STA257 Probability and Statistics I

Gun Ho Jang

Update on May 30, 2018

Note: This note is prepared for STA257. There might be numerous fault arguments/statements/typos. If you spot one, please contact the instructor or you may look up references which may contain errors too.

Expectation

Expectation of Random Variables

Example 61. Consider a die game. A dollar need to be paid to start a game. If an even number appears, you win three dollars. If an odd number appears, you lose the game. Are you going to play this game? Why?

Roughly saying, there will be an even number and an odd number in two tosses of a fair die in long term toss. Hence, you pay two dollars for two games and win three dollars in a game and gain nothing from a game. As a net, you win a dollar per two games or a half dollar per each game in long term play. If there is such a game, you should play the game for a long time. Unfortunately there is no such game because the game host will lose all the money very quickly.

In the above example, experiment is tossing a fair die and a random variable X be the net gain from a game. Then $X = 3 - 1$ for an even number and $X = 0 - 1$ for an odd number. The expected gain in the game could be

$$2 \times P(X = 2) + (-1) \times P(X = -1) = 2 \times 1/2 + (-1) \times 1/2 = 1/2.$$

This expected gain is the same to intuitive gain in the previous example. Hence an expectation of a random variable can be defined in a similar fashion.

The previous example also shows the intuition of expectation. Hence we define a new concept "expectation" of a random variable as follows.

Definition 33. The *expectation* (or *expected value* or *mean value*) of a discrete random variable is

$$\mathbb{E}X = \mathbb{E}(X) = \sum_x x \times P(X = x) = \sum_x x \times \text{pmf}_X(x)$$

when the sum is absolutely convergent.

This definition is only applicable for discrete random variables.

Example 62. Assume $X \sim \text{Bernoulli}(p)$, that is, $P(X = 1) = p$ and $P(X = 0) = 1 - p$. Then

$$\mathbb{E}(X) = 1 \times P(X = 1) + 0 \times P(X = 0) = p.$$

Example 63. Let X be the number of heads in 2 tosses of a fair coin. Since the coin is fair, it will show head half of time and tail half of time. In this reason an expected number of heads must be 1 in this experiment.

In probability sense, using the probabilities $P(X = 0) = 1/4, P(X = 1) = 1/2, P(X = 2) = 1/4$ we can define the expected number of head is

$$0 \times (1/4) + 1 \times (1/2) + 2 \times (1/4) = 1.$$

When $\sum_{x>0} x \times \text{pmf}_X(x) = \infty$ or $\sum_{x<0} x \times \text{pmf}_X(x) = -\infty$, the expectation may not be finite or may not be defined.

Example 64. Consider an integer valued random variable X with probability mass function $\text{pmf}_X(x) = 1/[2|x|(|x| + 1)]$ for $x \in \mathbb{Z} - \{0\}$. Then $\sum_{x\in\mathbb{N}} x \times 1/[2|x|(|x| + 1)] = (1/2)\sum_{n=1}^{\infty} 1/(n + 1) = \infty$ and $\sum_{-x\in\mathbb{N}} x \times 1/[2|x|(|x| + 1)] = -\infty$. Hence expectation of X is not defined.

The previous definition of expectation is only applicable for discrete random variables. Anyone will wonder how to define an expectation of a continuous random variable. The first step is to approximate continuous random variable by discrete random variables. Then compute expectation based on discrete version of expectation. Then take finer discretizations until expectations converge.

To make it easy, assume a continuous random variable X is bounded, that is, $P(a < X \leq b) = 1$ for some $a < b$. Let f and F be the probability density and the cumulative density functions of X. Consider a partition $a = x_0 \leq t_1 \leq x_1 \leq t_2 \leq x_2 \leq \cdots \leq x_{n-1} \leq t_n \leq x_n = b$. For each interval $\{x_{k-1} < X \leq x_k\}$, set $\tilde{X}_n = t_k$. Then $\tilde{X}_n$ is a discretized version of continuous random variable X and its expectation is

$$\begin{aligned}\mathbb{E}[\tilde{X}_n] &= \sum_{k=1}^{n} t_k P(\tilde{X}_n = t_k) = \sum_{k=1}^{n} t_k P(x_{k-1} < X \leq x_k) = \sum_{k=1}^{n} t_k (P(X \leq x_k) - P(X \leq x_{k-1})) \\ &= \sum_{k=1}^{\infty} t_k (F(x_k) - F(x_{k-1}))\end{aligned}$$

which is a Riemann-Stieltjes integral which converges to $\int x dF(x)$ or using the mean value theorem

$$\mathbb{E}[\tilde{X}_n] = \sum_{k=1}^{\infty} t_k (F(x_k) - F(x_{k-1})) \approx \sum_{k=1}^{\infty} t'_k f(t'_k)(x_k - x_{k-1}) \to \int x \cdot f(x)\, dx.$$

The approximated summation is Riemann sum which converges to the Riemann integral $\int x \times f(x)\ dx$. Hence the following definition expands expectation to continuous random variables.

Definition 34. The expectation of a continuous random variable X is defined by

$$\mathbb{E}(X) = \int x \times \text{pdf}_X(x)\, dx.$$

Example 65. A continuous random variable X has density $\text{pdf}_X(x) = 2x1(0 < x < 1)$. The expectation of

X is

$$\mathbb{E}(X) = \int_0^1 x \times 2x \, dx = 2x^3/3|_0^1 = 2/3.$$

Example 66. A continuous random variable X follows a *Cauchy* distribution if the density is

$$\mathrm{pdf}_X(x) = \frac{1}{\pi(1+x^2)}$$

Considering $\int_0^\infty x/[\pi(1+x^2)]\, dx \geq \int_1^\infty 1/[2\pi x]\, dx = \infty$ and $\int_{-\infty}^0 x/[\pi(1+x^2)]\, dx = -\infty$, no expectation is defined for Cauchy distribution.

There are random variables which are neither discrete nor continuous. For these random variables, Riemann-Stieltjes integral is still defined. Hence resulting in the next theorem.

Theorem 36. Assume a discrete random variable X is non-negative. Then

$$\mathbb{E}(X) = \int_0^\infty P(X > z)\, dz.$$

Proof. We draw a graph and prove from it.

`graph`

The area can be written as

$$\sum_{x:\mathrm{pmf}_X(x)>0} x \cdot \mathrm{pmf}_X(x) = \int_0^\infty P(X > x)\, dx$$

in two different interpretations.

□

Theorem 37. For any random variable X with finite expectation,

$$\mathbb{E}(X) = \int_0^\infty P(X > z)\, dz - \int_{-\infty}^0 P(X < z)\, dz.$$

Any random variable can be written as $X = X^+ - X^- = \max(0, X) - \max(0, -X)$. Applications of Theorems 36 and 40 result in the previous theorem, that is, under the linearity of expectation

$$\begin{aligned}\mathbb{E}(X) &= \mathbb{E}(X^+ - X^-) = \mathbb{E}(X^+) - \mathbb{E}(X^-) \\ &= \int_0^\infty P(X^+ > z)\, dz - \int_0^\infty P(X^- > z)\, dz = \int_0^\infty P(X > z)\, dz - \int_{-\infty}^0 P(X < z)\, dz.\end{aligned}$$

Without the linearity of expectation, the theorem can still be proved.

Proof. We show the second term is the same to the last term. Note that

$$\begin{aligned}\int_0^\infty P(X > z)\, dz &= \int_0^\infty (1 - F(z))\, dz = \int_0^\infty \int_z^\infty 1\, dF(x)\, dz \\ &= \int_0^\infty \int_0^\infty 1(x > z)\, dF(x)\, dz\end{aligned}$$

Two integrals are exchangeable due to Fatou's lemma

$$= \int_0^\infty \int_0^\infty 1(x > z)\, dz\, dF(x) = \int_0^\infty x\, dF(x)$$

Similarly, the second integral term becomes

$$\begin{aligned}-\int_{-\infty}^0 P(X < z)\, dz &= -\int_{-\infty}^0 \int_{-\infty}^0 1(x < z)\, dF(x)\, dz = -\int_{-\infty}^0 \int_{-\infty}^0 1(x < z)\, dz\, dF(x) = -\int_{-\infty}^0 (-x)\, dF(x) \\ &= \int_{-\infty}^0 x\, dF(x)\end{aligned}$$

Hence the theorem holds. □

Lemma 38. Let F be the cumulative distribution function of a random variable X. For an interval, $P(a < X \leq b) = \mathbb{E}[1(a < X \leq b)]$. In general, for each event A of X,

$$P(X \in A) = \mathbb{E}[1(X \in A)].$$

Proof. Let $Y = 1(a < X \leq b)$ so that $P(Y = 1) = P(a < X \leq b)$ and $P(Y = 0) = 1 - P(a < X \leq b)$. Then $\mathbb{E}[1(a < X \leq b)] = \mathbb{E}[Y] = 1 \times P(a < X \leq b) + 0 \times (1 - P(a < X \leq b)) = P(a < X \leq b)$. In general let $Z = 1(X \in A)$. Then $P(Z = 1) = P(X \in A)$. Hence $\mathbb{E}[1(X \in A)] = \mathbb{E}[Z] = P(Z = 1) = P(X \in A)$. □

Theorem 39. Let X be a random variable and g be a function on $\mathbb{R}$. If expectation of $Y = g(X)$ is defined,

$$\mathbb{E}(Y) = \int g(x)\, d\mathrm{cdf}_X(x) = \int_{-\infty}^\infty g(x) \cdot \mathrm{pdf}_X(x)\, dx \text{ or } \sum_x g(x) \cdot \mathrm{pmf}_X(x)$$

Proof. To simplify assume $g \geq 0$. Using Theorem 36,

$$\begin{aligned}\mathbb{E}(Y) &= \int_0^\infty P(Y > z)\, dz = \int_0^\infty P(g(X) > z)\, dz = \int_0^\infty \int 1(g(x) > z)\, d\mathrm{cdf}_X(x)\, dz \\ &= \int \int_0^\infty 1(g(x) > z)\, dz\, d\mathrm{cdf}_X(x) = \int g(x)\, d\mathrm{cdf}_X(x)\end{aligned}$$

Using the integration by parts, $\int g(x)\mathrm{pdf}_X(x)\, dx = \int g(x)\, d\mathrm{cdf}_X(x)$ for continuous cases. The Riemann-Stieltjes integral changes to summation for discrete cases. □

Example 67. Let $X \sim \text{uniform}(0,1)$ and $Y = X^2$.

$$\mathbb{E}(X) = \int_0^1 1(0 < x < 1)\, dx = 1 \text{ and } \mathbb{E}(Y) = \int_0^1 x^2 \cdot 1(0 < x < 1)\, dx = x^3/3\Big|_0^1 = \frac{1}{3}.$$

Using change of variable, the density of Y is

$$\text{pdf}_Y(y) = \text{pdf}_X(\sqrt{y})|\frac{d\sqrt{y}}{dy}| = 1(0 < y < 1)/(2\sqrt{y})$$

and its expectation by definition is

$$\mathbb{E}(Y) = \int y \cdot \text{pdf}_Y(y)\, dy = \int_0^1 y \cdot \frac{1}{2\sqrt{y}}\, dy = \frac{1}{3}y^{3/2}\Big|_0^1 = \frac{1}{3}.$$

The cumulative distribution function of Y is, for $0 < y < 1$,

$$\text{cdf}_Y(y) = P(Y \le y) = P(X \le \sqrt{x}) = \sqrt{y}$$

The Riemann-Stieltjes integral becomes

$$\mathbb{E}(Y) = \int y\, d\text{cdf}_Y(y) = \int_0^1 y d\sqrt{y} = \int_0^1 y \cdot \frac{1}{2\sqrt{y}}\, dy = \frac{1}{3}.$$

Hence Theorems 37 and 39 hold.

Properties of Expectations

Theorem 40 (Properties of Expectation)**.** (a) [linearity] Let $Y = aX + b$. Then $\mathbb{E}(Y) = a\mathbb{E}(X) + b$.
(b) If $X \ge 0$, that is, $P(X \ge 0) = 1$, then $\mathbb{E}(X) \ge 0$.
(c) $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.
(d) for constant random variable 1, $\mathbb{E}(1) = 1$.

Proof. (a) Assume $a > 0$ for simplicity. Consider Riemann-Stieltjes integral in Theorem 37, that is, using the distribution function $\text{cdf}_Y(y) = P(Y \le y) = P(aX + b \le y) = P(X \le (y-b)/a) = \text{cdf}_X((y-b)/a)$,

$$\int y\, d\text{cdf}_Y(y) = \int y\, d\text{cdf}_X((y-b)/a) = \int (az + b)\, d\text{cdf}_X(z) = a\int z\, d\text{cdf}_X(z) + b = a\mathbb{E}(X) + b.$$

(b) Theorem 36 implies

$$\mathbb{E}(X) = \int_0^\infty P(X > z)\, dz \ge 0.$$

(c) It is enough to prove for discrete cases. Assume X and Y are taking values at $x_1, x_2, \ldots$ and $y_1, y_2, \ldots$.

Then $Z = X + Y$ takes values on $\mathcal{Z} = \{x_i + y_j : i, j \geq 1\}$. Hence the expectation of Z is

$$\begin{aligned}
\mathbb{E}(Z) &= \sum_{z\in\mathcal{Z}} zP(X+Y=z) = \sum_{z\in\mathcal{Z}} z \sum_{i,j=1}^{\infty} P(X=x_i)P(Y=y_j)1(y_j = z - x_i) \\
&= \sum_{z\in\mathcal{Z}} \sum_{i,j=1}^{\infty} (x_i + y_j)P(X=x_i)P(Y=y_j)1(y_j = z - x_i) = \sum_{i,j=1}^{\infty} (x_i + y_j)P(X=x_i)P(Y=y_j) \\
&= \sum_{i,j=1}^{\infty} x_i P(X=x_i)P(Y=y_j) + \sum_{i,j=1}^{\infty} y_j P(X=x_i)P(Y=y_j) \\
&= \sum_{i=1}^{\infty} x_i P(X=x_i) \sum_{j=1}^{\infty} P(Y=y_j) + \sum_{j=1}^{\infty} y_j P(Y=y_j) \sum_{i=1}^{\infty} P(X=x_i) \\
&= \sum_{i=1}^{\infty} x_i P(X=x_i) + \sum_{j=1}^{\infty} y_j P(Y=y_j) = \mathbb{E}(X) + \mathbb{E}(Y).
\end{aligned}$$

(d) Let $X = 1$, then $P(X = 1) = 1$ and $\mathbb{E}(X) = 1 \times P(X = 1) = 1$. □

Exercise 16. When $P(a \leq X \leq b) = 1$, show that $a \leq \mathbb{E}(X) \leq b$.

Example 68. When $X \sim \text{uniform}(a, b)$, $P(a \leq X \leq b) = 1$ and $a \leq \mathbb{E}(X) \leq b$. Actually

$$\mathbb{E}(X) = \int_a^b x \cdot \frac{1}{b-a}\, dx = \frac{x^2}{2(b-a)}\Big|_a^b = \frac{a+b}{2}.$$

Theorem 41. Let X and Y be two independent random variables and g and h be real functions satisfying $g(X)$ and $h(Y)$ are random variables with finite expectations. Then

$$\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y))$$

Proof. Using Theorem 37 and the independence, we get

$$\begin{aligned}
\mathbb{E}(g(X)h(Y)) &= \int_{\mathbb{R}^2} g(x)h(y)\, d\text{cdf}_{X,Y}(x,y) = \int\int g(x)h(y)\, d\text{cdf}_X(x)\, d\text{cdf}_Y(y) = \int g(x)\, d\text{cdf}_X(x) \int h(y)\, d\text{cdf}_Y(y) \\
&= \mathbb{E}(g(X))\mathbb{E}(h(Y)).
\end{aligned}$$

□

Example 69. Let $X \sim \text{Bernoulli}(p)$ and $Y \sim \text{uniform}(0, 1)$ be independent random variables. Then

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y) = [0 \cdot (1-p) + 1 \cdot p] \times \int_0^1 y\, dy = p \cdot 1/2 = p/2$$

However the converse of Theorem 41 does not hold.

Example 70. Two discrete random variables X and Y satisfies $P(X = 1, Y = 0) = p$, $P(Y = -1, X = 0) = P(Y = 1, X = 0) = (1-p)/2$ with $0 < p < 1$ so that $X \sim \text{Bernoulli}(p)$. Then $\mathbb{E}(XY) = 0, \mathbb{E}(X) = p, \mathbb{E}(Y) = 0$ but $P(X = 1, Y = 0) = p \neq p^2 = P(X = 1)P(Y = 0)$. Hence a specific decomposition of expectation does not imply independence of random variables.

Example 71. Let $X \sim \text{binomial}(n, p)$.

$$\mathbb{E}(X) = \sum_{i=0}^{n} i \cdot \binom{n}{i} p^i (1-p)^{n-i} = \sum_{i=1}^{n} i \cdot \frac{n}{i} p \cdot \binom{n-1}{i-1} p^{i-1} (1-p)^{(n-1)-(i-1)} = np(p+1-p)^{n-1} = np.$$

Moments

Definition 35. For positive integer k, the *kth moment* of X is $\mathbb{E}(X^k)$ and the *kth central moment* is $\mathbb{E}[(X - \mathbb{E}(X))^k]$.

Exercise 17. If $\mathbb{E}|X| < \infty$ and density is symmetric around the mean, then the mean and median are the same.

Example 72. Assume $\mathbb{E}(X^2) < \infty$. The mean minimizes mean square error, that is, for $\mu = \mathbb{E}(X)$,

$$\mu = \arg\min_d \mathbb{E}(X-d)^2$$

Note that

$$\mathbb{E}(X-d)^2 = \mathbb{E}[(X-\mu+\mu-d)^2] = \mathbb{E}[(X-\mu)^2 + 2(X-\mu)(\mu-d) + (\mu-d)^2] = \mathbb{V}\text{ar}(X) + 0 + (\mu-d)^2 \geq \mathbb{V}\text{ar}(X).$$

The equality holds if and only if $d = \mu$.

Example 73. Assume $\mathbb{E}|X| < \infty$. The median minimizes mean absolute error, that is, for $m = \text{median}(X)$,

$$m = \arg\min_d \mathbb{E}|X-d|.$$

For simplicity assume $d \geq m$,

$$\begin{aligned}\mathbb{E}(|X-d| - |X-m|) &= \int_{-\infty}^{m} [(d-x)-(m-x)]\, dF(x) + \int_{m}^{d} [(d-x)-(x-m)]\, dF(x) + \int_{d}^{\infty} [(x-d)-(x-m)]\, dF(x) \\ &\geq (d-m)F(m) + \int_{m}^{d} (m-d)\, dF(x) + (m-d)(1-F(d)) = (d-m)(F(m)-1+F(m))\end{aligned}$$

Since $2F(m) - 1 \geq 0$, $\mathbb{E}|X-d|$ is minimized at $d = m$.

Example 74. A random variable X is said to follow a *exponential* distribution with parameter λ if its density is

$$\text{pdf}_X(x) = \lambda e^{-\lambda x} 1(x > 0).$$

Thus the cumulative distribution function is

$$\text{cdf}_X(x) = P(X \leq x) = \int_0^x \lambda e^{-\lambda z}\, dz = \int_0^{x/\lambda} e^{-z}\, dz = 1 - e^{-x/\lambda}.$$

The mean and median of X are

$$\begin{aligned}\mathbb{E}(X) &= \int_0^\infty x \cdot \lambda e^{-\lambda x}\, dx = \lambda^{-1} \int_0^\infty z e^{-z}\, dz = \frac{1}{\lambda}[-(z+1)e^{-z}]_0^\infty = \frac{1}{\lambda} \\ \text{median}(X) &= \text{cdf}_X^{-1}(1/2) = \lambda \log(2).\end{aligned}$$

Hence mean and median are different for exponential distributions.

Example 75. A random variable X follows *gamma* distribution with parameters α, β if it has density

$$\mathrm{pdf}_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} 1(x \geq 0)$$

where $\alpha, \beta > 0$ and $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx$ satisfying $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$. For $r > -\alpha$,

$$\mathbb{E}(X^r) = \int_0^\infty x^r \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}\, dx = \frac{1}{\beta^r \Gamma(\alpha)} \int_0^\infty z^{r+\alpha-1} e^{-z}\, dz = \frac{\Gamma(r+\alpha)}{\beta^r \Gamma(\alpha)} < \infty.$$

Hence gamma distribution has all (non-negative integer) moments.

Theorem 42. If $\mathbb{E}(|X|^t) < \infty$ for some $t > 0$, then $\mathbb{E}(|X|^s) < \infty$ for any $0 \leq s \leq t$.

Proof. Note that $|x|^s \leq [\max(1, |x|)]^{t-s} |x|^s \leq 1 + |x|^t$. Hence

$$\mathbb{E}(|X|^s) \leq 1 + \mathbb{E}(|X|^t) < \infty.$$

□

Definition 36. The *variance* of a random variable X is $\mathbb{E}[(X - \mathbb{E}(X))^2]$, that is, the expectation of squared deviation from the expectation $\mathbb{E}(X)$. The *covariance* between two random variables X and Y is $\mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$. Two random variables X and Y are *uncorrelated* if $\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 0$. The *correlation* of two random variables X and Y having finite second moment is

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathbb{V}\mathrm{ar}(X)\mathbb{V}\mathrm{ar}(Y)}}.$$

The standardized third and fourth moments are said to be *skewness* and *kurtosis*, that is, $\mathbb{E}[(X-\mu)^3]/\sigma^3$ and $\mathbb{E}[(X-\mu)^4]/\sigma^4$ where $\mu = \mathbb{E}(X)$ and $\sigma^2 = \mathbb{V}\mathrm{ar}(X)$.

The variance and covariance can be simplified using $\mu = \mathbb{E}(X)$ as follows.

$$\mathbb{V}\mathrm{ar}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}[(X-\mu)^2] = \mathbb{E}(X^2 - 2X\mu + \mu^2) = \mathbb{E}(X^2) - 2\mathbb{E}(X)\mu + \mu^2 = \mathbb{E}(X^2) - \mu^2 = \mathbb{E}(X^2) - (\mathbb{E}X)^2.$$

Similarly,

$$\begin{aligned}\mathrm{Cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] = \mathbb{E}[XY - X\mathbb{E}(Y) - \mathbb{E}(X)Y + \mathbb{E}(X)\mathbb{E}(Y)] \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(Y)\mathbb{E}(X) + \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).\end{aligned}$$

Example 76. Let $X \sim \mathrm{Bernoulli}(p)$.

$$\mathbb{E}(X^k) = 0^k P(X=0) + 1^k P(X=1) = P(X=1) = p.$$

Hence $\mathbb{E}(X) = 1$ and $\mathbb{V}\mathrm{ar}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = p - p^2 = p(1-p)$.

Example 77. Let $X \sim \text{binomial}(n, p)$.

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{k=0}^{n} k\binom{n}{k}p^k(1-p)^{n-k} = \sum_{k=1}^{n} np\binom{n-1}{k-1}p^{k-1}(1-p)^{(n-1)-(k-1)} = np(p+1-p)^{n-1} = np, \\
\mathbb{E}(X(X-1)) &= \sum_{k=0}^{n} k(k-1)\binom{n}{k}p^k(1-p)^{n-k} = \sum_{k=2}^{n} n(n-1)p^2\binom{n-2}{k-2}p^{k-2}(1-p)^{(n-2)-(k-2)} = n(n-1)p^2(p+1-p)^{n-2} \\
&= n(n-1)p^2, \\
\mathbb{E}(X^2) &= \mathbb{E}(X(X-1)+X) = n(n-1)p^2 + np, \\
\mathbb{V}\text{ar}(X) &= \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = n^2p^2 - np^2 + np - (np)^2 = np(1-p).
\end{aligned}$$

Example 78. Let X be a random variable having density symmetric to its mean. If there exists third moment, then skewness is zero.

$$\mathbb{E}(X-\mu)^3 = \int (x-\mu)^3 \text{pdf}_X(x)\, dx = \int z^3 \text{pdf}_X(z+\mu)\, dz = 0.$$

In the second last equality $z^3\text{pdf}_X(z+\mu)$ is odd function with finite integral.

Theorem 43 (Properties of variance)**.** (a) $\mathbb{V}\text{ar}(X) \geq 0$,
(b) $\mathbb{V}\text{ar}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$,
(c) $\mathbb{V}\text{ar}(aX+b) = a^2\mathbb{V}\text{ar}(X)$,
(d) $\mathbb{V}\text{ar}(X+Y) = \mathbb{V}\text{ar}(X) + \mathbb{V}\text{ar}(Y) + 2\text{Cov}(X,Y)$,
(e) $\mathbb{V}\text{ar}(X+Y) = \mathbb{V}\text{ar}(X) + \mathbb{V}\text{ar}(Y)$ if and only if X and Y are uncorrelated.

Proof. (a) $\mathbb{V}\text{ar}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] \geq \mathbb{E}(0) = 0$.
(b) $\mathbb{V}\text{ar}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}[X^2 - 2X\mathbb{E}(X) + \{\mathbb{E}(X)\}^2] = \mathbb{E}(X^2) - 2\{\mathbb{E}(X)\}^2 + \{\mathbb{E}(X)\}^2 = \mathbb{E}(X^2) - \{\mathbb{E}(X)\}^2$.
(c) Noting $\mathbb{E}(aX+b) = a\mathbb{E}(X)+b$, we get $\mathbb{V}\text{ar}(aX+b) = \mathbb{E}[(aX+b-(a\mathbb{E}(X)+b))^2] = \mathbb{E}[(aX - a\mathbb{E}(X))^2] = a^2\mathbb{E}[(X-\mathbb{E}(X))^2] = a^2\mathbb{V}\text{ar}(X)$.
(d) $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$ and $\mathbb{E}[(X+Y)^2] = \mathbb{E}(X^2+2XY+Y^2) = \mathbb{E}(X^2) + 2\mathbb{E}(XY) + \mathbb{E}(Y^2)$ implies $\mathbb{V}\text{ar}(X+Y) = \mathbb{E}[(X+Y)^2] - (\mathbb{E}(X+Y))^2 = \mathbb{E}(X^2) + \mathbb{E}(Y^2) + 2\mathbb{E}(XY) - ((\mathbb{E}(X))^2 + (\mathbb{E}(Y))^2 + 2\mathbb{E}(X)\mathbb{E}(Y)) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 + \mathbb{E}(Y^2) - (\mathbb{E}(Y))^2 + 2(\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)) = \mathbb{V}\text{ar}(X) + \mathbb{V}\text{ar}(Y) + 2\text{Cov}(X,Y)$,
(e) From part (d), $\mathbb{V}\text{ar}(X+Y) = \mathbb{V}\text{ar}(X) + \mathbb{V}\text{ar}(Y)$ if and only if $\text{Cov}(X,Y) = 0$ if and only if X and Y are uncorrelated. □

Exercise 18. Let $X_1, \ldots, X_n$ be independent random variables with finite second moments. Show that

$$\mathbb{V}\text{ar}(X_1 + \cdots + X_n) = \mathbb{V}\text{ar}(X_1) + \cdots + \mathbb{V}\text{ar}(X_n).$$

Exercise 19. Let $X_1, \ldots, X_n$ be random variables with finite second moments. Show that

$$\mathbb{V}\text{ar}(X_1 + \cdots + X_n) = \mathbb{V}\text{ar}(X_1) + \cdots + \mathbb{V}\text{ar}(X_n)$$

if $X_1, \ldots, X_n$ are pair-wise uncorrelated.

Theorem 44. (a) If a random variable X is bounded, then it must have finite variance.
(b) $\mathbb{V}\text{ar}(X) = 0$ if and only if $P(X = c) = 1$ for some $c \in \mathbb{R}$.

Proof. (a) Let $B > 0$ be a bound of X, that is, $P(|X| \le B) = 1$. Then $-B \le \mathbb{E}(X) \le B$ and $0 \le [X - \mathbb{E}(X)]^2 \le (2B)^2$. Hence $\mathbb{V}\mathrm{ar}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] \le 4B^2 < \infty$.
(b) Let $c = \mathbb{E}(X)$. Then $\mathbb{V}\mathrm{ar}(X) = \mathbb{E}[(X - c)^2] = 0$ and $P((X - c)^2 \ge 0) = 1$. Hence $P(X = c) = 1$. □

Inequalities

Theorem 45 (Chebychev's inequality)**.** Let X be a random variable with mean μ and variance σ^2. Then, for any $\alpha > 0$,

$$P(|X - \mu| \ge \alpha\sigma) \le \frac{1}{\alpha^2}.$$

Proof.

$$P(|X - \mu| \ge \alpha\sigma) = \int 1(|x - \mu| \ge \alpha\sigma)\, d\mathrm{cdf}_X(x) \le \int \Big(\frac{|x - \mu|}{\alpha\sigma}\Big)^2 1(|x - \mu| \ge \alpha\sigma)\, d\mathrm{cdf}_X(x) \le \mathbb{E}\Big(\frac{|X - \mu|}{\alpha\sigma}\Big)^2 = \frac{\sigma^2}{\alpha^2\sigma^2} = \frac{1}{\alpha^2}.$$

□

Example 79. Let X be a random variable having mean 3 and variance 1. What is the maximum of $P(X \ge 5)$?
Using Chebychev's inequality,

$$P(X \ge 5) = P(X - 3 \ge 2) \le P(|X - 3| \ge 2 \times 1) \le \frac{1}{2^2} = \frac{1}{4}.$$

Theorem 46 (Markov's inequality)**.** If $X \ge 0$ with $\mu = \mathbb{E}(X) < \infty$, then for any $\alpha > 0$

$$P(X \ge \alpha) \le \mu/\alpha.$$

Proof.

$$P(X \ge \alpha) = \int 1(x \ge \alpha)\, d\mathrm{cdf}_X(x) \le \int \frac{x}{\alpha} 1(x \ge \alpha)\, d\mathrm{cdf}_X(x) \le \mathbb{E}\Big(\frac{X}{\alpha}\Big) = \frac{\mu}{\alpha}.$$

□

Note. The Chebychev's inequality is a special case of Markov's inequality by considering $Y = (X - \mu)^2/\sigma^2$.

Theorem 47 (Cauchy-Schwartz' inequality)**.** Let X and Y be two random variables having finite second moment. Then

$$[\mathbb{E}(XY)]^2 \le \mathbb{E}(X^2)\mathbb{E}(Y^2)$$

where the equality holds if and only if $P(aX = bY) = 1$ for some $a, b \in \mathbb{R}$.

Proof. For any real number t,

$$0 \le \mathbb{E}[(tX - Y)^2] = t^2\mathbb{E}(X^2) - 2t\mathbb{E}(XY) + \mathbb{E}(Y^2) = \mathbb{E}(X^2)(t - \mathbb{E}(XY)/\mathbb{E}(X^2))^2 + \mathbb{E}(Y^2) - (\mathbb{E}(XY))^2/\mathbb{E}(X^2)$$

implies $\mathbb{E}(Y^2) \ge (\mathbb{E}(XY))^2/\mathbb{E}(X^2)$ with equality $P(tX + Y = 0) = 1$ for some t, that is, $aX = bY$ for some a, b. □

Example 80. The correlation between X and Y is between -1 and 1 inclusively. Note that $\mathrm{Corr}(X,Y) = \mathrm{Cov}(X,Y)/[\mathbb{V}\mathrm{ar}(X)\mathbb{V}\mathrm{ar}(Y)]^{1/2}$. By applying Cauchy-Schwartz' inequality for $X - \mathbb{E}(X)$ and $Y - \mathbb{E}(Y)$, we get

$$(\mathrm{Cov}(X,Y))^2 \le \mathbb{V}\mathrm{ar}(X)\mathbb{V}\mathrm{ar}(Y).$$

Hence $|\mathrm{Corr}(X,Y)| \le 1$ with equality holds if and only if $aX + bY = c$ for some a, b, c.

Example 81. If two random variables X and Y with finite second moment are independent, then X and Y are uncorrelated, that is, $\mathrm{Cov}(X,Y) = \mathrm{Corr}(X,Y) = 0$ because $\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(X)\mathbb{E}(Y) = 0$.

Theorem 48. Let X and Y be two random variables with finite second moment. Then $Y = aX + b$ for some a, b if and only if $|\mathrm{Corr}(X,Y)| = 1$.

Proof. ($\Longrightarrow$) Note $\mathrm{Cov}(X,Y) = \mathbb{E}(X(aX+b)) - \mathbb{E}(X)\mathbb{E}(aX+b) = a\mathbb{E}(X^2) + b\mathbb{E}(X) - [a(\mathbb{E}(X))^2 + b\mathbb{E}(X)] = a[\mathbb{E}(X^2) - (\mathbb{E}(X))^2] = a\mathbb{V}\mathrm{ar}(X)$ and $\mathbb{V}\mathrm{ar}(Y) = \mathbb{V}\mathrm{ar}(aX+b) = a^2\mathbb{V}\mathrm{ar}(X)$. Hence

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathbb{V}\mathrm{ar}(X)\mathbb{V}\mathrm{ar}(Y)}} = \frac{a\mathbb{V}\mathrm{ar}(X)}{\sqrt{\mathbb{V}\mathrm{ar}(X)a^2\mathbb{V}\mathrm{ar}(X)}} = \frac{a}{|a|} = \mathrm{sign}(a).$$

($\Longleftarrow$) From Cauchy-Schwartz' inequality, $|\mathrm{Corr}(X,Y)| = 1$ implies $P(a(X - \mathbb{E}(X)) = b(Y - \mathbb{E}(Y))) = 1$ for some a, b. Then $Y = (a/b)X + \mathbb{E}(Y) - (a/b)\mathbb{E}(X)$. Hence the theorem holds. □

Lemma 49 (Young's inequality)**.** For $p, q > 1$ with $1/p + 1/q = 1$ and two nonnegative real numbers $x, y \ge 0$,

$$xy \le x^p/p + y^q/q.$$

Proof. If $y = 0$, the inequality holds. Suppose $y > 0$. We prove $\varphi(x) = x^p/p - xy + y^q/q \ge 0$ for all $x \ge 0$. Note that $\varphi'(x) = x^{p-1} - y = 0$ has solution $x = y^{1/(p-1)} = y^{q-1}$. Hence $\varphi(x)$ has minimum at $x_0 = y^{q-1}$, that is,

$$\varphi(x) \ge \varphi(x_0) = \frac{y^{p(q-1)}}{p} - y^{q-1}y + \frac{y^q}{q} = \frac{y^q}{p} - y^q + \frac{y^q}{q} = y^q(\frac{1}{p} + \frac{1}{q} - 1) = 0.$$

Hence the inequality holds with equality if and only if $x^p = y^q$. □

Theorem 50 (Hölder's inequality)**.** For $p, q > 1$ with $1/p + 1/q = 1$, $\mathbb{E}|XY| \le ||X||_p ||Y||_q$ when the expectations exist and finite where $||X||_r = [\mathbb{E}(|X|^r)]^{1/r}$ for $r > 0$.

Proof. Let $U = X/||X||_p$ and $V = Y/||Y||_q$ so that $\mathbb{E}|U|^p = 1 = \mathbb{E}|V|^q$. Then,

$$\frac{\mathbb{E}|XY|}{||X||_p||Y||_q} = \mathbb{E}|UV| \le \mathbb{E}\Big(\frac{|U|^p}{p} + \frac{|V|^q}{q}\Big) = \frac{\mathbb{E}|U|^p}{p} + \frac{\mathbb{E}|V|^q}{q} = \frac{1}{p} + \frac{1}{q} = 1.$$

Hence the inequality holds with equality if and only if $|X|^p = c|Y|^q$ for some $c > 0$. □

Note. The Cauchy-Schwartz' inequality is a special case of Hölder's inequality ($p = q = 2$).

Definition 37. A function $\varphi : \mathbb{R} \to \mathbb{R}$ is said to be *convex* if $\varphi(\lambda x + (1-\lambda)y) \le \lambda\varphi(x) + (1-\lambda)\varphi(y)$ for any $0 \le \lambda \le 1$ and $x, y \in \mathbb{R}$. A function φ is said to be *concave* if $-\varphi$ is convex.

Example 82. For $x \geq 0$, functions $\varphi_p(x) = x^p$ is convex if $p \geq 1$ and concave if $p \leq 1$. The logarithm $\log(x)$ is concave while $\exp(x)$ is convex. Trigonometric functions $\sin(x)$, $\cos(x)$ and $\tan(x)$ are neither convex nor concave.

Exercise 20. If a function φ is both convex and concave, then it is a line of the form $\varphi(x) = ax + b$ for some $a, b \in \mathbb{R}$.

Exercise 21. If φ, ψ are convex and ψ is nondecreasing, then $\psi \circ \varphi$ is also convex.

Exercise 22. If f is twice differentiable and $f''(x) \geq 0$, then f is convex.

Exercise 23. Let φ be a convex function. (a) Show that there exist $a, b \in \mathbb{R}$ such that $\varphi(x) \geq ax + b$ for all $x \in \mathbb{R}$. Let $S = \{(a, b) \in \mathbb{R}^2 : \varphi(x) \geq ax + b \text{ for all } x \in \mathbb{R}\}$. (b) Show that $\varphi(x) = \sup_{(a,b)\in S}(ax + b)$.

Note. Roughly speaking, a convex function has the property that the function value of weighted average is not greater than the same weighted average of function values. This property can be generalized to random variables through the expectation.

Theorem 51 (Jensen's inequality)**.** For a convex function φ, $\varphi(\mathbb{E}(X)) \leq \mathbb{E}(\varphi(X))$ when expectations exist and are finite.

Proof. Note that for any $(a, b) \in S$, $\mathbb{E}[\varphi(X)] \geq \mathbb{E}[aX + b] = a\mathbb{E}(X) + b$. Hence $\mathbb{E}[\varphi(X)] \geq \sup_{(a,b)\in S}(a\mathbb{E}(X) + b) = \varphi(\mathbb{E}(X))$. □

Example 83 (Lyapounov's inequality)**.** If $\mathbb{E}(|X|^p) < \infty$ for $p > 0$, then $\mathbb{E}(|X|^q) \leq \{\mathbb{E}(|X|^p)\}^{q/p}$ for all $0 < q \leq p$. Note $\varphi(x) = x^{p/q}$ is a convex function. Hence,

$$\mathbb{E}(|X|^p) = \mathbb{E}(\varphi(|X|^q)) \geq \varphi(\mathbb{E}(|X|^q)) = \{\mathbb{E}(|X|^q)\}^{p/q}.$$

Exercise 24. Prove the following Minkowski's inequality.

Theorem (Minkowski's inequality)**.** For $p \geq 1$, $||X + Y||_p \leq ||X||_p + ||Y||_p$.

Conditional Expectation

Definition 38. The conditional expectation of Y given $X = x$ is defined by

$$\mathbb{E}(Y \,|\, X = x) = \int y \, d\mathrm{cdf}_{Y \,|\, X}(y \,|\, x).$$

Theorem 52. Assume $\mathbb{E}|Y| < \infty$. Then

$$\mathbb{E}(Y \,|\, X = x) = \int y \, d\mathrm{cdf}_{Y \,|\, X}(y \,|\, x) = \int_0^\infty P(Y > z \,|\, X = x)\, dz - \int_{-\infty}^0 P(Y < z \,|\, X = x)\, dz$$

and if Y is discrete, $\mathbb{E}(Y \,|\, X = x) = \sum_y y \times \mathrm{pmf}_{Y \,|\, X}(y \,|\, x)$; and if Y is continuous $\mathbb{E}(Y \,|\, X = x) = \int y \times \mathrm{pdf}_{Y \,|\, X}(y \,|\, x)\, dy$.

Proof of the theorem is very similar to unconditional cases and hence skipped.

The conditional expectation $\mathbb{E}(Y \,|\, X = x)$ is always a function of x, say $h(x)$. Then denote $h(X) = \mathbb{E}(Y \,|\, Y)$ is a random variable.

Theorem 53 (properties of conditional expectation)**.** (a) $\mathbb{E}(aY + b \mid X) = a\mathbb{E}(Y \mid X) + b$,
(b) If $P(Y \geq 0 \mid X) = 1$, then $\mathbb{E}(Y \mid X) \geq 0$.
(c) $\mathbb{E}(Y + Z \mid X) = \mathbb{E}(Y \mid X) + \mathbb{E}(Z \mid X)$.
(d) for constant random variable 1, $\mathbb{E}(1 \mid X) = 1$.

Proof of the theorem is again similar to unconditional cases and skipped at here.

Theorem 54. $\mathbb{E}[\mathbb{E}(Y \mid X)] = \mathbb{E}(Y)$.

Proof. Let $h(x) = \mathbb{E}(Y \mid X = x)$. Then

$$\begin{aligned}\mathbb{E}[\mathbb{E}(Y \mid X)] = \mathbb{E}[h(X)] &= \int h(x)\, d\mathrm{cdf}_X(x) = \int\int y\, d\mathrm{cdf}_{Y \mid X}(y \mid x)\, d\mathrm{cdf}_X(x) = \int_{\mathbb{R}^2} y\, d\mathrm{cdf}_{X,Y}(x, y) \\ &= \int\int y\, d\mathrm{cdf}_{X \mid Y}(x \mid y)\, d\mathrm{cdf}_Y(y) = \int y\, d\mathrm{cdf}_Y(y) = \mathbb{E}(Y).\end{aligned}$$

□

Definition 39. The conditional variance is given by

$$\mathbb{V}\mathrm{ar}(Y \mid X = x) = \mathbb{E}[(Y - \mathbb{E}(Y \mid X = x))^2 \mid X = x].$$

Exercise 25. Show that $\mathbb{V}\mathrm{ar}(Y \mid X) = \mathbb{E}(Y^2 \mid X) - [\mathbb{E}(Y \mid X)]^2$.

Exercise 26. Show that $\mathbb{E}[(Y - g(X))^2]$ is minimized by $g(x) = \mathbb{E}(Y \mid X = x)$.

Theorem 55. $\mathbb{V}\mathrm{ar}(Y) = \mathbb{E}[\mathbb{V}\mathrm{ar}(Y \mid X)] + \mathbb{V}\mathrm{ar}(\mathbb{E}(Y \mid X))$.

Proof. We show expansion of right hand side equals to left hand side. Let $h(x) = \mathbb{E}(Y \mid X = x)$. Then $\mathbb{E}(h(X)) = \mathbb{E}[\mathbb{E}(Y \mid X)] = \mathbb{E}(Y)$. Two terms in the right hand side are

$$\begin{aligned}\mathbb{E}[\mathbb{V}\mathrm{ar}(Y \mid X)] &= \mathbb{E}[\mathbb{E}(Y^2 \mid X) - (\mathbb{E}(Y \mid X))^2] = \mathbb{E}[\mathbb{E}(Y^2 \mid X)] - \mathbb{E}[(h(X))^2] = \mathbb{E}(Y^2) - \mathbb{E}[(h(X))^2], \\ \mathbb{V}\mathrm{ar}(\mathbb{E}(Y \mid X)] &= \mathbb{V}\mathrm{ar}(h(X)) = \mathbb{E}[(h(X))^2] - (\mathbb{E}(h(X)))^2 = \mathbb{E}[(h(X))^2] - (\mathbb{E}(Y))^2.\end{aligned}$$

Hence the right hand side becomes

$$\mathbb{E}[\mathbb{V}\mathrm{ar}(Y \mid X)] + \mathbb{V}\mathrm{ar}(\mathbb{E}(Y \mid X)) = \mathbb{E}(Y^2) - \mathbb{E}[(h(X))^2] + \mathbb{E}[(h(X))^2] - (\mathbb{E}(Y))^2 = \mathbb{E}(Y^2) - (\mathbb{E}(Y))^2 = \mathbb{V}\mathrm{ar}(Y).$$

Therefore the theorem holds. □

Probability Related Functions

Definition 40. Let X be a random variable. The *moment generating function (mgf)* is $\mathrm{mgf}_X(t) = \mathbb{E}(e^{tX})$ and the logarithm of moment generating function is called the *cumulant generating function*, that is, $\mathrm{cgf}_X(t) = \log \mathbb{E}(e^{tX})$. Similarly, the *probability generating function* is defined as $\mathrm{pgf}_X(z) = \mathbb{E}(z^X)$ and the *characteristic function* is defined as $\mathrm{chf}_X(t) = \mathbb{E}(e^{itX})$ where $i = \sqrt{-1}$ is the unit imaginary number.

Note. Characteristic function can be defined for any random variable while moment/cumulant/probability generating functions may not exists for some random variables like Cauchy distributions. Note that $\mathrm{chf}_X(t) = \mathrm{mgf}_X(it)$ and $\mathrm{pgf}_X(s) = \mathrm{mgf}_X(\log s)$.

Note. Since pgf/mgf/cgf/chf are defined through expectation, two random variables having the same distribution must have the same pgf/mgf/cgf/chf if it exists. Also the converse is true, but, a proof will appear later.

Example 84. Let Z be a random variable following the standard normal distribution. Then $\mathrm{pdf}_Z(x) = (2\pi)^{-1/2}\exp(-x^2/2)$, $\mathrm{cdf}_Z(x) = \Phi(x)$, $\mathrm{mgf}_Z(t) = \mathbb{E}e^{tZ} = e^{t^2/2}$, $\mathrm{cgf}_Z(t) = t^2/2$, $\mathrm{pgf}_Z(z) = e^{(\log z)^2/2} = z^{\log(z)/2}$ and $\mathrm{chf}_Z(t) = e^{-t^2/2}$.

Theorem 56. (a) $\mathrm{pgf}_X(1) = 1$, $\mathrm{mgf}_X(0) = 1$, $\mathrm{cgf}_X(0) = 0$, $\mathrm{chf}_X(0) = 1$.
(b) $\mathbb{E}(X(X-1)\cdots(X-k+1)) = \frac{d^k}{dz^k}\mathrm{pgf}_X(1)$ if it exists
(c) $\mathbb{E}(X^k) = \frac{d^k}{dt^k}\mathrm{mgf}_X(0)$ if it exists
(d) $\mathbb{E}(X^k) = (i)^{-k}\frac{d^k}{dt^k}\mathrm{chf}_X(0)$ if it exists
(e) If $\mathbb{E}(|X|^k) < \infty$, then, for $\mu_j = \mathbb{E}(X^j)$ where $j = 1, \ldots, k$,

$$\mathrm{mgf}_X(t) = 1 + \mu_1 t + \mu_2\frac{t^2}{2!} + \cdots + \mu_k\frac{t^k}{k!} + o(|t|^k) \quad \text{as well as}$$
$$\mathrm{chf}_X(t) = 1 + i\mu_1 t - \mu_2\frac{t^2}{2!} + \cdots + i^k\mu_k\frac{t^k}{k!} + o(|t|^k).$$

Proof. (a) $\mathrm{pgf}_X(1) = \mathbb{E}[1^X] = 1$, $\mathrm{mgf}_X(0) = \mathbb{E}[e^{0X}] = 1$, $\mathrm{cgf}_X(0) = \log\mathbb{E}[e^{0X}] = 0$, $\mathrm{chf}_X(0) = \mathbb{E}[e^{i0X}] = 1$.
(b) $\frac{d^k}{ds^k}\mathbb{E}(s^X) = \frac{d^{k-1}}{ds^{k-1}}\mathbb{E}(Xs^{X-1}) = \cdots = \mathbb{E}[X(X-1)\cdots(X-k+1)s^{X-k}]$. Hence $\frac{d^k\mathrm{pgf}_X}{ds^k}(1) = \mathbb{E}[X(X-1)\cdots(X-k+1)1^{X-k}]$.
(c) Similarly $\frac{d^k}{dt^k}\mathbb{E}(e^{tX}) = \frac{d^{k-1}}{dt^{k-1}}\mathbb{E}(Xe^{tX}) = \cdots = \mathbb{E}[X^k e^{tX}]$ implies $\frac{d^k\mathrm{mgf}_X}{dt^k}(0) = \mathbb{E}[X^k e^{0X}] = \mathbb{E}[X^k]$.
(d) Finally $\frac{d^k}{dt^k}\mathbb{E}[e^{itX}] = \mathbb{E}[(iX)^k e^{itX}]$ implies $\frac{d^k\mathrm{chf}_X}{dt^k}(0) = i^k\mathbb{E}[X^k e^{i\cdot 0\cdot X}] = i^k\mathbb{E}[X^k]$.
(e) The results can be obtained using mean value theorem stated below. □

Theorem 57 (Mean value theorem)**.** Let f be a k times continuously differentiable function. Then

$$f(a+b) = f(a) + f^{(1)}(a)b + \cdots + f^{(k)}(a)\frac{b^k}{k!} + \frac{b^k}{(k-1)!}\int_0^1 (1-x)^{k-1}[f^{(k)}(a+bx) - f^{(k)}(a)]\,dx.$$

Proof. Note that $f(a+b) - f(a) = \int_a^{a+b} f'(x)\,dx = b\int_0^1 f'(a+bx)\,dx = bf'(a) + b\int_0^1[f'(a+bx) - f'(a)]\,dx$. Inductively,

$$\begin{aligned}
\int_0^1 (1-x)^{m-1}[f^{(m)}(a+bx) - f^{(m)}(a)]\,dx &= \int_0^1 (1-x)^{m-1} b\int_0^x f^{(m+1)}(a+by)\,dy\,dx \\
&= b\int_0^1\int_y^1 (1-x)^{m-1}\,dx f^{(m+1)}(a+by)\,dy = \frac{b}{m}\int_0^1 (1-x)^m f^{(m+1)}(a+bx)\,dx \\
&= \frac{b}{m}[f^{(m+1)}(a)\int_0^1 (1-x)^m\,dx + \int_0^1 (1-x)^m[f^{(m+1)}(a+bx) - f^{(m+1)}(a)]\,dx] \\
&= \frac{b}{m(m+1)}f^{(m+1)}(a) + \frac{b}{m}\int_0^1 (1-x)^m[f^{(m+1)}(a+bx) - f^{(m+1)}(a)]\,dx.
\end{aligned}$$

Hence $f(a+b) = f(a) + bf'(a)/1! + \cdots + b^k f^{(k)}(a)/k! + (b^k/(k-1)!)\int_0^1(1-x)^{k-1}[f^{(k)}(a+bx) - f^{(k)}(a)]\,dx$. □

Example 85. If $\mathbb{E}(|X|^k) < \infty$, then for small t,

$$\text{mgf}(t) = 1 + \mu_1 t + \mu_2 \frac{t^2}{2!} + \cdots + \mu_k \frac{t^k}{k!} + \frac{t^k}{(k-1)!} \int_0^1 (1-x)^{k-1} [\text{mgf}^{(k)}(tw) - \text{mgf}^{(k)}(0)]\, dw.$$

Hence $\text{mgf}(t) = 1 + \mu_1 t + \mu_2 \frac{t^2}{2!} + \cdots + \mu_k \frac{t^k}{k!} + o(|t|^k)$.

It is expected that the existence of finite moments implies the uniqueness of the distribution. However, there are many counterexamples like Problem 5.12.43 in GS.

Example 86. Let $X \sim \text{Poisson}(\mu)$, that is $\text{pmf}_X(x) = e^{-\mu}\mu^x/x!$ and $\text{mgf}_X(t) = \mathbb{E}e^{tX} = \sum_{x=0}^{\infty} e^{tx} e^{-\mu}\mu^x/x! = e^{-\mu+\mu e^t}$.

Let $Y \sim \text{gamma}(\alpha, \beta)$, that is $\text{pdf}_Y(y) = y^{\alpha-1}e^{-y/\beta}/\Gamma(\alpha)\beta^\alpha$ and $\text{mgf}_Y(t) = \mathbb{E}e^{tY} = \int_0^\infty e^{ty} y^{\alpha-1}e^{-y/\beta}/\Gamma(\alpha)\beta^\alpha\, dy = (1/\beta - t)^{-\alpha}/\beta^\alpha = (1-t\beta)^\alpha$

Let $Z \sim N(\mu, \sigma^2)$, then $\text{mgf}_Z(t) = \int_{-\infty}^{\infty} e^{tx} e^{-(x-\mu)^2/2\sigma^2}/(2\pi\sigma^2)^{1/2}\, dz = e^{\mu t + \sigma^2 t^2/2}$.

Theorem 58 (Properties)**.** (a) $\text{mgf}_{aX+b}(t) = e^{bt}\text{mgf}_X(at)$ and $\text{chf}_{aX+b}(t) = e^{ibt}\text{chf}_X(at)$,
(b) If X and Y are independent, then $\text{mgf}_{X,Y}(s,t) = \text{mgf}_X(s)\text{mgf}_Y(t)$ as well as $\text{chf}_{X,Y}(s,t) = \text{chf}_X(s)\text{chf}_Y(t)$.

Proof. (a) $\text{mgf}_{aX+b}(t) = \mathbb{E}[e^{t(aX+b)}] = e^{bt}\mathbb{E}[e^{(at)X}] = e^{bt}\text{mgf}_X(at)$ and $\text{chf}_{aX+b}(t) = \text{mgf}_{aX+b}(it) = e^{ibt}\text{chf}_X(at)$.
(b) If X and Y are independent, then $\text{mgf}_{X,Y}(s,t) = \mathbb{E}[e^{sX+tY}] = \mathbb{E}[e^{sX}e^{tY}] = \mathbb{E}[e^{sX}]\mathbb{E}[e^{tY}] = \text{mgf}_X(s)\text{mgf}_Y(t)$.
Similarly $\text{chf}_{X,Y}(s,t) = \mathbb{E}[e^{isX+itY}] = \mathbb{E}[e^{isX}]\mathbb{E}[e^{itY}] = \text{chf}_X(s)\text{chf}_Y(t)$. □

Theorem 59. If two random variables X and Y have the same moment generating functions in an open neighborhood of 0, that is, $(-a, b)$ for $a, b > 0$, then X and Y are identically distributed.

This theorem will be proved later using characteristic functions.

Example 87. Let $X, X_1, \ldots, X_n \sim i.i.d.$ Bernoulli(p) and $Y \sim \text{binomial}(n, p)$. Let $Z = X_1 + \cdots + X_n$. The moment generating functions of X, Y, Z are

$$\begin{aligned}
\text{mgf}_X(t) &= \mathbb{E}(e^{tX}) = e^{t0}P(X=0) + e^{t1}P(X=1) = 1 - p + pe^t, \\
\text{mgf}_Y(t) &= \sum_{j=0}^{n} e^{tj} \cdot \binom{n}{j} p^j (1-p)^{n-j} = \sum_{j=0}^{n} \binom{n}{j} (pe^t)^j (1-p)^{n-j} = (1 - p + pe^t)^n, \\
\text{mgf}_Z(t) &= \mathbb{E}(e^{t(X_1+\cdots+X_n)}) = \mathbb{E}(e^{tX_1}) \cdots \mathbb{E}(e^{tX_n}) = [\mathbb{E}(e^{tX})]^n = (1 - p + pe^t)^n.
\end{aligned}$$

Hence Y and Z have the same moment generating function. Therefore binomial distributions are sum of i.i.d. Bernoulli distributions.

Exercise 27. Let $X_i \sim ind.$ binomial(n_i, p). Show that $X_1 + \cdots + X_k \sim \text{binomial}(n_1 + \cdots + n_k, p)$.

Exercise 28. Suppose two random variables X and Y have the same distribution. If $\mathbb{E}(|g(X)|) < \infty$, then $\mathbb{E}(g(X)) = \mathbb{E}(g(Y))$.

Exercise 29. Show that two random variables X and Y have the same distribution if $\mathbb{E}(g(X)) = \mathbb{E}(g(Y))$ for any bounded continuous function g.

Probabilities of random variables can be expressed through some functions like cumulative distribution functions. There are a few more summary functions like the probability mass/density functions.

Theorem 60. Let φ be a characteristic function of a random variable X.
(a) $\varphi(0) = 1$,
(b) $|\varphi(t)| \leq 1$ for all t,
(c) $\varphi(t)$ is uniformly continuous,
(d) for any $t_1, \ldots, t_n \in \mathbb{R}$ and $z_1, \ldots, z_n \in \mathbb{C}$, $\sum_{j,k} \varphi(t_j - t_k) z_j \overline{z}_k \geq 0$.

Proof. (a) $\varphi(0) = \mathbb{E}[e^{i0X}] = \mathbb{E}[e^0] = 1$.
(b) $|\varphi(t)| = |\mathbb{E}[e^{itX}]| \leq \mathbb{E}[|e^{itX}|] = \mathbb{E}[1] = 1$.
(c) Let $Y_h = |e^{ihX} - 1| \leq 2$. As $h \to 0$, $Y_h \to 0$ a.s. The difference becomes

$$\varphi(t+h) - \varphi(t)| = |\mathbb{E}[e^{i(t+h)X} - e^{itX}]| = |\mathbb{E}[e^{itX}(e^{ihX} - 1)]| \leq \mathbb{E}[|e^{ihX} - 1|] = \mathbb{E}[Y_h] \to 0$$

as $h \to 0$ by the bounded convergence theorem. Hence for any $\epsilon > 0$, there exists $\delta > 0$ such that $\mathbb{E}[Y_h] < \epsilon$ for $0 < h < \delta$. Thus $|\varphi(t+h) - \varphi(t)| \leq \mathbb{E}[Y_h] < \epsilon$ for any t. Which completes the proof
(d) The sum of terms can be expressed with respect to expectations like

$$\sum_{j,k=1}^{n} \varphi(t_j - t_k) z_j \overline{z}_k = \sum_{j,k=1}^{n} \mathbb{E}[e^{i(t_j - t_k)X} z_j \overline{z}_k] = \sum_{j,k=1}^{n} \mathbb{E}[e^{it_j X} z_j \overline{e^{it_k X} z_k}] = \mathbb{E}\Big[\sum_{j=1}^{n} e^{it_j X} z_j \overline{\sum_{k=1}^{n} e^{it_k X} z_k}\Big] = \mathbb{E}\Big[\Big|\sum_{j=1}^{n} e^{it_j X} z_j\Big|^2\Big] \geq 0.$$

□

Theorem 61. If a function $\varphi : \mathbb{R} \to \mathbb{C}$ satisfies (a)-(d) in Theorem 60, then there exists a random variable having φ as its characteristic function.

The proof is beyond our scope so it is skipped. If you are interested in, you may search Bochner's theorem.

Definition 41. The joint probability/moment/cumulant generating and characteristic functions of X and Y are $\mathrm{pgf}_{X,Y}(s,t) = \mathbb{E}[s^X t^Y]$, $\mathrm{mgf}_{X,Y}(s,t) = \mathbb{E}[e^{sX+tY}]$, $\mathrm{cgf}_{X,Y}(s,t) = \log \mathrm{mgf}_{X,Y}(s,t)$, and $\mathrm{chf}_{X,Y}(s,t) = \mathbb{E}[e^{isX+itY}]$.

Theorem 62 (Inversion Formula)**.** Let φ be a characteristic function of a random variable X. Then for any a, b,

$$P(a < X < b) + \{P(X = a) + P(X = b)\}/2 = \lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-iat} - e^{-ibt}}{it} \varphi(t)\, dt.$$

A proof of the above theorem requires the sine integral function given by

$$\frac{2}{\pi} \int_0^T \frac{\sin(tx)}{t}\, dt \to \mathrm{sign}(x) \quad \text{as } T \to \infty.$$

where $\mathrm{sign}(x) = +1$ if $x > 0$; $= -1$ if $x < 0$; $= 0$ if $x = 0$.

Proof. Note that

$$\frac{1}{\pi}\int_{-T}^{T}\frac{e^{-iat}-e^{-ibt}}{it}\varphi(t)\,dt=\frac{1}{\pi}\int_{-T}^{T}\frac{e^{-iat}-e^{-ibt}}{it}\int e^{itx}\,P(dx)\,dt=\int\frac{1}{\pi}\int_{-T}^{T}\frac{e^{-iat}-e^{-ibt}}{it}e^{itx}\,dt\,P(dx)$$
$$=\int\frac{1}{\pi}\int_{-T}^{T}\frac{e^{it(x-a)}-e^{it(x-b)}}{it}\,dt\,P(dx)=\int\frac{2}{\pi}\int_{0}^{T}\frac{\sin(t(x-a))-\sin(t(x-b))}{t}\,dt\,P(dx)$$
$$\to\int(\operatorname{sign}(x-a)-\operatorname{sign}(x-b))\,P(dx)=2P(a<X<b)+P(X=a)+P(X=b).$$

Hence the theorem follows. □

Theorem 63. Two random variables X and Y are independent if and only if one of the following holds
(a) $\mathrm{cdf}_{X,Y}(x,y)=\mathrm{cdf}_X(x)\mathrm{cdf}_Y(y)$,
(b) $\mathrm{pmf}_{X,Y}(x,y)=\mathrm{pmf}_X(x)\mathrm{pmf}_Y(y)$,
(c) $\mathrm{pdf}_{X,Y}(x,y)=\mathrm{pdf}_X(x)\mathrm{pdf}_Y(y)$,
(d) $\mathrm{pgf}_{X,Y}(s,t)=\mathrm{pgf}_X(s)\mathrm{pgf}_Y(t)$,
(e) $\mathrm{mgf}_{X,Y}(s,t)=\mathrm{mgf}_X(s)\mathrm{mgf}_Y(t)$,
(f) $\mathrm{cgf}_{X,Y}(s,t)=\mathrm{cgf}_X(s)+\mathrm{cgf}_Y(t)$,
(h) $\mathrm{chf}_{X,Y}(s,t)=\mathrm{chf}_X(s)\mathrm{chf}_Y(t)$.

Proof. It is already shown that the independence is equivalent to one of (a)-(c). The results for (d)-(h) holds as long as (h) holds. The independence of X and Y implies $\mathrm{chf}_{X,Y}(s,t)=\mathbb{E}[e^{isX+itY}]=\mathbb{E}[e^{isX}e^{itY}]=\mathbb{E}[e^{isX}]\mathbb{E}[e^{itY}]=\mathrm{chf}_X(s)\mathrm{chf}_Y(t)$. The converse requires inversion formula. Choose $x_1<x_2$ and $y_1<y_2$ with $P(X=x_1)=P(X=x_2)=P(Y=y_1)=P(Y=y_2)=0$. Then

$$P(x_1<X<x_2,y_1<Y<y_2)=\lim_{T\to\infty}\frac{1}{(2\pi)^2}\int_{-T}^{T}\int_{-T}^{T}\frac{e^{-ix_1s}-e^{-ix_2s}}{is}\frac{e^{-iy_1t}-e^{-iy_2t}}{it}\varphi_{X,Y}(s,t)\,ds\,dt$$
$$=\lim_{T\to\infty}\frac{1}{(2\pi)^2}\int_{-T}^{T}\int_{-T}^{T}\frac{e^{-ix_1s}-e^{-ix_2s}}{is}\frac{e^{-iy_1t}-e^{-iy_2t}}{it}\mathrm{chf}_X(s)\mathrm{chf}_Y(t)\,ds\,dt$$
$$=\lim_{T\to\infty}\frac{1}{2\pi}\int_{-T}^{T}\frac{e^{-ix_1s}-e^{-ix_2s}}{is}\mathrm{chf}_X(s)\,ds\lim_{T\to\infty}\frac{1}{2\pi}\int_{-T}^{T}\frac{e^{-iy_1t}-e^{-iy_2t}}{it}\mathrm{chf}_Y(t)\,dt$$
$$=P(x_1<X<x_2)P(y_1<Y<y_2).$$

Hence the decomposition of joint characteristic function implies the independence of random variables. □

Example 88. Let $X\sim\text{binomial}(n,p)$ and $Y\sim\text{binomial}(m,p)$ be independent. The characteristic functions are $\mathrm{chf}_X(t)=\mathbb{E}[e^{itX}]=\sum_{k=0}^{n}e^{itk}\binom{n}{k}p^k(1-p)^{n-k}=(1-p+pe^{it})^n$. Similarly, $\mathrm{chf}_Y(t)=(1-p+pe^{it})^m$. The independence implies

$$\mathrm{chf}_{X+Y}(t)=\mathbb{E}[e^{it(X+Y)}]=\mathbb{E}[e^{itX}e^{itY}]=\mathbb{E}[e^{itX}]\mathbb{E}[e^{itY}]=\mathrm{chf}_X(t)\mathrm{chf}_Y(t)=(1-p+pe^{it})^n(1-p+pe^{it})^m$$
$$=(1-p+pe^{it})^{n+m}.$$

Hence $X+Y\sim\text{binomial}(n+m,p)$.

Exercise 30. (a) Let $X_i\sim\text{Poisson}(\mu_i)$ independently. Show that $X_1+\cdots+X_n\sim\text{Poisson}(\mu_1+\cdots+\mu_n)$
(b) Let $X_i\sim\text{neg-bin}(n_i,p)$ independently. Show that $X_1+\cdots+X_n\sim\text{neg-bin}(n_1+\cdots+n_n,p)$

(c) Let $X_i \sim \text{gamma}(\alpha_i, \beta)$ independently. Show that $X_1 + \cdots + X_n \sim \text{gamma}(\alpha_1 + \cdots + \alpha_n, \beta)$
(d) Let $X_i \sim N(\mu_i, \sigma_i^2)$ independently. Show that $X_1 + \cdots + X_n \sim N(\mu_1 + \cdots + \mu_n, \sigma_1^2 + \cdots + \sigma_n^2)$.

Chernoff Bound

Theorem 64 (Chernoff Bound)**.** Let X a random variable having moment generating function. For any constant x,

$$P(X \geq x) \leq \inf_{t>0} e^{-xt} \text{mgf}_X(t).$$

Proof. The event $X \geq x$ is equivalent to $tX \geq tx$ for $t > 0$ which is again equivalent to $e^{tX} \geq e^{tx}$ for $t > 0$. Using Markov's inequality, we get

$$P(X \geq x) = P(e^{tX} \geq e^{tx}) \leq \mathbb{E}(e^{tX})/e^{tx} = e^{-tx} \text{mgf}_X(t).$$

The positive constant t is arbitrary. Hence t is chosen to obtain smallest upper bound. □

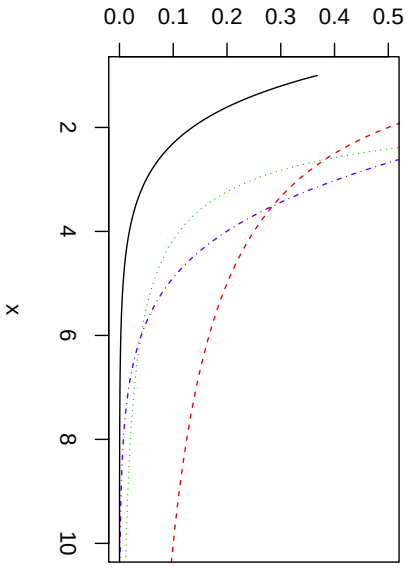
Example 89. Let $X \sim$ Exponential(λ). Then $\text{mgf}_X(t) = \mathbb{E}[e^{tX}] = \int_0^\infty e^{tx} \lambda e^{-\lambda x}\, dx = \lambda/(\lambda - t) = (1 - t/\lambda)^{-1}$. For any $x > 1/\lambda$ and $0 < t < \lambda$,

$$P(X \geq x) = P(e^{tX} \geq e^{tx}) \leq \inf_{0<t<\lambda} e^{-tx}(1 - t/\lambda)^{-1}.$$

By solving $\frac{d}{dt} \log[e^{-tx}(1 - t/\lambda)^{-1}] = -x - (1 - t/\lambda)^{-1}(-1/\lambda = 0$, the minimizing t is $t = \lambda - 1/x$ and the corresponding upper bound is $(e\lambda x)e^{-\lambda x}$ which is close to the true value $e^{-\lambda x}$.

Markov's inequality implies $P(X \geq x) \leq \mathbb{E}(X)/x = 1/(\lambda x)$ and Chebychev's inequality is $P(X \geq$

$x - 1/\lambda + 1/\lambda) \leq \mathbb{V}\mathrm{ar}(X)/(x - 1/\lambda)^2 = (\lambda x - 1)^{-2}$.



black solid : true
red dashed : Markov
green dotted : Chebychev
blue dot-dash : Chernoff

Markov's inequality uses the first moment while Chebychev's inequality uses both first and second moments. That is why Chebychev's inequality is more accurate at the tail. Since moment generating function contains all moments, Chernoff's inequality (or bound) uses all moment and much more accurate in the tail.

Exercise 31. Let $X \sim N(\mu, \sigma^2)$. For any $x > \mu$, compute and compare the upper bounds of Chebychev's/Markov's/Chernoff's inequalities/bounds.

Exercise 32. For $X \sim \text{Binomial}(n, p)$ and $0 < \alpha < 1/p - 1$, show that

$$P(X \geq (1+\alpha)np) \leq \left[\left(\frac{p}{\beta}\right)^{\beta}\left(\frac{1-p}{1-\beta}\right)^{1-\beta}\right]^n$$

where $\beta = (1+\alpha)p < 1$.

Exercises. (DS) 4.1.4, 4.1.5, 4.1.6, 4.1.8, 4.1.11, 4.1.15, 4.2.2, 4.2.6, 4.2.8, 4.2.9, 4.2.11, 4.5.3, 4.5.6, 4.5.12, 4.5.16, 4.5.17, 4.5.18, 4.3.4, 4.3.5, 4.3.6, 4.3.7, 4.3.9, 4.3.12, 4.6.2, 4.6.5, 4.6.8, 4.6.10, 4.6.13, 4.6.18, 4.7.3,

4.7.6, 4.7.13, 4.7.16, 4.4.4, 4.4.5, 4.4.9, 4.4.11, 4.4.14, 4.4.16, 4.9.1, 4.9.9, 4.9.10, 4.9.14, 4.9.15, 4.9.18, 4.9.26; (GS) 3.3.1, 3.3.5, 3.4.7, 3.4.9, 3.5.2, 3.5.3, 3.6.3, 3.6.4, 3.6.5, 3.6.8, 3.7.1, 3.7.4, 3.7.6, 3.7.9, 3.8.3, 3.8.6, 3.9.1, 3.9.2, 3.10.2, 3.11.2, 3,11.5, 3.11.6, 3.11.7, 3.11.8, 3.11.13, 3.11.16, 3.11.26, 3.11.32, 4.2.2, 4.2.4, 4.3.1, 4.3.3, 4.4.1, 4.4.5, 4.4.8, 4.5.1, 4.5.5, 4.5.,7, 4.5.8, 4.6.3, 4.6.4, 4.6.7, 4.6.8, 4.6.9, 4.6.10, 4.7.1, 4.7.5, 4.7.7, 4.7.10, 4.8.3, 4.8.6, 4.8.7, 4.14.3, 4.14.4, 4.14.5, 4.14.7, 4.14.12, 4.14.19, 4.14.22, 4.14.39, 4.14.58, 5.1.2, 5.1.3, 5.1.4, 5.2.3, 5.2.8, 5.4.1, 5.6.2, 5.6.4, 5.7.3, 5.7.7, 5.8.2, 5.8.5, 5.8.7; (Ri) 2.43, 2.44, 2.46, 2.48, 2.49, 2.52, 2.54, 2.55, 2.57, 2.60, 2.61, 2.62, 2.66, 68 2.69, 2.73, 2.76, 2.80, 2.84, 2.85, 2.86, 3.8, 3.9, 3.10, 3.11, 3.12, 3.13, 3.15, 3.16, 3.18, 3.19, 3.21, 3.25, 3.30, 3.36, 3.37, 3.43, 3.46, 3.47, 3.56, 3.58; (RM) 4.7.5, 4.7.7, 4.7.8, 4.7.12, 4.7.13, 4.7.14, 4.7.18, 4.7.19, 4.7.22, 4.7.25, 4.7.26, 4.7.27, 4.7.30, 4.7.31, 4.7.34, 4.7.39, 4.7.42, 4.7.43, 4.7.46, 4.7.48, 4.7.49, 4.7.50, 4.7.52, 4.7.53, 4.7.54, 4.7.60, 4.7.61, 4.7.64, 4.7.68, 4.7.70, 4.7.74, 4.7.77, 4.7.80, 4.7.89, 4.7.94, 4.7.96, 4.7.97, 4.7.99, 4.7.101, 4.7.103;