaussian. Then the stochastic process given by linear system

$$\mathbf{X}(k+1) = F(k)\mathbf{X}(k) + \mathbf{V}(k)$$

  is called a **Gauss-Markov process**. Here $F(0), F(1), \ldots$ are known, non-random $n \times n$ matrices.

  Some properties:

  - every $\mathbf{X}(k)$ is Gaussian, since linear transformations preserve normality
  - the process is Markovian: for all $k \geq j$, $\mathbf{X}(k)$ depends only on $\mathbf{X}(j)$ and $\mathbf{V}(j), \mathbf{V}(j+1), \ldots, \mathbf{V}(k-1)$

  NB. If $\mathbf{X}(0) = \mathbf{0}$ with probability 1, and if $F(k) = I_n$ for all k, then the Gauss-Markov process reduces to a random walk.

Part II
Statistics

17 Numerical Data Summaries

Definitions

- Given an *iid* sample $X_1, \ldots, X_n$, the **median of absolute deviations (MAD)** of the sample is

$$\text{MAD}(X_1, \ldots, X_n) = \text{Med}(|X_1 - \text{Med}(X_1, \ldots, X_n)|, \ldots, |X_n - \text{Med}(X_1, \ldots, X_n)|)$$

  where the **median** $\text{Med}(X_1, \ldots, X_n)$ is the middle order statistic of the arguments.

- The **sample variance** of the sample $X_1, \ldots, X_n$ is

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2.$$

  Why divide by $n-1$ instead of n? Because the factor $n-1$ leads to an unbiased estimator, whereas n does not. See Chapter 19 for more details.

- The **sample standard deviation** of the sample $X_1, \ldots, X_n$ is

$$S_n = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2}.$$

  NB. The sample standard deviation is a biased estimator of the true standard deviation σ.

18 Basic Statistical Models

Definitions

- A **sample** or **random sample** is a collection of *iid* RVs $X_1, \ldots, X_n$. In most of statistics, we view our data, $x_1, \ldots, x_n$ as realizations of a random sample.
- A **statistic** (or **sample statistic**) is any function of the sample $X_1, \ldots, X_n$.

  Examples: sample mean $\bar{X}_n$, sample maximum $M_n = \max\{X_1, \ldots, X_n\}$, $\text{Med}(X_1, \ldots, X_n)$, $\text{MAD}(X_1, \ldots, X_n)$, etc.
- The **empirical distribution function** $F_n(a)$ of the sample $X_1, \ldots, X_n$ is

$$F_n(a) = \frac{\text{number of } X_i \text{ in } (-\infty, a]}{n}.$$

  The LLN tells us that for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbf{P}(|F_n(a) - F(a)| > \epsilon) = 0,$$

  i.e. for large n the empirical distribution function approaches the true *cdf*.

Results

- Let $(x-h, x+h]$ be a bin of width $2h$ in the **histogram** of the sample $X_1, \ldots, X_n$. Then

$$\frac{\text{number of } X_i \text{ in } (x-h, x+h]}{2hn} \approx f_X(x),$$

  i.e. the height of the histogram approximates the value of the *pdf* at the midpoint of the bin, and the approximation gets better as n increases and as the bin width decreases.

  We have a similar result for the **kernel density estimate**. For large n, it approaches the true *pdf*.
- The **relative frequency** of realizations from a discrete, *iid* sample $X_1, \ldots, X_n$ approximates the *pmf*:

$$\text{relative frequency} = \frac{\text{number of } X_i \text{ equal to } a}{n} \approx p_X(a).$$

- Some common estimators of true features of the distribution:

$$\begin{aligned}
\mathbf{E}\, X_i &\approx \bar{X}_n \\
\text{Var}(X_i) &\approx S_n^2 \\
\sigma_{X_i} &\approx S_n \\
F_X^{-1}(0.5) &\approx \text{Med}(X_1, \ldots, X_n) \\
F_X^{-1}(q) &\approx q^{\text{th}} \text{ empirical quantile} \\
F_X^{-1}(0.75) - F_X^{-1}(0.5) &\approx \text{MAD}(X_1, \ldots, X_n) \quad \text{(for symmetric } F_X(x_i)\text{)}.
\end{aligned}$$

19 Unbiased Estimators

In this chapter, we discuss how to use sample statistics to recover information about the true distribution (model parameters, *cdf*, *pdf*, *pmf*, etc).

Definitions

- A **point estimator** $\hat{\Theta} = \hat{\Theta}(X_1, \ldots, X_n)$ is a statistic, suitably designed to estimate population parameter θ. A **point estimate** $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$ is a realization of a point estimator.

- The **bias** of an estimator is defined by

$$\text{Bias}(\hat{\theta}) = \mathbf{E}\left[\hat{\theta}\right] - \theta.$$

  If $\text{Bias}(\hat{\theta}) = 0$, then we call $\hat{\theta}$ **unbiased**. A biased estimator has a systematic tendency to overestimate or underestimate the true parameter θ.

  Note that the naively normalized estimator of the variance,

$$\tilde{S}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2,$$

  is biased. This is why we used the unbiased estimator S_n^2, normalized by $\frac{1}{n-1}$.

- For a vector-valued estimator $\hat{\boldsymbol{\theta}}$ of parameter vector $\boldsymbol{\theta}$, with estimation error $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}$, the bias is

$$\text{Bias}(\hat{\boldsymbol{\theta}}) = \mathbf{E}\,\tilde{\boldsymbol{\theta}}$$

  (Note that the expectation may be taken over different distributions, in the Bayesian and non-Bayesian approaches.)

- An estimator $\hat{\theta}$ is **consistent** if, for large n, $\hat{\theta} \approx \theta$. More precisely, $\hat{\theta}$ is consistent if, for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbf{P}(\left|\hat{\theta} - \theta\right| > \epsilon) = 0,$$

  i.e. $\hat{\theta} \to \theta$ **in probability**.

  Note: if $\hat{\theta}$ is consistent, then as $n \to \infty$, $\text{Bias}(\hat{\theta}) \to 0$.

Results

- Even if $\text{Bias}(\hat{\theta}) = 0$, the transformed estimator $g(\hat{\theta})$ may be a biased estimator of the transformed parameter $g(\theta)$. Example: even though S_n^2 is an unbiased estimator of σ^2, $S_n = \sqrt{S_n^2}$ is a biased estimator of σ. Thus, transformation does not in general preserve unbiasedness. A special case where unbiasedness *is* preserved is under linear transformation:

$$\text{Bias}(\hat{\theta}) = 0 \Rightarrow \text{Bias}(a\hat{\theta} + b) = 0.$$

20 Precision of an Estimator

This chapter investigates how, given a portfolio of estimators for an unknown parameter θ, we can choose the "best" one (by some agreed-upon metric). One such metric is the bias. Others are efficiency, variance, MSE, mean absolute deviation, etc. Various norms can be used to evaluate the precision of an estimator.

Definitions

- Let $\hat{\theta_1}(X_1, \ldots, X_n)$ and $\hat{\theta_2}(X_1, \ldots, X_n)$ be unbiased estimators of θ. Then $\hat{\theta_1}(X_1, \ldots, X_n)$ is more **efficient** than $\hat{\theta_2}(X_1, \ldots, X_n)$ if, for all θ,

$$\text{Var}(\hat{\theta_1}(X_1, \ldots, X_n)) < \text{Var}(\hat{\theta_2}(X_1, \ldots, X_n)).$$

  "An estimator with smaller variance is more efficient."

- The **mean squared error** of an estimator $\hat{\theta}$ of a parameter θ is

$$\begin{aligned}\textbf{MSE}(\hat{\theta}) &= \mathbf{E}\left[(\hat{\theta} - \theta)^2\right] \\ &= \text{Var}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2.\end{aligned}$$

  Let $\hat{\boldsymbol{\theta}} \in \mathbf{R}^n$ be an estimator of the parameter vector $\boldsymbol{\theta} \in \mathbf{R}^n$. Let $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}$ be the estimation error. Then

$$\text{Var}(\hat{\boldsymbol{\theta}}) = \mathbf{E}\left[(\tilde{\boldsymbol{\theta}} - \mathbf{E}\,\tilde{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}} - \mathbf{E}\,\tilde{\boldsymbol{\theta}})^T\right] \in \mathbf{R}^{n \times n}$$

  and

$$\textbf{MSE}(\hat{\boldsymbol{\theta}}) = \mathbf{E}\left[\tilde{\boldsymbol{\theta}}\tilde{\boldsymbol{\theta}}^T\right]$$

  The MSE of an estimator measures its spread around the true parameter value. We say that $\hat{\theta_1}(X_1, \ldots, X_n)$ is better than $\hat{\theta_2}(X_1, \ldots, X_n)$ in the MSE sense if

$$\textbf{MSE}(\hat{\theta_1}(X_1, \ldots, X_n)) < \textbf{MSE}(\hat{\theta_2}(X_1, \ldots, X_n)).$$

  Note that in some cases, a biased estimator with a small MSE is preferable to an unbiased estimator with a larger MSE.

  It is almost never possible to find the estimator that's optimal in the MSE sense, i.e. the one with the smallest possible MSE. We can, however, sometimes find the estimator with the smallest variance.

- The **mean absolute deviation** of an estimator $\hat{\theta}$ of a parameter θ is

$$\text{MAD}(\hat{\theta}) = \mathbf{E}\left|\hat{\theta} - \theta\right|$$

  Why use MAD over MSE? If you have noisy observations or modeling errors, then MAD may be preferable. However, about 95% of the time, MSE is the metric for measuring the precision of an estimator.

- The **minimum variance unbiased estimator** (MVUE) of a parameter θ is the unbiased estimator with the smallest possible variance.

  Example: the sample mean is the MVUE of the true mean, when the sample distribution is normal.

Results

- **The Cramér-Rao bound.** Let sample $X_1, \ldots, X_n$ have *pdf* $f_{X|\Theta}(x_i|\theta)$ (assumed to be smooth), for some parameter θ. Then the variance of any estimator $\hat{\theta}$ of θ is bounded below, for all θ, by

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{n\,\mathbf{E}\left[(\frac{\partial}{\partial\theta}\ln f_{X|\Theta}(X_i|\theta))^2\right]}.$$

21 Maximum Likelihood Estimation

This section presents one method for producing "good" estimators: maximum likelihood estimation. There are other methods, such as Bayesian estimation and estimation by the Method of Moments. ML estimators are typically very good, but can be hard to construct because they require optimization.

Definitions

- The **likelihood function** of a dataset $\mathbf{x} = (x_1, \ldots, x_n)$ sampled from $p_{X|\boldsymbol{\Theta}}(x|\boldsymbol{\theta})$ or $f_{X|\boldsymbol{\Theta}}(x|\boldsymbol{\theta})$ with unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ is

$$L(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p_X(x_i|\boldsymbol{\theta}) \quad \text{(discrete case)}$$
$$L(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f_X(x_i|\boldsymbol{\theta}) \quad \text{(continuous case)}$$

- The **loglikelihood function** of a dataset $\mathbf{x} = (x_1, \ldots, x_n)$ sampled from $pmf\, p_{X|\boldsymbol{\Theta}}(x|\boldsymbol{\theta})$ or $pdf\, f_{X|\boldsymbol{\Theta}}(x|\boldsymbol{\theta})$ with unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ is

$$l(\mathbf{x}|\boldsymbol{\theta}) = \ln(L(\mathbf{x}|\boldsymbol{\theta})).$$

  A value $\boldsymbol{\theta}^*$ maximizes $l(\mathbf{x}|\boldsymbol{\theta})$ if and only if it maximizes $L(\mathbf{x}|\boldsymbol{\theta})$.

  When finding the MLE of θ involves differentiation, the loglikelihood function is often easier to work with than the likelihood function, because the log of the product is the sum of the logs.

- The **maximum likelihood estimator** of a parameter vector $\hat{\boldsymbol{\theta}}$ is

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{ML} &= \arg\max_{\boldsymbol{\theta}} L(\mathbf{x}|\boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{\theta}} l(\mathbf{x}|\boldsymbol{\theta}).\end{aligned}$$

  In general, $\hat{\boldsymbol{\theta}}_{ML}$ may be biased or unbiased. $\hat{\boldsymbol{\theta}}_{ML}$ is always consistent (and therefore asymptotically unbiased, as $n \to \infty$). As $n \to \infty$, $\hat{\boldsymbol{\theta}}_{ML}$ has the lowest possible MSE among unbiased estimators (i.e. $\hat{\boldsymbol{\theta}}_{ML}$ attains the Cramér-Rao bound).

Results

- If $h(\cdot)$ is a one-to-one function, and if $\hat{\boldsymbol{\theta}}_{ML}$ is the MLE of $\hat{\boldsymbol{\theta}}$, then $h(\hat{\boldsymbol{\theta}}_{ML})$ is the MLE of $h(\hat{\boldsymbol{\theta}})$.

22 Method of Moments Estimation

Definitions

- The **k^{th} sample moment** of a sample $x_1, \ldots, x_n$ is

$$\hat{m}_k = \frac{1}{n} \sum_{i=1}^{n} x_i^k, \text{ for } k = 1, 2, \ldots$$

- The **k^{th} population moment** of a population $X_1, \ldots, X_n$ is

$$\mu_k(\boldsymbol{\theta}) = \mathbf{E}\,[X_1^k], \text{ for } k = 1, 2, \ldots$$

  so $\mu_1 = \mathbf{E}\,X_1$ is the mean. In estimation problems, the population moments depend on the unknown parameters $\boldsymbol{\theta}$.

Results

- **Recipe for Method of Moments estimation.** If k parameters need to be estimated,
  1. For each $i \in \{1, \ldots, k\}$, write $\hat{m}_i = \mu_i(\boldsymbol{\theta})$
  2. Solve the resulting system for $\theta_1, \ldots, \theta_k$.
- MoM estimation is easier to perform than ML estimation, because solving an algebraic system is usually easier than solving an optimization problem.
- Tradeoff: MoM estimators, unlike ML estimators, don't always give sensical results. The most common use of MoM estimators is as the initial guess for numerical ML estimation.

23 Bayesian Estimation

Bayesian estimation is another approach to the problem of using data to estimate an unknown parameter from the sample distribution. While MLEs are consistent and asymptotically unbiased but hard to compute, and MoM estimators are easy to compute but sometimes nonsensical, Bayesian estimators have the advantage of allowing us to incorporate our prior beliefs into the process of estimation. Bayesian estimation is somewhat different, on a philosophical level, from ML and MoM estimation. Why? Because in ML and MoM estimation, we view θ as some unknown constant, but in Bayesian estimation we view θ as a RV with its own probability distribution. This worldview is not universally accepted in the statistics community.

As in the previous chapters, our task is to use the realizations $\mathbf{x} = (x_1, \ldots, x_n)$ of the sample $\mathbf{X} = (X_1, \ldots, X_n)$ to compute an estimator $\hat{\Theta}$ (or a particular estimate $\hat{\theta}$) of the parameter θ.

Bayesian estimation works with either discrete or continuous distributions on both the parameter and the sample. Bayesian estimation preserves the nature of the parameter: if the prior distribution on θ is continuous (or discrete), then the posterior distribution on θ is also continuous (or discrete).

Definitions

- The **prior distribution** $p(\theta)$ encodes our a priori beliefs about θ. The prior distribution is an assumption; Bayesian estimation is generally quite sensitive to this a priori assumption. $p(\theta)$ may be discrete or continuous.
- The uniform distribution is called the **non-informative prior**. It's used when we have no prior beliefs about θ.
- The prior distribution $p(\theta) = 1$ is called the **improper prior**. Although it's not a legitimate density, it's sometimes used because it's easy and it works.
- The **posterior distribution** $f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$ (or $p_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$, if $p(\theta)$ is a *pmf*) uses Bayes' Rule to update the distribution on θ based on the data $\mathbf{x}$.

Results

- **Recipe for Bayesian estimation.**
  1. Postulate prior distribution $p(\theta)$
  2. Use Bayes' Rule and the observations $\mathbf{x}$ to compute the posterior distribution on θ:
$$f_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) = \frac{p(\theta)\prod_{i=1}^{n} f_{X|\Theta}(x_i|\theta)}{f_{\mathbf{X}}(\mathbf{x})}$$

where the denominator can be (but doesn't always need to be) computed with the LTP for RVs:

$$f_{\mathbf{X}}(\mathbf{x}) = \int_{-\infty}^{\infty} p(\theta) \prod_{i=1}^{n} f_{X|\Theta}(x_i|\theta) \mathrm{d}\theta$$

3. Construct either the **mean** Bayesian estimator,

$$\hat{\Theta}_{\text{mean}} = \mathbf{E}\left[\Theta|\mathbf{X}\right] = \int_{-\infty}^{\infty} \theta f_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) \mathrm{d}\theta$$

or the **mode** Bayesian estimator,

$$\hat{\Theta}_{\text{mode}} = \arg\max_{\theta} f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$$

Note that if the mode estimator is used, the denominator $f_{\mathbf{X}}(\mathbf{x})$ never needs to be computed, because it's constant with respect to θ and therefore doesn't affect the maximization.

- **Normal sample and prior.** Let the sample distribution be $\mathcal{N}(\theta, \sigma_0^2)$, where θ is the parameter to be estimated and σ_0^2 is known. Let the prior distribution be $\mathcal{N}(\mu, \sigma_0^2)$, where μ and σ^2 are both known. Then both the mean and mode Bayesian estimators of θ are

$$\hat{\theta}_{\text{mode}} = \hat{\theta}_{\text{mean}} = \frac{1}{1 + n\frac{\sigma^2}{\sigma_0^2}}\mu + \frac{n\frac{\sigma^2}{\sigma_0^2}}{1 + n\frac{\sigma^2}{\sigma_0^2}}\bar{X}_n,$$

i.e. the estimator is a weighted sum of our prior beliefs (μ) and the data ($\bar{X}_n$), where the weight on μ reflects how confident we are in our prior beliefs, and the weight on $\bar{X}_n$ reflects how much we trust the data.

24 Least Squares Estimation

Sometimes we have a bunch of data $\mathbf{x} = (x_1, \ldots, x_n)$, but we don't know what distribution it's sampled from. We only know (or assume) that the data are linear functions of θ:

$$x_j = \gamma_j \theta + w_j, \quad j \in \{1, \ldots, n\}$$

for some scalars $\gamma_1, \ldots, \gamma_n$ and some sequence of independent noises $W_1, \ldots, W_n$ with unknown distributions but known variances $\mathrm{Var}(W_k) = \sigma^2$ for all k. (This is the *linear* least squares problem; there's a nonlinear version too.) Define the cost

$$C_{\mathrm{LS}}(\theta|\mathbf{x}) = \frac{1}{2\sigma^2} \sum_{j=1}^{k} w_j^2 = \frac{1}{2\sigma^2} \sum_{j=1}^{k} (x_j - \gamma_j \theta)^2$$

Then the (linear) least squares estimator (LSE) of θ is

$$\hat{\theta}_{\mathrm{LS}}(\mathbf{x}) = \arg\min_{\theta} C_{\mathrm{LS}}(\theta|\mathbf{x}) = \frac{\sum_{j=1}^{n} \gamma_j x_j}{\sum_{j=1}^{k} \gamma_j^2}$$

Linear least squares estimation finds the θ that minimizes the square of the distance between each measurement x_j and a line through the data.

Under the assumptions that the noises w_j are *iid* Gaussian RVs, the LSE and MLE are equivalent.

25 Minimum Mean Squared Error Estimation

This is another estimator that follows the Bayesian worldview of considering the parameter θ as **random** with a prior density $p(\theta)$. Let $\mathbf{x} = (x_1, \ldots, x_n)$ be the data. Define the cost

$$C_{\text{MSE}}(\hat{\theta}|\mathbf{x}) = \mathbf{E}\left[(\hat{\Theta} - \Theta)^2|\mathbf{x}\right] = \int_{-\infty}^{\infty} (\hat{\theta} - \theta)^2 f_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) \mathrm{d}x$$

Note that the posterior density $f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$ is calculated by the same process as in Bayesian estimation.

The minimum mean squared error estimator (MMSEE) of θ is

$$\hat{\theta}_{\text{MMSE}} = \arg\min_{\hat{\theta}} C(\hat{\theta}|\mathbf{x}) = \mathbf{E}\left[\Theta|\mathbf{x}\right]$$

26 Hypothesis Testing

Hypothesis testing, like estimation, deals with using data to figure stuff out about an unknown parameter θ of the sample distribution. In estimation, we try to come up with a "good" value of θ (through ML, MoM, or Bayesian estimation). In hypothesis testing, we want to choose between two different statements about θ. For instance, $\theta = 5$ vs. $\theta = 10$, or $0 \leq \theta \leq 7$ vs. $\theta > 10$. This chapter discusses how to construct good hypothesis tests, and how to compare the "quality" (mostly *size* and *power*) of two tests.

As in the previous sections, $\mathbf{X} = (X_1, \ldots, X_n)$ is a sample (and therefore *iid*) whose distribution depends on the parameter θ.

Definitions

- The **null hypothesis**, $H_0 : \theta \in \Omega_0$, models our current beliefs about θ.
- The **alternative hypothesis**, $H_1 : \theta \in \Omega_1$, models some different beliefs about θ. Generally, hypothesis testing is *conservative*, in that we require very strong evidence to reject H_0 in favor of H_1. Our inclination is to stick with the status quo.
- A hypothesis with $\Omega_i = \{\theta_i\}$, where $\theta_i \in \mathbf{R}$, is called **simple**. A **composite** hypothesis is one that's not simple (i.e. Ω_i contains more than one element).
- We call the set of all possible parameter values Ω. Ω_0 and Ω_1 are disjoint subsets of Ω.
- A **test statistic** $T(\mathbf{X})$ is some function of the sample that we use to check whether or not we reject H_0.
- The **critical region** C is the set of all values that cause us to reject H_0. Given a test statistic,

$$\begin{aligned} C &= \{\mathbf{x} \mid \text{if } \mathbf{X} = \mathbf{x}, \text{ we reject } H_0\} \\ &= \{\mathbf{x} \mid T(\mathbf{x}) \in R\} \end{aligned}$$

  where the set R, and therefore the critical region C, is constructed "somehow" by the tester. The quality of a test depends on the choice of C.

  We call the rule "reject H_0 if $\mathbf{x} \in C$" $\iff$ "reject H_0 if $T(\mathbf{x}) \in R$" the **decision rule**.

- A **Type 1 error** is a false rejection of H_0. Formally, the probability of Type 1 error is

$$\alpha = \mathbf{P}(\text{reject } H_0 \mid H_0 \text{ is true})$$

  If the null hypothesis is simple, then α is a constant. If the null hypothesis is composite, then $\alpha = \alpha(\theta)$.

- A **Type 2 error** is a false acceptance of H_0. Formally, the probability of Type 2 error is

$$\beta = \mathbf{P}(\text{accept } H_0 \mid H_1 \text{ is true})$$

If the alternative hypothesis is simple, then β is a constant. If the alternative hypothesis is composite, then $\beta = \beta(\theta)$.

- Since we're conservative, controlling Type 1 error (i.e. making α small) is most important.

- The **significance level** or **size** of a test, denoted $\bar{\alpha}$, is an upper bound on α. The significance level is a *specification* set by the tester: "This test must have a significance level no higher than $\bar{\alpha}$." Typical values of $\bar{\alpha}$ are $0.05, 0.01$, or 0.005.

  For a simple null hypothesis, we choose C such that $\alpha = \bar{\alpha}$.

  For a composite null hypothesis, we choose C such that $\alpha(\theta) \leq \bar{\alpha}$ for all $\theta \in \Omega_0$.

- A test with a significance level $\bar{\alpha}$ is called a **size $\bar{\alpha}$ test**.

- The **power function** of a test is the probability of rejecting the null hypothesis, given that the alternative hypothesis is true. Formally,

$$\begin{aligned}\eta(\theta) &= \mathbf{P}(\text{reject } H_0 \mid H_1 \text{ is true}) \\ &= 1 - \beta(\theta)\end{aligned}$$

  When considering two tests of size $\bar{\alpha}$, the test with the higher power is better. If the power of a size $\bar{\alpha}$ test is at least as high as the power of any other size $\bar{\alpha}$ test, then it's called a **best test of size $\bar{\alpha}$**.

- Consider the test of size $\bar{\alpha}$, $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$ (i.e. simple hypotheses), with critical region $C = \{\mathbf{x} \mid T(\mathbf{x}) \geq \gamma_{\bar{\alpha}}\}$ such that $\mathbf{P}(T(\mathbf{x}) \geq \gamma_{\bar{\alpha}} \mid H_0 \text{ is true}) = \bar{\alpha}$. Then the ***p*-value** of this test is

$$p\text{-value} = \mathbf{P}(T(\mathbf{X}) \geq T(\mathbf{x}) \mid H_0 \text{ is true})$$

  Why do we care about the p-value? Well, we can use it to make our decision: we reject the null hypothesis if the p-value is smaller than $\bar{\alpha}$, and we accept the null hypothesis if the p-value exceeds $\bar{\alpha}$.

Results

- **Recipe for likelihood ratio testing with simple hypotheses and continuous likelihood ratios.**

  1. Compute the likelihood under H_0:

$$L_0(\mathbf{x}) = \prod_{i=1}^{n} f_{X|\Theta_0}(x_i|\theta_0)$$

2. Compute the likelihood under H_1:

$$L_1(\mathbf{x}) = \prod_{i=1}^{n} f_{X|\Theta_1}(x_i|\theta_1)$$

3. Compute the likelihood ratio $L_0(\mathbf{x})/L_1(\mathbf{x})$.
4. Apply the Neyman-Pearson Lemma: the best test of size $\bar{\alpha}$ has critical region

$$C = \{\mathbf{x} \mid \frac{L_0(\mathbf{x})}{L_1(\mathbf{x})} \leq \gamma_{\bar{\alpha}}\}$$

Thus, we reject the null hypothesis when the likelihood ratio is too small. Often, we can reformulate the critical region in terms of a test statistic $T(\mathbf{x})$ that's easier to work with than $L_0(\mathbf{x})/L_1(\mathbf{x})$.

- **Recipe for likelihood ratio testing with simple hypotheses and discrete likelihood ratios.** If $L_0(\mathbf{x})/L_1(\mathbf{x})$ is a discrete RV, then the best test of size $\bar{\alpha}$ is **randomized**.

1. Compute the likelihood under H_0:

$$L_0(\mathbf{x}) = \prod_{i=1}^{n} p_{X|\Theta_0}(x_i|\theta_0)$$

2. Compute the likelihood under H_1:

$$L_1(\mathbf{x}) = \prod_{i=1}^{n} p_{X|\Theta_1}(x_i|\theta_1)$$

3. Compute the likelihood ratio $L_0(\mathbf{x})/L_1(\mathbf{x})$.
4. Reject H_0 if the LR is too small.

Note that in this case, the LR takes only discrete values, e.g. 1, 2, 3, 4, and 5. The "LR too small" criterion may tell us to reject H_0 if the LR is 1 or 2, to stick with H_0 if the LR is 4 or 5, but the case $LR = 3$ may be "degenerate" in the sense that if we assign $LR = 3$ to one hypothesis or the other, we end up with a test of size $\alpha \neq \bar{\alpha}$. This is no longer the best test of size $\bar{\alpha}$, so we use a **randomization**.

How does this work? We introduce an intermediate RV X (e.g. a Ber(p) RV), and split the $LR = 3$ case into two subcases: $LR = 3$ and $X = 0$ vs. $LR = 3$ and $X = 1$. Then we figure out what value of p to use to get $\alpha = \bar{\alpha}$, i.e. to exactly meet the spec.

- **Recipe for likelihood ratio testing with composite hypotheses.** When testing $H_0 : \theta \in \Omega_0$ vs. $H_1 : \theta \in \Omega_1$, best tests are not generally available. However, likelihood ratio tests are still very powerful.

1. Compute the maximum likelihood under H_0:

$$L_0^*(\mathbf{x}) = \max_{\theta \in \Omega_0} \prod_{i=1}^{n} f_{X|\Theta_0}(x_i|\theta_0)$$

2. Compute the maximum likelihood under H_0 *or* H_1:

$$L^*(\mathbf{x}) = \max_{\theta \in \Omega_0 \cup \Omega_1} \prod_{i=1}^{n} f_{X|\Theta_1}(x_i|\theta_1)$$

3. Compute the maximum likelihood ratio $L_0^*(\mathbf{x})/L^*(\mathbf{x})$
4. Reject H_0 if $L_0^*(\mathbf{x})/L^*(\mathbf{x}) \leq \gamma_{\bar{\alpha}}$ for an appropriate $\gamma_{\bar{\alpha}}$.

In many cases, the threshold $\gamma_{\bar{\alpha}}$ is difficult or impossible to compute exactly. In these cases, we can use the following approximation

$$n \gg 1 \quad \Rightarrow \quad -2\ln(\frac{L_0^*(\mathbf{x})}{L^*(\mathbf{x})}) \sim \chi_m^2 \text{ (approximately), where}$$
$$m = \dim(\Omega_0 \cup \Omega_1) - \dim(\Omega_0)$$

Our null rejection criterion therefore becomes

$$\text{reject } H_0 \text{ if } -2\ln(\frac{L_0^*(\mathbf{x})}{L^*(\mathbf{x})}) > \chi_m^2(\bar{\alpha})$$

- **Locally most powerful tests.** Given the hypothesis test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta > \theta_0$, the locally most powerful test of size α is

$$\text{reject } H_0 \text{ if } \Big[\frac{\partial}{\partial\theta} \ln f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta)\Big]\Big|_{\theta_0} > \gamma_\alpha$$

where the threshold γ_α satisfies

$$\mathbf{P}\left(\Big[\frac{\partial}{\partial\theta} \ln f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta)\Big]\Big|_{\theta_0} > \gamma_\alpha \;\Big|\; \theta = \theta_0\right) = \alpha$$

LMP tests perform very well in the neighborhood of the null hypothesis, i.e. for values of θ near θ_0.