

11.2: The Binomial Distribution

Basic Theory

Definitions

Our random experiment is to perform a sequence of Bernoulli trials $\mathbf{X} = (X_1, X_2, \dots)$. Recall that \mathbf{X} is a sequence of independent, identically distributed indicator random variables, and in the usual language of reliability, 1 denotes success and 0 denotes failure. The common probability of success $p = \mathbb{P}(X_i = 1)$, is the basic parameter of the process. In statistical terms, the first n trials (X_1, X_2, \dots, X_n) form a random sample of size n from the Bernoulli distribution.

In this section we will study the random variable that gives the number of successes in the first n trials and the random variable that gives the proportion of successes in the first n trials. The underlying distribution, the binomial distribution, is one of the most important in probability theory, and so deserves to be studied in considerable detail. As you will see, some of the results in this section have two or more proofs. In almost all cases, note that the proof from Bernoulli trials is the simplest and most elegant.

For $n \in \mathbb{N}$, the number of successes in the first n trials is the random variable

$$Y_n = \sum_{i=1}^n X_i, \quad n \in \mathbb{N} \quad (11.2.1)$$

The distribution of Y_n is the *binomial distribution* with trial parameter n and success parameter p .

Note that $\mathbf{Y} = (Y_0, Y_1, \dots)$ is the partial sum process associated with the Bernoulli trials sequence \mathbf{X} . In particular, $Y_0 = 0$, so point mass at 0 is considered a degenerate form of the binomial distribution.

Distribution Functions

The probability density function f_n of Y_n is given by

$$f_n(y) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y \in \{0, 1, \dots, n\} \quad (11.2.2)$$

Proof

Recall that if $(x_1, x_2, \dots) \in \{0, 1\}^n$ with $\sum_{i=1}^n x_i = y$ (that is, a bit string of length n with 1 occurring exactly y times), then by independence,

$$\mathbb{P}[(X_1, X_2, \dots, X_n) = (x_1, x_2, \dots, x_n)] = p^y (1-p)^{n-y} \quad (11.2.3)$$

Moreover, the number of bit strings of length n with 1 occurring exactly y times is the binomial coefficient $\binom{n}{y}$. By the additive property of probability

$$\mathbb{P}(Y_n = y) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y \in \{0, 1, \dots, n\} \quad (11.2.4)$$

Check that f_n is a valid PDF

Clearly $f_n(y) \geq 0$ for $y \in \{0, 1, \dots, n\}$. From the binomial theorem

$$\sum_{y=0}^n f_n(y) = \sum_{y=0}^n \binom{n}{y} p^y (1-p)^{n-y} = [p + (1-p)]^n = 1 \quad (11.2.5)$$

In the binomial coin experiment, vary n and p with the scrollbars, and note the shape and location of the probability density function. For selected values of the parameters, run the simulation 1000 times and compare the relative frequency function to the probability density function.

The binomial distribution is unimodal: For $k \in \{1, 2, \dots, n\}$,

1. $f_n(k) > f_n(k-1)$ if and only if $k < (n+1)p$.
2. $f_n(k) = f_n(k-1)$ if and only if $k = (n+1)p$ is an integer between 1 and n .

Proof

1. $f_n(k) > f_n(k-1)$ if and only if $\binom{n}{k} p^k (1-p)^{n-k} > \binom{n}{k-1} p^{k-1} (1-p)^{n-k+1}$ if and only if $\frac{p}{k} > \frac{1-p}{n-k+1}$ if and only if $k < (n+1)p$
2. As in (a), $f_n(k) = f_n(k-1)$ if and only if $k = (n+1)p$, which must be an integer.

Thus, the density function at first increases and then decreases, reaching its maximum value at $\lfloor (n+1)p \rfloor$. This integer is a *mode* of the distribution. In the case that $m = (n+1)p$ is an integer between 1 and n , there are two consecutive modes, at $m-1$ and m .

Now let F_n denote the distribution function of Y_n , so that

$$F_n(y) = \mathbb{P}(Y_n \leq y) = \sum_{k=0}^y f_n(k) = \sum_{k=0}^y \binom{n}{k} p^k (1-p)^{n-k}, \quad y \in \{0, 1, \dots, n\} \quad (11.2.6)$$

The distribution function F_n and the quantile function F_n^{-1} do not have simple, closed forms, but values of these functions can be computed from mathematical and statistical software.

Open the special distribution calculator and select the binomial distribution and set the view to CDF. Vary n and p and note the shape and location of the distribution/quantile function. For various values of the parameters, compute the median and the first and third quartiles.

The binomial distribution function also has a nice relationship to the beta distribution function.

The distribution function F_n can be written in the form

$$F_n(k) = \frac{n!}{(n-k-1)!k!} \int_0^{1-p} x^{n-k-1} (1-x)^k dx, \quad k \in \{0, 1, \dots, n\} \quad (11.2.7)$$

Proof

Let $G_n(k)$ denote the expression on the right. Substitution and simple integration shows that $G_n(0) = (1-p)^n = f_n(0) = F_n(0)$. For $k \in \{1, 2, \dots, n\}$, integrating by parts with $u = (1-x)^k$ and $dv = x^{n-k-1} dx$ gives

$$G_n(k) = \binom{n}{k} p^k (1-p)^{n-k} + \frac{n!}{(n-k)!(k-1)!} \int_0^{1-p} x^{n-k} (1-x)^k dx = f_n(k) + G_n(k-1) \quad (11.2.8)$$

It follows that $G_n(k) = \sum_{j=0}^k f_n(j) = F_n(k)$ for $k \in \{0, 1, \dots, n\}$.

The expression on the right in the previous theorem is the beta distribution function, with left parameter $n-k$ and right parameter $k+1$, evaluated at $1-p$.

Moments

The mean, variance and other moments of the binomial distribution can be computed in several different ways. Again let $Y_n = \sum_{i=1}^n X_i$ where $\mathbf{X} = (X_1, X_2, \dots)$ is a sequence of Bernoulli trials with success parameter p .

The mean and variance of Y_n are

1. $E(Y_n) = np$
2. $\text{var}(Y_n) = np(1-p)$

Proof from Bernoulli trials

1. Recall that $\mathbb{E}(X_i) = p$ for each i . Hence from the additive property of expected value,

$$\mathbb{E}(Y_n) = \sum_{i=1}^n \mathbb{E}(X_i) = \sum_{i=1}^n p = np \quad (11.2.9)$$

2. Recall that $\text{var}(X_i) = p(1-p)$ for each i . Hence from the additive property of variance for *independent* variables,

$$\text{var}(Y_n) = \sum_{i=1}^n \text{var}(X_i) = \sum_{i=1}^n p(1-p) = np(1-p) \quad (11.2.10)$$

Direct proof of (a)

Recall the identity $y \binom{n}{y} = n \binom{n-1}{y-1}$ for $n, y \in \mathbb{N}_+$. Using the binomial theorem,

$$\mathbb{E}(Y_n) = \sum_{y=0}^n y \binom{n}{y} p^y (1-p)^{n-y} = \sum_{y=1}^n y \binom{n}{y} p^y (1-p)^{n-y} \quad (11.2.11)$$

$$= \sum_{y=1}^n n \binom{n-1}{y-1} p^y (1-p)^{n-y} = np \sum_{y=1}^n \binom{n-1}{y-1} p^{y-1} (1-p)^{(n-1)-(y-1)} \quad (11.2.12)$$

$$= np \sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{n-1-k} = np [p + (1-p)]^{n-1} = np \quad (11.2.13)$$

A similar, but more complicated proof can be used for part (b).

The expected value of Y_n also makes intuitive sense, since p should be approximately the proportion of successes in a large number of trials. We will discuss the point further in the subsection below on the [proportion of successes](#). Note that the graph of $\text{var}(Y_n)$ as a function of $p \in [0, 1]$ is parabola opening downward. In particular the maximum value of the variance is $n/4$ when $p = 1/2$, and the minimum value is 0 when $p = 0$ or $p = 1$. Of course, in the last two cases, Y_n is deterministic, taking just the value 0 if $p = 0$ and just the value n when $p = 1$.

In the binomial coin experiment, vary n and p with the scrollbars and note the location and size of the mean \pm standard deviation bar. For selected values of the parameters, run the simulation 1000 times and compare the sample mean and standard deviation to the distribution mean and standard deviation.

The probability generating function of Y_n is $P_n(t) = \mathbb{E}(t^{Y_n}) = [(1-p) + pt]^n$ for $t \in \mathbb{R}$.

Proof from Bernoulli trials

Recall that the probability generating function of a sum of independent variables is the product of the probability generating functions of the terms. Recall also that the PGF of X_i is $P(t) = \mathbb{E}(t^{X_i}) = (1-p) + pt$ for each i . Hence $P_n(t) = P^n(t)$.

Direct Proof

Using the binomial theorem yet again,

$$\mathbb{E}(t^{Y_n}) = \sum_{y=0}^n t^y \binom{n}{y} p^y (1-p)^{n-y} = \sum_{y=0}^n \binom{n}{y} (pt)^y (1-p)^{n-y} = [pt + (1-p)]^n \quad (11.2.14)$$

Recall that for $x \in \mathbb{R}$ and $k \in \mathbb{N}$, the *falling power* of x of order k is $x^{(k)} = x(x-1) \cdots (x-k+1)$. If X is a random variable and $k \in \mathbb{N}$, then $\mathbb{E}[X^{(k)}]$ is the *factorial moment* of X of order k . The probability generating function provides an easy way to compute the factorial moments of the binomial distribution.

$$\mathbb{E}[Y_n^{(k)}] = n^{(k)} p^k \text{ for } k \in \mathbb{N}.$$

Proof

Recall that $P_n^{(k)}(1) = \mathbb{E}[Y_n^{(k)}]$ where $P_n^{(k)}$ denotes the k th derivative of P_n . By simple calculus, $P_n^{(k)}(t) = n^{(k)} [(1-p) + pt]^{n-k} p^k$, so $P_n^{(k)} = n^{(k)} p^k$.

Our next result gives a recursion equation and initial conditions for the moments of the binomial distribution.

Recursion relation and initial conditions

1. $\mathbb{E}(Y_n^k) = np \mathbb{E}[(Y_{n-1} + 1)^{k-1}]$ for $n, k \in \mathbb{N}_+$
2. $\mathbb{E}(Y_n^0) = 1$ for $n \in \mathbb{N}$
3. $\mathbb{E}(Y_0^k) = 0$ for $k \in \mathbb{N}_+$

Proof

Recall again the identity $y \binom{n}{y} = n \binom{n-1}{y-1}$.

$$\mathbb{E}(Y_n^k) = \sum_{y=0}^n y^k \binom{n}{y} p^y (1-p)^{n-y} = \sum_{y=1}^n y^{k-1} y \binom{n}{y} p^y (1-p)^{n-y} \quad (11.2.15)$$

$$= \sum_{y=1}^n y^{k-1} n \binom{n-1}{y-1} p^y (1-p)^{n-y} = np \sum_{y=1}^n y^{k-1} \binom{n-1}{y-1} p^{y-1} (1-p)^{(n-1)-(y-1)} \quad (11.2.16)$$

$$= np \sum_{j=0}^{n-1} (j+1)^{k-1} \binom{n-1}{j} p^j (1-p)^{n-1-j} = np \mathbb{E}[(Y_{n-1} + 1)^{k-1}] \quad (11.2.17)$$

The ordinary (raw) moments of Y_n can be computed from the factorial moments and from the recursion relation. Here are the first four, which will be needed below.

The first four moments of Y_n are

1. $\mathbb{E}(Y_n) = np$
2. $\mathbb{E}(Y_n^2) = n^{(2)} p^2 + np$
3. $\mathbb{E}(Y_n^3) = n^{(3)} p^3 + 3n^{(2)} p^2 + np$
4. $\mathbb{E}(Y_n^4) = n^{(4)} p^4 + 6n^{(3)} p^3 + 7n^{(2)} p^2 + np$

Our final moment results gives the skewness and kurtosis of the binomial distribution.

For $p \in (0, 1)$, the skewness of Y_n is

$$\text{skew}(Y_n) = \frac{1-2p}{\sqrt{np(1-p)}} \quad (11.2.18)$$

1. $\text{skew}(Y_n) > 0$ if $p < \frac{1}{2}$, $\text{skew}(Y_n) < 0$ if $p > \frac{1}{2}$, and $\text{skew}(Y_n) = 0$ if $p = \frac{1}{2}$

2. For fixed n , $\text{skew}(Y_n) \rightarrow \infty$ as $p \downarrow 0$ and as $p \uparrow 1$
3. For fixed p , $\text{skew}(Y_n) \rightarrow 0$ as $n \rightarrow \infty$

Proof

These result follow from the standard computational formulas for skewness and kurtosis and the first three moments of the binomial distribution.

Open the binomial timeline experiment. For each of the following values of n , vary p from 0 to 1 and note the shape of the probability density function in light of the previous results on skewness.

1. $n = 10$
2. $n = 20$
3. $n = 100$

For $p \in (0, 1)$, the kurtosis of Y_n is

$$\text{kurt}(Y_n) = 3 - \frac{6}{n} + \frac{1}{np(1-p)} \quad (11.2.19)$$

1. For fixed n , $\text{kurt}(Y_n)$ decreases and then increases as a function of p , with minimum value $3 - \frac{2}{n}$ at the point of symmetry $p = \frac{1}{2}$
2. For fixed n , $\text{kurt}(Y_n) \rightarrow \infty$ as $p \downarrow 0$ and as $p \uparrow 1$
3. For fixed p , $\text{kurt}(Y_n) \rightarrow 3$ as $n \rightarrow \infty$

Proof

These result follow from the standard computational formulas for skewness and kurtosis and the first four moments of the binomial distribution.

Note that the *excess kurtosis* is $\text{kurt}(Y_n) - 3 = \frac{1}{np(1-p)} - \frac{6}{n} \rightarrow 0$ as $n \rightarrow \infty$. This is related to the [convergence of the binomial distribution to the normal](#), which we will discuss below.

The Partial Sum Process

Several important properties of the random process $\mathbf{Y} = (Y_0, Y_1, Y_2, \dots)$ stem from the fact that it is a partial sum process corresponding to the sequence $\mathbf{X} = (X_1, X_2, \dots)$ of independent, identically distributed indicator variables.

\mathbf{Y} has *stationary, independent increments*:

1. If m and n are positive integers with $m \leq n$ then $Y_n - Y_m$ has the same distribution as Y_{n-m} , namely binomial with parameters $n - m$ and p .
2. If $n_1 \leq n_2 \leq n_3 \leq \dots$ then $(Y_{n_1}, Y_{n_2} - Y_{n_1}, Y_{n_3} - Y_{n_2}, \dots)$ is a sequence of independent variables.

Proof

Every partial sum process corresponding to a sequence of independent, identically distributed variables has stationary, independent increments.

The following result gives the finite dimensional distributions of \mathbf{Y} .

The joint probability density functions of the sequence \mathbf{Y} are given as follows:

$$\mathbb{P}(Y_{n_1} = y_1, Y_{n_2} = y_2, \dots, Y_{n_k} = y_k) = \binom{n_1}{y_1} \binom{n_2 - n_1}{y_2 - y_1} \dots \binom{n_k - n_{k-1}}{y_k - y_{k-1}} p^{y_k} (1-p)^{n_k - y_k} \quad (11.2.20)$$

where $n_1, n_2, \dots, n_k \in \mathbb{N}_+$ with $n_1 < n_2 < \dots < n_k$ and where $y_1, y_2, \dots, y_k \in \mathbb{N}$ with $0 \leq y_j - y_{j-1} \leq n_j - n_{j-1}$ for each $j \in \{1, 2, \dots, k\}$.

Proof

From the stationary and independent increments properties,

$$\mathbb{P}(Y_{n_1} = y_1, Y_{n_2} = y_2, \dots, Y_{n_k} = y_k) = f_{n_1}(y_1) f_{n_2 - n_1}(y_2 - y_1) \dots f_{n_k - n_{k-1}}(y_k - y_{k-1}) \quad (11.2.21)$$

The result then follows from substitution and simplification.

Transformations that Preserve the Binomial Distribution

There are two simple but important transformations that preserve the binomial distribution.

If U is a random variable having the binomial distribution with parameters n and p , then $n - U$ has the binomial distribution with parameters n and $1 - p$.

Proof from Bernoulli trials

Recall that if (X_1, X_2, \dots) is a Bernoulli trials sequence with parameter p , then $(1 - X_1, 1 - X_2, \dots)$ is a Bernoulli trials sequence with parameter $1 - p$. Also U has the same distribution as $\sum_{i=1}^n X_i$ (binomial with parameters n and p) so $n - U$ has the same distribution as $\sum_{i=1}^n (1 - X_i)$ (binomial with parameters n and $1 - p$).

Proof from density functions

Note that $\mathbb{P}(n - U = k) = \mathbb{P}(U = n - k) = \binom{n}{k} p^{n-k} (1-p)^k$ for $k \in \{0, 1, \dots, n\}$

The sum of two independent binomial variables *with the same success parameter* also has a binomial distribution.

Suppose that U and V are independent random variables, and that U has the binomial distribution with parameters m and p , and V has the binomial distribution with parameters n and p . Then $U + V$ has the binomial distribution with parameters $m + n$ and p .

Proof from Bernoulli trials

Let (X_1, X_2, \dots) be a Bernoulli trials sequence with parameter p , and let $Y_k = \sum_{i=1}^k X_i$ for $k \in \mathbb{N}$. Then U has the same distribution as Y_m and V has the same distribution as $Y_{m+n} - Y_m$. Since Y_m and $Y_{m+n} - Y_m$ are independent, $U + V$ has the same distribution as $Y_m + (Y_{m+n} - Y_m) = Y_{m+n}$.

Proof from convolution powers

Let f denote the PDF of an indicator variable X with parameter p , so that $f(x) = p^x (1-p)^{1-x}$ for $x \in \{0, 1\}$. The binomial distribution with parameters $k \in \mathbb{N}_+$ and p has PDF $f_k = f^{*k}$, the k -fold convolution power of f . In particular, U has PDF f^{*m} , V has PDF f^{*n} and hence $U + V$ has PDF $f^{*m} * f^{*n} = f^{*(m+n)}$.

Proof from generating functions

U and V have PGFs $P_m(t) = (1 - p + pt)^m$ and $P_n(t) = (1 - p + pt)^n$ for $t \in \mathbb{R}$, respectively. Hence by independence, $U + V$ has PGF $P_m P_n = P_{m+n}$.

Sampling and the Hypergeometric Distribution

Suppose that we have a *dichotomous population*, that is a population of two types of objects. Specifically, suppose that we have m objects, and that r of the objects are type 1 and the remaining $m - r$ objects are type 0. Thus $m \in \mathbb{N}_+$ and $r \in \{0, 1, \dots, m\}$. We select n objects at random from the population, so that all samples of size n are equally likely. If the sampling is *with replacement*, the sample size n can be any positive integer. If the sampling is *without replacement*, then we must have $n \in \{1, 2, \dots, m\}$.

In either case, let X_i denote the type of the i 'th object selected for $i \in \{1, 2, \dots, n\}$ so that $Y = \sum_{i=1}^n X_i$ is the number of type 1 objects in the sample. As noted in the Introduction, if the sampling is with replacement, (X_1, X_2, \dots, X_n) is a sequence of Bernoulli trials, and hence Y has the binomial distribution parameters n and $p = r/m$. If the sampling is without replacement, then Y has the *hypergeometric distribution* with parameters m , r , and n . The hypergeometric distribution is studied in detail in the chapter on Finite Sampling Models. For reference, the probability density function of Y is given by

$$\mathbb{P}(Y = y) = \frac{\binom{r}{y} \binom{m-r}{n-y}}{\binom{m}{n}} = \binom{n}{y} \frac{r^{(y)} (m-r)^{(n-y)}}{m^{(n)}}, \quad y \in \{0, 1, \dots, n\} \quad (11.2.22)$$

and the mean and variance of Y are

$$\mathbb{E}(Y) = n \frac{r}{m}, \quad \text{var}(Y) = n \frac{r}{m} \left(1 - \frac{r}{m}\right) \frac{m-n}{m-1} \quad (11.2.23)$$

If the population size m is large compared to the sample size n , then the dependence between the indicator variables is slight, and so the hypergeometric distribution should be close to the binomial distribution. The following theorem makes this precise.

Suppose that $r_m \in \{0, 1, \dots, m\}$ for each $m \in \mathbb{N}_+$ and that $r_m/m \rightarrow p \in [0, 1]$ as $m \rightarrow \infty$. Then for fixed $n \in \mathbb{N}_+$ the hypergeometric distribution with parameters m , r_m and n converges to the binomial distribution with parameters n and p as $m \rightarrow \infty$.

Proof

The hypergeometric PDF has the form

$$g_m(y) = \binom{n}{y} \frac{r_m^{(y)} (m - r_m)^{(n-y)}}{m^{(n)}}, \quad y \in \{0, 1, \dots, n\} \quad (11.2.24)$$

Note that the fraction above has n factors in the numerator and n factors in the denominator. We can group these, in order, to form a product of n fractions. The first y fractions have the form

$$\frac{r_m - i}{m - i} \quad (11.2.25)$$

where $i \in \{0, 1, \dots, y-1\}$. Each of these converges to p as $m \rightarrow \infty$. The remaining $n - y$ fractions have the form

$$\frac{m - r_m - j}{m - y - j} \quad (11.2.26)$$

where $j \in \{0, 1, \dots, n - y - 1\}$. For fixed y and n , each of these converges to $1 - p$ as $m \rightarrow \infty$. Hence $g_m(y) \rightarrow \binom{n}{y} p^y (1-p)^{n-y}$ as $m \rightarrow \infty$ for each $y \in \{0, 1, \dots, n\}$

Under the conditions in the previous theorem, the mean and variance of the hypergeometric distribution converge to the mean and variance of the limiting binomial distribution:

1. $n \frac{r_m}{m} \rightarrow np$ as $m \rightarrow \infty$
2. $n \frac{r_m}{m} \left(1 - \frac{r_m}{m}\right) \frac{m-n}{m-1} \rightarrow np(1-p)$ as $m \rightarrow \infty$

Proof

By assumption $r_m/m \rightarrow p$ as $m \rightarrow \infty$ and n is fixed, so also $(m-n)/(m-1) \rightarrow 1$ as $n \rightarrow \infty$

In the ball and urn experiment, vary the parameters and switch between sampling without replacement and sampling with replacement. Note the difference between the graphs of the hypergeometric probability density function and the binomial probability density function. In particular, note the similarity when m is large and n small. For selected values of the parameters, and for both sampling modes, run the experiment 1000 times.

From a practical point of view, the convergence of the hypergeometric distribution to the binomial means that if the population size m is “large” compared to the sample size, then the hypergeometric distribution with parameters m , r and n is well approximated by the binomial distribution with parameters n and $p = r/m$. This is often a useful result, because the binomial distribution has fewer parameters than the hypergeometric distribution (and often in real problems, the parameters may only be known approximately). Specifically, in the approximating binomial distribution, we do not need to know the population size m and the number of type 1 objects r individually, but only in the ratio r/m . Generally, the approximation works well if m is large compared to n that $\frac{m-n}{m-1}$ is close to 1. This ensures that the variance of the hypergeometric distribution is close to the variance of the approximating binomial distribution.

Now let's return to our usual sequence of Bernoulli trials $\mathbf{X} = (X_1, X_2, \dots)$, with success parameter p , and to the binomial variables $Y_n = \sum_{i=1}^n X_i$ for $n \in \mathbb{N}$. Our next result shows that given k successes in the first n trials, the trials on which the successes occur is simply a random sample of size k chosen without replacement from $\{1, 2, \dots, n\}$.

Suppose that $n \in \mathbb{N}_+$ and $k \in \{0, 1, \dots, n\}$. Then for $(x_1, x_2, \dots, x_n) \in \{0, 1\}^n$ with $\sum_{i=1}^n x_i = k$,

$$\mathbb{P}[(X_1, X_2, \dots, X_n) = (x_1, x_2, \dots, x_n) \mid Y_n = k] = \frac{1}{\binom{n}{k}} \quad (11.2.27)$$

Proof

From the definition of conditional probability,

$$\mathbb{P}[(X_1, X_2, \dots, X_n) = (x_1, x_2, \dots, x_n) \mid Y_n = k] = \frac{\mathbb{P}[(X_1, X_2, \dots, X_n) = (x_1, x_2, \dots, x_n)]}{\mathbb{P}(Y_n = k)} = \frac{p^k(1-p)^{n-k}}{\binom{n}{k}p^k(1-p)^{n-k}} = \frac{1}{\binom{n}{k}} \quad (11.2.28)$$

Note in particular that the conditional distribution above does not depend on p . In statistical terms, this means that relative to (X_1, X_2, \dots, X_n) , random variable Y_n is a *sufficient statistic* for p . Roughly, Y_n contains all of the information about p that is available in the entire sample (X_1, X_2, \dots, X_n) . Sufficiency is discussed in more detail in the chapter on Point Estimation. Next, if $m \leq n$ then the conditional distribution of Y_m given $Y_n = k$ is hypergeometric, with population size n , type 1 size m , and sample size k .

Suppose that $p \in (0, 1)$ and that $m, n, k \in \mathbb{N}_+$ with $m \leq n$ and $k \leq n$. Then

$$\mathbb{P}(Y_m = j \mid Y_n = k) = \frac{\binom{m}{j} \binom{m-n}{k-j}}{\binom{n}{k}}, \quad j \in \{0, 1, \dots, k\} \quad (11.2.29)$$

Proof from the previous result

Given $Y_n = k$, the trial numbers of the successes form a random sample of size k chosen without replacement from $\{1, 2, \dots, n\}$. Designate trials $\{1, 2, \dots, m\}$ as type 1 and trials $\{m+1, m+2, \dots, n\}$ as type 0. Then Y_m is the number of type 1 trials in the sample, and hence (given $Y_n = k$) has the hypergeometric distribution with population size n , type 1 size m , and sample size k .

Direct proof

From the definition of conditional probability,

$$\mathbb{P}(Y_m = j \mid Y_n = k) = \frac{\mathbb{P}(Y_m = j, Y_n = k)}{\mathbb{P}(Y_n = k)} = \frac{\mathbb{P}(Y_m = j, Y_n - Y_m = k - j)}{\mathbb{P}(Y_n = k)} \quad (11.2.30)$$

But Y_m and $Y_n - Y_m$ are independent. Both variables have binomial distributions; the first with parameters m and p , and the second with parameters $n - m$ and p . Hence

$$\mathbb{P}(Y_m = j \mid Y_n = k) = \frac{\binom{m}{j} p^j (1-p)^{m-j} \binom{n-m}{k-j} p^{k-j} (1-p)^{n-m-(k-j)}}{\binom{n}{k} p^k (1-p)^{n-k}} = \frac{\binom{m}{j} \binom{m-n}{k-j}}{\binom{n}{k}} \quad (11.2.31)$$

Once again, note that the conditional distribution is independent of the success parameter p .

The Poisson Approximation

The *Poisson process* on $[0, \infty)$, named for Simeon Poisson, is a model for random points in continuous time. There are many deep and interesting connections between the Bernoulli trials process (which can be thought of as a model for random points in *discrete* time) and the Poisson process. These connections are explored in detail in the chapter on the Poisson process. In this section we just give the most famous and important result—the convergence of the binomial distribution to the Poisson distribution.

For reference, the Poisson distribution with rate parameter $r \in (0, \infty)$ has probability density function

$$g(n) = e^{-r} \frac{r^n}{n!}, \quad n \in \mathbb{N} \quad (11.2.32)$$

The parameter r is both the mean and the variance of the distribution. In addition, the probability generating function is $t \mapsto e^{t(r-1)}$.

Suppose that $p_n \in (0, 1)$ for $n \in \mathbb{N}_+$ and that $np_n \rightarrow r \in (0, \infty)$ as $n \rightarrow \infty$. Then the binomial distribution with parameters n and p_n converges to the Poisson distribution with parameter r as $n \rightarrow \infty$.

Proof from density functions

Let f_n denote the binomial PDF with parameters n and p_n . Then for $k \in \{0, 1, \dots, n\}$

$$f_n(k) = \binom{n}{k} p_n^k (1 - p_n)^{n-k} = \frac{1}{k!} [np_n] [(n-1)p_n] \cdots [(n-k+1)p_n] \left(1 - \frac{p_n}{n}\right)^{n-k} \quad (11.2.33)$$

But $(n-j)p_n \rightarrow r$ as $n \rightarrow \infty$ for fixed j . Also, using a basic theorem from calculus, $(1 - p_n/n)^{n-k} \rightarrow e^{-r}$ as $n \rightarrow \infty$. Hence $f_n(k) \rightarrow e^{-r} \frac{r^k}{k!}$ as $n \rightarrow \infty$.

Proof from generating functions

For $t \in \mathbb{R}$, using the same basic limit from calculus,

$$[(1 - p_n) + p_n t]^n = \left[1 + n \frac{p_n}{n} (t - 1)\right]^n \rightarrow e^{r(t-1)} \text{ as } n \rightarrow \infty \quad (11.2.34)$$

The left side is the PGF of the binomial distribution with parameters n and p_n , while the right side is the PGF of the Poisson distribution with parameter r .

Under the same conditions as the previous theorem, the mean and variance of the binomial distribution converge to the mean and variance of the limiting Poisson distribution, respectively.

1. $np_n \rightarrow r$ as $n \rightarrow \infty$
2. $np_n(1 - p_n) \rightarrow r$ as $n \rightarrow \infty$

Proof

By assumption, $np_n \rightarrow r$ as $n \rightarrow \infty$, and so it also follows that $p_n \rightarrow 0$ as $n \rightarrow \infty$.

Compare the Poisson experiment and the binomial timeline experiment.

1. Open the Poisson experiment and set $r = 1$ and $t = 5$. Run the experiment a few times and note the general behavior of the random points in time. Note also the shape and location of the probability density function and the mean \pm standard deviation bar.
2. Now open the binomial timeline experiment and set $n = 100$ and $p = 0.05$. Run the experiment a few times and note the general behavior of the random points in time. Note also the shape and location of the probability density function and the mean \pm standard deviation bar.

From a practical point of view, the convergence of the binomial distribution to the Poisson means that if the number of trials n is “large” and the probability of success p “small”, so that np^2 is small, then the binomial distribution with parameters n and p is well approximated by the Poisson distribution with parameter $r = np$. This is often a useful result, because the Poisson distribution has fewer parameters than the binomial distribution (and often in real problems, the parameters may only be known approximately). Specifically, in the approximating Poisson distribution, we do not need to know the number of trials n and the probability of success p *individually*, but only in the *product* np . The condition that np^2 be small means that the variance of the binomial distribution, namely $np(1 - p) = np - np^2$ is approximately r , the variance of the approximating Poisson distribution.

The Normal Approximation

Open the binomial timeline experiment. For selected values of $p \in (0, 1)$, start with $n = 1$ and successively increase n by 1. For each value of n , Note the shape of the probability density function of the number of successes and the proportion of successes. With $n = 100$, run the experiment 1000 times and compare the empirical density function to the probability density function for the number of successes and the proportion of successes

The characteristic bell shape that you should observe in the previous exercise is an example of the central limit theorem, because the binomial variable can be written as a sum of n independent, identically distributed random variables (the indicator variables).

The standard score Z_n of Y_n is the same as the standard score of M_n :

$$Z_n = \frac{Y_n - np}{\sqrt{np(1-p)}} = \frac{M_n - p}{\sqrt{p(1-p)/n}} \quad (11.2.35)$$

The distribution of Z_n converges to the standard normal distribution as $n \rightarrow \infty$.

This version of the central limit theorem is known as the *DeMoivre-Laplace theorem*, and is named after Abraham DeMoivre and Simeon Laplace. From a practical point of view, this result means that, for large n , the distribution of Y_n is approximately normal, with mean np and standard deviation $\sqrt{np(1-p)}$ and the distribution of M_n is approximately normal, with mean p and standard deviation $\sqrt{p(1-p)/n}$. Just how large n needs to be for the normal approximation to work well depends on the value of p . The rule of thumb is that we need $np \geq 5$ and $n(1-p) \geq 5$ (the first condition is the significant one when $p \leq \frac{1}{2}$ and the second condition is the significant one when $p \geq \frac{1}{2}$). Finally, when using the normal approximation, we should remember to use the continuity correction, since the binomial is a discrete distribution.

General Families

For a fixed number of trials n , the binomial distribution is a member of two general families of distributions. First, it is a general exponential distribution.

Suppose that Y has the binomial distribution with parameters n and p , where $n \in \mathbb{N}_+$ is fixed and $p \in (0, 1)$. The distribution of Y is a one-parameter exponential family with natural parameter $\ln\left(\frac{p}{1-p}\right)$ and natural statistic Y .

Proof

This follows from the definition of the general exponential family. The support set $\{0, 1, \dots, n\}$ does not depend on p , and for y in this set,

$$f_n(y) = \binom{n}{y} p^y (1-p)^{n-y} = \binom{n}{y} (1-p)^n \left(\frac{p}{1-p}\right)^y = \binom{n}{y} (1-p)^n \exp\left[y \ln\left(\frac{p}{1-p}\right)\right] \quad (11.2.36)$$

Note that the natural parameter is the logarithm of the *odds ratio* corresponding to p . This function is sometimes called the *logit* function. The binomial distribution is also a power series distribution

Suppose again that Y has the binomial distribution with parameters n and p , where $n \in \mathbb{N}_+$ is fixed and $p \in (0, 1)$. The distribution of Y is a power series distribution in the parameter $\theta = \frac{p}{1-p}$, corresponding to the function $\theta \mapsto (1+\theta)^n$.

Proof

This follows from the definition of the power series distribution. As before, for $y \in \{0, 1, \dots, n\}$

$$f_n(y) = \binom{n}{y} p^y (1-p)^{n-y} = \binom{n}{y} (1-p)^n \left(\frac{p}{1-p}\right)^y = \frac{1}{(1+\theta)^n} \binom{n}{y} \theta^y \quad (11.2.37)$$

where $\theta = \frac{p}{1-p}$. This is the power series distribution in θ , with coefficients $\binom{n}{y}$, corresponding to the function $\theta \mapsto (1+\theta)^n$.

The Proportion of Successes

Suppose again that $\mathbf{X} = (X_1, X_2, \dots)$ is a sequence of Bernoulli trials with success parameter p , and that as usual, $Y_n = \sum_{i=1}^n X_i$ is the number of successes in the first n trials for $n \in \mathbb{N}$. The *proportion of successes* in the first n trials is the random variable

$$M_n = \frac{Y_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i \quad (11.2.38)$$

In statistical terms, M_n is the sample mean of the random sample (X_1, X_2, \dots, X_n) . The proportion of successes M_n is typically used to estimate the probability of success p when this probability is unknown.

It is easy to express the probability density function of the proportion of successes M_n in terms of the probability density function of the number of successes Y_n . First, note that M_n takes the values k/n where $k \in \{0, 1, \dots, n\}$.

The probability density function of M_n is given by

$$\mathbb{P}\left(M_n = \frac{k}{n}\right) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k \in \{0, 1, \dots, n\} \quad (11.2.39)$$

Proof

Trivially, $M_n = k/n$ if and only if $Y_n = k$ for $k \in \{0, 1, \dots, n\}$.

In the binomial coin experiment, select the proportion of heads. Vary n and p with the scroll bars and note the shape of the probability density function. For selected values of the parameters, run the experiment 1000 times and compare the relative frequency function to the probability density function.

The mean and variance of the proportion of successes M_n are easy to compute from the mean and variance of the number of successes Y_n .

The mean and variance of M_n are

1. $\mathbb{E}(M_n) = p$
2. $\text{var}(M_n) = \frac{1}{n}p(1-p)$.

Proof

From the scaling properties of expected value and variance,

1. $\mathbb{E}(M_n) = \frac{1}{n}\mathbb{E}(Y_n) = \frac{1}{n}np = p$
2. $\text{var}(M_n) = \frac{1}{n^2}\text{var}(Y_n) = \frac{1}{n^2}np(1-p) = \frac{1}{n}p(1-p)$

In the binomial coin experiment, select the proportion of heads. Vary n and p and note the size and location of the mean \pm standard deviation bar. For selected values of the parameters, run the experiment 1000 times and compare the empirical moments to the distribution moments.

Recall that skewness and kurtosis are standardized measures. Since M_n and Y_n have the same standard score, the skewness and kurtosis of M_n are the same as the [skewness](#) and [kurtosis](#) of Y_n given above.

In statistical terms, part (a) of the moment result above means that M_n is an unbiased estimator of p . From part (b) note that $\text{var}(M_n) \leq \frac{1}{4n}$ for any $p \in [0, 1]$. In particular, $\text{var}(M_n) \rightarrow 0$ as $n \rightarrow \infty$ and the convergence is uniform in $p \in [0, 1]$. Thus, the estimate improves as n increases; in statistical terms, this is known as *consistency*.

For every $\epsilon > 0$, $\mathbb{P}(|M_n - p| \geq \epsilon) \rightarrow 0$ as $n \rightarrow \infty$ and the convergence is uniform in $p \in [0, 1]$.

Proof

This follows from the last result and Chebyshev's inequality.

The last result is a special case of the weak law of large numbers and means that $M_n \rightarrow p$ as $n \rightarrow \infty$ in probability. The strong law of large numbers states that the convergence actually holds with probability 1.

The *proportion* of successes M_n has a number of nice properties as an estimator of the *probability* of success p . As already noted, it is unbiased and consistent. In addition, since Y_n is a [sufficient statistic](#) for p , based on the sample (X_1, X_2, \dots, X_n) , it follows that M_n is sufficient for p as well. Since $\mathbb{E}(X_i) = p$, M_n is trivially the method of moments estimator of p . Assuming that the parameter space for p is $[0, 1]$, it is also the maximum likelihood estimator of p .

The likelihood function for $p \in [0, 1]$, based on the observed sample $(x_1, x_2, \dots, x_n) \in \{0, 1\}^n$, is $L_y(p) = p^y(1-p)^{n-y}$, where $y = \sum_{i=1}^n x_i$. The likelihood is maximized at $m = y/n$.

Proof

By definition, the likelihood function is simply the joint PDF of (X_1, X_2, \dots, X_n) thought of as a function of the parameter p , for fixed (x_1, x_2, \dots, x_n) . Thus the form of the likelihood function follows from the joint PDF given in the Introduction. If $y = 0$, L_y is decreasing and hence is maximized at $p = 0 = y/n$. If $y = n$, L_y is increasing and is maximized at $p = 1 = y/n$. If $0 < y < n$, the log-likelihood function is

$$\ln[L_y(p)] = y \ln(p) + (n-y) \ln(1-p) \quad (11.2.40)$$

and the derivative is

$$\frac{d}{dp} \ln[L_y(p)] = \frac{y}{p} - \frac{n-y}{1-p} \quad (11.2.41)$$

There is a single critical point at y/n . The second derivative of the log-likelihood function is negative, so the maximum on $(0, 1)$ occurs at the critical point.

See Estimation in the Bernoulli Model in the chapter on Set Estimation for a different approach to the problem of estimating p .

Examples and Applications

Simple Exercises

A student takes a multiple choice test with 20 questions, each with 5 choices (only one of which is correct). Suppose that the student blindly guesses. Let X denote the number of questions that the student answers correctly. Find each of the following:

1. The probability density function of X .
2. The mean of X .
3. The variance of X .
4. The probability that the student answers at least 12 questions correctly (the score that she needs to pass).

Answer

1. $\mathbb{P}(X = x) = \binom{20}{x} \left(\frac{1}{5}\right)^x \left(\frac{4}{5}\right)^{20-x}$ for $x \in \{0, 1, \dots, 20\}$
2. $\mathbb{E}(X) = 4$
3. $\text{var}(X) = \frac{16}{5}$

4. $\mathbb{P}(X \geq 12) \approx 0.000102$ She has no hope of passing.

A certain type of missile has failure probability 0.02. Let Y denote the number of failures in 50 tests. Find each of the following:

1. The probability density function of Y .
2. The mean of Y .
3. The variance of Y .
4. The probability of at least 47 successful tests.

Answer

1. $\mathbb{P}(Y = y) = \binom{50}{y} \left(\frac{1}{50}\right)^y \left(\frac{49}{50}\right)^{50-y}$ for $y \in \{0, 1, \dots, 50\}$
2. $\mathbb{E}(Y) = 1$
3. $\text{var}(Y) = \frac{49}{50}$
4. $\mathbb{P}(Y \leq 3) \approx 0.9822$

Suppose that in a certain district, 40% of the registered voters prefer candidate A . A random sample of 50 registered voters is selected. Let Z denote the number in the sample who prefer A . Find each of the following:

1. The probability density function of Z .
2. The mean of Z .
3. The variance of Z .
4. The probability that Z is less than 19.
5. The normal approximation to the probability in (d).

Answer

1. $\mathbb{P}(Z = z) = \binom{50}{z} \left(\frac{2}{5}\right)^z \left(\frac{3}{5}\right)^{50-z}$ for $z \in \{0, 1, \dots, 50\}$
2. $\mathbb{E}(Z) = 20$
3. $\text{var}(Z) = 12$
4. $\mathbb{P}(Z < 19) = 0.3356$
5. $\mathbb{P}(Z < 19) \approx 0.3330$

Coins and Dice

Recall that a *standard* die is a six-sided die. A *fair* die is one in which the faces are equally likely. An *ace-six flat* die is a standard die in which faces 1 and 6 have probability $\frac{1}{4}$ each, and faces 2, 3, 4, and 5 have probability $\frac{1}{8}$.

A standard, fair die is tossed 10 times. Let N denote the number of aces. Find each of the following:

1. The probability density function of N .
2. The mean of N .
3. The variance of N .

Answer

1. $\mathbb{P}(N = k) = \binom{10}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{10-k}$ for $k \in \{0, 1, \dots, 10\}$
2. $\mathbb{E}(N) = \frac{5}{3}$
3. $\text{var}(N) = \frac{25}{18}$

A coin is tossed 100 times and results in 30 heads. Find the probability density function of the number of heads in the first 20 tosses.

Answer

Let Y_n denote the number of heads in the first n tosses.

$$\mathbb{P}(Y_{20} = y \mid Y_{100} = 30) = \frac{\binom{20}{y} \binom{80}{30-y}}{\binom{100}{30}}, \quad y \in \{0, 1, \dots, 20\} \quad (11.2.42)$$

An ace-six flat die is rolled 1000 times. Let Z denote the number of times that a score of 1 or 2 occurred. Find each of the following:

1. The probability density function of Z .
2. The mean of Z .
3. The variance of Z .
4. The probability that Z is at least 400.
5. The normal approximation of the probability in (d)

Answer

1. $\mathbb{P}(Z = z) = \binom{1000}{z} \left(\frac{3}{8}\right)^z \left(\frac{5}{8}\right)^{1000-z}$ for $z \in \{0, 1, \dots, 1000\}$
2. $\mathbb{E}(Z) = 375$
3. $\text{var}(Z) = 1875/8$
4. $\mathbb{P}(Z \geq 400) \approx 0.0552$
5. $\mathbb{P}(Z \geq 400) \approx 0.0550$

In the binomial coin experiment, select the proportion of heads. Set $n = 10$ and $p = 0.4$. Run the experiment 100 times. Over all 100 runs, compute the square root of the average of the squares of the errors, when M used to estimate p . This number is a measure of the quality of the estimate.

In the binomial coin experiment, select the number of heads Y , and set $p = 0.5$ and $n = 15$. Run the experiment 1000 times with and compute the following:

1. $\mathbb{P}(5 \leq Y \leq 10)$
2. The relative frequency of the event $\{5 \leq Y \leq 10\}$
3. The normal approximation to $\mathbb{P}(5 \leq Y \leq 10)$

Answer

1. $\mathbb{P}(5 \leq Y \leq 10) = 0.8815$
3. $\mathbb{P}(5 \leq Y \leq 10) \approx 0.878$

In the binomial coin experiment, select the proportion of heads M and set $n = 30$, $p = 0.6$. Run the experiment 1000 times and compute each of the following:

1. $\mathbb{P}(0.5 \leq M \leq 0.7)$
2. The relative frequency of the event $\{0.5 \leq M \leq 0.7\}$
3. The normal approximation to $\mathbb{P}(0.5 \leq M \leq 0.7)$

Answer

1. $\mathbb{P}(0.5 \leq M \leq 0.7) = 0.8089$
3. $\mathbb{P}(0.5 \leq M \leq 0.7) \approx 0.808$

Famous Problems

In 1693, Samuel Pepys asked Isaac Newton whether it is more likely to get at least one 6 in 6 rolls of a die, or at least two 6's in 12 rolls, or at least three 6's in 18 rolls. This problem is known as a *Pepys' problem*; naturally, Pepys had fair dice in mind.

Solve Pepys' problem using the binomial distribution.

Answer

Let Y_n denote the number of 6's in n rolls of a fair die, so that Y_n has the binomial distribution with parameters n and $p = \frac{1}{6}$. Using the binomial PDF (and complements to simplify the computations),

1. $\mathbb{P}(Y_6 \geq 1) = 1 - \mathbb{P}(Y_6 = 0) = 0.6651$
2. $\mathbb{P}(Y_{12} \geq 2) = 1 - \mathbb{P}(Y_{12} \leq 1) = 0.6187$
3. $\mathbb{P}(Y_{18} \geq 3) = 1 - \mathbb{P}(Y_{18} \leq 2) = 0.5973$

So the first event, at least one 6 in 6 rolls, is the most likely, and in fact the three events, in the order given, decrease in probability.

With fair dice run the simulation of the dice experiment 500 times and compute the relative frequency of the event of interest. Compare the results with the theoretical probabilities above.

1. At least one 6 with $n = 6$.
2. At least two 6's with $n = 12$.
3. At least three 6's with $n = 18$.

It appears that Pepys had some sort of misguided linearity in mind, given the comparisons that he wanted. Let's solve the general problem.

For $k \in \mathbb{N}_+$, let Y_k denote the number of 6's in $6k$ rolls of a fair die.

1. Identify the distribution of Y_k .
2. Find the mean and variance of Y_k .
3. Give an exact formula for $\mathbb{P}(Y_k \geq k)$, the probability of at least k 6's in $6k$ rolls of a fair die.
4. Give the normal approximation of the probability in (c).
5. Find the limit of the probability in (d) as $k \rightarrow \infty$.

Proof

1. Y_k has the binomial distribution with parameters $n = 6k$ and $p = \frac{1}{6}$.

2. $\mathbb{E}(Y_k) = (6k)\frac{1}{6} = k$, $\text{var}(Y_k) = (6k)\frac{1}{6}\frac{5}{6} = \frac{5}{6}k$
3. $\mathbb{P}(Y_k \geq k) = 1 - \mathbb{P}(Y \leq k-1) = 1 - \sum_{j=0}^{k-1} \binom{6k}{j} \left(\frac{1}{6}\right)^j \left(\frac{5}{6}\right)^{6k-j}$
4. Using the continuity correction,

$$\mathbb{P}(Y_k \geq k) = \mathbb{P}\left(Y_k \geq k - \frac{1}{2}\right) = \mathbb{P}\left(\frac{Y_k - k}{\sqrt{(5/6)k}} \geq \frac{-1/2}{\sqrt{(5/6)k}}\right) = 1 - \Phi\left(\frac{-1/2}{\sqrt{(5/6)k}}\right) \quad (11.2.43)$$

5. $1 - \Phi\left(\frac{-1/2}{\sqrt{(5/6)k}}\right) \rightarrow 1 - \Phi(0) = \frac{1}{2}$ as $k \rightarrow \infty$

So on average, the number of 6's in $6k$ rolls of a fair die is k , and that fact might have influenced Pepys thinking. The next problem is known as *DeMere's problem*, named after Chevalier De Mere

Which is more likely: at least one 6 with 4 throws of a fair die or at least one double 6 in 24 throws of two fair dice? .

Answer

Let Y_n denote the number of 6's in n rolls of a fair die, and let Z_n denote the number of double 6's in n rolls of a pair of fair dice. Then Y_n as the binomial distribution with parameters n and $p = \frac{1}{6}$, and Z_n has the binomial distribution with parameters n and $p = \frac{1}{36}$. Using the binomial PDF,

1. $\mathbb{P}(Y_4 \geq 1) = 0.5177$
2. $\mathbb{P}(Z_{24} \geq 1) = 0.4914$

Data Analysis Exercises

In the cicada data, compute the proportion of males in the entire sample, and the proportion of males of each species in the sample.

Answer

1. $m = 0.433$
2. $m_0 = 0.636$
3. $m_1 = 0.259$
4. $m_2 = 0.5$

In the M&M data, pool the bags to create a large sample of M&Ms. Now compute the sample proportion of red M&Ms.

Answer

$m_{\text{red}} = 0.168$

The Galton Board

The *Galton board* is a triangular array of pegs. The rows are numbered by the natural numbers $\mathbb{N} = \{0, 1, \dots\}$ from top downward. Row n has $n+1$ pegs numbered from left to right by the integers $\{0, 1, \dots, n\}$. Thus a peg can be uniquely identified by the ordered pair (n, k) where n is the row number and k is the peg number in that row. The Galton board is named after Francis Galton.

Now suppose that a ball is dropped from above the top peg $(0, 0)$. Each time the ball hits a peg, it bounces to the right with probability p and to the left with probability $1-p$, independently from bounce to bounce.

The number of the peg that the ball hits in row n has the binomial distribution with parameters n and p .

In the Galton board experiment, select random variable Y (the number of moves right). Vary the parameters n and p and note the shape and location of the probability density function and the mean \pm standard deviation bar. For selected values of the parameters, click single step several times and watch the ball fall through the pegs. Then run the experiment 1000 times and watch the path of the ball. Compare the relative frequency function and empirical moments to the probability density function and distribution moments, respectively.

Structural Reliability

Recall the discussion of *structural reliability* given in the last section on Bernoulli trials. In particular, we have a system of n similar components that function independently, each with reliability p . Suppose now that the system as a whole functions properly if and only if at least k of the n components are good. Such a systems is called, appropriately enough, a *k out of n system*. Note that the *series* and *parallel* systems considered in the previous section are n out of n and 1 out of n systems, respectively.

Consider the k out of n system.

1. The state of the system is $\mathbf{1}(Y_n \geq k)$ where Y_n is the number of working components.
2. The reliability function is $r_{n,k}(p) = \sum_{i=k}^n \binom{n}{i} p^i (1-p)^{n-i}$.

In the binomial coin experiment, set $n = 10$ and $p = 0.9$ and run the simulation 1000 times. Compute the empirical reliability and compare with the true reliability in each of the following cases:

1. 10 out of 10 (series) system.
2. 1 out of 10 (parallel) system.
3. 4 out of 10 system.

Consider a system with $n = 4$ components. Sketch the graphs of $r_{4,1}$, $r_{4,2}$, $r_{4,3}$, and $r_{4,4}$ on the same set of axes.

An n out of $2n - 1$ system is a *majority rules* system.

1. Compute the reliability of a 2 out of 3 system.
2. Compute the reliability of a 3 out of 5 system
3. For what values of p is a 3 out of 5 system more reliable than a 2 out of 3 system?
4. Sketch the graphs of $r_{3,2}$ and $r_{5,3}$ on the same set of axes.
5. Show that $r_{2n-1,n}(\frac{1}{2}) = \frac{1}{2}$.

Answer

1. $r_{3,2}(p) = 3p^2 - 2p^3$
2. $r_{5,3}(p) = 10p^3 - 15p^4 + 6p^5$
3. 3 out of 5 is better for $p \geq \frac{1}{2}$

In the binomial coin experiment, compute the empirical reliability, based on 100 runs, in each of the following cases. Compare your results to the true probabilities.

1. A 2 out of 3 system with $p = 0.3$
2. A 3 out of 5 system with $p = 0.3$
3. A 2 out of 3 system with $p = 0.8$
4. A 3 out of 5 system with $p = 0.8$

Reliable Communications

Consider the transmission of bits (0s and 1s) through a noisy channel. Specifically, suppose that when bit $i \in \{0, 1\}$ is transmitted, bit i is received with probability $p_i \in (0, 1)$ and the complementary bit $1 - i$ is received with probability $1 - p_i$. Given the bits transmitted, bits are received correctly or incorrectly independently of one-another. Suppose now, that to increase reliability, a given bit I is repeated $n \in \mathbb{N}_+$ times in the transmission. *A priori*, we believe that $\mathbb{P}(I = 1) = \alpha \in (0, 1)$ and $\mathbb{P}(I = 0) = 1 - \alpha$. Let X denote the number of 1s received when bit I is transmitted n times.

Find each of the following:

1. The conditional distribution of X given $I = i \in \{0, 1\}$
2. the probability density function of X
3. $\mathbb{E}(X)$
4. $\text{var}(X)$

Answer

1. Give $I = 1$, X has the binomial distribution with parameters n and p_1 . Given $I = 0$, X has the binomial distribution with parameters n and $1 - p_0$.
2. $\mathbb{P}(X = k) = \binom{n}{k} [\alpha p_1^k (1 - p_1)^{n-k} + (1 - \alpha)(1 - p_0)^k p_0^{n-k}]$ for $k \in \{0, 1, \dots, n\}$.
3. $\mathbb{E}(X) = n [\alpha p_1 + (1 - \alpha)(1 - p_0)]$
4. $\text{var}(X) = \alpha [np_1(1 - p_1) + n^2 p_1^2] + (1 - \alpha) [np_0(1 - p_0) + n^2 (1 - p_0)^2] - n^2 [\alpha p_1 + (1 - \alpha)(1 - p_0)]^2$

Simplify the results in the last exercise in the symmetric case where $p_1 = p_0 =: p$ (so that the bits are equally reliable) and with $\alpha = \frac{1}{2}$ (so that we have no prior information).

Answer

1. Give $I = 1$, X has the binomial distribution with parameters n and p . Given $I = 0$, X has the binomial distribution with parameters n and $1 - p$.
2. $\mathbb{P}(X = k) = \frac{1}{2} \binom{n}{k} [p^k (1 - p)^{n-k} + (1 - p)^k p^{n-k}]$ for $k \in \{0, 1, \dots, n\}$.
3. $\mathbb{E}(X) = \frac{1}{2} n$
4. $\text{var}(X) = np(1 - p) + \frac{1}{2} n^2 [p^2 + (1 - p)^2] - \frac{1}{4} n^2$

Our interest, of course, is predicting the bit transmitted given the bits received.

Find the posterior probability that $I = 1$ given $X = k \in \{0, 1, \dots, n\}$.

Answer

$$\mathbb{P}(I = 1 \mid X = k) = \frac{\alpha p_1^k (1 - p_1)^{n-k}}{\alpha p_1^k (1 - p_1)^{n-k} + (1 - \alpha)(1 - p_0)^k p_0^{n-k}} \quad (11.2.44)$$

Presumably, our decision rule would be to conclude that 1 was transmitted if the posterior probability in the previous exercise is greater than $\frac{1}{2}$ and to conclude that 0 was transmitted if this probability is less than $\frac{1}{2}$. If the probability equals $\frac{1}{2}$, we have no basis to prefer one bit over the other.

Give the decision rule in the symmetric case where $p_1 = p_0 =: p$, so that the bits are equally reliable. Assume that $p > \frac{1}{2}$, so that we at least have a better than even chance of receiving the bit transmitted.

Answer

Give $X = k$, we conclude that bit 1 was transmitted if

$$k > \frac{n}{2} - \frac{1}{2} \frac{\ln(\alpha) - \ln(1 - \alpha)}{\ln(p) - \ln(1 - p)} \quad (11.2.45)$$

and we conclude that bit 0 was transmitted if the reverse inequality holds.

Not surprisingly, in the symmetric case with no prior information, so that $\alpha = \frac{1}{2}$, we conclude that bit i was transmitted if a majority of bits received are i .

Bernstein Polynomials

The *Weierstrass Approximation Theorem*, named after Karl Weierstrass, states that any real-valued function that is continuous on a closed, bounded interval can be uniformly approximated on that interval, to any degree of accuracy, with a polynomial. The theorem is important, since polynomials are simple and basic functions, and a bit surprising, since continuous functions can be quite strange.

In 1911, Sergei Bernstein gave an explicit construction of polynomials that uniformly approximate a given continuous function, using Bernoulli trials. Bernstein's result is a beautiful example of the *probabilistic method*, the use of probability theory to obtain results in other areas of mathematics that are seemingly unrelated to probability.

Suppose that f is a real-valued function that is continuous on the interval $[0, 1]$. The *Bernstein polynomial* of degree n for f is defined by

$$b_n(p) = \mathbb{E}_p[f(M_n)], \quad p \in [0, 1] \quad (11.2.46)$$

where M_n is the proportion of successes in the first n Bernoulli trials with success parameter p , as defined earlier. Note that we are emphasizing the dependence on p in the expected value operator. The next exercise gives a more explicit representation, and shows that the Bernstein polynomial is, in fact, a polynomial

The Bernstein polynomial of degree n can be written as follows:

$$b_n(p) = \sum_{k=0}^n f\left(\frac{k}{n}\right) \binom{n}{k} p^k (1-p)^{n-k}, \quad p \in [0, 1] \quad (11.2.47)$$

Proof

This follows from the change of variables theorem for expected value.

The Bernstein polynomials satisfy the following properties:

1. $b_n(0) = f(0)$ and $b_n(1) = f(1)$
2. $b_1(p) = f(0) + [f(1) - f(0)]p$ for $p \in [0, 1]$.
3. $b_2(p) = f(0) + 2[f(\frac{1}{2}) - f(0)]p + [f(1) - 2f(\frac{1}{2}) + f(0)]p^2$ for $p \in [0, 1]$

From part (a), the graph of b_n passes through the endpoints $(0, f(0))$ and $(1, f(1))$. From part (b), the graph of b_1 is a line connecting the endpoints. From (c), the graph of b_2 is parabola passing through the endpoints and the point $(\frac{1}{2}, \frac{1}{4}f(0) + \frac{1}{2}f(\frac{1}{2}) + \frac{1}{4}f(1))$.

The next result gives *Bernstein's theorem* explicitly.

$b_n \rightarrow f$ as $n \rightarrow \infty$ uniformly on $[0, 1]$.

Proof

Since f is continuous on the closed, bounded interval $[0, 1]$ it is bounded on this interval. Thus, there exists a constant C such that $|f(p)| \leq C$ for all $p \in [0, 1]$. Also, f is uniformly continuous on $[0, 1]$. Thus, for any $\epsilon > 0$ there exists $\delta > 0$ such that if $p, q \in [0, 1]$ and $|p - q| < \delta$ then $|f(p) - f(q)| < \epsilon$. From basic properties of expected value,

$$|b_n(p) - f(p)| \leq \mathbb{E}_p[|f(M_n) - f(p)|, |M_n - p| < \delta] + \mathbb{E}_p[|f(M_n) - f(p)|, |M_n - p| \geq \delta] \quad (11.2.48)$$

Hence $|b_n(p) - f(p)| \leq \epsilon + 2C\mathbb{P}_p(|M_n - p| \geq \delta)$ for any $p \in [0, 1]$. But by [weak law of large numbers](#) above, $\mathbb{P}_p(|M_n - p| \geq \delta) \rightarrow 0$ as $n \rightarrow \infty$ uniformly in $p \in [0, 1]$.

Compute the Bernstein polynomials of orders 1, 2, and 3 for the function f defined by $f(x) = \cos(\pi x)$ for $x \in [0, 1]$. Graph f and the three polynomials on the same set of axes.

Answer

1. $b_1(p) = 1 - 2p$
2. $b_2(p) = 1 - 2p$
3. $b_3(p) = 1 - \frac{3}{2}p - \frac{3}{2}p^2 + p^3$

Use a computer algebra system to compute the Bernstein polynomials of orders 10, 20, and 30 for the function f defined below. Use the CAS to graph the function and the three polynomials on the same axes.

$$f(x) = \begin{cases} 0, & x = 0 \\ x \sin\left(\frac{\pi}{x}\right), & x \in (0, 1] \end{cases} \quad (11.2.49)$$

This page titled [11.2: The Binomial Distribution](#) is shared under a [CC BY 2.0](#) license and was authored, remixed, and/or curated by [Kyle Siegrist \(Random Services\)](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.