Miscellaneous Results from Applied Probability

Dennis Bricker
Dept of Mechanical & Industrial Engineering
The University of Iowa
dennis-bricker@uiowa.edu

Probability of simultaneous occurrence of events

If A and B are two independent events, then by definition:

$$P\{A \text{ AND } B\} = P\{A \cap B\} = P\{A\} \times P\{B\}$$

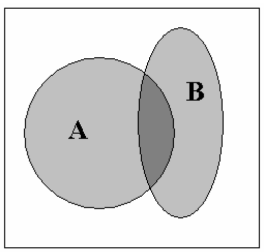
If A and B are dependent, then

$$\begin{aligned} P\{A \cap B\} &= P\{B \mid A\} \times P\{A\} \\ &= P\{A \mid B\} \times P\{B\} \end{aligned}$$

Note: *This follows from the definition of conditional probability:*

$$P\{A \mid B\} \equiv \frac{P\{A \cap B\}}{P\{B\}} \quad \text{and} \quad P\{B \mid A\} \equiv \frac{P\{A \cap B\}}{P\{A\}}$$

Probability of occurrence of *at least one* of two events



If A and B are two events, then the probability of A OR B is

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

Note:

- *If A and B are mutually exclusive, then this is equivalent to* $$P\{A \cup B\} = P\{A\} + P\{B\}$$
- *If A and B are independent, then this is equivalent to*

$$P\{A \cup B\} = P\{A\} + P\{B\} - P\{A\} \times P\{B\}$$

Law of Total Probability for computation of the probability of an event A.

Suppose that B_1, B_2, …B_n are mutually exclusive events, i.e.,

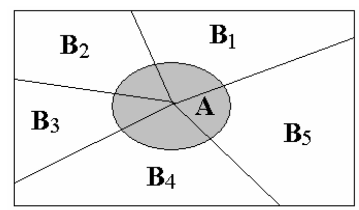
$$B_i \cap B_j = \varnothing, i \neq j$$

such that $A \subseteq \bigcup_i B_i$

Then

$$P(A) = \sum_i P(A \mid B_i) P(B_i)$$

"chain rule"



In particular, we can condition upon the value of a discrete random variable Y:

$$P(A) = \sum_y P(A \mid Y = y) P(Y = y)$$

Example:

A certain product is manufactured in three plants:

Plant #	*market share*	*defective rate*
1	30%	5%
2	25%	4%
3	45%	3%

Suppose that we sample a unit of this product from a retail shop shelf.

What is the probability that it is defective?

Let Y= plant which manufactured the unit.

A = event that product is defective.

$$P(\text{unit is defective}) = \sum_{i=1}^{3} P(defect \mid Y=i)P(Y=i)$$
$$= 0.05\times 0.30 + 0.04\times 0.25 + 0.03\times 0.45$$
$$= 0.0385$$

That is, the probability that we have sampled a defective unit is 3.85%.

Law of Total Expectation

For any two random variables X and Y,

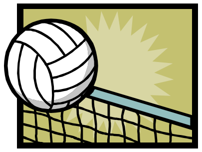
$$E(X) = \sum_{y} E(X \mid Y=y)P(Y=y)$$

where Y has discrete values and we assume that the expectations exist.

Example:

Consider a series of competitions between two evenly-matched teams, in which the series ends as soon as one team has won **3 consecutive games**.

What is the expected length of the series?



Solution:

Define $\mathbf{X_i}$ = number of games to follow after one of the teams has won the last i consecutive games

What is the value of $E(X_0)$?

Define **Y** =1 if the team which is ahead wins the next game, 0 otherwise.

Then by the *Law of Total Expectation*,

$$E(X_i) = E(X_i \mid Y=0)P(Y=0) + E(X_i \mid Y=1)P(Y=1)$$

Now, $E\{X_i \mid Y=0\} = 1 + E\{X_1\}$,

since, if the leading team loses, then the opposing team has a "winning streak" of one game,

and $E\{X_i \mid Y=1\} = 1 + E\{X_{i+1}\}$,

since, if the leading team wins, then it has a "winning streak" of i+1 games.

Therefore,

$$E\{X_i\} = E\{X_i \mid Y=0\}P\{Y=0\} + E\{X_i \mid Y=1\}P\{Y=1\}$$
$$= \frac{1}{2}\left(1 + E\{X_1\}\right) + \frac{1}{2}\left(1 + E\{X_{i+1}\}\right)$$
$$= \left[\frac{1}{2} + \frac{1}{2}E\{X_1\}\right] + \left[\frac{1}{2} + \frac{1}{2}E\{X_{i+1}\}\right] = 1 + \frac{1}{2}E\{X_1\} + \frac{1}{2}E\{X_{i+1}\}$$

for i=0, 1, and 2, where $E(X_3) \equiv 0$.

$$\boxed{E\{X_i\} = 1 + \frac{1}{2}E\{X_1\} + \frac{1}{2}E\{X_{i+1}\} \quad \text{for } i = 0,1,2}$$

Case *i=0*: $E\{X_0\} = 1 + \frac{1}{2}E\{X_1\} + \frac{1}{2}E\{X_1\} \Rightarrow E\{X_0\} - E\{X_1\} = 1$

Case *i=1*: $E\{X_1\} = 1 + \frac{1}{2}E\{X_1\} + \frac{1}{2}E\{X_2\} \Rightarrow \frac{1}{2}E\{X_1\} - \frac{1}{2}E\{X_1\} = 1$

Case *i=2*: since $E\{X_3\} = 0$,

$$E\{X_2\} = 1 + \frac{1}{2}E\{X_1\} + \frac{1}{2}E\{X_3\} \Rightarrow -\frac{1}{2}E\{X_1\} + E\{X_2\} = 1$$

We thus obtain the system of *linear* equations $\begin{cases} EX_0 - EX_1 = 1 \\ 0.5EX_1 - 0.5EX_2 = 1 \\ -0.5EX_1 + EX_2 = 1 \end{cases}$

which has the solution (using Gauss elimination, for example):

$$\boxed{EX_0 = 7, EX_1 = 6, EX_2 = 4}$$

*That is, the series is expected to end after **7** games.*

Notice that when one team has already won two consecutive games, we expect a total of four games, i.e., two additional games to be played!

The k^{th} **Moment** of a distribution is defined to be

$$E\left[X^k\right]$$

The k^{th} ***Central* Moment** of a distribution is

$$M_k = E\left[X - EX\right]^k$$

In particular, the 2^{nd} central moment is the *variance*,

the 3^{rd} central moment is the *skewness*, and

the 4^{th} central moment is the *kurtosis*.

Variance is the *second* central moment: $Var\{X\} = \sigma^2 \equiv M_2$

Therefore,

$$Var\{X\} = E\left[X - E(X)\right]^2$$

Sometimes it is more convenient to compute the variance using the relationship:

$$\begin{aligned} Var\{X\} &= E\left[X^2 - 2X \times E(X) + \left[E(X)\right]^2\right] \\ &= E\left[X^2\right] - E\left[2X \times E(X)\right] + E\left[E^2(X)\right] \\ &= E\left[X^2\right] - 2E(X) \times E(X) + E^2(X) \end{aligned}$$

$$\boxed{Var\{X\} = E\left(X^2\right) - E^2(X)}$$

Other relationships: if c is a constant,

$$\begin{aligned} &Var\{X + c\} = V\{X\} \\ &Var\{cX\} = c^2 Var\{X\} \\ &Var\{X + Y\} = Var\{X\} + Var\{Y\} + 2E\{(X - EX)(Y - EY)\} \end{aligned}$$

The term

$$E\{(X - EX)(Y - EY)\}$$

is called the *covariance* of X & Y, or Cov{X,Y}

A more useful formula for computation is

$$\boxed{Cov\{X, Y\} = E\{XY\} - E\{X\} \times E\{Y\}}$$

In general,

the expectation of a sum = sum of expectations, i.e.,

$$E\left(\sum_{i=1}^{N} X_i\right) = \sum_{i=1}^{N} E(X_i)$$

and, if the random variables X_i are *independent*,

the variance of a sum = sum of variances, i.e.,

$$\text{var}\left(\sum_{i=1}^{N} X_i\right) = \sum_{i=1}^{N} \text{var}(X_i).$$

Thus, if the random variables X_i are *independent* & *identically-distributed*,

$$E\left(\sum_{i=1}^{N} X_i\right) = NE(X_1) \text{ and } \text{var}\left(\sum_{i=1}^{N} X_i\right) = N\,\text{var}(X_1)$$

Suppose that the number of terms N is itself a random variable.

If the random event $\{N=n\}$ is independent of $X_{n+1}, X_{n+2}, \dots$ for each $n\geq 1$:

Wald's Equation:

$$E\left(\sum_{i=1}^{N} X_i\right) = E(N)E(X_1)$$

If in addition, the event $\{N=n\}$ is independent of *all* X_i, $i=1,2,3,\dots$.

$$\operatorname{var}\left(\sum_{i=1}^{N} X_i\right) = E(N)\operatorname{var}(X_1) + \operatorname{var}(N)E^2(X_1)$$

Suppose that $X_1, X_2, X_3, \dots$ is a sequence of *independent, identically-distributed* random variables with mean and variance which exist:

$$E\{X_i\} = \mu \quad \text{and} \quad Var\{X_i\} = \sigma^2 > 0$$

and let Y_n be the n^{th} partial sum, i.e.,

$$Y_n = \sum_{i=1}^{n} X_i$$

so that the expected value of Y_n is $n\mu$, and its variance is $n\sigma^2$.

Central Limit Theorem: The distribution of the *standardized* form of Y_n,

i.e., $$\frac{Y_n - n\mu}{\sigma\sqrt{n}},$$

is, in the limit, the *normal* distribution with mean 0 and variance 1.

Normal Approximation to Binomial Distribution

Suppose that $\{X_n\}$ is a Bernouilli process,

i.e., $X_1, X_2, \dots$ is a sequence of *Bernouilli* random variables, with $P\{X_i=1\}=p$.

Then the n^{th} partial sum Y_n has binomial distribution with parameters n & p, with mean np and variance $np(1-p)$.

Therefore, by the *Central Limit Theorem*, the cumulative distribution function (CDF) of Y_n, i.e.,

$$P\{Y_n \leq t\} = \sum_{k=0}^{t} \binom{n}{k} p^k (1-p)^{n-k}$$

should have, for "large" n, *approximately* normal distribution with mean np and standard deviation $\sqrt{np(1-p)}$

Example: In a true-false examination, What is the probability that a student can guess the correct answers to 15 or more out of 27 questions, i.e., a score of at least 55.5%?

Solution: Assuming p = 50%, e.g., the student chooses his/her answer by flipping an unbiased coin, the exact value is

$$P\{Y_n \leq t\} = \sum_{k=0}^{t} \binom{n}{k} p^k (1-p)^{n-k}$$

These values are shown on the following page, where we see that

$$P\{Y_{27} \geq 15\} = 1 - P\{Y_{27} \leq 14\} \approx 1 - 0.65 = 35\%$$

x	P{x}	P{X<=x}	P{X>x}
0	0.00000001	0.00000001	0.99999999
1	0.00000020	0.00000021	0.99999979
2	0.00000262	0.00000282	0.99999718
3	0.00002179	0.00002462	0.99997538
4	0.00013076	0.00015537	0.99984463
5	0.00060149	0.00075686	0.99924314
6	0.00220545	0.00296231	0.99703769
7	0.00661634	0.00957865	0.99042135
8	0.01654085	0.02611949	0.97388051
9	0.03491957	0.06103906	0.93896094
10	0.06285522	0.12389428	0.87610572
11	0.09713989	0.22103417	0.77896583
12	0.12951985	0.35055402	0.64944598
13	0.14944598	0.50000000	0.50000000
14	0.14944598	0.64944598	0.35055402
15	0.12951985	0.77896583	0.22103417
16	0.09713989	0.87610572	0.12389428
17	0.06285522	0.93896094	0.06103906
18	0.03491957	0.97388051	0.02611949
19	0.01654085	0.99042135	0.00957865
20	0.00661634	0.99703769	0.00296231
21	0.00220545	0.99924314	0.00075686
22	0.00060149	0.99984463	0.00015537
23	0.00013076	0.99997538	0.00002462
24	0.00002179	0.99999718	0.00000282
25	0.00000262	0.99999979	0.00000021
26	0.00000020	0.99999999	0.00000001
27	0.00000001	1.00000000	0.00000000

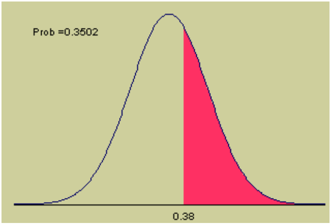
Binomial(27, 0.5)

According to the *Central Limit Theorem*, Y_{27} has approximately a *normal* distribution with mean 13.5 and standard deviation

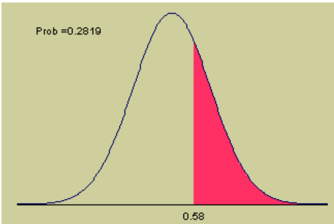
$$\sqrt{27\times(0.5)(0.5)} = 0.5\sqrt{27} = 2.59808$$

Therefore, according to tables for the Normal distribution or other resource (e.g., http://psych.colorado.edu/~mcclella/java/normal/normz.html)

$$P\{Y_{27}\geq 14.5\} = P\left\{\frac{Y_{27}-13.5}{2.59808}\geq\frac{14.5-13.5}{2.59808}\right\}\approx P\{Z\geq 0.3849\} = 0.3502$$



Note: If we had computed $P\{Y_{27}\geq 15\}$ *(instead of* $P\{Y_{27}\geq 14.5\}$*), we would have obtained a less accurate value of about 28.2%.*



Normal Distributions

A continuous random variable X has a *Normal* distribution with the two parameters:

- the mean μ and
- the standard deviation σ>0,

if its probability density function is of the form

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

This Normal distribution, N(μ,σ), is symmetric about the mean μ, i.e., its *skewness* is 0.

The *"Standard"* Normal distribution N(0,1) has mean μ=0 and standard deviation σ=1, with pdf

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\}$$

If the random variable X has N(μ,σ) distribution,

then $\frac{X-\mu}{\sigma}$ has the standard N(0,1) distribution.

The Cumulative Distribution Function (CDF) of the N(0,1) distribution cannot be expressed in closed form:

$$F(z) = P\{Z \le t\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} \exp\left\{-\frac{t^2}{2}\right\} dt$$

Reproductive Property of Normal Distribution

The Normal distribution has an important "reproductive" property:

> *The sum of (independent) Normally distributed random variables has a Normal distribution!*

Specifically, if $X_i \sim N(\mu_i, \sigma_i)$,

then $T = \sum_{i=1}^{k} a_i X_i$ has N(μ,σ) distribution,

with mean $\mu = \sum_{i=1}^{k} a_i \mu_i$ and variance $\sigma^2 = \sum_{i=1}^{k} a_i^2 \sigma_i^2$.

Note: The binomial, gamma, and a few other distributions also have this reproductive property!

Random Generation of Normally Distributed Random Variables

Generate two random numbers u_i and v_i, uniformly distributed in [0,1].

Perform the transformations

$$z_i = \sqrt{-2\ln u_i} \times \sin(2\pi v_i)$$

and

$$z_{i+1} = \sqrt{-2\ln u_i} \times \cos(2\pi v_i)$$

to generate a *pair* of random variables having N(0,1) distribution.

(This is known as the Box/Muller technique.

Cf. http://mathworld.wolfram.com/Box-MullerTransformation.html)

Note: a less efficient and less accurate method, based upon the central limit theorm, would be to sum a "large" number (15?) uniformly-distributed random variables.