Probability & Statistics Review

Dave Goldsman

Updated 8/21/12

1 Getting Started — The Gambler's Ruin

Each time a gambler plays, he wins \$1 with probability p and loses \$1 with probability $1 - p = q$. Each play is independent. Suppose he starts with $\$i$. Find the probability that his fortune will hit $\$N$ (i.e., he breaks the bank) before it hits \$0 (i.e., he is ruined).

Let X_n denote his fortune at time n. It turns out that $X_1, X_2, \ldots$ is a *Markov chain* — a stochastic process where the next state depends only the current state (more on this later).

To get things going, we'll use a common trick — a so-called *one-step analysis*. Let

$$\begin{aligned}
P_i &\equiv \mathsf{Pr}(\text{Eventually hit } \$N | X_0 = i) \\
&= \mathsf{Pr}(\text{Event. hit } N | X_1 = i+1 \textbf{ and } X_0 = i)\, \mathsf{Pr}(X_1 = i+1 | X_0 = i) \\
&\quad + \mathsf{Pr}(\text{Event. hit } N | X_1 = i-1 \textbf{ and } X_0 = i)\, \mathsf{Pr}(X_1 = i-1 | X_0 = i) \\
&= \mathsf{Pr}(\text{Event. hit } N | X_1 = i+1) p + \mathsf{Pr}(\text{Event. hit } N | X_1 = i-1) q \\
&= pP_{i+1} + qP_{i-1}, \quad i = 1, 2, \ldots, N-1.
\end{aligned}$$

Since $p + q = 1$, we have

$$pP_i + qP_i = pP_{i+1} + qP_{i-1}$$

iff

$$p(P_{i+1} - P_i) = q(P_i - P_{i-1})$$

iff

$$P_{i+1} - P_i = \frac{q}{p}(P_i - P_{i-1}), \quad i = 1, 2, \ldots, N-1.$$

Since $P_0 = 0$, we have

$$\begin{aligned}
P_2 - P_1 &= \frac{q}{p} P_1 \\
P_3 - P_2 &= \frac{q}{p}(P_2 - P_1) = \left(\frac{q}{p}\right)^2 P_1 \\
&\vdots \\
P_i - P_{i-1} &= \frac{q}{p}(P_{i-1} - P_{i-2}) = \left(\frac{q}{p}\right)^{i-1} P_1.
\end{aligned}$$

Summing up the LHS terms and the RHS terms,

$$\sum_{j=2}^{i}(P_j - P_{j-1}) \;=\; P_i - P_1 \;=\; \sum_{j=1}^{i-1}\left(\frac{q}{p}\right)^j P_1.$$

This implies that

$$P_i \;=\; P_1\sum_{j=0}^{i-1}\left(\frac{q}{p}\right)^j \;=\; \begin{cases} \frac{1-(q/p)^i}{1-(q/p)}\,P_1 & \text{if } q \neq p \ (p \neq 1/2) \\ \\ iP_1 & \text{if } q = p \ (p = 1/2) \end{cases}.$$

In particular, note that

$$1 \;=\; P_N \;=\; \begin{cases} \frac{1-(q/p)^N}{1-(q/p)}\,P_1 & \text{if } p \neq 1/2 \\ \\ NP_1 & \text{if } p = 1/2 \end{cases}.$$

Thus,

$$P_1 \;=\; \begin{cases} \frac{1-(q/p)}{1-(q/p)^N} & \text{if } p \neq 1/2 \\ \\ 1/N & \text{if } p = 1/2 \end{cases},$$

so that

$$P_i \;=\; \begin{cases} \frac{1-(q/p)^i}{1-(q/p)^N} & \text{if } p \neq 1/2 \\ \\ i/N & \text{if } p = 1/2 \end{cases}. \quad \diamondsuit$$

By the way, as $N \to \infty$,

$$P_i \;\to\; \begin{cases} 1-(q/p)^i & \text{if } p > 1/2 \\ \\ 0 & \text{if } p \leq 1/2 \end{cases}. \quad \diamondsuit$$

Example: A guy can somehow win any blackjack hand w.p. 0.6. If he wins, he fortune increases by \$100; a loss costs him \$100. Suppose he starts out with \$500, and that he'll quit playing as soon as his fortune hits \$0 or \$1500. What's the probability that he'll eventually hit \$1500?

$$P_5 \;=\; \frac{1-(0.4/0.6)^5}{1-(0.4/0.6)^{15}} \;=\; 0.870. \quad \diamondsuit$$

2 Probability Review

2.1 Preliminaries

To start with, I'll assume that you know the following things:

- The set of all possible outcomes of an experiment is the *sample space* Ω.
- Any subset E of a sample space Ω is an *event*.
- The set of all possible events is denoted by $\mathcal{F}$, which is called a *sigma field* of Ω.

  For example, if $\Omega = \{H, T\}$, then $\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$.

  A sigma field must satisfy the following:

  1. $A \in \mathcal{F}$ implies the complement $\bar{A} \in \mathcal{F}$.
  2. $A_1, A_2, \ldots \in \mathcal{F}$ implies $\cup_{j=1}^{\infty} A_j \in \mathcal{F}$.

- The *probability* function $P(\cdot)$ must satisfy 3 axioms:

  1. For any event $E \in \mathcal{F}$, we must have $0 \le P(E) \le 1$
  2. $P(\Omega) = 1$
  3. For any *disjoint* sequence of events $E_1, E_2, \ldots$ (i.e., $E_i \cap E_j = \phi$ if $i \neq j$), we have $P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$.

- A *probability space* is the triple $(\Omega, \mathcal{F}, P)$.

2.2 Conditional Probability and Independence

Definition: If $P(B) > 0$, then $P(A|B) \equiv P(A \cap B)/P(B)$ is the *conditional probability of A given B*.

Example: Toss a fair die. Let $A = \{1, 2, 3\}$ and $B = \{3, 4, 5, 6\}$. Then

$$P(A|B) \;=\; \frac{P(A \cap B)}{P(B)} \;=\; \frac{1/6}{4/6} \;=\; 1/4. \quad \diamondsuit$$

Definition: If $P(A \cap B) = P(A)P(B)$, then A and B are *independent* events.

Theorem: If A and B are independent, then $P(A|B) = P(A)$.

Proof: Easy. $\diamondsuit$

Example: Toss two dice. Let $A =$ "Sum is 7" and $B =$ "First die is 4". Then $P(A) = 1/6$, $P(B) = 1/6$, and $P(A \cap B) = P((4,3)) = 1/36 = P(A)P(B)$; so A and B are independent. $\diamondsuit$

2.3 Random Variables

Definition: A *random variable* (RV) X is a function from Ω to the real line R, i.e., $X : \Omega \to \mathsf{R}$.

Example: Let X be the sum of two dice rolls. Then $X((4,6)) = 10$. In addition,

$$P(X = x) = \begin{cases} 1/36 & \text{if } x = 2 \\ 2/36 & \text{if } x = 3 \\ \vdots & \\ 1/36 & \text{if } x = 12 \\ 0 & \text{otherwise} \end{cases} \qquad \diamondsuit$$

Definition: If the number of possible values of a RV X is finite or countably infinite, then X is a *discrete* RV. Its *probability mass function* (pmf) is $f(x) \equiv P(X = x)$. Note that $\sum_x f(x) = 1$.

Example: Flip 2 coins. Let X be the number of heads.

$$f(x) = \begin{cases} 1/4 & \text{if } x = 0 \text{ or } 2 \\ 1/2 & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases} \qquad \diamondsuit$$

Examples: Here are some well-known discrete RV's that you should review: Bernoulli(p), Binomial(n,p), Geometric(p), Negative Binomial, Poisson(λ), etc.

Definition: A *continuous* RV is one with probability zero at every individual point. A RV is continuous if there exists a *probability density function* (pdf) $f(x)$ such that $P(X \in A) = \int_A f(x)\,dx$ for every set A. Note that $\int_x f(x)\,dx = 1$.

Example: Pick a random number between 3 and 7. Then

$$f(x) = \begin{cases} 1/4 & \text{if } 3 \le x \le 7 \\ 0 & \text{otherwise} \end{cases} \qquad \diamondsuit$$

Examples: Here are some well-known continuous RV's that you should review: Uniform(a, b), Exponential(λ), Normal(μ, σ^2), etc.

Definition: For any RV X (discrete or continuous), the *cumulative distribution function* (cdf) is

$$F(x) \equiv P(X \le x) = \begin{cases} \sum_{y \le x} f(y) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{x} f(y)\, dy & \text{if } X \text{ is continuous} \end{cases}$$

For convenience, we'll henceforth write $F(x) = \int_{-\infty}^{x} dF(y)$ to denote *both* the discrete and continuous cases. Note that $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.

Example: Flip 2 coins. Let X be the number of heads.

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1/4 & \text{if } 0 \le x < 1 \\ 3/4 & \text{if } 1 \le x < 2 \\ 1 & \text{if } x \ge 2 \end{cases} \qquad \diamondsuit$$

Example: Suppose $X \sim \text{Exp}(\lambda)$ (i.e., X has the exponential distribution with parameter λ). Then $f(x) = \lambda e^{-\lambda x}$, $x \ge 0$, and the cdf is $F(x) = 1 - e^{-\lambda x}$, $x \ge 0$. $\diamondsuit$

2.4 Expectation

Definition: The *expected value* (or *mean*) of a RV X is

$$\mathsf{E}[X] \equiv \int_{\mathbb{R}} x\, dF(x) = \begin{cases} \sum_x x P(X = x) & \text{if } X \text{ is discrete} \\ \int_{\mathbb{R}} x f(x)\, dx & \text{if } X \text{ is continuous} \end{cases}$$

Example: Suppose that $X \sim \text{Bernoulli}(p)$. Then

$$X = \begin{cases} 1 & \text{with prob. } p \\ 0 & \text{with prob. } 1 - p \end{cases}$$

and we have $\mathsf{E}[X] = \sum_x x f(x) = p$. $\diamondsuit$

Example: Suppose that $X \sim \text{Uniform}(a, b)$. Then

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}$$

and we have $\mathsf{E}[X] = \int_{\mathbb{R}} x f(x)\, dx = (a + b)/2$. $\diamondsuit$

"Definition" (the "law of the unconscious statistician"): Suppose that $g(X)$ is some function of the RV X. Then $\mathsf{E}[g(X)] \equiv \int_{\mathbb{R}} g(x)\, dF(x)$.

Example: Suppose X is the following discrete RV:

x	2	3	4
$f(x)$	0.3	0.6	0.1

Then $\mathsf{E}[X^3] = \sum_x x^3 f(x) = 8(0.3) + 27(0.6) + 64(0.1) = 25$. $\diamondsuit$

Example: Suppose $X \sim \mathrm{U}(0,2)$. Then $\mathsf{E}[X^n] = \int_{\mathbb{R}} x^n f(x)\, dx = 2^n/(n+1)$. $\diamondsuit$

Definitions: $\mathsf{E}[X^n]$ is the nth *moment* of X. $\mathsf{E}[(X - \mathsf{E}[X])^n]$ is the nth *central moment* of X. $\mathsf{Var}(X) \equiv \mathsf{E}[(X - \mathsf{E}[X])^2]$ is the *variance* of X.

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

Proof: Easy. $\diamondsuit$

Example: Suppose $X \sim \mathrm{Bern}(p)$. Recall that $\mathsf{E}[X] = p$. Further,

$$\mathsf{E}[X^2] = \sum_x x^2 f(x) \;=\; 0^2(1-p) + 1^2 p \;=\; p$$

and

$$\mathsf{Var}(X) \;=\; \mathsf{E}[X^2] - (\mathsf{E}[X])^2 \;=\; p - p^2 \;=\; p(1-p). \quad \diamondsuit$$

Example: Suppose $X \sim \mathrm{U}(0,2)$. By previous examples, $\mathsf{E}[X] = 1$ and $\mathsf{E}[X^2] = 4/3$. So $\mathsf{Var}(X) = \mathsf{E}[X^2] - (\mathsf{E}[X])^2 = 1/3$. $\diamondsuit$

Theorem: $\mathsf{E}[aX + b] = a\mathsf{E}[X] + b$ and $\mathsf{Var}(aX + b) = a^2\mathsf{Var}(X)$.

Proof: Easy. $\diamondsuit$

Definition: The *moment generating function* (mgf) of X is $M_X(t) \equiv \mathsf{E}[e^{tX}]$. For now, we'll assume that this expectation is finite in a neighborhood of $t = 0$.

Example: Suppose $X \sim \mathrm{Bern}(p)$. $M_X(t) = \sum_x e^{tx} f(x) = pe^t + q$. $\diamondsuit$

Example: Suppose $X \sim \mathrm{Exp}(\lambda)$. $M_X(t) = \int_0^\infty e^{tx} \lambda e^{-\lambda x}\, dx = \lambda/(\lambda - t)$, $t < \lambda$. $\diamondsuit$

Theorem (why we call them moment generating functions): Assuming that the mgf exists in a neighborhood around $t = 0$,

$$\mathsf{E}[X^k] \;=\; \frac{d^k}{dt^k} M_X(t)|_{t=0}, \quad k = 1, 2, \ldots .$$

"Proof" We'll just do the first moment. (The others are similar.)

$$\frac{d}{dt} M_X(t) \;=\; \frac{d}{dt}\mathsf{E}[e^{tX}] \text{ ``='' } \mathsf{E}\left[\frac{d}{dt} e^{tX}\right] \text{ ``='' } \mathsf{E}[Xe^{tX}].$$

If you believe the above steps, then

$$\frac{d}{dt} M_X(t)|_{t=0} \;=\; \mathsf{E}[X]. \quad \diamondsuit$$

Theorem: If X and Y have the same mgf, then they have the *same distribution*. Note that there are problems if the mgf doesn't exist around $t = 0$.

Bonus Definition (which is sometimes useful): The *probability generating function* (pgf) of a random variable X is $P_X(s) \equiv \mathsf{E}[s^X]$. It can be shown that

$$\frac{d^k}{ds^k} P_X(s)|_{s=1} \;=\; \mathsf{E}[X(X-1)\cdots(X-k+1)], \; k = 1, 2, \ldots .$$

2.5 Functions of a RV

Problem: Suppose we have a RV X with p.d.f./p.m.f. $f(x)$. Let $Y = h(X)$. Find $g(y)$, the p.d.f./p.m.f. of Y.

Discrete case: If X is discrete, then Y will be discrete, in which case

$$g(y) \;=\; \mathsf{Pr}(Y = y) \;=\; \mathsf{Pr}[h(X) = y] \;=\; \mathsf{Pr}\{x : h(x) = y\} \;=\; \sum_{x:h(x)=y} f(x).$$

Example: Let X denote the number of H's from two coin tosses. We want the p.m.f. for $Y = X^2 - X$.

x	0	1	2
$f(x)$	1/4	1/2	1/4
$y = x^2 - x$	0	0	2

This implies that $g(0) = \mathsf{Pr}(Y = 0) = \mathsf{Pr}(X = 0 \text{ or } 1) = 3/4$ and $g(2) = \mathsf{Pr}(Y = 2) = 1/4$. In other words,

$$g(y) \;=\; \begin{cases} 3/4 & \text{if } y = 0 \\ 1/4 & \text{if } y = 2 \\ 0 & \text{otherwise} \end{cases} \;. \quad \diamondsuit$$

Continuous Case: We'll assume that if X is continuous, then so is Y. The usual method is to first compute the c.d.f. of Y,

$$G(y) \;=\; \mathsf{Pr}(Y \le y) \;=\; \mathsf{Pr}[h(X) \le y] \;=\; \int_{\{x:h(x)\le y\}} f(x)\,dx,$$

and then take the derivative, $g(y) = G'(y)$.

Example: Suppose X has p.d.f. $f(x) = |x|$, $-1 \le x \le 1$. Find the p.d.f. of $Y = X^2$. First of all, the c.d.f. of Y is

$$\begin{aligned} G(y) \;&=\; \mathsf{Pr}(Y \le y) \;=\; \mathsf{Pr}(X^2 \le y) \;=\; \mathsf{Pr}(-\sqrt{y} \le X \le \sqrt{y}) \\ &=\; \int_{-\sqrt{y}}^{\sqrt{y}} |x|\,dx \;=\; y, \quad 0 < y < 1. \end{aligned}$$

Thus, the p.d.f. of Y is $g(y) = G'(y) = 1$, $0 < y < 1$, indicating that $Y \sim \mathsf{Unif}(0,1)$. $\diamondsuit$

Example: Here is a great result sometimes called the Inverse Transform Theorem. Suppose X is a continuous random variable having c.d.f. $F(x)$. Then, amazingly, $F(X) \sim \text{Unif}(0,1)$.

Proof: Let $Y = F(X)$. Then the c.d.f. of Y is

$$\mathsf{Pr}(Y \le y) \;=\; \mathsf{Pr}(F(X) \le y) \;=\; \mathsf{Pr}(X \le F^{-1}(y)) \;=\; F(F^{-1}(y)) \;=\; y,$$

which is the c.d.f. of the Unif(0,1). $\diamondsuit$

Here is a *more-direct* method for dealing with functions of RV's...

Theorem: Suppose that X has p.d.f. $f(x)$, $a \le x \le b$. Let $Y = h(X)$ be a *monotone* function (either increasing or decreasing) of X. Then the p.d.f. of Y is

$$g(y) \;=\; f(h^{-1}(y)) \left| \frac{d}{dy} h^{-1}(y) \right|, \quad h(a) \le y \le h(b) \;\; (\text{or } \; h(b) \le y \le h(a)).$$

Remarks: (i) Warning: You can only use this method of $h(x)$ if *monotone*! (The p.d.f. $f(x)$ doesn't have to be monotone.) (ii) Think of the inverse function $h^{-1}(y) = x$, and the quantity in the $|\cdot|$ as the Jacobian of the transformation.

Example: Suppose that $f(x) = 3x^2$, $0 \le x \le 1$. Find the p.d.f. of $Y = h(X) = X^2$. Note that $f(x)$ is only defined on the domain $0 \le x \le 1$; and on this range, $h(x)$ is monotone increasing — so it's OK to use the wonderful theorem.

First, we have $x = h^{-1}(y) = \pm\sqrt{y} = \sqrt{y}$ (since we're only concerned with positive x's). The theorem then implies that

$$\begin{aligned} g(y) &= f(\sqrt{y})\left|\frac{d}{dy}\sqrt{y}\right|, \quad h(0) \le y \le h(1) \\ &= 3y \times \frac{1}{2\sqrt{y}} = \frac{3}{2}\sqrt{y}, \quad 0 \le y \le 1. \quad \diamondsuit \end{aligned}$$

Remark: We can also look at functions of ≥ 2 RV's, but this takes more work. See any probability text for more info on this important topic.

2.6 Jointly Distributed RV's

Definition: The *joint cdf* of X and Y is $F(x,y) \equiv P(X \le x, Y \le y)$, for all x, y.

Remark: The *marginal cdf* of X is $F_X(x) = F(x,\infty)$. (We use the X subscript to remind us that it's just the cdf of X all by itself.) Similarly, the marginal cdf of Y is $F_Y(y) = F(\infty, y)$.

Definition: If X and Y are discrete, then the *joint pmf* of X and Y is $f(x,y) \equiv P(X = x, Y = y)$.

Remark: The *marginal pmf* of X is

$$f_X(x) = P(X = x) = \sum_y f(x,y).$$

The marginal pmf of Y is

$$f_Y(y) = P(Y = y) = \sum_x f(x,y).$$

Example: Suppose the following table gives the joint pmf of X and Y, along with the accompanying marginals.

	$X=2$	$X=3$	$X=4$	$f_Y(y)$
$Y=4$	0.3	0.2	0.1	0.6
$Y=6$	0.1	0.2	0.1	0.4
$f_X(x)$	0.4	0.4	0.2	1

Definition: If X and Y are continuous, then the *joint pdf* of X and Y is $f(x,y) \equiv \frac{\partial^2}{\partial x \partial y} F(x,y)$.

Remark: The *marginal pdf* of X is

$$f_X(x) \;=\; \int_{\mathbb{R}} f(x,y)\,dy.$$

The *marginal pdf* of Y is

$$f_Y(y) \;=\; \int_{\mathbb{R}} f(x,y)\,dx.$$

Example: This example shows that you have to be careful about "funny" limits when computing marginals. Suppose the joint pdf is

$$f(x,y) \;=\; \frac{21}{4}x^2 y, \quad x^2 \le y \le 1.$$

Then the marginal pdf's are:

$$f_X(x) \;=\; \int_{\mathbb{R}} f(x,y)\,dy \;=\; \int_{x^2}^{1} \frac{21}{4}x^2 y\,dy \;=\; \frac{21}{8}x^2(1-x^4), \quad -1 \le x \le 1$$

and

$$f_Y(y) \;=\; \int_{\mathbb{R}} f(x,y)\,dx \;=\; \int_{-\sqrt{y}}^{\sqrt{y}} \frac{21}{4}x^2 y\,dx \;=\; \frac{7}{2}y^{5/2}, \quad 0 \le y \le 1. \quad \diamondsuit$$

2.7 Independent RV's

Definition: X and Y are *independent* RV's if $f(x,y) = f_X(x) f_Y(y)$ for all x, y.

Examples: If $f(x,y) = cxy$ for $0 \le x \le 2$, $0 \le y \le 3$, then X and Y are independent. If $f(x,y) = \frac{21}{4}x^2 y$ for $x^2 \le y \le 1$, then X and Y are *not* independent. If $f(x,y) = c/(x+y)$ for $1 \le x \le 2$, $1 \le y \le 3$, then X and Y are *not* independent. $\diamondsuit$

Definition: If $f_X(x) > 0$, then $f(y|x) \equiv f(x,y)/f_X(x)$ is the *conditional pdf* (or *pmf*) of Y given $X = x$.

Example: Suppose $f(x,y) = \frac{21}{4}x^2 y$ for $x^2 \le y \le 1$. By a previous example, we find that

$$f(y|x) \;=\; \frac{f(x,y)}{f_X(x)} \;=\; \frac{\frac{21}{4}x^2 y}{\frac{21}{8}x^2(1-x^4)} \;=\; \frac{2y}{1-x^4}, \quad x^2 \le y \le 1. \quad \diamondsuit$$

"Definition": Suppose that $h(X,Y)$ is some function of the RV's X and Y. Then

$$\mathsf{E}[h(X,Y)] = \begin{cases} \sum_x \sum_y h(x,y)f(x,y) & \text{if } (X,Y) \text{ is discrete} \\ \int_{\mathbb{R}} \int_{\mathbb{R}} h(x,y)f(x,y)\,dx\,dy & \text{if } (X,Y) \text{ is continuous} \end{cases}$$

Example/Theorem: Whether or not X and Y are independent, we have $\mathsf{E}[X+Y] = \mathsf{E}[X] + \mathsf{E}[Y]$. In fact, if $X_1, X_2, \ldots$ are RV's, then $\mathsf{E}[\sum_i X_i] = \sum_i \mathsf{E}[X_i]$.

Theorem: If X and Y are *independent*, then $\mathsf{E}[XY] = \mathsf{E}[X]\mathsf{E}[Y]$ and $\mathsf{Var}(X+Y) = \mathsf{Var}(X) + \mathsf{Var}(Y)$.

Proof: Easy algebra. $\diamondsuit$

Theorem: Suppose that $X_1, \ldots, X_n$ are *independent* RV's. If $Y = \sum_{i=1}^n X_i$, then

$$M_Y(t) = \mathsf{E}[e^{tY}] = \mathsf{E}\left[e^{t\sum X_i}\right] = \prod_{i=1}^n \mathsf{E}[e^{tX_i}] = \prod_{i=1}^n M_{X_i}(t).$$

Definition: $X_1, \ldots, X_n$ form a *random sample* from $f(x)$ is

1. $X_1, \ldots, X_n$ are independent, and
2. Each X_i has the same pdf (or pmf) $f(x)$.

Notation: $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} f(x)$. (The term "iid" reads *independent and identically distributed*)

Corollary: $X_1, \ldots, X_n$ iid implies that $M_Y(t) = [M_{X_i}(t)]^n$.

Example: Suppose $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} \text{Bern}(p)$. Then $M_{\sum_i X_i}(t) = (pe^t + q)^n$. It turns out that this is the mgf for the $\text{Bin}(n,p)$ distribution. Thus, by a previous theorem, we have $\sum_{i=1}^n X_i \sim \text{Bin}(n,p)$. $\diamondsuit$

Example: If $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} f(x)$ and $\bar{X} \equiv \sum_{i=1}^n X_i/n$, then $\mathsf{E}[\bar{X}] = \mathsf{E}[X_i]$ and $\mathsf{Var}(\bar{X}) = \mathsf{Var}(X_i)/n$. Thus, the variance *decreases*. $\diamondsuit$

2.8 Covariance and Correlation

Definition: The *covariance* between X and Y is $\mathsf{Cov}(X,Y) \equiv \mathsf{E}[(X - \mathsf{E}[X])(Y - \mathsf{E}[Y])]$. Note that $\mathsf{Var}(X) = \mathsf{Cov}(X,X)$.

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

Proof: Easy. $\diamondsuit$

Theorem: If X and Y are independent RV's, then $\mathsf{Cov}(X,Y) = 0$.

Proof: Since X and Y are independent, we have $\mathsf{E}[XY] = \mathsf{E}[X]\mathsf{E}[Y]$. $\diamondsuit$

Remark: $\mathsf{Cov}(X,Y) = 0$ does *not* imply that X and Y are independent!

Example: Suppose $X \sim \text{Unif}(-1,1)$ and $Y = X^2$. Then X and Y are clearly dependent. However,

$$\mathsf{Cov}(X,Y) \;=\; \mathsf{E}[X^3] - \mathsf{E}[X]\mathsf{E}[X^2] \;=\; \mathsf{E}[X^3] \;=\; \int_{-1}^{1} \frac{x^3}{2}\,dx \;=\; 0. \quad \diamondsuit$$

Theorem: If a and b are constants, the $\mathsf{Cov}(aX,bY) = ab\mathsf{Cov}(X,Y)$.

Definition: The *correlation* between X and Y is

$$\rho \;\equiv\; \frac{\mathsf{Cov}(X,Y)}{\sqrt{\mathsf{Var}(X)\mathsf{Var}(Y)}}.$$

Theorem: $-1 \le \rho \le 1$.

Proof: Follows from the Cauchy-Schwarz inequality. $\diamondsuit$

Remark: If $\rho \approx 1$, we say that X and Y have "high positive" correlation. If $\rho \approx 0$, X and Y have "low" correlation. If $\rho \approx -1$, there is "high negative" correlation.

Example: Suppose that X is the average yards per carry gained by a University of Georgia fullback and Y is his IQ. Further suppose that the joint pmf $f(x,y)$ is given in the following table.

	$X=2$	$X=3$	$X=4$	$f_Y(y)$
$Y=40$	0.00	0.20	0.10	0.3
$Y=50$	0.15	0.10	0.05	0.3
$Y=60$	0.30	0.00	0.10	0.4
$f_X(x)$	0.45	0.30	0.25	1

Then we have $\mathsf{E}[X] = 2.8$, $\mathsf{Var}(X) = 0.66$, $\mathsf{E}[Y] = 51$, $\mathsf{Var}(Y) = 69$, $\mathsf{E}[XY] = \sum_x \sum_y xyf(x,y) = 140$, and

$$\rho = \frac{\mathsf{E}[XY] - \mathsf{E}[X]\mathsf{E}[Y]}{\sqrt{\mathsf{Var}(X)\mathsf{Var}(Y)}} = -0.415. \quad \diamondsuit$$

Theorem: $\mathsf{Var}(\sum_{i=1}^n a_i X_i) = \sum_{i=1}^n a_i^2 \mathsf{Var}(X_i) + 2\sum\sum_{i<j} a_i a_j \mathsf{Cov}(X_i, X_j)$.

Corollary: $\mathsf{Var}(X+Y) = \mathsf{Var}(X) + \mathsf{Var}(Y) + 2\mathsf{Cov}(X,Y)$.

Corollary: $\mathsf{Var}(X-Y) = \mathsf{Var}(X) + \mathsf{Var}(Y) - 2\mathsf{Cov}(X,Y)$.

2.9 Some Fun Distributions

First, some discrete distributions…

2.9.1 Bernoulli

$X \sim$ Bernoulli(p).

$$f(x) = \begin{cases} p & \text{if } x = 0 \\ 1-p & \text{if } x = 1 \end{cases}$$

$\mathsf{E}[X] = p$, $\mathsf{Var}(X) = p(1-p)$, $M_X(t) = pe^t + q$.

If $X_1, X_2, \ldots, X_n$ are i.i.d. Bern(p), we say that they form a series of *Bernoulli(p) trials*.

2.9.2 Binomial

$X \sim$ Binomial(n,p).

$$f(x) = \binom{n}{k} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n.$$

$\mathsf{E}[X] = np$, $\mathsf{Var}(X) = np(1-p)$, $M_X(t) = (pe^t + q)^n$. If $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim}$ Bern(p), then $\sum_{i=1}^n X_i \sim$ Bin(n,p).

2.9.3 Geometric

$X \sim$ Geom(p) is the number of Bern(p) trials until a success occurs. For example, "FFFS" implies that $X = 4$.

$$f(x) = (1-p)^{x-1} p, \quad x = 1, 2, \ldots.$$

$\mathsf{E}[X] = 1/p$, $\mathsf{Var}(X) = q/p^2$.

2.9.4 Negative Binomial

$X \sim \text{NegBin}(r, p)$ is the sum of r i.i.d. Geom(p) RV's, i.e., the time until the rth success occurs. For example, "FFFSSFS" implies that $\text{NegBin}(3, p) = 7$.

$$f(x) = \binom{x-1}{r-1}(1-p)^{x-r}p^r, \quad x = r, r+1, \ldots.$$

$\mathsf{E}[X] = r/p$, $\mathsf{Var}(X) = qr/p^2$.

2.9.5 Poisson

A *counting process* $N(t)$ tallies the number of "arrivals" observed in $[0, t]$. A *Poisson process* is a counting process satisfying the following.

i. Arrivals occur one-at-a-time.

ii. Independent increments, i.e., the numbers of arrivals in disjoint time intervals are independent.

iii. Stationary increments, i.e., the distribution of the number of arrivals only depends on the length of the time interval under observation.

$X \sim \text{Pois}(\lambda)$ is the number of arrivals that a Poisson processes experiences in one time unit, i.e., $N(1)$.

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, \ldots.$$

$\mathsf{E}[X] = \lambda = \mathsf{Var}(X)$.

Now, some continuous distributions...

2.9.6 Uniform

$X \sim \text{Unif}(a, b)$.

$$f(x) = \frac{1}{b-a}, \quad a \le x \le b.$$

$\mathsf{E}[X] = \frac{a+b}{2}$, $\mathsf{Var}(X) = \frac{(b-a)^2}{12}$, $M_X(t) = \frac{e^{tb}-e^{ta}}{t}$.

2.9.7 Exponential

$X \sim \text{Exp}(\lambda)$.

$$f(x) \;=\; \lambda e^{-\lambda x}, \quad x \geq 0.$$

$\mathsf{E}[X] = 1/\lambda$, $\mathsf{Var}(X) = 1/\lambda^2$, $M_X(t) = \frac{\lambda}{\lambda - t}$, $t < \lambda$.

Theorem. The exponential distribution has the *memoryless property*, i.e., for $s, t > 0$,

$$\mathsf{Pr}(X > s + t | X > s) \;=\; \mathsf{Pr}(X > t).$$

By the way, the $\text{Exp}(\lambda)$ is the only continuous distribution with this property.

Example: Suppose that a light bulb has a lifetime that is exponential with mean 1000 hours. Suppose it has already survived 500 hours. Then the probability that it makes it to 2000 is

$$\mathsf{Pr}(X > 2000 | X > 500) \;=\; \mathsf{Pr}(X > 1500) \;=\; e^{-\lambda t} \;=\; e^{-1500/1000}. \quad \diamondsuit$$

2.9.8 Gamma

$X \sim \text{Gamma}(\alpha, \lambda)$.

$$f(x) \;=\; \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}, \quad x \geq 0,$$

where the gamma function is

$$\Gamma(z) \;\equiv\; \int_0^\infty t^{z-1} e^{-t}\, dt.$$

$\mathsf{E}[X] = \alpha/\lambda$, $\mathsf{Var}(X) = \alpha/\lambda^2$, $M_X(t) = (\frac{\lambda}{\lambda - t})^\alpha$. If $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$, then $Y \equiv \sum_{i=1}^n X_i \sim \text{Gamma}(n, \lambda)$. The $\text{Gamma}(n, \lambda)$ is also called the $\text{Erlang}_n(\lambda)$. It has c.d.f.

$$F_Y(y) \;=\; 1 - e^{-\lambda y} \sum_{j=0}^{n-1} \frac{(\lambda y)^j}{j!}, \quad y \geq 0.$$

2.9.9 Normal

$X \sim \mathsf{Nor}(\mu, \sigma^2)$.

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

$\mathsf{E}[X] = \mu$, $\mathsf{Var}(X) = \sigma^2$, $M_X(t) = \exp\{\mu t + \frac{1}{2}\sigma^2 t^2\}$.

Theorem (Additive Property of Normals): Suppose that $X_1, X_2, \ldots, X_n$ are *independent* with $X_i \sim \mathsf{Nor}(\mu_i, \sigma_i^2)$, $i = 1, 2\ldots, n$. Then

$$Y \;=\; \sum_{i=1}^{n} a_i X_i + b \;\sim\; \mathsf{Nor}\left(\sum_{i=1}^{n} a_i \mu_i + b,\, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right).$$

Proof: Use m.g.f.'s.

$$\begin{aligned}
M_Y(t) &= \mathsf{E}(e^{tY}) \;=\; \mathsf{E}\left(\exp\left\{t\left(\sum_{i=1}^{n} a_i X_i + b\right)\right\}\right) \\
&= e^{tb}\mathsf{E}\left(\exp\left\{\sum_{i=1}^{n}(a_i t)X_i\right\}\right) \\
&= e^{tb}\prod_{i=1}^{n}\mathsf{E}\left(e^{(a_i t)X_i}\right) \quad \text{(by independence)} \\
&= e^{tb}\prod_{i=1}^{n} M_{X_i}(a_i t) \\
&= e^{tb}\prod_{i=1}^{n}\exp\left\{\mu_i(a_i t) + \frac{1}{2}\sigma_i^2 (a_i t)^2\right\} \\
&= \exp\left\{\left(\sum_{i=1}^{n}\mu_i a_i + b\right)t + \frac{1}{2}\left(\sum_{i=1}^{n} a_i^2\sigma_i^2\right)t^2\right\}. \quad \diamondsuit
\end{aligned}$$

Example: Suppose $X \sim \mathsf{Nor}(3, 4)$, $Y \sim \mathsf{Nor}(4, 6)$, and X and Y are independent. Then

$$2X - 3Y + 1 \;\sim\; \mathsf{Nor}(2\mathsf{E}[X] - 3\mathsf{E}[Y] + 1, 4\mathsf{Var}(X) + 9\mathsf{Var}(Y)) \;\sim\; \mathsf{Nor}(-5, 70). \quad \diamondsuit$$

Corollary: If $X \sim \mathsf{Nor}(\mu, \sigma^2)$, then $aX + b \sim \mathsf{Nor}(a\mu + b, a^2\sigma^2)$.

Corollary: If $X \sim \mathsf{Nor}(\mu, \sigma^2)$, then $Z \equiv \frac{X-\mu}{\sigma} \sim \mathsf{Nor}(0, 1)$, the standard normal distribution.

Notation: The standard normal's p.d.f. is $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$, and the c.d.f. is $\Phi(x)$, which is usually tabled. For example, $\Phi(1.96) \doteq 0.975$.

2.10 A First Look at Some Limit Theorems

Corollary (of theorem on linear combinations of normals from previous subsection): If $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Nor}(\mu, \sigma^2)$, then the sample mean

$$\bar{X} \;\equiv\; \frac{1}{n}\sum_{i=1}^{n} X_i \;\sim\; \mathsf{Nor}(\mu, \sigma^2/n).$$

This is a special case of the *Law of Large Numbers*, which says that $\bar{X}$ approximates μ well as n becomes large.

Markov's Inequality: If X is a non-negative RV, then for all $\epsilon > 0$, we have

$$\mathsf{Pr}(X \geq \epsilon) \;\leq\; \mathsf{E}[X]/\epsilon.$$

Proof: Since X is non-negative,

$$\mathsf{E}[X] \;=\; \int_0^\infty x f(x)\,dx \;\geq\; \int_\epsilon^\infty x f(x)\,dx \;\geq\; \epsilon \int_\epsilon^\infty f(x)\,dx \;=\; \epsilon \mathsf{Pr}(X \geq \epsilon). \quad \diamondsuit$$

Chebychev's Inequality: For any RV X and for all $\epsilon > 0$, we have

$$\mathsf{Pr}(|X - \mathsf{E}[X]| \geq \epsilon) \leq \mathsf{Var}(X)/\epsilon^2.$$

Proof: Uses Markov's Inequality; see any probability test. $\diamondsuit$

Bonus Generalization: $\mathsf{Pr}(|X| \geq \epsilon) \leq \mathsf{E}[|X|^r]/\epsilon^r$.

Remark: These inequalities are usually pretty crude!

Example: Suppose that $X \sim \mathsf{Unif}(0,1)$. Then the probability that X deviates from its mean by at least 1/4 is exactly

$$\begin{aligned} \mathsf{Pr}\left(\left|X - \frac{1}{2}\right| \geq \frac{1}{4}\right) \;&=\; 1 - \mathsf{Pr}\left(\left|X - \frac{1}{2}\right| < \frac{1}{4}\right) \;=\; 1 - \mathsf{Pr}\left(-\frac{1}{4} < X - \frac{1}{2} < \frac{1}{4}\right) \\ &=\; 1 - \mathsf{Pr}\left(\frac{1}{4} < X < \frac{3}{4}\right) \;=\; \frac{1}{2}. \end{aligned}$$

Meanwhile, by Chebychev (with $\mathsf{E}[X] = 1/2$, $\mathsf{Var}(X) = 1/12$, and $\epsilon = 1/4$), we have

$$\mathsf{Pr}\left(\left|X - \frac{1}{2}\right| \geq \frac{1}{4}\right) \;\leq\; \frac{\mathsf{Var}(X)}{\epsilon^2} \;=\; \frac{16}{12} \;=\; \frac{4}{3},$$

which is a very crude upper bound indeed! $\diamondsuit$

Definition: The sequence of random variables $Y_1, Y_2, \ldots$ with respective c.d.f.'s $F_{Y_1}(y), F_{Y_2}(y), \ldots$ *converges in distribution* to the random variable Y having c.d.f. $F_Y(y)$ if $\lim_{n\to\infty} F_{Y_n}(y) = F_Y(y)$ for all y belonging to the continuity set of Y (i.e., the set of all points y at which $F_Y(y)$ is continuous). Notation: $Y_n \xrightarrow{\mathcal{D}} Y$. (Also sometimes called *convergence in law* or *weak convergence*.)

Idea: If $Y_n \xrightarrow{\mathcal{D}} Y$, then you would expect to be able to approximate the distribution of Y_n by the limiting distribution of Y, at least for large enough n.

Central Limit Theorem: If $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} f(x)$ with mean μ and variance σ^2, then

$$Z_n \equiv \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{\mathcal{D}} \mathsf{Nor}(0,1),$$

where $\bar{X}_n$ is the sample mean. Thus, the c.d.f. of Z_n approaches that of the standard normal as n increases. The CLT usually works pretty well if the pdf/pmf is fairly symmetric and $n \geq 15$.

Example: Suppose that $X_1, X_2, \ldots, X_{100} \overset{\text{iid}}{\sim} \mathsf{Exp}(1)$. Then

$$\begin{aligned} \mathsf{Pr}\left(90 \leq \sum_{i=1}^{100} X_i \leq 110\right) &= \mathsf{Pr}\left(\frac{90-100}{\sqrt{100}} \leq Z_{100} \leq \frac{110-100}{\sqrt{100}}\right) \\ &= \mathsf{Pr}(-1 \leq Z_{100} \leq 1) \approx \mathsf{Pr}(-1 \leq \mathsf{Nor}(0,1) \leq 1) = 2\Phi(1) - 1 \approx 0.683. \quad \diamondsuit \end{aligned}$$

Definition: The sequence of random variables $Y_1, Y_2, \ldots$ is said to *converge in probability* to Y (often a constant) if for all $\epsilon > 0$, $\mathsf{Pr}(|Y_n - Y| > \epsilon) \to 0$ as $n \to \infty$. Notation: $Y_n \xrightarrow{\mathcal{P}} Y$.

Theorem: $Y_n \xrightarrow{\mathcal{P}} Y$ implies $Y_n \xrightarrow{\mathcal{D}} Y$. In other words, convergence in probability is a bit stronger than convergence in distribution.

Weak Law of Large Numbers: If $X_1, X_2, \ldots$ are i.i.d. with mean μ, then $\bar{X}_n \xrightarrow{\mathcal{P}} \mu$. Why is this called the *weak* LLN? Simply because there's a stronger one coming up later.

Continuous Mapping Theorem: If $Y_n \xrightarrow{\mathcal{P}} Y$ and $g(\cdot)$ is a nice, continuous function, then $g(Y_n) \xrightarrow{\mathcal{P}} g(Y)$. The CMT is often useful for characterizing the convergence of nasty functions of the Y_i's.

Definition: The sequence of random variables $Y_1, Y_2, \ldots$ is said to *converge in rth mean* to Y (often a constant) if $\mathsf{E}[|Y_n - Y|^r] \to 0$ as $n \to \infty$. Notation: $Y_n \xrightarrow{r} Y$.

Theorem: $Y_n \xrightarrow{r} Y$ implies $Y_n \xrightarrow{\mathcal{P}} Y$. In other words, convergence in rth mean is a bit stronger than convergence in probability.

Proof: Follows immediately from bonus version of Chebychev. $\diamondsuit$

Definition: The sequence of random variables $Y_1, Y_2, \ldots$ *converges almost surely* (or *with probability one*) to Y if $\mathsf{Pr}(Y_n$ converges to $Y) = 1$ as $n \to \infty$. Notation: $Y_n \overset{a.s.}{\to} Y$.

Theorem: $Y_n \overset{a.s.}{\to} Y$ implies $Y_n \overset{\mathcal{P}}{\to} Y$. In other words, convergence almost surely is a bit stronger than convergence in probability.

Strong Law of Large Numbers: If $X_1, X_2, \ldots$ are i.i.d. with mean μ, then $\bar{X}_n \overset{a.s.}{\to} \mu$. It turns out that the SLLN implies the WLLN.

How do almost sure and rth mean convergence relate to each other?

Dominated Convergence Theorem: If $Y_n \overset{a.s.}{\to} Y$ and there exists a random variable W such that $\mathsf{Pr}(|Y_n| \leq W) = 1$ for every n, then $\mathsf{E}[Y_n] \to \mathsf{E}[Y]$.