Topic 8: The Expected Value*

September 27 and 29, 2011

Among the simplest summary of quantitative data is the sample mean. Given a random variable, the corresponding concept is given a variety of names, the **distributional mean**, the **expectation** or the **expected value**. We begin with the case of discrete random variables where this analogy is more apparent. The formula for continuous random variables is obtained by approximating with a discrete random and noticing that the formula for the expected value is a Riemann sum. Thus, expected values for continuous random variables are determined by computing an integral.

1 Discrete Random Variables

Recall for a data set taking values $x_1, x_2, \ldots, x_n$, one of the methods for computing sample mean of a function of the data is accomplished by evaluating

$$\overline{h(x)} = \sum_x h(x)p(x),$$

where $p(x)$ is the proportion of observations taking the value x.

For a finite sample space $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$ and a probability P on Ω, we can define the **expectation** or the **expected value** of a random variable X by an analogous average,

$$EX = \sum_{j=1}^{N} X(\omega_j)P\{\omega_j\}. \tag{1}$$

More generally for a function g of the random variable X, we have the formula

$$Eg(X) = \sum_{j=1}^{N} g(X(\omega_j))P\{\omega_j\}.$$

Notice that even though we have this analogy, the two formulas come from very different starting points. The value of $\overline{h(x)}$ is derived from **data** whereas no data are involved in computing $Eg(X)$. The starting point for the expected value is a **probability model**.

Example 1. *Roll one die. Then $\Omega = \{1, 2, 3, 4, 5, 6\}$. Let X be the value on the die. So, $X(\omega) = \omega$. If the die is fair, then the probability model has $P\{\omega\} = 1/6$ for each outcome ω and the expected value*

$$\begin{aligned} EX &= 1 \cdot P\{1\} + 2 \cdot P\{2\} + 3 \cdot P\{3\} + 4 \cdot P\{4\} + 5 \cdot P\{5\} + 6 \cdot P\{6\} \\ &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{21}{6} = \frac{7}{2}. \end{aligned}$$

An example of an unfair dice would be the probability with $P\{1\} = P\{2\} = P\{3\} = 1/4$ and $P\{4\} = P\{5\} = P\{6\} = 1/12$. In this case, the expected value

$$EX = 1 \cdot \frac{1}{4} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{4} + 4 \cdot \frac{1}{12} + 5 \cdot \frac{1}{12} + 6 \cdot \frac{1}{12} = \frac{11}{4}.$$

Exercise 2. *Find EX^2 for these two examples.*

Two properties of expectation are immediate from the formula for EX in (1):

1. If $X(\omega) \geq 0$ for every outcome $\omega \in \Omega$, then every term in the sum in (1) is nonnegative and consequently their sum $EX \geq 0$.

2. Let X_1 and X_2 be two random variables and c_1, c_2 be two real numbers, then by using the distributive property in (1), we find out that
$$E[c_1X_1 + c_2X_2] = c_1EX_1 + c_2EX_2.$$

The first of these properties states that nonnegative random variables have nonnegative expected value. The second states that expectation is a linear operation. Taking these two properties together, we say that the operation of taking an expectation
$$X \mapsto EX$$
is a **positive linear functional**. We have studied extensively another example of a postive linear functional, namely, the definite integral
$$g \mapsto \int_a^b g(x)\,dx$$
that takes a continuous positive function and gives the area between the graph of g and the x-axis between the vertical lines $x = a$ and $x = b$. For this example, these two properties become:

1. If $g(x) \geq 0$ for every $x \in [a, b]$, then $\int_a^b g(x)\,dx \geq 0$.

2. Let g_1 and g_2 be two continuous functions and c_1, c_2 be two real numbers, then
$$\int_a^b (c_1g_1(x) + c_2g_2(x))\,dx = c_1\int_a^b g_1(x)\,dx + c_2\int_a^b g_2(x)\,dx.$$

This analogy will be useful to keep in mind when considering the properties of expectation.

Example 3. *If X_1 and X_2 are the values on two rolls of a fair die, then the expected value of the sum*
$$E[X_1 + X_2] = EX_1 + EX_2 = \frac{7}{2} + \frac{7}{2} = 7.$$

Because sample spaces can be extraordinarily large even in routine situations, we rarely use the probability space Ω as the basis to compute the expected value. We illustrate this with the example of tossing a coin three times. Let X denote the number of heads. To compute the expected value EX, we can proceed as described in (1). For the table below, we have grouped the outcomes ω that have a common value $x = 3, 2, 1$ or 0 for $X(\omega)$.

ω	$X(\omega)$	x	$P\{\omega\}$	$P\{X = x\}$	$X(\omega)P\{\omega\}$	$xP\{X = x\}$
HHH	3	3	$P\{HHH\}$	$P\{X = 3\}$	$X(HHH)P\{HHH\}$	$3P\{X = 3\}$
HHT	2		$P\{HHT\}$		$X(HHT)P\{HHT\}$	
HTH	2	2	$P\{HTH\}$	$P\{X = 2\}$	$X(HTH)P\{HTH\}$	$2P\{X = 2\}$
THH	2		$P\{THH\}$		$X(THH)P\{THH\}$	
HTT	1		$P\{HTT\}$		$X(HHT)P\{HHT\}$	
TTH	1	1	$P\{THT\}$	$P\{X = 1\}$	$X(HTH)P\{HTH\}$	$1P\{X = 1\}$
THT	1		$P\{TTH\}$		$X(THH)P\{THH\}$	
TTT	0	0	$P\{TTT\}$	$P\{X = 0\}$	$X(TTT)P\{TTT\}$	$0P\{X = 0\}$

Note, for example, that, three outcomes HHT, HTH and THH each give a value of 2 for X. Because these outcomes are disjoint, we can add probabilities

$$P\{HHT\} + P\{HTH\} + P\{THH\} = P\{HHT, HTH, THH\}$$

But, the event

$$\{HHT, HTH, THH\} \text{ can also be written as the event } \{X = 2\}.$$

This is shown for each value of x in moving from column 4 and column 5 in the table above.

Thus, by combining outcomes that result in the same value for the random variable, we simplify, as shown in the rightmost column of the chart, the computation of the expected value.

$$EX = 0 \cdot P\{X = 0\} + 1 \cdot P\{X = 1\} + 2 \cdot P\{X = 2\} + 3 \cdot P\{X = 3\}.$$

As in the discussion above, we can, in general, find $Eg(X)$ by partitioning the sample space Ω into the outcomes ω that result in the same value x for the random variable $X(\omega)$. As the equality indicated between the fourth and fifth column in the table above indicates, we find, for each possible value of x, the probability $P\{X = x\}$ by collecting the outcomes ω that satisfy $X(\omega) = x$ and sum these probabilities.

In symbols, this can be written

$$\sum_{\omega; X(\omega)=x} P\{\omega\} = P\{X = x\}.$$

For these particular outcomes, $g(X(\omega)) = g(x)$ and

$$\sum_{\omega; X(\omega)=x} g(X(\omega))P\{\omega\} = \sum_{\omega; X(\omega)=x} g(x)P\{\omega\} = g(x)P\{X = x\}.$$

Now, sum over all possible value for X for each side of this equation.

$$Eg(X) = \sum_{\omega} g(X(\omega))P\{\omega\} = \sum_{x} g(x)P\{X = x\} = \sum_{x} g(x)f_X(x)$$

where $f_X(x) = P\{X = x\}$ is the probability mass function for X.

The identity

$$Eg(X) = \sum_{x} g(x)f_X(x) \tag{2}$$

is the most frequent method used to compute expectation of discrete random variables.

Example 4. *Flip a biased coin twice and let X be the number of heads. Then, to compute the expected value of X and X^2 we construct a table to prepare to use (2).*

x	$f_X(x)$	$xf_X(x)$	$x^2f_X(x)$
0	$(1-p)^2$	0	0
1	$2p(1-p)$	$2p(1-p)$	$2p(1-p)$
2	p^2	$2p^2$	$4p^2$
sum	1	$2p$	$2p + 2p^2$

Thus, $EX = 2p$ and $EX^2 = 2p + 2p^2$.

Exercise 5. *Draw 5 cards from a standard deck. Let X be the number of hearts. Use* R *to find EX and EX^2.*

A similar formula to (2) holds if we have a vector of random variables $X = (X_1, X_2, \ldots, X_n)$, f_X, the joint probability mass function and g a real-valued function of $x = (x_1, x_2, \ldots, x_n)$. In the two dimensional case, this takes the form

$$Eg(X_1, X_2) = \sum_{x_1} \sum_{x_2} g(x_1, x_2) f_{X_1,X_2}(x_1, x_2). \tag{3}$$

We will return to (3) in computing the distributional covariance of two random variables.

2 Bernoulli Trials

Bernoulli trials are the simplest and among the most common models for an experimental procedure. Each trial has two possible outcomes, variously called,

heads-tails, yes-no, up-down, left-right, win-lose, female-male, green-blue, dominant-recessive, or **success-failure**.

depending on the circumstances. We will use the principles of counting and the properties of expectation to analyze Bernoulli trials. From the point of view of statistics, the data have an **unknown** success parameter p. Thus, the goal of statistical inference is to make as precise a statement as possible for the value of p behind the production of the data. Consequently, any experimenter that uses Bernoulli trials as a model ought to mirrow its properties closely.

Example 6 (Bernoulli trials)**.** *Random variables $X_1, X_2, \ldots, X_n$ are called a sequence of* **Bernoulli trials** *provided that:*

1. *Each X_i takes on two values, namely, 0 and 1. We call the value 1 a* **success** *and the value 0 a* **failure***.*
2. *Each trial has the same probability for success, i.e., $P\{X_i = 1\} = p$ for each i.*
3. *The outcomes on each of the trials is independent.*

For each trial i, the expected value

$$EX_i = 0 \cdot P\{X_i = 0\} + 1 \cdot P\{X_i = 1\} = 0 \cdot (1-p) + 1 \cdot p = p$$

is the same as the success probability. Let $S_n = X_1 + X_2 + \cdots + X_n$ be the total number of successes in n Bernoulli trials. Using the linearity of expectation, we see that

$$ES_n = E[X_1 + X_2 \cdots + X_n] = p + p + \cdots + p = np,$$

the expected number of successes in n Bernoulli trials is np.

In addition, we can use our ability to count to determine the probability mass function for S_n.
Beginning with a concrete example, let $n = 8$, and the outcome

success, fail, fail, success, fail, fail, success, fail.

Using the independence of the trials, we can compute the probability of this outcome:

$$p \times (1-p) \times (1-p) \times p \times (1-p) \times (1-p) \times p \times (1-p) = p^3(1-p)^5.$$

Moreover, any of the possible $\binom{8}{3}$ particular sequences of 8 Bernoulli trials having 3 successes also has probability $p^3(1-p)^5$. Each of the outcomes are mutually exclusive, and, taken together, their union is the event $\{S_8 = 3\}$. Consequently, by the axioms of probability, we find that

$$P\{S_8 = 3\} = \binom{8}{3} p^3(1-p)^5.$$

Returning to the general case, we replace 8 by n and 3 by x to see that any particular sequence of n Bernoulli trials having x successes has probability

$$p^x(1-p)^{n-x}.$$

In addition, we know that we have

$$\binom{n}{x}$$

mutually exclusive sequences of n Bernoulli trials that have x successes. Thus, we have the mass function

$$f_{S_n}(x) = P\{S_n = x\} = \binom{n}{x} p^x(1-p)^{n-x}, \quad x = 0, 1, \ldots, n.$$

The fact that the sum

$$\sum_{x=0}^{n} f_{S_n}(x) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} = (p + (1-p))^n = 1^n = 1$$

follows from the binomial theorem. Consequently, S_n is called a **binomial random variable**.

In the exercise above where X is the number of hearts in 5 cards, let $X_i = 1$ if the i-th card is a heart and 0 if it is not a heart. Then, the X_i are not Bernoulli trials because the chance of obtaining a heart on one card depends on whether or not a heart was obtained on other cards. Still,

$$X = X_1 + X_2 + X_3 + X_4 + X_5$$

is the number of hearts and

$$EX = EX_1 + EX_2 + EX_3 + EX_4 + EX_5 = 1/4 + 1/4 + 1/4 + 1/4 + 1/4 = 5/4.$$

3 Continuous Random Variables

For X a continuous random variable with density f_X, consider the discrete random variable $\tilde{X}$ obtained from X by rounding down. Say, for example, we give lengths by rounding down to the nearest millimeter. Thus, $\tilde{X} = 2.134$ meters for any lengths X satisfying 2.134 meters $< X \leq 2.135$ meters.

The random variable $\tilde{X}$ is discrete. To be precise about the rounding down procedure, let Δx be the spacing between values for $\tilde{X}$. Then, $\tilde{x}$, an integer multiple of Δx, represents a possible value for $\tilde{X}$, then this rounding becomes

$$\tilde{X} = \tilde{x} \quad \text{if and only if} \quad \tilde{x} < X \leq \tilde{x} + \Delta x.$$

With this, we can give the mass function

$$f_{\tilde{X}}(\tilde{x}) = P\{\tilde{X} = \tilde{x}\} = P\{\tilde{x} < X \leq \tilde{x} + \Delta x\}.$$

Now, by the property of the density function,

$$P\{\tilde{x} \leq X < \tilde{x} + \Delta x\} \approx f_X(x)\Delta x. \tag{4}$$

In this case, we need to be aware of a possible source of confusion due to the similarity in the notation that we have for both the mass function $f_{\tilde{X}}$ for the discrete random variable $\tilde{X}$ and a density function f_X for the continuous random variable X.

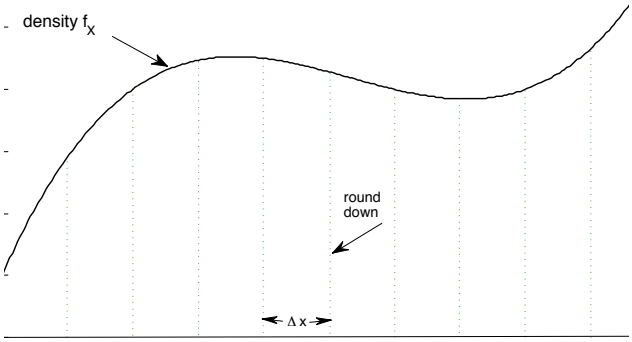


Figure 1: The discrete random variable $\tilde{X}$ is obtained by rounding down the continuous random variable X to the nearest multiple of Δx. The mass function $f_{\tilde{X}}(\tilde{x})$ is the integral of the density function from $\tilde{x}$ to $\tilde{x} + \Delta x$ indicated at the area under the density function between two consecutive vertical lines.

For this discrete random variable $\tilde{X}$, we can use identity (2) and the approximation in (4) to compute the expected value.

$$\begin{aligned} Eg(\tilde{X}) &= \sum_{\tilde{x}} g(\tilde{x}) f_{\tilde{X}}(\tilde{x}) = \sum_{\tilde{x}} g(\tilde{x}) P\{\tilde{x} \leq X < \tilde{x} + \Delta x\} \\ &\approx \sum_{\tilde{x}} g(\tilde{x}) f_x(\tilde{x}) \Delta x. \end{aligned}$$

This last sum is a Riemann sum and so taking limits as $\Delta x \to 0$ yields the definite integral

$$Eg(X) = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx. \tag{5}$$

As in the case of discrete random variables, a similar formula to (5) holds if we have a vector of random variables $X = (X_1, X_2, \ldots, X_n)$, f_X, the joint probability density function and g a real-valued function of the vector $x = (x_1, x_2, \ldots, x_n)$. The expectation in this case is an n-dimensional Riemann integral.

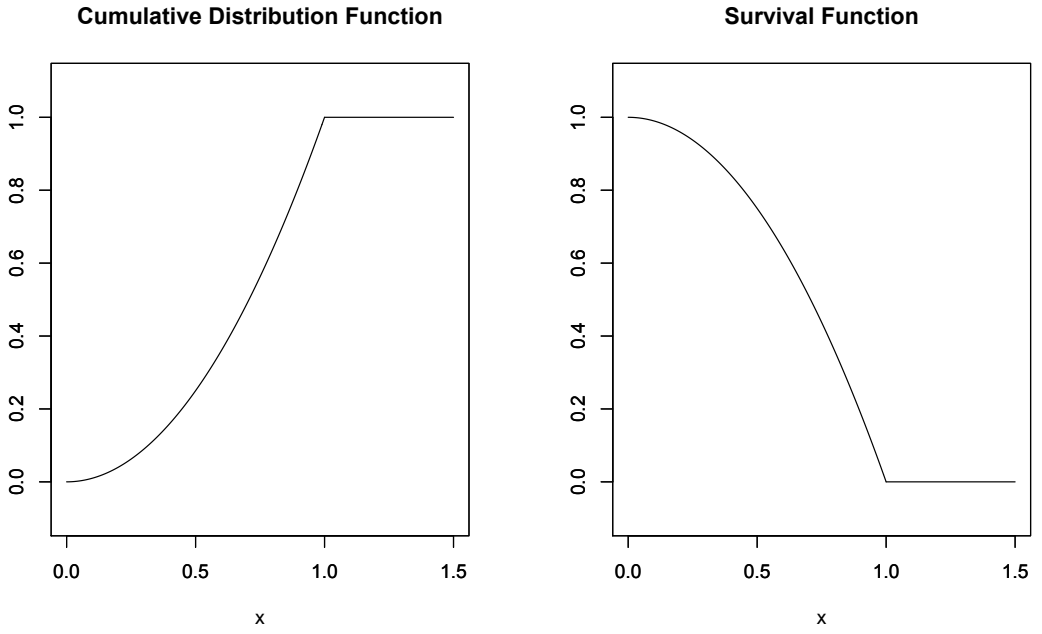


Figure 2: The cumulative distribution function $F_X(x)$ and the survival function $\bar{F}_X(x) = 1 - F_X(x)$ for the dart board example. Using the expression (6), we see that the expected value $EX = 2/3$ is the area under the survival function.

Example 7. *For the dart example, the density $f_X(x) = 2x$ on the interval $[0, 1]$, Thus,*

$$EX = \int_0^1 x \cdot 2x \, dx = \int_0^1 2x^2 \, dx = \frac{2}{3}x^3\Big|_0^1 = \frac{2}{3}.$$

Exercise 8. *If X is a nonnegative random variable, then $F_X(0) = 0$.*

If we were to compute the mean of T, an exponential random variable,

$$ET = \int_0^\infty t f_T(t) \, dt = \int_0^\infty t\lambda e^{-\lambda t} \, dt,$$

then our first step is to integrate by parts. This situation occurs with enough regularity that we will benefit in making the effort to see how integration by parts gives an alternative to computing expectation. In the end, we will see an analogy between the mean with the survival function $P\{X > x\} = 1 - F_X(x)$, and the sample mean with the empirical survival function.

Let X be a positive random variable, then the expectation is the improper integral

$$EX = \int_0^\infty x f_X(x) \, dx$$

(The unusual choice for v is made to simplify some computations and to anticipate the appearance of the survival function.)

$$\begin{array}{ll} u(x) = x & v(x) = -(1 - F_X(x)) = -\bar{F}_X(x) \\ u'(x) = 1 & v'(x) = f_X(x) = -\bar{F}'_X(x). \end{array}$$

First in integrate from 0 to b and take the limit as $b \to \infty$. Then, because $F_X(0) = 0$, $\bar{F}_X(0) = 1$ and

$$\begin{aligned}\int_0^b x f_X(x)\,dx &= -x\bar{F}_X(x)\Big|_0^b + \int_0^b (1 - F_X(x))\,dx \\ &= -b\bar{F}_X(b) + \int_0^b \bar{F}_X(x)\,dx\end{aligned}$$

The product term in the integration by parts formula converges to 0 as $b \to \infty$. Thus, we can take a limit to obtain the identity,

$$EX = \int_0^\infty P\{X > x\}\,dx. \tag{6}$$

Exercise 9. *Show that the product term in the integration by parts formula does indeed converge to 0 as $b \to \infty$.*

In words, the expected value is the area between the cumulative distribution function and the line $y = 1$ or the area under the survival function. For the case of the dart board, we see that the area under the distribution function between $y = 0$ and $y = 1$ is $\int_0^1 x^2 dx = 1/3$, so the area below the survival function $EX = 2/3$.

Example 10. *Let T be an exponential random variable, then for some λ, the survival function $\bar{F}_T(t) = P\{T > t\} = \exp(-\lambda t)$. Thus,*

$$ET = \int_0^\infty P\{T > t\}\,dt = \int_0^\infty \exp(-\lambda t)\,dt = -\frac{1}{\lambda}\exp(-\lambda t)\Big|_0^\infty = 0 - (-\frac{1}{\lambda}) = \frac{1}{\lambda}.$$

Exercise 11. *Generalize the identity (6) above to X be a positive random variable and g a* non-decreasing *function to show that the expectation*

$$Eg(X) = \int_0^\infty g(x) f_X(x)\,dx = g(0) + \int_0^\infty g'(x) P\{X > x\}\,dx.$$

The most important density function we shall encounter is

$$\phi(z) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{z^2}{2}), \quad z \in \mathbb{R}.$$

for Z, the standard normal random variable. Because the function ϕ has no simple antiderivative, we must use a numerical approximation to compute the cumulative distribution function, denoted Φ for a standard normal random variable.

Exercise 12. *Show that ϕ is increasing for $z < 0$ and decreasing for $z > 0$. In addition, show that ϕ is concave down for z between -1 and 1 and concave up otherwise.*

Example 13. *The expectation of a standard normal random variable,*

$$EZ = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty z\exp(-\frac{z^2}{2})\,dz = 0$$

because the integrand is an odd function. Next to evaluate

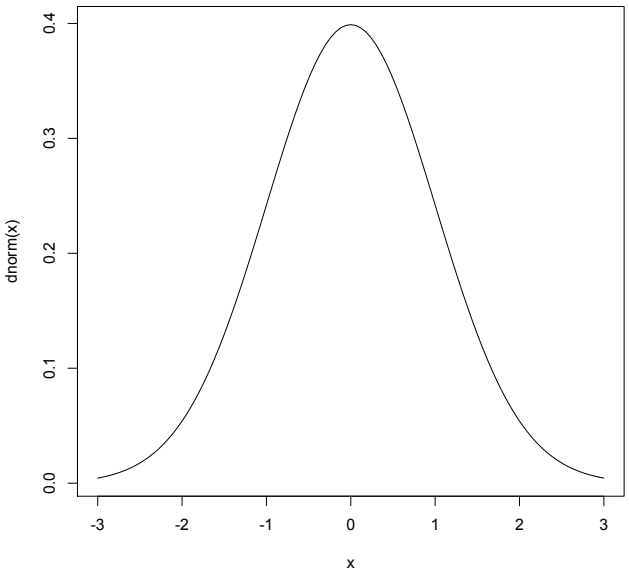


Figure 3: The density of a standard normal density, drawn in R using the command `curve(dnorm(x),-3,3)`.

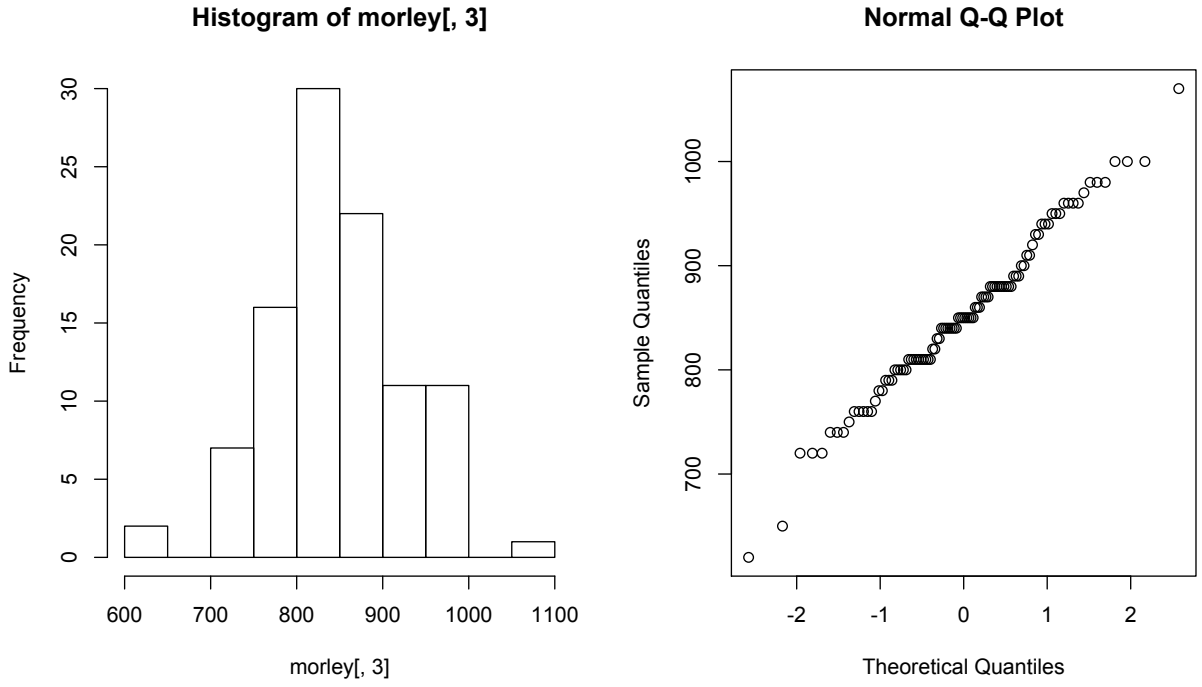


Figure 4: Histogram and normal probability plot of Morley's measurements of the speed of light.

$$EZ^2 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 \exp(-\frac{z^2}{2})\, dz,$$

we integrate by parts. (Note the choices of u *and* v'*.)*

$$\begin{array}{ll} u(z) = z & v(z) = -\exp(-\frac{z^2}{2}) \\ u'(z) = 1 & v'(z) = z\exp(-\frac{z^2}{2}) \end{array}$$

Thus,

$$EZ^2 = \frac{1}{\sqrt{2\pi}} \left(-z\exp(-\frac{z^2}{2})\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \exp(-\frac{z^2}{2})\, dz \right) = 1.$$

Use l'Hôpital's rule to see that the first term is 0. The fact that the integral of a probability density function is 1 shows that the second term equals 1.

Exercise 14. *For* Z *a standard normal random variable, show that* $EZ^3 = 0$ *and* $EZ^4 = 3$*.*

4 Quantile Plots and Probability Plots

We have seen the quantile-quantile or Q-Q plot provides a visual method way to compare two quantitative data sets. A more common comparison is between quantitative data and the quantiles of the probability distribution of a continuous random variable. We will demonstrate the properties with an example.

Example 15. *As anticipated by Galileo, errors in independent accurate measurements of a quantity follow approximately a sample from a normal distribution with mean equal to the true value of the quantity. The standard deviation gives information on the precision of the measuring devise. We will learn more about this aspect of measurements when we study the central limit theorem. Our example is Morley's measurements of the speed of light, found in the*

third column of the data set `morley`*. The values are the measurements of the speed of light minus 299,000 kilometers per second.*

```
> length(morley[,3])
[1] 100
> mean(morley[,3])
[1] 852.4
> sd(morley[,3])
[1] 79.01055
> par(mfrow=c(1,2))
> hist(morley[,3])
> qqnorm(morley[,3])
```

The histogram has the characteristic bell shape of the normal density. We can obtain a clearer picture of the closeness of the data to a normal distribution by drawing a **Q-Q plot***. (In the case of the normal distribution, the Q-Q plot is often called the* **normal probability plot***.) One method of making this plot begins by ordering the measurements from smallest to largest:*

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$$

If these are independent measurements from a normal distribution, then these values should be close to the quantiles of the evenly space values

$$\frac{1}{n+1}, \frac{2}{n+1}, \cdots, \frac{n}{n+1}$$

(For the Morley data, $n = 100$*). Thus, the next step is to find the values in the standard normal distribution that have these quantiles. We can find these values by applying* Φ^{-1}*, the inverse distribution function for the standard normal (*`qnorm` *in* R*), applied to* n *values listed above. Then the Q-Q plot is the scatterplot of the pairs*

$$\left(x_{(1)}, \Phi^{-1}\left(\frac{1}{n+1}\right)\right), \left(x_{(2)}, , \Phi^{-1}\left(\frac{2}{n+1}\right)\right), \ldots, \left(x_{(n)}\Phi^{-1}\left(\frac{n}{n+1}\right)\right)$$

Then a good fit of the data and a normal is distribution can be seen in how well the plot follows a straight line. Such a plot can be seen in Figure 4.

Exercise 16. *Describe the normal probability plot in the case in which the data* X *are skewed right.*

5 Summary

	distribution function $F_X(x) = P\{X \le x\}$	
discrete	random variable	continuous
mass function $f_X(x) = P\{X = x\}$		density function $f_X(x)\Delta x \approx P\{x \le X < x + \Delta x\}$
$f_X(x) \ge 0$ $\sum_{\text{all } x} f_X(x) = 1$	properties	$f_X(x) \ge 0$ $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$
$Eg(X) = \sum_{\text{all } x} g(x) f_X(x)$	expectation	$Eg(X) = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$

6 Names for $Eg(X)$.

Several choice for g have special names. We shall later have need for several of these expectations. Others are included to create a comprehensive reference list.

1. If $g(x) = x$, then $\mu = EX$ is called variously the **mean**, and the **first moment.**
2. If $g(x) = x^k$, then EX^k is called the k-**th moment**. These names were made in analogy to a similar concept in physics. The second moment in physics is associated to the moment of inertia.
3. For integer valued random variables, if $g(x) = (x)_k$, where $(x)_k = x(x-1)\cdots(x-k+1)$, then $E(X)_k$ is called the k-**th factorial moment.** For random variable taking values in the natural numbers $x = 0, 1, 2, \ldots$, factorial moments are typically easier to compute than moments for these random variables.
4. If $g(x) = (x-\mu)^k$, then $E(X-\mu)^k$ is called the k-**th central moment.**
5. The most frequently-used central moment is the second central moment $\sigma^2 = E(X-\mu)^2$ commonly called the **variance.** Note that

$$\sigma^2 = \text{Var}(X) = E(X-\mu)^2 = EX^2 - 2\mu EX + \mu^2 = EX^2 - 2\mu^2 + \mu^2 = EX^2 - \mu^2.$$

   This gives a frequently used alternative to computing the variance. In analogy with the corresponding concept with quantitative data, we call σ the **standard deviation**.

   Exercise 17. *Find the variance of a single Bernoulli trial.*

   Exercise 18. *Compute the variance for the two types of dice in Exercise 1.*

   Exercise 19. *Compute the variance for the dart example.*

   Note that

$$Z = \frac{X-\mu}{\sigma}$$

   has mean 0 and variance 1. Z is called the **standardized version** of X.
6. The third moment of the standardized random variable

$$E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$

   is called the **skewness**.
7. The fourth moment of the standard normal random variable is 3. The **kurtosis** compares the fourth moment of the standardized random variable to this value

$$E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] - 3.$$

8. For d-dimensional vectors $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_d)$ define the **standard inner product**,

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{d} x_i y_i.$$

   If X is $\mathbb{R}^d$-valued and $g(x) = e^{i\langle \theta, x\rangle}$, then $\phi_X(\theta) = Ee^{i\langle \theta, X\rangle}$ is called the **Fourier transform** or the **characteristic function**. The characteristic function receives its name from the fact that the mapping

$$F_X \mapsto \phi_X$$

from the distribution function to the characteristic function is one-to-one. Consequently, if we have a function that we know to be a characteristic function, then it can only have arisen from one distribution. In this way, ϕ_X characterizes that distribution.

9. Similarly, if X is $\mathbb{R}^d$-valued and $g(x) = e^{\langle \theta, x \rangle}$, then $M_X(\theta) = Ee^{\langle \theta, X \rangle}$ is called the **Laplace transform** or the **moment generating function**. The moment generating function also gives a one-to-one mapping. However, not every distribution has a moment generating function. To justify the name, consider the one-dimensional case $M_X(\theta) = Ee^{\theta X}$. Then, by noting that
$$\frac{d^k}{d\theta^k} e^{\theta x} = x^k e^{\theta x},$$
we substitute the random variable X for x, take expectation and evaluate at $\theta = 0$.
$$\begin{array}{ll} M_X'(\theta) = EXe^{\theta X} & M_X'(0) = EX \\ M_X''(\theta) = EX^2 e^{\theta X} & M_X''(0) = EX^2 \\ \vdots & \vdots \\ M_X^{(k)}(\theta) = EX^k e^{\theta X} & M_X^{(k)}(0) = EX^k. \end{array}$$

10. Let X have the natural numbers for its state space and $g(x) = z^x$, then $\rho_X(z) = Ez^X = \sum_{x=0}^{\infty} P\{X = x\} z^x$ is called the **(probability) generating function**. For these random variables, the probability generating function allows us to use ideas from the analysis of the complex variable power series.

Exercise 20. *Show that the moment generating function for an exponential random variable is*
$$M_X(t) = \frac{\lambda}{\lambda - t}.$$
Use this to find $\text{Var}(X)$.

Exercise 21. *For the probability generating function, show that* $\rho_X^{(k)}(1) = E(X)_k$. *This gives an instance that shows that falling factorial moments are easier to compute for natural number valued random variables.*

Particular attention should be paid to the next exercise.

Exercise 22. $\text{Var}(aX + b) = a^2 \text{Var}(X)$.

The variance is meant to give a sense of the spread of the values of a random variable. Thus, the addition of a constant b should not change the variance. If we write this in terms of standard deviation, we have that
$$\sigma_{aX+b} = |a| \sigma_X.$$
Thus, multiplication by a factor a spreads the data, as measured by the standard deviation, by a factor of $|a|$. For example
$$\text{Var}(X) = \text{Var}(-X).$$
These identities are identical to those for a sample variance s^2 and sample standard deviation s.

7 Independence

Expected values in the case of more than one random variable is based on the same concepts as for a single random variable. For example, for two discrete random variables X_1 and X_2, the expected value is based on the **joint mass function** $f_{X_1,X_2}(x_1, x_2)$. In this case the expected value is computed using a double sum seen in the identity (3).

We will not investigate this in general, but rather focus on the case in which the random variables are independent. Here, we have the factorization identity $f_{X_1,x_2}(x_1,x_2) = f_{X_1}(x_1)f_{X_2}(x_2)$ Now, apply identity (3) to the the product of functions $g(x_1,x_2) = g_1(x_1)g_2(x_2)$ to find that

$$\begin{aligned} E[g_1(X_1)g_2(X_2)] &= \sum_{x_1}\sum_{x_2} g_1(x_1)g_2(x_2)f_{X_1,X_2}(x_1,x_2) = \sum_{x_1}\sum_{x_2} g_1(x_1)g_2(x_2)f_{X_1}(x_1)f_{X_2}(x_2) \\ &= \left(\sum_{x_1} g_1(x_1)f_{X_1}(x_1)\right)\left(\sum_{x_2} g_2(x_2)f_{X_2}(x_2)\right) = E[g_1(X_1)]\cdot E[g_2(X_2)] \end{aligned}$$

A similar identity that the expectation of the product of two independent random variables equals to the product of the expectation holds for continuous random variables.

A very important example begins by taking X_1 and X_2 random variables with respective means μ_1 and μ_2, then by the definition of variance

$$\begin{aligned} \text{Var}(X_1+X_2) &= E[((X_1+X_2)-(\mu_1+\mu_2))^2] \\ &= E[((X_1-\mu_1)+(X_2-\mu_2))^2] \\ &= E[(X_1-\mu_1)^2] + 2E[(X_1-\mu_1)(X_2-\mu_2)] \\ &\quad + E[(X_2-\mu_2)^2] \\ &= \text{Var}(X_1) + 2\text{Cov}(X_1,X_2) + \text{Var}(X_2). \end{aligned}$$

where the **covariance** $\text{Cov}(X_1,X_2) = E[(X_1-\mu_1)(X_2-\mu_2)]$.

As you can see, the definition of covariance is analogous to that for a sample covariance. The analogy continues to hold for the **correlation** ρ, defined by

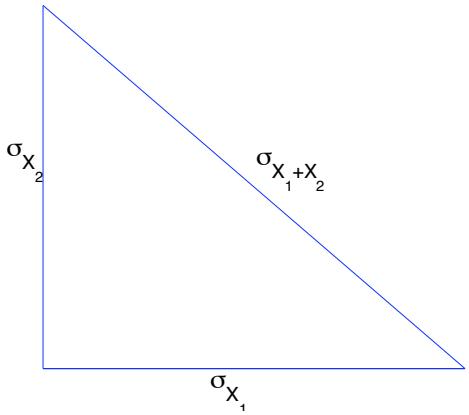


Figure 5: For independent random variables, the standard deviations σ_{X_1} and σ_{X_2} satisfy the Pythagorean theorem $\sigma^2_{X_1+X_2} = \sigma^2_{X_1} + \sigma^2_{X_2}$.

$$\rho(X_1,X_2) = \frac{\text{Cov}(X_1,X_2)}{\sqrt{\text{Var}(X_1)}\sqrt{\text{Var}(X_2)}}.$$

We can also use the computation for sample covariance to see that distributional covariance is also between -1 and 1. Correlation 1 occurs only when X and Y have a perfect positive linear association. Correlation -1 occurs only when X and Y have a perfect negative linear association.

If X_1 and X_2 are independent, then $\text{Cov}(X_1,X_2) = E[X_1-\mu_1]\cdot E[X_2-\mu_2] = 0$ and the variance of the sum is the sum of the variances. This identity and its analogy to the Pythagorean theorem is shown in Figure 4.

We can extend this to a generalized Pythagorean identity for n independent random variable $X_1, X_2, \ldots, X_n$ each having a finite variance. Then, for constants $c_1, c_2, \ldots, c_n$, we have the identity

$$\text{Var}(c_1X_1 + c_2X_2 + \cdots c_nX_n) = c_1^2\text{Var}(X_1) + c_2^2\text{Var}(X_2) + \cdots + c_n^2\text{Var}(X_n).$$

We will see several opportunities to apply this identity. If we drop the assumption of independence, we have the following. For example, if we take $c_1 = c_2 \cdots = c_n = 1$, then we have that for ndependent random variables

$$\text{Var}(X_1 + X_2 + \cdots X_n) = \text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n),$$

the variance of the sum is the sum of the variances.

Exercise 23. *Find the variance of a binomial random variable based on n trials with success parameter p.*

Exercise 24. *For random variables $X_1, X_2, \ldots, X_n$ with finite variance and constants $c_1, c_2, \ldots, c_n$*

$$\text{Var}(c_1X_1 + c_2X_2 + \cdots c_nX_n) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_ic_j\text{Cov}(X_i,X_j).$$

Recall that $\text{Cov}(X_i,X_i) = \text{Var}(X_i)$. If the random variables are independent, then $\text{Cov}(X_i,X_j) = 0$ and the identity above give the generalized Pythagorean identity.

7.1 Equivalent Conditions for Independence

We can summarize the discussions of independence to present the following 4 equivalent conditions for independent random variables $X_1, X_2, \ldots, X_n$.

1. For events $A_1, A_2, \ldots A_n$,
$$P\{X_1 \in A_1, X_2 \in A_2, \ldots X_n \in A_n\} = P\{X_1 \in A_1\}P\{X_2 \in A_2\} \cdots P\{X_n \in A_n\}.$$
2. The joint distribution function equals to the product of marginal distribution function.
$$F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = F_{X_1}(x_1)F_{X_2}(x_2) \cdots F_{X_n}(x_n).$$
3. The joint density (mass) function equals to the product of marginal density (mass) function.
$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1)f_{X_2}(x_2) \cdots f_{X_n}(x_n).$$
4. For bounded functions $g_1, g_2, \ldots, g_n$, the expectation of the product of the random variables equals to the product of the expectation.
$$E[g_1(X_1)g_2(X_2) \cdots g_n(X_n)] = Eg_1(X_1) \cdot Eg_2(X_2) \cdots Eg_n(X_n).$$

We will have many opportunities to use each of these conditions.

8 Answers to Selected Exercises

2. For the fair die
$$EX^2 = 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + 3^2 \cdot \frac{1}{6} + 4^2 \cdot \frac{1}{6} + 5^2 \cdot \frac{1}{6} + 6^2 \cdot \frac{1}{6} = (1 + 4 + 9 + 16 + 25 + 36) \cdot \frac{1}{6} = \frac{91}{6}.$$
For the unfair dice
$$EX^2 = 1^2 \cdot \frac{1}{4} + 2^2 \cdot \frac{1}{4} + 3^2 \cdot \frac{1}{4} + 4^2 \cdot \frac{1}{12} + 5^2 \cdot \frac{1}{12} + 6^2 \cdot \frac{1}{12} = (1 + 4 + 9) \cdot \frac{1}{4} + (16 + 25 + 36) \cdot \frac{1}{12} = \frac{119}{12}.$$

5. The random variable X can take on the values 0, 1, 2, 3, 4, and 5. Thus,
$$EX = \sum_{x=0}^{5} x f_X(x) \text{ and } EX^2 = \sum_{x=0}^{5} x^2 f_X(x).$$
The R commands and output follow.

```
> hearts<-c(0:5)
> f<-choose(13,hearts)*choose(39,5-hearts)/choose(52,5)
> sum(f)
[1] 1
> prod<-hearts*f
> prod2<-hearts^2*f
> data.frame(hearts,f,prod,prod2)
  hearts            f       prod      prod2
1      0 0.2215336134 0.00000000 0.00000000
2      1 0.4114195678 0.41141957 0.41141957
```

```
3      2 0.2742797119 0.54855942 1.09711885
4      3 0.0815426170 0.24462785 0.73388355
5      4 0.0107292917 0.04291717 0.17166867
6      5 0.0004951981 0.00247599 0.01237995
> sum(prod);sum(prod2)
[1] 1.25
[1] 2.426471
```

Look in the text for an alternative method to find EX.

8. If X is a non-negative random variable, then $P\{X > 0\} = 1$. Taking complements, we find that

$$F_X(0) = P\{X \leq 0\} = 1 - P\{X > 0\} = 1 - 1 = 0.$$

9. The convergence can be seen by the following argument.

$$0 \leq b(1 - F_X(b)) = b\int_b^\infty f_X(x)\,dx = \int_b^\infty bf_X(x)\,dx \leq \int_b^\infty xf_X(x)\,dx$$

Use the fact that $x \geq b$ in the range of integration to obtain the inequality in the line above.. Because, $\int_0^\infty xf_X(x)\,dx < \infty$ we have that $\int_b^\infty xf_X(x)\,dx \to 0$ as $b \to \infty$. Consequently, $0 \leq b(1 - F_X(b) \to 0$ as $b \to \infty$ b the squeeze theorem.

11. The expectation is the integral

$$Eg(X) = \int_0^\infty g(x)f_X(x)\,dx.$$

It will be a little easier to look at $h(x) = g(x) - g(0)$. Then

$$Eg(X) = g(0) + Eh(X).$$

For integration by parts, we have

$$\begin{array}{ll} u(x) = h(x) & v(x) = -(1 - F_X(x)) = -\bar{F}_X(x) \\ u'(x) = h'(x) = g'(x) & v'(x) = f_X(x) = -\bar{F}'_X(x). \end{array}$$

Again, because $F_X(0) = 0$, $\bar{F}_X(0) = 1$ and

$$\begin{aligned} Eh(X) = \int_0^b h(x)f_X(x)\,dx &= -h(x)\bar{F}_X(x)\Big|_0^b + \int_0^b h'(x)(1 - F_X(x))\,dx \\ &= -h(b)\bar{F}_X(b) + \int_0^b g'(x)\bar{F}_X(x)\,dx \end{aligned}$$

To see that the product term in the integration by parts formula converges to 0 as $b \to \infty$, note that, similar to Exercise 8,

$$0 \leq h(b)(1 - F_X(b)) = h(b)\int_b^\infty f_X(x)\,dx = \int_b^\infty h(b)f_X(x)\,dx \leq \int_b^\infty h(x)f_X(x)\,dx$$

The first inequality uses the assumption that $h(b) \geq 0$. The second uses the fact that h is non-decreaasing. Thus, $h(x) \geq h(b)$ if $x \geq b$. Now, because $\int_0^\infty h(x)f_X(x)\,dx < \infty$, we have that $\int_b^\infty h(x)f_X(x)\,dx \to 0$ as $b \to \infty$. Consequently, $h(b)(1 - F_X(b) \to 0$ as $b \to \infty$ by the squeeze theorem.

12. For the density function ϕ, the derivative

$$\phi'(z) = \frac{1}{\sqrt{2\pi}}(-z)\exp(-\frac{z^2}{2}).$$

Thus, the sign of $\phi'(z)$ is opposite to the sign of z, i.e.,

$$\phi'(z) > 0 \text{ when } z < 0 \quad \text{and} \quad \phi'(z) < 0 \text{ when } z > 0.$$

Consequently, ϕ is increasing when z is negative and ϕ is decreasing when z is positive. For the second derivative,

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

Thus,

$$\phi \text{ is concave down} \quad \text{if and only if} \quad \phi''(z) < 0 \quad \text{if and only if} \quad z^2 - 1 < 0.$$

This occurs if and only if z is between -1 and 1.

14. As argued above,

$$EZ^3 = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} z^3\exp(-\frac{z^2}{2})\,dz = 0$$

because the integrand is an odd function. For EZ^4, we again use integration by parts,

$$\begin{array}{ll} u(z) = z^3 & v(z) = -\exp(-\frac{z^2}{2}) \\ u'(z) = 3z^2 & v'(z) = z\exp(-\frac{z^2}{2}) \end{array}$$

Thus,

$$EZ^4 = \frac{1}{\sqrt{2\pi}}\left(-z^3\exp(-\frac{z^2}{2})\Big|_{-\infty}^{\infty} + 3\int_{-\infty}^{\infty} z^2\exp(-\frac{z^2}{2})\,dz\right) = 3EZ^2 = 3.$$

Use l'Hôpital's rule several times to see that the first term is 0. The integral is EZ^2 which we have previously found to be equal to 1.

15. Let X be a Bernoulli random variable. $\mu = EX = p$. Note that because X takes on only the values 0 and 1, $X = X^2$ and so $EX^2 = p$. Thus,

$$\text{Var}(X) = EX^2 - \mu^2 = p - p^2 = p(1-p).$$

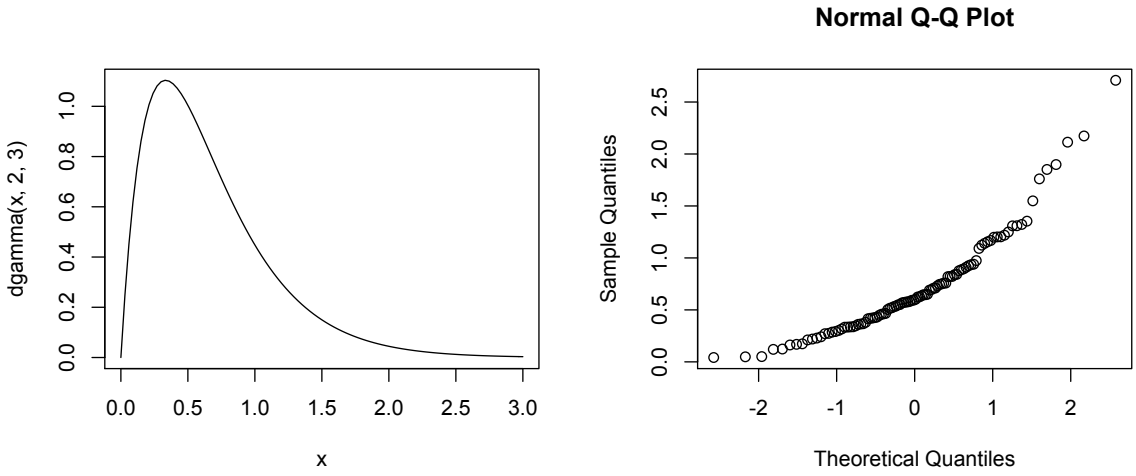
16. For the larger order statistics, $z_{(k)}$ for the standardized version of the observations, the values are larger than what one would expect when compared to observations of a standard normal random variable. Thus, the probability plot will have a concave upward shape. As an example, we let X have the density shown below. Beside this is the probability plot for X based on 100 samples. (X is a $\Gamma(2,3)$ random variable. We will encounter these random variables soon.)

17. For the fair die, the mean $\mu = EX = 7/2$ and the second moment $EX^2 = 91/6$. Thus,

$$\text{Var}(X) = EX^2 - \mu^2 = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{182-147}{12} = \frac{35}{12}.$$

For the unfair die, the mean $\mu = EX = 11/4$ and the second moment $EX^2 = 119/12$. Thus,

$$\text{Var}(X) = EX^2 - \mu^2 = \frac{119}{12} - \left(\frac{11}{4}\right)^2 = \frac{476-363}{48} = \frac{113}{48}.$$



18. For the dart, we have that the mean $\mu = EX = 2/3$.

$$EX^2 = \int_0^1 x^2 \cdot 2x \, dx = \int_0^1 2x^3 \, dx = \frac{2}{4}x^4\Big|_0^1 = \frac{1}{2}.$$

Thus,

$$\text{Var}(X) = EX^2 - \mu^2 = \frac{1}{2} - \left(\frac{2}{3}\right)^2 = \frac{1}{18}.$$

19. If $t < \lambda$, we have that $e^{(t-\lambda)x} \to 0$ as $x \to \infty$ and so

$$M_X(t) = Ee^{tX} = \lambda \int_0^\infty e^{tx} e^{-\lambda x} \, dx = \lambda \int_0^\infty e^{(t-\lambda)x} \, dx = \frac{\lambda}{t-\lambda} e^{(t-\lambda)x}\Big|_0^\infty = \frac{\lambda}{\lambda - t}$$

Thus,

$$M'(t) = \frac{\lambda}{(\lambda - t)^2}, \quad EX = M'(0) = \frac{1}{\lambda},$$

and

$$M''(t) = \frac{2\lambda}{(\lambda - t)^3}, \quad EX = M''(0) = \frac{2}{\lambda^2}.$$

Thus,

$$\text{Var}(X) = EX^2 - (EX)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

20. $\rho_X(z) = Ez^X = \sum_{x=0}^\infty P\{X = x\} z^x$ The k-th derivative of z^x with respect to z is

$$\frac{d^k}{dz^k} z^x = (x)_k z^{x-k}.$$

Evaluating at $z = 1$, we find that

$$\frac{d^k}{dz^k} z^x\Big|_{z=1} = (x)_k.$$

Thus the k-th derivative of ρ,

$$\rho_X^{(k)}(z) = \sum_{x=0}^{\infty}(x)_k P\{X = x\}z^{x-k} \text{ and, thus,}$$
$$\rho_X^{(k)}(1) = \sum_{x=0}^{\infty}(x)_k P\{X = x\} = E(X)_k.$$

21. Let $EX = \mu$. Then the expected value $E[aX + b] = a\mu + b$ and the variance

$$\text{Var}(aX + b) = E[((aX + b) - (a\mu + b))^2] = E[(a(X - \mu))^2] = a^2E[(X - \mu)^2] = a^2\text{Var}(X).$$

22. This binomial random variable is the sum on n independent Bernoulli random variable. Each of these random variables has variance $p(1 - p)$. Thus, the binomial random variable has variance $np(1 - p)$.