Chap 2 Note: Random Variables

1 Random Variables and CDF

Example Suppose that independent trials, each of which results in any of m possible outcomes with respective probabilities $p_1, p_2, \cdots, p_m$, $\sum_{i=1}^{m} p_i = 1$, are continually performed. Let X denote the number of trials needed until each outcome has occurred at least once.

Explain Rather than directly considering $P\{X = n\}$ we will first determine $P\{X > n\}$, the probability that at least one of the outcomes has not yet occurred after n trails. Letting A_i denote the event that outcome i has not yet occurred after the first n trails, $i = 1, \cdots, m$, then

$$\begin{aligned} P\{x > n\} =& P(\cup_{i=1}^{m} A_i) \\ =& \sum_{i=1}^{m} P(A_i) - \sum\sum_{i<j} P(A_i A_j) + \cdots \\ &+ \sum\sum\sum_{i<j<k} P(A_i A_j A_k) - \cdots + (-1)^{m+1} P(A_1 \cdots A_n) \end{aligned}$$

$P(A_i)$ is the probability that each of the first n trails results in a non-i outcome, and so by independence $P(A_i) = (1 - p_i)^n$

Similarly, $P(A_i A_j)$ is the probability that the first n trails all result in a non-i and non-j outcome, and so

$$P(A_i A_j) = (1 - p_i - p_j)^n$$

As all of the other probabilities are similar, we see that

$$\begin{aligned} P\{X > n\} =& \sum_{i=1}^{m} (1 - p_i)^n - \sum\sum_{i<j} (1 - p_i - p_j)^n \\ &+ \sum\sum\sum_{i<j<k} (1 - p_i - p_j - p_k)^n - \cdots \end{aligned}$$

Since $P\{X = n\} = P\{X > n - 1\} - P\{x > n\}$, we see, upon using the algebraic identity $(1 - a)^{n-1} - (1 - a)^n = a(1 - a)^{n-1}$, that

$$\begin{aligned} P\{X = n\} =& \sum_{i=1}^{m} p_i * (1 - p_i)^{n-1} - \sum\sum_{i<j} (p_i + p_j) * (1 - p_i - p_j) \\ &+ \sum\sum\sum_{i<j<k} (p_i + p_j + p_k)(1 - p_i - p_j - p_k)^{n-1} - \cdots \end{aligned}$$

CDF The cumulative distribution function(cdf) $F(.)$ of the random variable X is defined for any real number b, $-\infty < b < \infty$, by

$$F(b) = P\{X \leq b\}$$

We have

$$P\{a < X \leq b\} = F(b) - F(a), \text{for all } a < b$$
$$P\{X < b\} = \lim_{h \to 0^+} P\{X \leq b - h\} = \lim_{h \to 0^+} F(b - h)$$

2 Discrete Random Variables

2.1 The Bernoulli Random Variable

Suppose that a trail, or an experiment, whose outcome can be classified as either a "success" or as a "failure" is performed. If we let X equal 1 if the outcome is a success and 0 if it is a failure, then the probability mass function of X is given by

$$p(0) = P\{X = 0\} = 1 - p$$
$$p(1) = P\{X = 1\} = p$$

where $p, 0 \leq p \leq 1$, is the probability that the trail is a "success".

A random variable X is said to be a *Bernoulli random variable* if its probability mass function is given by the above equation for $p \in (0, 1)$.

2.2 The Binomial Random Variable

Suppose that n independent trails, each of which results in a "success" with probability p and in a "failure" with probability $1 - p$, are to be performed. If X represents the number of successes that occur in the n trails, then X is said to be a *binomial random variable* with parameters (n, p).

The probability mass function of a binomial random variable having parameters (n, p) is given by

$$p(i) = \binom{n}{i} p^i * (1 - p)^{n-i}, i = 0, 1, \cdots, n$$

where

$$\binom{n}{i} = \frac{n!}{(n - i)! * i!}$$

Note that, by the binomial theorem, the probabilities sum to one, that is,

$$\sum_{i=0}^{\infty} p(i) = \sum_{i=0}^{n} p^i * (1 - p)^{n-i} = (1 + (1 - p))^n = 1$$

2.3 The Geometric Random Variable

Suppose that independent trails, each having probability p of being a success, are performed until a success occurs. If we let X be the number of trails required until the first success, then X is said to be *geometric random variable* with parameter p. Its probability mass function is given by

$$p(n) = P\{X = n\} = (1 - p)^{n-1} * p, n = 1, 2, \cdots$$

To check that $p(n)$ is a probability mass function, we note that

$$\sum_{n=1}^{\infty} p(n) = \sum_{n=1}^{\infty} (1 - p)^{n-1} = 1$$

2.4 The Poisson Random Variable

A random variable X, taking on one of the values $0, 1, 2, \cdots$ is said to be a Poisson random variable with parameter λ, if for some $\lambda \geq 0$,

$$p(i) = P[X = i] = e^{-\lambda} \frac{\lambda^i}{i!}, i = 0, 1, \cdots$$

The above equation defines a probability mass function since

$$\sum_{i=0}^{\infty} p(i) = e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = e^{-\lambda} * e^{\lambda} = 1$$

Approximate a Binomial Random Variable by Poisson An important property of the Poisson random variable is that it may e used to approximate a binomial random variable when the binomial parameter n is large and p is small. To see this, suppose that X is a binomial random variable with parameters (n, p) and let $\lambda = n * p$. Then

$$\begin{aligned}
PX = i =& \frac{n!}{(n-i)! * i!} p^i * (1-p)^{n-i} \\
=& \frac{n!}{(n-i)! * i!} * (\lambda/n)^i (1 - \lambda/n)^{n-i} \\
=& \frac{n * (n-1) * \cdots * (n-i+1)}{n^i} * \frac{\lambda^i}{i!} * \frac{(1-\lambda/n)^n}{(1-\lambda/n)^i}
\end{aligned}$$

For n large and p small,

$$\begin{aligned}
&(1-\lambda/n)^n \approx e^{-\lambda} \\
&\frac{n * (n-1) * \cdots * (n-i+1)}{n^i} \approx 1 \\
&(1-\lambda/n)^i \approx 1
\end{aligned}$$

Hence for n large and p small, we have

$$P\{X = i\} \approx e^{-\lambda} * \frac{\lambda^i}{i!}$$

3 Continuous Random Variables

Probability Density Function $f(x)$ is called the probability density function, which is a derivative of cumulative distribution function(CDF). We have

$$\begin{aligned}
&P\{a \leq X \leq b\} = \int_a^b f(x)dx \\
&P\{X = a\} = \int_a^a f(x)dx = 0 \\
&P\{a - \epsilon/2 \leq X \leq a + \epsilon/2\} = \int_{a-\epsilon/2}^{a+\epsilon/2} f(x)dx \approx \epsilon * f(a)
\end{aligned}$$

In other words, the probability that X will be contained in an interval of length ϵ around the point a is approximately $\epsilon * f(a)$. From this, we see that $f(a)$ is a measure of how likely it is that the random variable will be near a.

3.1 The Uniform Random Variable

A random variable is said to be uniformly distributed over the interval $(0, 1)$ if its probability density is given by

$$f(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & otherwise. \end{cases}$$

Note that the preceding is a density function since $f(x) \geq 0$ and

$$\int_{-\infty}^{\infty} f(x)dx = \int_{0}^{1} = 1$$

In general, we say that X is a uniform random variable on the interval (α, β) if its probability density function is given by

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

3.2 Exponential Random Variable

A continuous random variable whose probability density function is given, for some $\lambda > 0$, by

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

is said to be an exponential random variable with parameter λ.

3.3 Gamma Random Variables

A continuous random variable whose density is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} (\lambda x)^{\alpha - 1} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

for some $\lambda > 0, \alpha > 0$ is said to be a gamma random variable with parameter α, λ. The quantity $\Gamma(\alpha)$ is called the gamma function and is defined by

$$\Gamma(\alpha) = \int_{0}^{\infty} e^{-x} x^{\alpha - 1} dx$$

It is easy to show by induction that for integral α, say $\alpha = n$,

$$\Gamma(n) = (n - 1)!$$

3.4 Normal Random Variables

We say that X is a normal random variable(or simply that X is normal distributed) with parameters μ and σ^2 of X is given by

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

The density function is a bell-shaped curve that is symmetric around μ.

An important fact about normal random variables is that if X is normally distributed with parameter μ and σ^2 then $Y = \alpha * X + \beta$ is normally distributed with parameters $\alpha * \mu + \beta$ and $\alpha^2 \sigma^2$.

Proof. Suppose first that $\alpha > 0$ and note that $F_Y(.)$, the cumulative distribution function of the random variable Y, is given by

$$\begin{aligned}
F_Y(a) =& P\{Y \leq a\} \\
=& P\{\alpha X + \beta \leq a\} \\
=& P\{X \leq \frac{a-\beta}{\alpha}\} \\
=& F_X(\frac{a-\beta}{\alpha}) \\
=& \int_{-\infty}^{(a-\beta)/\alpha} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} dx \\
=& \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}\alpha\sigma} exp\{\frac{-(v-(\alpha\mu+\beta))^2}{2\alpha^2\sigma^2}\} dv
\end{aligned}$$

where the last equality is obtained by the change in variables $v = \alpha x + \beta$. However, since $F_Y(a) = \int_{-\infty}^{a} f_Y(v)dv$, it follows from the Equation that the profit density function $f_Y(.)$ is given by

$$f_Y(v) = \frac{1}{\sqrt{2\pi}\alpha\sigma} exp\{\frac{-(v-(\alpha\mu+\beta))^2}{2(\alpha\sigma)^2}\}, -\infty < v < \infty$$

Hence, Y is normally distributed with parameters $\alpha\mu + \beta$ and $(\alpha\mu)^2$. A similar result is also true when $\alpha < 0$. $\square$

Implementation One implementation of the preceding result is that if X is normally distributed with parameters μ and σ^2 then $Y = \frac{X-\mu}{\sigma}$ is normally distributed with parameters 0 and 1. Such a random variable Y is said to have the standard or unit normal distribution.

4 Expectation of a Random Variable

Expected Value a weighted average of the possible value that X can take on, each value being weighted by the probability that X could be.

4.1 The Discrete Case

- **Bernoulli Random Variable** p
- **Binomial Random Variable** np
- **Geometric Random Variable** $\frac{1}{p}$
- **Poisson Random Variable** λ

4.2 The Continuous Case

Example(Expectation of an Exponential Random Variable) Let X be exponentially distributed with parameter λ, calculated $E[X]$.

Solution:

$$E[X] = \int_0^\infty x\lambda e^{\lambda X} dx$$

Integrating by parts ($dv = \lambda e^{\lambda x}$, $u = x$) yields,

$$\begin{aligned} E[X] &= -xe^{-\lambda x}|_0^\infty + \int_0^\infty e^{-\lambda x} dx \\ &= 0 - \frac{e^{-\lambda x}}{\lambda}|_0^\infty \\ &= \frac{1}{\lambda} \end{aligned}$$

4.3 Expectation of a Function of a Random Variable

Example Let X be uniformly distributed over $(0, 1)$. Calculate $E[X^3]$.

Solution:

$$E[X^3] = \int_0^1 x^3 dx = \frac{1}{4}$$

Moment The expected value of a random variable X, $E[X]$, is also referred to as the mean or the first moment of X. The quantity $E[X^n], n \geq 1$, is called the n-th moment of X. We have

$$E[X^n] = \begin{cases} \sum_{x:p(x)>0} x^n p(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x^n f(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

Variance

$$Var(X) = E[X^2] - (E[x])^2$$

5 Jointly Discrete Random Variables

5.1 Joint Distribution Functions

Example Calculate the expected sum obtained when three fair dice are rolled.

Solution: Let X denote the sum obtained. Then $X = X_1 + X_2 + X_3$ where X_i represents the value of the i-th die. Thus,

$$E[X] = E[X_1] + E[X_2] + E[X_3] = 3 * (7/2) = 21/2$$

Example At a party n men throw their hats into the center of a room. The hats are mixed up and each man randomly selects one. Find the expected number of men who select their own hats.

Solution: Letting X denote the number of men that selects their own hats, we can best compute $E[X]$ by noting that

$$X = X_1 + X_2 + \cdots + X_N$$

where

$$X_i = \begin{cases} 1 & \text{if the ith man selects his own hat} \\ 0 & \text{otherwise} \end{cases}$$

Now, because the i-th man is equally likely to select any of the N hats, it follows that

$$P\{X_i = 1\} = P\{\text{the ith man selects his own hat}\} = \frac{1}{N}$$

and so

$$E[X_i] = 1 * Pr\{X_i = 1\} + 0 * Pr\{X_i = 0\}$$

Hence, we have $E[X] = 1$. Hence now matter how many people are at the party, on the average exactly one of the men will select his own hat.

Example Suppose that there are 25 different types of coupons and suppose that each time one obtains a coupon, it is equally likely to be any one of the 25 types. Compare the expected number of different types that are contained in a set of 10 coupons.

Solution: Let X denote the number of different types in the set of 10 coupons. We compare $E[X]$ by using the representation $E[X] = E[X_1] + \cdots + E[X_{25}]$.

where

$$X_i = \begin{cases} 1 & \text{if at least one type } i \text{ coupon is in the set of 10} \\ 0 & \text{otherwise} \end{cases}$$

Now,

$$\begin{aligned} E[X_i] &= P\{X_i =\} \\ &= P\{\text{At least one type } i \text{ coupon is in the set of 10}\} \\ &= 1 - P\{\text{no type } i \text{ coupons are in the set of 10}\} \\ &= 1 - (\frac{24}{25})^2 \end{aligned}$$

Therefore, $E[X] = 25 * (1 - (\frac{24}{25})^2 5)$.

5.2 Covariance and Variance of Sums of Random Variables

The covariance of any two random variables X and Y, denoted by $cov(X, T)$, is defined by

$$\begin{aligned} Cov(X, Y) &= E[(x - E[X])(Y - E[Y])] \\ &= E[XY - XE[Y] - YE[X] + E[X]E[Y]] \\ &= E[XY] - E[X]E[Y] - E[Y]E[X] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

Note that if X and Y are independent, then it follows that $Cov(X, Y) = 0$.

Properties of Covariance For any random variables X, Y, Z and constant c

- $Cov(X, X) = Var(X)$
- $Cov(cX, Y) = cCov(X, Y)$
- $Cov(X, Y + Z) = Cov(X, Y) + Cov(X, Z)$

For the last property, we can easily generalizes to give the following result

$$cov(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j) = \sum_{i=1}^{n}\sum_{j=1}^{m} Cov(X_i, Y_j)$$

A useful expression for the variance of the sum of random variables can be obtained as follows

$$\begin{aligned} var(\sum_{i=1}^{n} X_i) &= Cov(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_i, X_j) \\ &= \sum_{i=1}^{n} Cov(X_i, X_i) + \sum_{i=1}^{n}\sum_{j\neq i} Cov(X_i, X_j) \\ &= \sum_{i=1}^{n} Var(X_i) + 2\sum_{i=1}^{n}\sum_{j<i} Cov(X_i, X_j) \end{aligned}$$

If $X_i, i = 1, 2, \cdots, n$ are independent random variables, then we can get

$$Var(\sum_{i=1}^{n} x_i) = \sum_{i=1}^{n} Var(X_i)$$

Proposition 1. *Suppose that $X_1, \cdots, X_n$ are independent and identically distributed with expected value μ and variance σ^2. Then*
(a) $E[\bar{X}] = \mu$
(b) $Var(\bar{X}] = \sigma^2/n$
(c) $Cov(\bar{X}, X_i - \bar{X}) = 0, i = 1, \cdots, n$.

Example Compute the variance of a Binomial Random Variable X with parameters n and p.

Solution: Since such a random variable represents the number of successes in n independent trails when each trail has a common probability p of being a success, we may write

$$X = X_1 + X_2 + \cdots X_n$$

where X_i are independent Bernoulli random variables such that

$$X_i = \begin{cases} 1 & \text{if the ithe trial is a success} \\ 0 & \text{otherwise} \end{cases}$$

Hence we have

$$Var(X) = Var(X_1) + Var(X_2) + \cdots + Var(X_n)$$

since $Var(X_i) = E[(X_i)^2] - (E[X_i])^2 = p - p^2$ since $(X_i)^2 = X_i$.
And thus $Var(X) = np(1-p)$.

Hypergeometric The random variable $\sum_{i=1}^{n} X_i$ can be thought of representing the number of white balls obtained when n balls are randomly selected from a population consisting of Np white and $N - Np$ black balls. Such a random variable is called hypergeometric and has a probability mass function given by

$$P\{\sum_{i=1}^{n} X_i = k\} = \frac{\binom{Np}{k}\binom{N-Np}{n-k}}{\binom{N}{n}}$$

5.3 Distribution of X+Y

We consider the distribution of $X + Y$ from the distributions of X and Y when X and Y are independent. Suppose first that X and Y are continuous, X having probability density f and Y having probability density g. Then letting $F_{X+Y}(a)$ be the cumulative distribution function of $X + Y$, we have

$$\begin{aligned} F_{X+Y}(a) &= PX + Y \leq a \\ &= \int\int +X + Y \leq af(x)g(y)dxdy \\ &= \int_{\infty}^{\infty}\int_{\infty}^{a-y} f(x)g(y)dxdy \\ &= \int_{-\infty}^{\infty}(\int_{-\infty}^{a-y} dx)g(y)dy \\ &= \int_{-\infty}^{\infty} F_X(a-y)g(y)dy \end{aligned}$$

The cumulative distribution function $F + X + Y$ is called the convolution of the distributions F_X and F_T(the cumulative distributions functions of X and Y respectively).

By differentiating the equation, we obtain the probability density function $f_{X+Y}(a)$ of $X+Y$ given by

$$\begin{aligned} f+X+Y)(a) &= \frac{d}{da}\int_{-\infty}^{\infty} F_X(a-y)g(y)dy \\ &= \int_{-\infty}^{\infty} \frac{d}{da}(F_X(a-y))g(y)dy \\ &= \int_{-\infty}^{\infty} f(a-y)g(y)dy \end{aligned}$$

Example If X and Y are independent random variables both uniformly distributed on $(0,1)$, then calculate the probability of $X+Y$.

Solution: since

$$f(a) = g(a) = \begin{cases} 1 & 0 ¡ a ¡ 1 \\ 0 & otherwise \end{cases}$$

We obtain

$$f_{X+Y}(a) = \int_0^1 f(a-y)g(y)dy = \int_0^a f(a-y)dy$$

For $0 \leq a \leq 1$, this yields

$$f_{X+Y}(a) = \int_0^a dy = a$$

For $1 < a < 2$, we get

$$f_{X+Y}(a) = \int_{(a-1)}^1 dy = 2-a$$

Hence, we have

$$f_{X+Y}(a) = \begin{cases} a & 0 \leq a \leq 1 \\ 2-a & \text{1 ¡a ¡2} \\ 0 & \text{otherwise} \end{cases}$$

Sum of Independent Poisson Random Variables Let X and Y be independent Poisson random variables with respective means λ_1 and λ_2. Then X_1+X_2 has a Poisson distribution with mean $\lambda_1+\lambda_2$.

to-do#1: Read Page 55-56, Joint Probability Distribution of Functions of Random Variables

6 Moment Generating Functions

Definition The moment generating function $\phi(t)$ of the random variable X is defined for all values t by

$$\phi(t) = E[e^{tX}] = \begin{cases} \sum_x e^{tx}p(x) & \text{if X is discrete} \\ \sum_{-\infty}^{\infty} e^{tx}f(x)dx\text{if } X \text{ is continous} \end{cases}$$

We call $\phi(t)$ the moment generating function because all of the moments of X can be obtained by successively differentiating $\phi(t)$. For example

$$\phi'(t) = \frac{d}{dt}E[e^{tX}] = E[Xe^{tX}]$$

Hence, $\phi'(0) = E[X]$

Similarly, $\phi''(t) = E[X^2e^{tX}]$ and so $\phi''(0) = E[X^2]$.

In general, the n-th derivative of $\phi(t)$ evaluated at $t=0$ equals $E[X^n]$, that is

$$\phi^n(t) = E[X^n], n \geq 1$$

Example-The Binomial Distribution with Parameters n and p

$$\begin{aligned}\phi(t) &= \sum_{k=0}^{n} e^{tk}\binom{n}{k}p^k(1-p)^{n-k} \\ &= \sum_{k=0}^{n}\binom{n}{k}(pe^t)^k(1-p)^{n-k} \\ &= (pe^t+1-p)^n\end{aligned}$$

thus $\phi'(t) = n([e^t+1-p)^{n-1}pe^t$, and so $E[X] = \phi'(0) = np$. and by calculate $\phi''(t)$, we get $E[X^2] = \phi''(0) = n(n-1)p^2 + np$.

Example: The Normal Distribution with Parameters μ and σ^2 The moment generating function for a standard normal random variable is obtained as follows

$$\begin{aligned}E[e^{tZ}] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{tx}e^{-x^2/2}dx \\ &= e^{t^2/2}\end{aligned}$$

If Z is a standard normal, then $X = \sigma Z + \mu$ is normal with parameters μ and σ^2, therefore

$$\begin{aligned}\phi(t) = E[e^{tX}] &= E[e^{t(\sigma Z+\mu}] \\ &= e^{t\mu}E[e^{t\sigma Z}] \\ &= exp\{\frac{\sigma^2t^2}{2}+\mu t\}\end{aligned}$$

Sum of Independent Random Variables Moment generating function of the sum of independent random variables is just the product of the individual moment generating functions.

To see this, suppose that X and Y are independent and have moment generating function $\phi_X(t)$ and $\phi_Y(t)$, respectively. Then $\phi_{X+Y}(t)$, the moment generating function of $X+Y$, is given by

$$\phi_{X+Y}(t) = E[e^{t(X+Y)}] = E[e^{tX}e^{tY}] = E[e^{tX}]E[e^{tY}] = \phi_X(t)\phi_Y(t)$$

Laplace transform For a nonnegative random variable X, it is often convenient to define its *Laplace transform* $g(t), t \geq 0$, by

$$g(t) = \phi(-t) = E[e^{-tX}]$$

This is, the Laplace transform evaluated at t is just the moment generating function evaluated at $-t$. The advantage of dealing with the Laplace transform, rather than the moment generating function, when the random variable is nonnegative is that if $X \geq 0$ and $t \geq 0$, then

$$0 \leq e^{-tX} \leq 1$$

That is, the Laplace transform is always between 0 and 1. As in the case of moment generating functions, it remains true that nonnegative random varaibles that have the same Laplace transform must also have the same distribution.

6.1 The Joint Distribution of the Sample Mean and Sample Variance from a Normal Distribution

Let $X_1, \cdots, X_n$ be independent and identically distributed random variables, each with mean μ and variance σ^2. The random variable S^2 is defined by

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

is called the sample variance of the data. To compute $E[S^2]$, we use the identity

$$\sum_n i = 1(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2$$

which is proven as follows

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \bar{X}) &= \sum_{i=1}^{n}(X_i - \mu + \mu - \bar{X})^2 \\ &= \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2\end{aligned}$$

Using this identity, we have

$$\begin{aligned}E[(n-1)S^2] &= \sum_{i=1}^{n} E[(X_i - \mu)^2] - nE[(\bar{X} - \mu)^2] \\ &= n\sigma^2 - nVar(\bar{X}) \\ &= n\sigma^2 - n * \frac{\sigma^2}{n} \\ &= (n-1)\sigma^2\end{aligned}$$

Thus, we obtain the preceding that

$$E[S^2] = \sigma^2$$

Chi-square random variable If $Z_1, \cdots, Z_n$ are independent standard normal random variables, then the random variable $\sum_{i=1}^{n}(Z_i)^2$ is said to be a chi-squared random variable with n degrees of freedom.

to-do#2: review chi-square

Proposition 2. *If $X_1, \cdots, X_n$ are independent and identically distributed normal random variables with mean μ and variance σ^2, then the sample mean $\bar{X}$ and the sample variance S^2 are independent. $\bar{X}$ is a normal random variable with mean μ and variance σ^2/n; $\frac{(n-1)\S^2}{\sigma^2}$ is a chi-squared random variable with $(n-1)$ degrees of freedom.*

7 The Distribution of the Number of Events that Occur

7.1 Define A, B, B^c

Consider arbitrary events $A_1, \cdots, A_n$, and let X denote the number of these events that occur. We will determine the probability mass function of X. To begin with, $1 \le k \le n$, let

$$S_k = \sum_{i1 < \ldots < ik} P(A_{i1} \ldots A_{ik})$$

equal the sum of the probabilities of all the $\binom{n}{k}$ intersections of k distinct events, and note that the inclusion-exclusion identity states that

$$P(X > 0) = P(\cup_{i=1}^{n} A_i) = S_1 - S_2 + S_3 - \cdots + (-1)^{n+1} S_n$$

Now, fix k of the n events, say $A_{i1}, \cdots, A_{ik}$, and let

$$A = \cap_{j=1}^{k} A_{ij}$$

be the event that k of these events occur. Also let

$$B = \cap_{j \notin i_1, \cdots, i_k} A_j^c$$

be the events that none of the other $(n-k)$ events occur. Consequently, AB is the event that $A_{i1}, \cdots, A_{ik}$ are the only events to occur. Because

$$A = AB \cup AB^c$$

Because B^c is at least one of the events $A_j, j \neq i_1, \cdots, i_k$, occur, we see that

$$B^c = \cup_{j \notin i_1, \cdots, i_k} A_j$$

Thus

$$P(AB^c) = P(A \cup_{j \notin \{i_1, \cdots, i_k\}} A_j) = P(\cup_{j \notin \{i_1, \cdots, i_k\}} AA_j)$$

7.2 Calculate $P(X = k)$

Applying the inclusion-exclusion identity gives

$$\begin{aligned} P(AB^c) = \sum_{j \notin \{i_1, \cdots, i_k\}} P(AA_j) - \sum_{j_1 < j_2 \notin \{i_1, \cdots, i_k\}} P(AA_{j_1} A_{j_2}) \\ + \sum_{j_1 < j_2 < j_3 \notin \{i_1, \cdots, i_k\}} P(AA_{j_1} A_{j_2} A_{j_3}) \end{aligned}$$

using that $A = \cap_{j=1}^{k} A_{ij}$, the preceding shows that the probability that the k events $A_{i1}, \cdots, A_{ik}$ are the only events to occur is

$$\begin{aligned} P(A) - P(AB^c) = P(A_{i1}, \cdots, A_{ik}) - \sum_{j \notin \{i_1, \cdots, i_k\}} P(A_{i1}, \cdots, A_{ik} A_j) \\ + \sum_{j_1 < j_2 \notin \{i_1, \cdots, i_k\}} P(A_{i1}, \cdots, A_{ik} A_{j_1} A_{j_2}) - \sum_{j_1 < j_2 < j_3 \notin \{i_1, \cdots, i_k\}} P(A_{i1}, \cdots, A_{ik} A_{j_1} A_{j_2} A_{j_3}) + \cdots \end{aligned}$$

Summing the preceding over all sets of k distinct indices yields

$$\begin{aligned} P(X = k) = \sum_{i_1 < \cdots < i_k} P(A_{i1}, \cdots, A_{ik}) - \sum_{i_1 < \cdots < i_k} \sum_{j \notin \{i_1, \cdots, i_k\}} P(A_{i1}, \cdots, A_{ik} A_j) \\ + \sum_{i_1 < \cdots < i_k} \sum_{j_1 < j_2 \notin \{i_1, \cdots, i_k\}} P(A_{i1}, \cdots, A_{ik} A_{j_1} A_{j_2}) - \cdots \end{aligned}$$

First, note that

$$\sum_{i_1, \cdots, i_k} P(A_{i1} \cdots A_{ik}) = S_k$$

The probability of every intersection of $k+1$ distinct events $A_{m1} \cdots A_{mk+1}$ will appear $\binom{k+1}{k}$ times in this multiple summation. This is so because each of k of of its indices to play the role of $i_1, \cdots, i_k$ and the other to play the role of j results in the addition of therm $P(A_{m1} \cdots A_{mk+1})$, hence

$$\begin{aligned} \sum_{i_1 < \cdots i_k} \sum_{j \notin \{i_1, \cdots, i_k\}} P(A_{i1} \cdots A_{ik} A_j) &= \binom{k+1}{k} \sum_{m_1 < \cdots < mk+1} P(A_{m1} < \cdots < A_{mk+1}) \\ &= \binom{k+1}{k} S_{k+1} \end{aligned}$$

Repeating this argument for the rest of the multiple summations yields the result

$$P(X=k)=S_k-\binom{k+1}{k}S_{k+1}+\binom{k+2}{k}S_{k+2}-\cdots+(-1)\binom{n}{k}S_n$$

The preceding can be written as

$$P(X=k)=\sum_{j=k}^{n}(-1)^{k+j}\binom{j}{k}S_j$$

7.3 Calculate $P(X \geq k)$

We will prove

$$P(X\geq k)=\sum_{j=k}^{n}(-1)^{k+j}\binom{j-1}{k-1}S_j$$

This proof uses a backwards mathematical induction that starts with $k=n$. Now when $k=n$ the proceeding identity states that

$$P(X=n)=S_n$$

which is true. So assume that

$$P(X\geq k+1)\sum_{j=k+1}^{n}(-1)^{k+1+j}\binom{j-1}{k}S_j$$

But then

$$\begin{aligned}P(X\geq k)&=P(X=k)+P(X\geq k+1)\\&=\sum_{j=k}^{n}(-1)^{k+j}\binom{j}{k}S_j+\sum_{j=k+1}^{n}(-1)^{k+1+j}\binom{j-1}{k}S_j\\&=\sum_{j=k}^{n}(-1)^{k+j}\binom{j-1}{k-1}S_j\end{aligned}$$

8 Limit Theorems

Markov's Inequality If X is a random variable that takes only nonnegative values, then for any value $a>0$

$$P\{X\geq a\}\leq\frac{E[X]}{a}$$

Proof. We give a proof for the case where X is continuous with density f.

$$\begin{aligned}E[X]&=\int_0^\infty xf(x)dx\\&=\int_0^a x(f)dx+\int_a^\infty xf(x)dx\\&\geq\int_a^\infty xf(x)dx\\&\geq\int_a^\infty af(x)dx\\&=a\int_a^\infty f(x)dx\\&=aPX\geq a\end{aligned}$$

□

Chebyshev's Inequality If X is a random variable with mean μ and variance σ^2, for any value $k > 0$,

$$P\{|X-\mu| \geq k\} \leq \frac{\sigma^2}{k}$$

Proof. Since $(X-\mu)^2$ is a nonnegative random variable, we can apply Markov's inequality(with $a = k^2$) to obtain

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

But since $(X-\mu)^2 \geq k^2$ if an only if $|X-\mu| \geq k$, the preceding is equivalent to

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

□

Importance The importance of Markov's and Chebyshev's inequalities is that they enable us to derive bounds on probabilities when only the mean, or both the mean and the variance, of the probability distribution are known. Of course, if the actual distribution were known, then the desired probabilities can be exactly computed, and we would not need to resort bounds.

Strong Law of Large Numbers It states that the average of a sequence of independent random variables having the same distribution will, with probability 1, converge to the mean of that distribution.

Central Limit Theorem Let $X_1, X_2, \cdots$ be a sequence of independent, identically distributed random variables, each with mean μ and variance σ^2. Then the distribution of

$$\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

tends to the standard normal as $n \to \infty$. That is

$$P\{\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}\}$$