Intro to Probability

Spring 2018

Contents

1 Random Variables

Definition 1.1. A probability model is a mathematical description of an uncertain situation. It is composed of

a) a sample space Ω which is the set of all possible outcomes. A subset of Ω is called an event, and the set of all possible events is denoted by $\mathcal{F}$.

b) a probability measure $P : \mathcal{F} \to \mathbb{R}$ satisfying
   - $P(A) \geq 0$ for all $A \in \mathcal{F}$;
   - $P(\Omega) = 1$ and $P(\varnothing) = 0$; and
   - if $A_1, A_2, \ldots$ is a sequence of disjoint events, then $P(\cup A_i) = \sum P(A_i)$.

The triple $(\Omega, \mathcal{F}, P)$ is called a **probability space**.

Theorem 1.2. For any events A, B,

- $P(A^C) = 1 - P(A)$.
- If $A \subseteq B$, then $P(A) \leq P(B)$.
- $P(A \cup B) \leq P(A) + P(B)$.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Definition 1.3. A **random variable** is a function $X : \Omega \to \mathbb{R}$. We say X is a discrete random variable if the range of X is countable, otherwise, X is a continuous random variable. The **probability distribution** of a random variable X defines

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

In particular, if X is a discrete, then the p.m.f. of X, denoted p_X, is defined by $p_X(k) = P(X = k)$ for all $k \in \text{range}(X)$.

Proposition 1.4. The probability distribution of a discrete random variable is completely determined by its p.m.f.

2 Conditional Probability

Definition 2.1 (Conditional Probability). Let A, B be events with $P(B) \neq 0$, then

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

Proposition 2.2. If Ω has finitely many equally likely outcomes, then $P(A|B) = \frac{\#(A \cap B)}{\#B}$.

Proposition 2.3. For events $A_1, \ldots, A_n$ having nonzero probability,

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

Theorem 2.4 (Law of Total Probability). Let $B_1, B_2, \ldots$ be a sequence of events that partitions Ω. Then for any event A, we have $A \cap B_1, A \cap B_2, \ldots$ are disjoint, $A = \cup_i (A \cap B_i)$, and

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

2.1 Bayes' Formula

Theorem 2.5. For events A, B,

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

If $A_1, A_2, \ldots,$ is a sequence of events that partitions Ω, then

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_i P(B|A_i)P(A_i)}.$$

Definition 2.6. Events $A_1, \ldots, A_n$ are **independent** if

$$P\left(\bigcap_i A_i\right) = \prod_{i \in S} P(A_i),$$

for all $S \subseteq \{1, \ldots, n\}$. The infinite set of events $A_1, A_2, \ldots$ is independent if any finite subset is independent.

Theorem 2.7. The following are equivalent:

- A and B are independent;
- A^C and B are independent;
- A and B^C are independent;
- A^C and B^C are independent.

Definition 2.8. The random variables $X_1, \ldots, X_n$ are **independent** if

$$P(X_1 \in B_1, \ldots, X_n \in B_n) = \prod_{i=1}^{n} P(X_i \in B_i),$$

for all $B_i \subset R$.

- If $X_1, \ldots, X_n$ are discrete, then they are independent if

$$P(X_1 = c_1, \ldots, X_n = c_n) = \prod_{i=1}^{n} P(X_i = c_i),$$

for all $c_i \in R$.

- If $X_1, \ldots, X_n$ are continuous, then they are independent if

$$P(X_1 \leq c_1, \ldots, X_n \leq c_n) = \prod_{i=1}^{n} P(X_i \leq c_i),$$

for all $c_i \in R$.

3 Independent Trials

Definition 3.1 (Bernoilli Distribution). Let $0 \leq p \leq 1$. A random variable X has the Bernoulli distribution with success parameter p if X is $\{0, 1\}$-valued and $P(X = 1) = p$. We write $X \sim Ber(p)$.

Definition 3.2 (Binomial Distribution). Let $0 \leq p \leq 1$. A random variable X has the binomial distribution with parameters (n, p) if $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$. We write $X \sim Bin(n, p)$.

Models the number of successes in a sample of size n drawn with replacement with success probability p.

Definition 3.3 (Geometric Distribution). Let $0 < p < 1$. A random variable X has the geometric distribution with success parameter p if $P(X = k) = p(1-p)^{k-1}$. We write $X \sim Geom(n, p)$.

Gives the probability distribution of the number X of Bernoulli trials needed to get one success.

Definition 3.4 (Hypergeometric Distribution). A random variable X has the hypergeometric distribution with parameters (N, K, n) if $P(X = k) = \binom{K}{k}\binom{N-k}{k}/\binom{N}{n}$. We write $X \sim Hypergeo(n, p)$.

Describes the probability of k successes in n draws, without replacement, from a finite population of size N that contains exactly K objects with a desired feature, wherein each draw is either a success or a failure.

Definition 3.5 (Cumulative Distribution Function). Let X be a random variable. The cdf of X is a function F_X defined by

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

Theorem 3.6.

1) F is non-decreasing
2) F is right-continuous
3) $\lim_{t \to -\infty} F(t) = 0$ and $\lim_{t \to \infty} F(t) = 1$.

If X is a discrete random variable with pmf p_X, then

$$F_X(t) = \sum_{\substack{x \in X(\Omega) \\ x \leqslant t}} p_X(x),$$

and if X is continuous with pdf f_x, then

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

Given a finite interval $[c, d]$. Let X be a random variable with pdf $f(x) = \frac{1}{d-c}$ if $x \in [c, d]$, and 0 otherwise. Then $X \sim Unif[c, d]$.

4 Expectation

For a discrete random variable X with pmf p, the **expectation** of X is

$$\mathbb{E}(X) = \sum kp(k),$$

where k ranges over $X(\Omega)$. If X is a continuous random variable with pdf f, then the expectation of X is

$$\mathbb{E}(X) = \int_R xf(x)\,\mathrm{d}x.$$

Examples.

- If $X \sim Ber(p)$, then $\mathbb{E}(X) = p$.
- If $X \sim Bin(n, p)$, then $\mathbb{E}(X) = np$.
- If $X \sim Unif[a, b]$, then $\mathbb{E}(X) = \frac{1}{2}(a + b)$.

Law of the Unconscious Statistician. Given a function $g : R \to R$, $\mathbb{E}(g(X)) = \sum g(k)p(X = k)$, if X is a discrete rv, or $\mathbb{E}(g(X)) = \int_R g(x)f(x)\mathrm{d}x$, if X is a continuous rv.

Let X be a random variable with expectation $\mu = E(X)$. Then the **variance** of X is

$$\operatorname{Var}(X) = \mathbb{E}((X - \mu)^2).$$

Proposition 4.1. $\operatorname{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$.

Example. If $X \sim Unif[a, b]$, then $\operatorname{Var}(X) = \frac{1}{12}(b - a)^2$.

Proposition 4.2. Let X, Y be random variables and let $a, b \in R$.

1) $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$.
2) $\operatorname{Var}(aX + b) = a^2 \operatorname{Var}(X)$.
3) $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.
4) $\operatorname{Var}(X + Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X, Y)$,

where $\operatorname{Cov}(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))$, the covariance of X and Y.

Distribution	Expectation	Variance	$M_X(t)$
$Ber(p)$	p	$p(1-p)$	$pe^t + 1 - p$
$Bin(n,p)$	np	$np(1-p)$	$(pe^t + 1 - p)^n$
$Geom(p)$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^t}{1-(1-p)e^t}$, $t < -\ln(1-p)$
$Unif[a,b]$	$\frac{1}{2}(a+b)$	$\frac{1}{12}(b-a)^2$	$\frac{1}{t(b-a)}(e^{bt} - e^{at})$, $t \neq 0$
$Poisson(\lambda)$	λ	λ	$e^{\lambda(e^t-1)}$
$Exp(\lambda)$	$1/\lambda$	$1/\lambda^2$	$\frac{\lambda}{\lambda-t}$, $t < \lambda$

4.1 Gaussian Distribution

Definition 4.3. Let $\mu \in \mathbb{R}$ and $\sigma^2 \geq 0$. A random variable X with pdf

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2},$$

has the **Gaussian (normal) distribution** with parameters μ and σ^2. We write $X \sim \mathcal{N}(\mu, \sigma^2)$. If $Z \sim \mathcal{N}(0,1)$, then we say that Z has the **standard normal distribution**. We denote its pdf and cdf by ϕ and Φ, respectively.

Theorem 4.4. Let $X = \sigma Z + \mu$. Then, $X \sim \mathcal{N}(\mu, \sigma^2)$ if and only if $Z \sim \mathcal{N}(0,1)$.

Proof.

$$\begin{aligned}
P(Z \leq t) &= P(X \leq \sigma t - \mu) \\
&= \int_{-\infty}^{\sigma t + \mu} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \, \mathrm{d}x \\
&= \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-y^2/2} \sigma \, \mathrm{d}y \qquad (y = \frac{x-\mu}{\sigma}) \\
&= \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \, \mathrm{d}y.
\end{aligned}$$

∎

Let $Z \sim \mathcal{N}(0,1)$ and $X = \sigma Z + \mu$. Then $\mathbb{E}Z = 0$, $\mathrm{Var}(Z) = 1$ and $\mathbb{E}X = \mu$, $\mathrm{Var}(X) = \sigma^2$.

4.2 Binomial Approximation

Let $S_n \sim Bin(n,p)$. Recall that $\mathbb{E}S_n = np$ and $\mathrm{Var}(S_n) = np(1-p)$. Thus, $\mathbb{E}\left(\frac{S_n - np}{\sqrt{np(1-p)}}\right) = 0$ and $\mathrm{Var}\left(\frac{S_n - np}{\sqrt{np(1-p)}}\right) = 1$. When $np(1-p)$ is sufficiently large (at least > 10), we have

$$P\left(\frac{S_n - np}{\sqrt{np(1-p)}} \leq t\right) \approx \Phi(t).$$

In particular, we have

$$\left| P\left(\frac{S_n - np}{\sqrt{np(1-p)}} \le t\right) - \Phi(t) \right| \le \frac{3}{\sqrt{np(1-p)}}. \tag{4.1}$$

4.2.1 Continuity Correction

Example. Let $S_n \sim Bin(720, \frac{1}{6})$. Suppose we want to estimate $P(S_n = 113)$. We use a continuity correction to allow us to approximate the probability by normalization.

$$\begin{aligned} P(S_n = 113) &= P(112.5 \le S_n \le 113.5) \\ &\approx P(-0.75 \le Z \le -0.65) \\ &= \Phi(0.75) - \Phi(0.65) = 0.312... \end{aligned}$$

4.3 Confidence Intervals

Example. Suppose we have a possibly biased coin with $P(h) = p$. Let $S_n = \#$ of heads $\sim Bin(n, p)$. We estimate p by $\hat{p} = S_n/n$.

Fact. Let $\varepsilon > 0$. Then

$$P(|p - \hat{p}| < \varepsilon) \ge 2\Phi(2\sqrt{n}\varepsilon) - 1.$$

1) How many times should we flip the coin such that $\hat{p}$ is within 0.05 of p with probability at least 0.99? Using the fact, we have $P(|p - \hat{p}| < 0.05) \ge 0.99 \leftrightarrow \Phi(0.10\sqrt{n}) \ge 0.995 \leftrightarrow n \ge 665.64$. So we should flip the coin 666 times.

2) Find the smallest interval around $\hat{p}$ that contains p with probability 0.95. Again, applying the fact, we have $P(|p - \hat{p}| < \varepsilon) \ge 0.95 \leftrightarrow \varepsilon \ge \frac{0.98}{\sqrt{n}}$. So the smallest such interval is $(\hat{p} - \varepsilon_0, \hat{p} + \varepsilon_0)$, where $\epsilon_0 = \frac{0.98}{\sqrt{n}}$. This interval is called a 95% confidence interval for p.

5 Poisson Distribution

Example. A computer server receives 3 request/sec on average. Estimate the probability the server receives 5 requests in any given second.

For example, we can divide the interval into 10 seconds and define $S_{10} = \#$ of requests $\sim Bin(10, 3/10)$. Then $P(S_{10} = 5) = 0.1029$. In general, for $S_n \sim Bin(n, p)$ where $\mathbb{E}S_n = np = \lambda$ is a constant, we have

$$\lim_{n\to\infty} P(S_n = k) = \frac{e^{-\lambda}\lambda^k}{k!}.$$

Proof. Substitute $p = \frac{\lambda}{n}$. Factor out constants and use $(1 + \frac{x}{n})^n \to e^x$ as $n \to \infty$. Show the rest reduces to 1. ∎

Definition 5.1 (Poisson Distribution). Let $\lambda > 0$. The random variable Y has Poisson distribution if

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

Write $Y \sim \text{Poisson}(\lambda)$. We may approximate a binomial distribution with the Poisson distribution ($\lambda = np$) when $np^2 < 1$. We have the following

$$\left|P(S_n = k) - P(Y = k)\right| \leq np^2.$$

5.1 Exponential Distribution

Definition 5.2 (Exponential Distribution). Let $\lambda > 0$. A random variable X with p.d.f.

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

has the exponential distribution. We write $X \sim Exp(\lambda)$.

The exponential distribution has c.d.f.

$$F(t) = 1 - e^{-\lambda t} \text{ for } t \geq 0.$$

Proposition 5.3. If $X \sim Exp(\lambda)$ then for any $s, t > 0$

$$P(X > s + t | X > t) = P(X > s).$$

Remark. If T_n is a random variable such that $nT_n \sim Geom(\frac{\lambda}{n})$, then $\lim_{n\to\infty} P(T_n < t) = 1 - e^{-\lambda t}$.

6 Moment Generating Function

Definition 6.1. The **moment generating function** of a random variable X is

$$M_X(t) = \mathbb{E}(e^{Xt}).$$

Examples.

- $X \sim Ber(p)$. Then $M_X(t) = pe^t + 1 - p$.
- $X \sim Poisson(\lambda)$. Then $M_X(t) = e^{\lambda(e^t - 1)}$.
- $X \sim \mathcal{N}(0, 1)$. Then $M_X(t) = e^{t^2/2}$.
- $X \sim Exp(\lambda)$. Then $M_X(t) = \begin{cases} \frac{\lambda}{\lambda - t} & t < \lambda \\ \infty & t \geq \lambda \end{cases}$.

The n^{th} **moment** of a random variable X is $\mathbb{E}X^n$. Assuming the m.g.f. $M_X(t)$ is well-behaved around the origin, we have

$$M_X^{(n)}(0) = \mathbb{E}X^n.$$

7 Joint Distributions

Discrete Case. The **joint pdf** of discrete random variables X and Y is

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

Continuous Case. If X and Y are continuous random variables and $f : \mathbb{R}^2 \to \mathbb{R}$ is a function such that

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_c^d \int_a^b f(x,y)\ \mathrm{d}x\mathrm{d}y$$

for all $a, b, c, d \in \mathbb{R}$, then X and Y are jointly continuous and f is the **joint pmf** of X and Y.

We can recover the **marginal pmf/pdfs** of X and Y. For example

$$p_x(x) = \sum_{y \in \text{ range}(Y)} p_{X,Y}(x,y).$$

If $X_1, \ldots, X_n$ are random variables, then $(X_1, \ldots, X_n)$ is called a **random vector**. The random vector $(X_1, \ldots, X_n)$ has the **multinomial distribution** with parameters $n, r, p_1, \ldots, p_r$ with $p_1 + \ldots + p_r = 1$ if the joint pmf is

$$P(X_1 = k_1, \ldots, X_r = k_r) = \binom{n}{k_1, \ldots, k_r} p_1^{k_1} \ldots p_r^{k_r}$$

for non-negative integers $k_1, \ldots, k_r$ with $k_1 + \ldots + k_r = n$. Write $(X_1, \ldots, X_n) \sim \mathit{Multi}(n, r, p_1, \ldots, p_r)$. The motivation for this distribution is an experiment with n trials and r possible outcomes per trial.

7.1 Independence

Discrete Case. $X_1, \ldots, X_n$ are independent if and only if $p_{X_1,\ldots,X_n}(k_1, \ldots, k_n) = p_{X_1}(k_1) \ldots p_{X_n}(k_n)$.

Continuous Case. $X_1, \ldots, X_n$ are independent if and only if $f_{X_1,\ldots,X_n}(k_1, \ldots, k_n) = f_{X_1}(k_1) \ldots f_{X_n}(k_n)$.

Expectation of a Function of Two RVs.

$$\mathbb{E}(g(X,Y)) = \begin{cases} \sum_{k \in R(X)} \sum_{l \in R(Y)} g(k,l) p_{X,Y}(k,l) \\ \iint_{\mathbb{R}^2} g(x,y) f_{X,Y}(x,y)\ \mathrm{d}x\mathrm{d}y. \end{cases} \tag{7.1}$$

Indicator Random variables. Given an event A, the indicator random variable of A is

$$I_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \omega \notin A \end{cases}$$

Note $\mathbb{E}[I_A] = P(A)$ and $I_A \sim \mathit{Ber}(P(A))$.

> **Example.** Suppose we draw 5 cards from a standard deck. Let X be the number of aces. Label the aces $i = 1, \ldots, 4$. Let A_i be the event that ace i is drawn. Then $X = X_{A_1} + \ldots + X_{A_4}$. We can now easily compute $\mathbb{E}[X] = 4P(A_i) = \frac{5}{13}$.

Note. Know how to prove linearity of independence for continuous and discrete cases of 2 or 3 random variables.

7.2 Covariance

Recall that for random variables X and Y,

$$\operatorname{Var}(X) = \mathbb{E}[(X - \mu_X)^2]$$

and

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

We have the following:

1) $\operatorname{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$
2) $\operatorname{Cov}(aX + b, Y) = a \operatorname{Cov}(X, Y) = \operatorname{Cov}(X, aY + b)$
3) For random variables X_i, Y_j, $\operatorname{Cov}(\sum_i^m X_i, \sum_j^n Y_j) = \sum_{i,j} \operatorname{Cov}(X_i, Y_j)$
4) $\operatorname{Cov}(\sum_i^m (a_i X_i + c_i), \sum_j^n (b_j Y_j + d_j)) = \sum_{i,j} a_i b_j \operatorname{Cov}(X_i, Y_j)$
5) $\operatorname{Var}(X + Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2 \operatorname{Cov}(X, Y)$.

7.3 Correlation

We say $X_1, \ldots, X_n$ are **uncorrelated** if $\operatorname{Cov}(X_i, X_j) = 0$ whenever $i \neq j$. If $X_1, \ldots, X_n$ are uncorrelated, then $\operatorname{Var}(\sum_i X_i) = \sum_i(\operatorname{Var}(X_i))$.

Definition 7.1 (Correlation).

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

Proposition 7.2.

1) $-1 \leq \operatorname{Corr}(X, Y) \leq 1$
2) $\operatorname{Corr}(X, Y) = 1$ if and only if $Y = aX + b$ for $a > 0$ and $b \in \mathbb{R}$.
3) $\operatorname{Corr}(X, Y) = -1$ if and only if $Y = aX + b$ for $a < 0$ and $b \in \mathbb{R}$.

7.4 Independence Revisited

Proposition 7.3. If $X_1, \ldots, X_n$ are independent, then $\mathbb{E}(\prod X_i) = \prod \mathbb{E}(X_i)$.

Corollary. If X and Y are independent, then $\operatorname{Cov}(X, Y) = 0$. Note the converse statement does not necessarily hold.

Proposition 7.4. If X and Y are independent, then $M_{X+Y}(t) = M_X(t) \cdot M_Y(t)$.

Proposition 7.5. X and Y are equal in distribution if and only if $M_X(t) = M_Y(t)$ for all t in some open interval containing 0.

7.5 Convolution

If X and Y are discrete

$$p_{X+Y}(n) = \sum_{k \in R(X)} p_{X,Y}(k, n-k) = \sum_{l \in R(Y)} p_{X,Y}(n-l, l).$$

If X and Y are continuous

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

8 Poisson Process

Let $k \in \mathbb{N}$ and $0 < p < 1$. A random variable X has the **negative binomial distribution** with parameters (k, p) and support $\{k, k+1, \dots\}$ if

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

for all $n \geq k$. We write $X \sim Negbin(k, p)$. This distribution models the number of trials needed for k successes, where the probability of success in every trial is p.

A Poisson process models a sequence of events occurring randomly over a continuous time period starting a time $t = 0$. Let I denote the time interval, e.g. $I = [a, b]$, and $|I|$ denote the length of the interval. Let $N(I)$ be the number of occurrences in interval I.

In a **Poisson process** with intensity rate λ

- $N(I) \sim Poisson(\lambda|I|)$ for any bounded $I \subset [0, \infty]$;
- if $I_1, \dots, I_n$ are disjoint (except possibly at their endpoints), then $N(I_1), \dots, N(I_n)$ are independent.

Note it follows that $\mathbb{E}(N(I)) = \lambda|I|$.

8.1 Waiting Times

Let T_k be the time of the kth occurrence. Define $W_1 = T_1$ and $W_k = T_k - T_{k-1}$ for $k \geq 2$, so that $T_k = W_1 + \ldots + W_k$. Then W_i are i.i.d. exponential random variables with parameter λ. Note, by definition,

$$P(T_k > t) = P(N[0, t] < k).$$

Definition 8.1 (Gamma Distribution). Let $\lambda > 0$ and $k \in \mathbb{N}$. We say $X \sim Gamma(\lambda, k)$ if its pdf is

$$f(t) = \begin{cases} 0, & \text{if } x < 0; \\ \frac{\lambda^k t^{k-1} e^{-\lambda t}}{(k-1)!} & \text{if } x \geq 0. \end{cases}$$

Note this distribution can be generalized such to any real number $k > 0$.

Proposition 8.2. If $X \sim Gamma(\lambda, k)$ then $M_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^k$, when $t < k$. In fact, if $X_1, \ldots, X_n$ are independent and each $X_i \sim Gamma(\lambda, k_i)$, then $X_1 + \ldots + X_n \sim Gamma(\lambda, k)$ where $k = k_1 + \ldots + k_n$.

Therefore, we see that since $W_k \sim Exp(\lambda) \sim Gamma(\lambda, 1)$, so $T_k \sim Gamma(\lambda, k)$.

9 Tail Inequalities

If $X \geq 0$, then $\mathbb{E}[X] \geq 0$ and if $X \geq Y$, then $\mathbb{E}[X] \geq \mathbb{E}[Y]$.

Theorem 9.1 (Markov). If X is a nonnegative random variable and $a > 0$, then

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Proof.

$$\mathbb{E}[X] \geq \int_t^\infty x f(x)\,\mathrm{d}x \geq t \int_t^\infty f(x)\,\mathrm{d}x = tP(X \geq t).$$

∎

Theorem 9.2 (Chebyshev). Let X be a random variable with finite mean μ and finite variance σ^2. Then for any $t > 0$

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2} \qquad \text{and} \qquad P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Proof. By Markov's inequality,

$$P(|X - \mu| \geq t) = P(|X - \mu|^2 \geq t^2) \leq \frac{\sigma^2}{t^2},$$

the second part follows by setting $t = k\sigma$. ∎

Appendix A - Estimators

Let X be a random variable with probability distribution depending on parameter θ. The **maximum likelihood function** $\mathcal{L}(\theta; x)$ is the probability that $X = x$ for parameter θ. The **maximum likelihood estimate** (MLE) is

$$\hat{\theta} = \{\arg\max_{\theta\in\Theta} \mathcal{L}(\theta; x)\}.$$

In practice we often take the logarithm of the likelihood function, called log-likelihood

$$l = \ln \mathcal{L}(\theta; x)$$

or the average log-likelihood

$$\hat{l} = \frac{1}{n} \ln \mathcal{L}(\theta; x).$$

Example. Consider Unif$[a, b]$ with unknown parameters $a < b$. Suppose we want the relative error estimate $\hat{c}$ of $c := b - a$ to satisfy

$$P(|c - \hat{c}| < \varepsilon c) \geq 1 - \delta \text{ for some } \delta \in (0, 1). \tag{9.1}$$

Suppose we sample n times, $X_1, \dots, X_n \sim$ Unif$[0, c]$. Let $x_{(1)}, \dots, x_{(n)}$ denote the order statistics. We have $\mathcal{L}(\theta; x) = \frac{1}{\theta^n}$, and $\frac{\mathrm{d} \ln \mathcal{L}(\theta,x)}{\mathrm{d}\theta} = -\frac{n}{\theta}$. So $\mathcal{L}$ is decreasing for $\theta \geq x_{(n)}$, so $\mathcal{L}$ is maximized at $x_{(n)} = \max\{X_1, \dots, X_n\}$. We have

$$P(|c - \hat{c}| < \epsilon c) = 1 - (1 - \varepsilon)^n,$$

so we require

$$n \geq \frac{\ln \delta}{\ln(1 - \varepsilon)}.$$

Thus, given δ and ε we should sample $\lceil \frac{\ln \delta}{\ln(1-\varepsilon)} \rceil$ times and return the maximum value.

Appendix B - Law of Large Numbers

Assume $X_1, \dots, X_n$ are i.i.d., with finite mean $\mathbb{E}[X_i] = \mu$ and finite variance $\mathrm{Var}(X_i) = \sigma^2$. Let $S_n = X_1 + \dots + X_n$. Then the sample mean $\overline{X_n} = \frac{1}{n} S_n$. Note that $\mathbb{E}[\overline{X_n}] = \mu$ and $\mathrm{Var}(\overline{X_n}) = \frac{\sigma^2}{n}$.

(Weak) Law of Large Numbers. For any $\varepsilon > 0$, we have

$$\lim_{n\to\infty} P\left(|\overline{X_n} - \mu| < \varepsilon\right) = 1.$$

That is, the sample mean converges to the true mean with probability one.

Central Limit Theorem. If $X_1, \ldots, X_n$ are i.i.d. with finite expectation μ and variance σ^2, then

$$\lim_{n\to\infty} P\left(\left|\frac{\overline{X_n} - \mu}{\sigma}\right| \leq t\right) = \Phi(t).$$