# 2 Bernoulli Trials

## 2.1 Introduction

A Bernoulli random variable is the simplest variable one can think of. It takes only two values, 0 and 1, which are commonly identified with *failure* and *success*. The probability function for a Bernoulli variable $X$ is determined by the probability of success, which is usually denoted by $p$:

$$P(X = 1) = p.$$

Since *success* and *failure* are complementary events, the probability of failure is

$$P(X = 0) = 1 - P(X = 1) = 1 - p$$

and we frequently use the notation $q = 1 - p$.

If $A$ is an event, we define the *indicator function* of $A$ by

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A. \end{cases}$$

This random variable has value 1 when $A$ happens and 0 when it doesn't, i.e. it indicates when $A$ occurs. Therefore $\mathbf{1}_A$ has a Bernoulli distribution with $p = P(A)$.

We can easily calculate the moments of this distribution:

$$\mathrm{E}(X) = 1 \times p + 0 \times q = p. \tag{2.1}$$

Since $X = X^2$, because this variable only takes the values 0 and 1, we have that $\mathrm{E}(X^2) = p$ and therefore

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = p - p^2 = p(1 - p) = pq. \tag{2.2}$$

## 2.2 Distributions Related to Bernoulli Trials

### 2.2.1 Binomial Distribution

For $n \in \mathbb{N}$ we perform a series of independent Bernoulli trials with the same probability $p$ of success and denote by $S_n$ the total number of successes in the $n$ trials. Every trial corresponds to a Bernoulli variable $X_i, 1 \leq i \leq n$ and

$$S_n = X_1 + X_2 + \cdots + X_n. \tag{2.3}$$

The outcome of a series of $n$ Bernoulli trials can be represented by an $n$-dimensional vector $(X_1, X_2, \ldots, X_n)$ with entries equal to 0 or 1 according to the result of the corresponding trial. For instance, $(1, 1, 0)$ corresponds to a series of three trials in which the first two resulted in success and the third in failure.

To find the distribution of $S_n$ we start by answering the following question: What is the probability of getting a given vector $(i_1, i_2, \ldots, i_n)$ in a sequence of $n$ trials? Since the variables are independent

$$P(X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n) = P(X_1 = i_1)P(X_2 = i_2)\cdots P(X_n = i_n).$$

The probabilities we are multiplying on the right-hand side can only take values $p$ or $q$, depending on whether they correspond to a success or a failure. Therefore, if there are $S_n$ successes in the $n$ trials, no matter where they are placed, this probability will be equal to $p^{S_n}q^{n-S_n}$:

$$P(X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n) = p^{S_n}q^{n-S_n}.$$

To find $P(S_n = k)$ for $0 \leq k \leq n$, we have to multiply $p^k q^{n-k}$ times the number of results ($n$-dimensional vectors) that have exactly $k$ successes. This is equal to the number of ways of selecting $k$ places among $n$, which is $\binom{n}{k}$. Thus,

$$P(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k} = p_{k,n} \qquad (k = 0, 1, \ldots, n). \tag{2.4}$$

If a discrete random variable $X$ has this distribution function we say that $X$ has a *binomial distribution* with parameters $n$ and $p$ and we use the notation $X \sim b(n, p)$

If $p = 1/2$ the probability function is symmetric with respect to $n/2$, since in this case $P(d_n = k) = P(d_n = n - k)$.

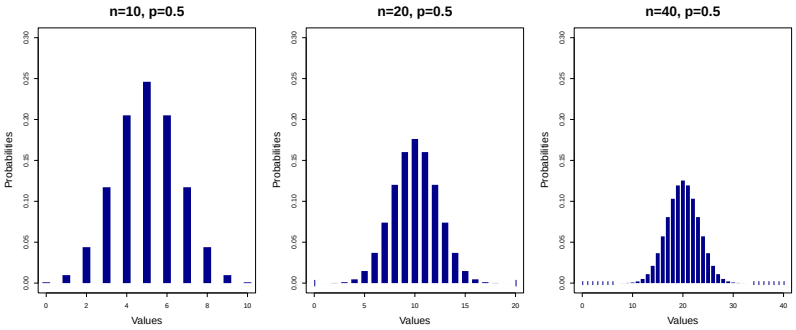


Figure 2.1: Binomial distribution for $n = 20$ and three values of $p$.

**Problem 2.1**
Five cards are drawn with replacement from a full deck. Let $X$ be the number of diamonds in the sample. Find the probability that there are exactly two diamonds among the five cards. What is the probability that there are at most two diamonds?

▶ To answer the first question we want to calculate $P(X = 2)$, and since the probability of getting a diamond in each draw is $1/4$ we have

$$P(X = 2) = \binom{5}{2}\left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right)^3 = 0.264$$

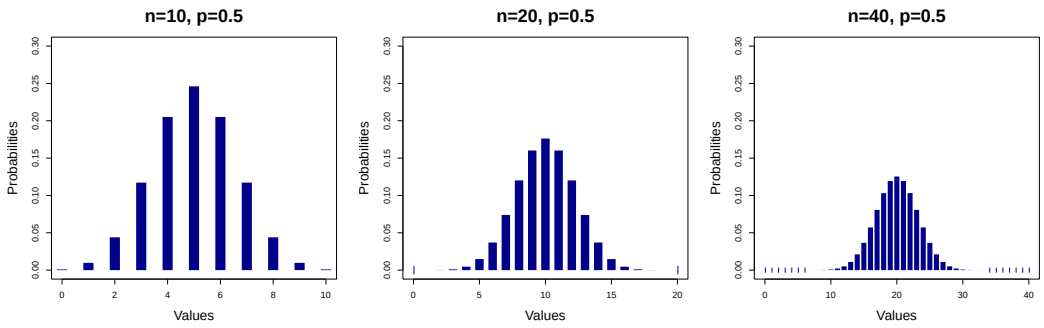


Figure 2.2: Binomial distribution for $p = 0.5$ and three values of $n$.

For the second question we have that

$$\begin{aligned} P(X \leq 2) &= P(X=0) + P(X=1) + P(X=2) \\ &= \binom{5}{0}\left(\frac{3}{4}\right)^5 + \binom{5}{1}\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)^4 + \binom{5}{2}\left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right)^3 \\ &= 0.237 + 0.396 + 0.264 \\ &= 0.897 \end{aligned}$$

▲

### Recursive relation

It is possible to get a recursive relation for the probability function. If $X \sim b(n,p)$ we have

$$\begin{aligned} P(X=k+1) &= \binom{n}{k+1} p^{k+1}(1-p)^{n-k-1} \\ &= \frac{n!}{(k+1)!(n-k-1)!} p^{k+1}(1-p)^{n-k-1} \\ &= \frac{n-k}{k+1}\frac{n!}{k!(n-k)!}\left(\frac{p}{1-p}\right)p^k(1-p)^{n-k} \\ &= \frac{n-k}{k+1}\left(\frac{p}{1-p}\right)P(X=k). \end{aligned} \tag{2.5}$$

We can use this relation starting with $P(X=0) = (1-p)^n$ or with $P(X=n) = p^n$ to obtain all the values of the probability function.

### Moments

Starting with (2.3) we get

$$\mathrm{E}(S_n) = \mathrm{E}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathrm{E}(X_i) = np.$$

Moreover, since the variables $X_1, X_2, \ldots, X_n$ are independient

$$\mathrm{Var}(S_n) = \mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathrm{Var}(X_i)$$

and

$$\mathrm{Var}(X_i) = \mathrm{E}(X_i^2) - (\mathrm{E}(X_i))^2 = p - p^2 = p(1-p).$$

Replacing in the previous equation, we get

$$\mathrm{Var}(S_n) = np(1-p).$$

### 2.2.2 Poisson Distribution

We say that the random variable $X$ follows a *Poisson distribution* with parameter $\lambda$, $(\lambda > 0)$ if

$$P(X = n) = \frac{\lambda^n}{n!} e^{-\lambda} \qquad (n = 0, 1, 2, \dots).$$

To verify that this equation defines a probability function we may use the power series expansion for the exponential function to obtain

$$\sum_{n=0}^{\infty} P(X = n) = e^{-\lambda} \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} = e^{-\lambda} e^{\lambda} = 1.$$

This is an example of a random variable taking values in an (infinite) countable set. We use the notation $X \sim \mathcal{P}(\lambda)$. Figure 2.3 shows three examples of a Poisson distribution.

This distribution has numerous applications and is interesting in itself, but it is also useful as an approximation to the binomial distribution when $n$ is big and $p$ small. To see this, consider the binomial distribution as $n$ grows and $p_n$ goes to zero so that the product $np_n$ remains constant. Let $np_n = \lambda > 0$. The binomial probability function is

$$p_{k,n} = \binom{n}{k} p_n^k (1-p_n)^{n-k} = \frac{n(n-1)\cdots(n-k+1)}{k!} p_n^k (1-p_n)^{n-k}.$$

Multiplying numerator and denominator by $n^k$ we get

$$\begin{aligned} p_{k,n} &= \frac{n(n-1)\cdots(n-k+1)}{n^k\ k!} (np_n)^k (1-p_n)^{n-k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{n^k} \frac{\lambda^k}{k!} (1-p_n)^{n-k} \\ &= \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) \dots \left(1 - \frac{k-1}{n}\right) \frac{\lambda^k}{k!} (1-p_n)^{n-k} \\ &= \frac{\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) \dots \left(1 - \frac{k-1}{n}\right)}{(1-p_n)^k} \frac{\lambda^k}{k!} (1-p_n)^n \end{aligned} \tag{2.6}$$

But we can write

$$(1-p_n)^n = [(1-p_n)^{-1/p_n}]^{-np_n} = [(1-p_n)^{-1/p_n}]^{-\lambda}$$

and by the definition of $e$ we know that

$$\lim_{z \to 0} (1-z)^{1/z} = e^{-1}.$$

Therefore, putting $z = -p_n$ we get

$$\lim_{p \to 0} (1-p_n)^n = \lim_{p \to 0} [(1-p_n)^{-1/p_n}]^{-\lambda} = e^{-\lambda}.$$

Moreover

$$\lim_{n\to\infty} \frac{\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{k-1}{n}\right)}{(1-p_n)^k} = 1$$

since we have assumed that $p_n \to 0$ as $n \to \infty$ and $np_n = \lambda$ remains constant. Using these two results in (2.6) we get

$$\lim_{n\to\infty} p_{k,n} = \frac{e^{-\lambda}\lambda^k}{k!}.$$

We have proved the following theorem:

**Theorem 2.1 (Poisson Approximation)** *Let $X_n \sim b(n, p_n)$ and assume that when $n \to \infty$, $p_n \to 0$ so that $np_n$ remains constant and is equal to $\lambda$. Then, as $n \to \infty$*

$$p_{k,n} = P(X_n = k) \to e^{-\lambda}\frac{\lambda^k}{k!}.$$

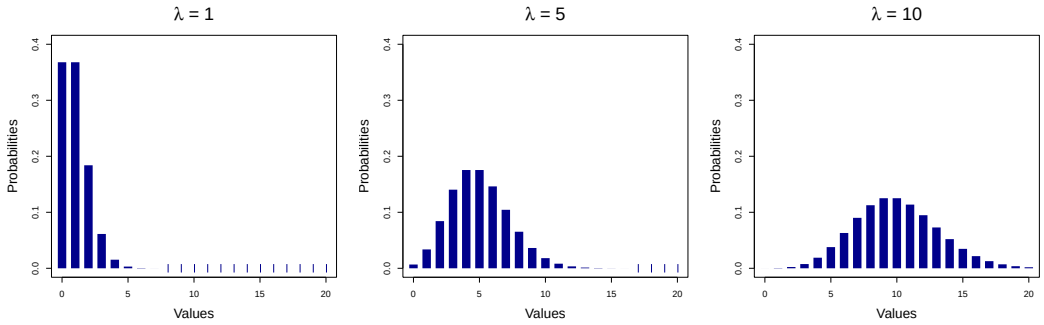


Figure 2.3: Poisson distribution for three values of $\lambda$.

**Problem 2.2**
The number of calls arriving at a switchboard have a Poisson distribution with parameter $\lambda = 4$ calls per minute. If the switchboard can handle up to six calls per minute, find the probability that the switchboard cannot to handle the calls arriving at any given minute.

▶ Let $X$ be the number of calls arriving in a period of one minute, then

$$P(X \le 6) = \sum_{i=0}^{6} P(X = i) = \sum_{i=1}^{6} \frac{e^{-4}4^i}{i!} = 0.889,$$

and therefore

$$P(X > 6) = 1 - P(X \le 6) = 1 - 0.889 = 0.11.$$

▲

**Problem 2.3**
Consider a sample of 400 fuses produced in a factory that, on average, produces 1% of faulty fuses. Find the probability that there are, at most, 5 defective fuses in the sample.

▶ Let $X$ be the number of defective fuses in the sample. We know that $X$ has a binomial distribution with $n = 400$, $p = 0.01$ and we want

$$P(X \le 5) = \sum_{i=0}^{5} P(X = i) = \sum_{i=0}^{5} \binom{400}{i} (0.01)^i (0.99)^{400-i}.$$

This can be calculated using R, as we shall se in the R section of this chapter but for now we will use the Poisson approximation with parameter

$$\lambda = np = 400 \times 0.01 = 4,$$

$$\begin{aligned} P(X \leq 5) &\approx \sum_{i=0}^{5} \frac{e^{-4}4^i}{i!} = e^{-4}\left(1 + 4 + \frac{4^2}{2} + \frac{4^3}{6} + \frac{4^4}{24} + \frac{4^5}{120}\right) \\ &\approx 0.785. \end{aligned}$$

▲

**Recursive relation**

There is also a recursive relation for the Poisson distribution that is useful to calculate its values. If $X \sim \mathcal{P}(\lambda)$ then

$$\frac{P(X = i+1)}{P(X = i)} = \frac{e^{-\lambda}\lambda^{i+1}/(i+1)!}{e^{-\lambda}\lambda^i/i!} = \frac{\lambda}{i+1}$$

i.e.

$$P(X = i+1) = \frac{\lambda}{i+1}P(X = i), \quad i \geq 0. \tag{2.7}$$

**Moments**

The expected value of a Poisson random variable is

$$\begin{aligned} \mathrm{E}(X) &= \sum_{k=0}^{\infty} kP(X = k) = \sum_{k=0}^{\infty} k\frac{\lambda^k}{k!}e^{-\lambda} \\ &= \lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda}\sum_{j=0}^{\infty}\frac{\lambda^j}{j!} \\ &= \lambda e^{-\lambda}e^{\lambda} = \lambda, \end{aligned}$$

where we have used the expansion in power series for the exponential function $e^{\lambda} = \sum_{j=0}^{\infty}\frac{\lambda^j}{j!}$. Hence

$$\mathrm{E}(X) = \lambda.$$

Also

$$\begin{aligned} \mathrm{E}(X(X-1)) &= \sum_{k=0}^{\infty} k(k-1)\frac{\lambda^k}{k!}e^{-\lambda} = \lambda^2 e^{-\lambda}\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!} \\ &= \lambda^2 e^{-\lambda}\sum_{j=0}^{\infty}\frac{\lambda^j}{j!} = \lambda^2 \end{aligned}$$

and therefore

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \mathrm{E}(X(X-1)) + \mathrm{E}(X) - (\mathrm{E}(X))^2 \\ &= \lambda^2 + \lambda - \lambda^2 = \lambda. \end{aligned}$$

Thus

$$\mathrm{Var}(X) = \lambda.$$

### 2.2.3 Geometric Distribution

Consider an electronic component that does not deteriorate as it ages but may fail due to changes in the electric current that occur randomly but at a homogeneous rate in time. The component is used every day and $X$ denotes the number of days that the component is in working order, assuming that at day zero the component is new. We want to find the probability distribution of $X$.

The idea that the component does not age can be expressed precisely as follows: if we know that the component has not failed by day $n$, i.e. $X > n$, the probability it will not fail until after day $n + m$, $P(X > n + m|X > n)$ should be equal to the probability that a new component installed on day $n$, will not fail for the next $m$ days. Since we assumed that the variations in the electric current occur homogeneously in time , this probability depends only on the number of days that have passed since the component was installed, $m$, and not on the day the component was installed, $n$. Therefore we have the equation

$$P(X > n + m|X > n) = P(X > m)$$

and using the definition of conditional probability we can write this equation as

$$P(X > n + m) = P(X > n)P(X > m) \quad n, m = 0, 1, 2, \ldots \tag{2.8}$$

Putting $n = m = 0$ we get

$$P(X > 0) = (P(X > 0))^2$$

and therefore $P(X > 0) = 0$ or 1. If $P(X > 0) = 0$ then $P(X = 0) = 1$, which corresponds to a defective component and is not interesting. Hence, $P(X > 0) = 1$.

Let $p = P(X = 1)$, then

$$P(X > 1) = 1 - p$$

and using (2.8) with $m = 1$ we get

$$P(X > n + 1) = (1 - p)P(X > n).$$

Iterating in $n$ we get that

$$P(X > n) = (1 - p)^n$$

and thus

$$\begin{aligned} P(X = n) &= P(X > n - 1) - P(X > n) \\ &= (1 - p)^{n-1} - (1 - p)^n \\ &= p(1 - p)^{n-1} \end{aligned}$$

for $n \geq 1$.

**Definition 2.1** We say that the random variable $X$ has *geometric distribution* if its probability function is given by

$$P(Y = n) = \begin{cases} p(1 - p)^{n-1} & n = 0, 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases}$$

with $0 < p < 1$. We use the notation $X \sim \mathcal{G}(p)$.

Figure 2.4 shows three examples of the geometric distribution with $p = 0.5, 0.3$ and $0.1$.

Consider now a sequence of Bernoulli trials with probability of success $p$ and let $X$ be the random variables that counts the number of trials needed to get the first success. For the event $\{X = n\}$ to occur, we need to have failures in the first $n - 1$ trials and a (first) success at trial $n$. Since the probability of failure is $q$ we have

$$P(X = k) = q^{n-1}p$$

and we see that $X$ has a geometric distribution.

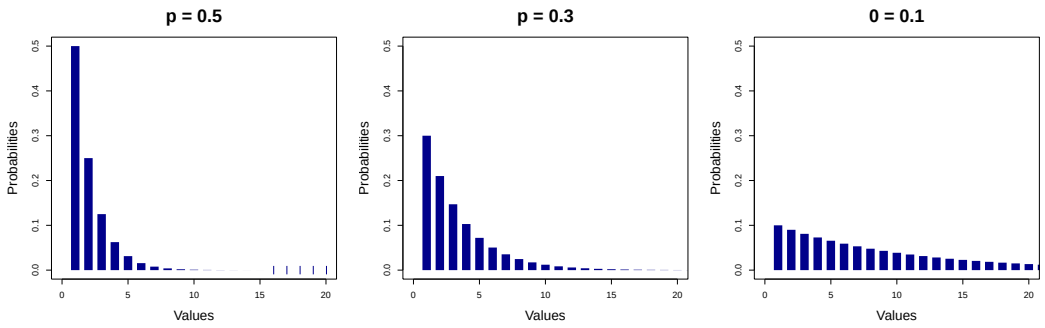


Figure 2.4: Geometric distribution for three values of $p$.

### Moments

The expected value is

$$\mathrm{E}(X) = \sum_{n=1}^{\infty} nP(X = n) = \sum_{n=1}^{\infty} np(1-p)^{n-1}. \tag{2.9}$$

To calculate this sum, recall that if

$$\sum_{n=0}^{\infty} c_n z^n$$

is a power series and $R$ is the radius of convergence (which means that if $|z| < R$ the series converges and if $R < \infty$ and $|z| > R$ the series does not converge), if we denote by $f(z)$ the sum of the series for $|z| < R$:

$$f(z) = \sum_{n=0}^{\infty} c_n z^n \qquad (|z| < R),$$

then $f$ is differentiable and

$$f'(z) = \sum_{n=0}^{\infty} n c_n z^{n-1} \qquad (|z| < R).$$

This means that for $|z| < R$, we may calculate the derivative of $f(z)$ as if it was a finite sum, i.e., adding the derivatives for each of the terms.

Let us use this result for the power series

$$\sum_{n=0}^{\infty} z^n.$$

First, we verify that the radius of convergence for this series is 1. If $|z| > 1$ then $|z|^n \to \infty$ which means that the general term of the series does not go to zero and the series does not converge; if $|z| < 1$, the partial sum of order $N$ is

$$\sum_{n=0}^{N} z^n = 1 + z + z^2 + \cdots + z^N = \frac{1 - z^{N+1}}{1 - z} \overset{N\to\infty}{\to} \longrightarrow \frac{1}{1-z}$$

which means that the series converges when $|z| < 1$ and its sum is $\frac{1}{1-z}$, i.e.

$$\sum_{n=0}^{\infty} z^n = \frac{1}{1-z}, \qquad |z| < 1.$$

Taking derivatives term by term as we have mentioned, we get

$$\frac{1}{(1-z)^2} = \sum_{n=0}^{\infty} nz^{n-1} = \sum_{n=1}^{\infty} nz^{n-1}, \qquad |z| < 1. \tag{2.10}$$

Going back to the problem we were considering, replacing $z$ by $1-p$ in (2.10) we get

$$\mathrm{E}(X) = p\sum_{n=1}^{\infty} n(1-p)^{n-1} = p\frac{1}{(1-(1-p))^2} = \frac{1}{p}.$$

To calculate the variance for this distribution start with

$$\mathrm{Var}(X) = \mathrm{E}(X(X-1)) + \mathrm{E}(X) - (\mathrm{E}(X))^2, \tag{2.11}$$

and

$$\begin{aligned}\mathrm{E}(X(X-1)) &= \sum_{n=1}^{\infty} n(n-1)P(X=n) = \sum_{n=1}^{\infty} n(n-1)p(1-p)^{n-1} \\ &= p(1-p)\sum_{n=2}^{\infty} n(n-1)(1-p)^{n-2}.\end{aligned}$$

To calculate the value of this series we proceed as before, calculating $f''(z)$ by taking derivatives term by term in the series that defines $f'(z)$, i.e. for $|z| < 1$,

$$\frac{1}{1-z} = \sum_{n=0}^{\infty} z^n \Rightarrow \frac{1}{(1-z)^2} = \sum_{n=1}^{\infty} nz^{n-1} \Rightarrow \frac{2}{(1-z)^3} = \sum_{n=2}^{\infty} n(n-1)z^{n-2}.$$

Therefore, replacing $z$ by $1-p$, we get

$$\mathrm{E}(X(X-1)) = p(1-p)\frac{2}{(1-(1-p))^3} = \frac{2(1-p)}{p^2}$$

and going back to (2.11)

$$\mathrm{Var}(X) = \frac{2(1-p)}{p^2} + \frac{1}{p} - \left(\frac{1}{p}\right)^2 = \frac{1-p}{p^2}.$$

**Problem 2.4**
You take a sample from the stock of a factory that produces, on average, 2% of defective items. What is the probability that the first defective object appears in the first five objects sampled. What the expected number of item you have to inspect to find the first defective?

▶ Let $X$ denote the number or items inspected until the first defective is found. Then

$$\begin{aligned}P(X \le 5) &= 1 - P(X \ge 5) \\ &= 1 - (0.98)^5 \\ &\approx 0.096\end{aligned}$$

On the other hand, we know that the expected value is the inverse of the success probability, i.e. $\mathrm{E}(X) = 1/0.02) = 50$.

▲

### 2.2.4 Negative Binomial Distribution.

Also known as Pascal's distribution, it is connected to sequences of independent Bernoulli trials with success probability $p$ and to the geometric distribution. Here we look for the distribution of the number of trials needed to get $k$ successes.

Let $X$ be the random variable that counts the number of trials until the $k$-th success: $X = n$ if and only if the $k$-th success occurs in the $n$-th trial. For this to happen, we need to have $k-1$ successes in the first $n-1$ trials and a success in the $n$-th trial. This last event has a probability $p$ while the probability of having $k-1$ successes in $n-1$ trials is a binomial distribution:

$$\binom{n-1}{k-1} p^{k-1} q^{n-k}.$$

Since the trials are independent, the probability $P(X = n)$ is the product of the two previous expressions, i.e.

$$P(X = n) = \binom{n-1}{k-1} p^k q^{n-k}$$

for $n \geq 1, 1 \leq k \leq n$.

Figure 2.5 presents three instances of this distribution for $p = 1/2$ and $k = 2, 5$ and 10.

**Problem 2.5**
A fisherman goes every day to the dock where where he stays fishing until he catches a fish or until two hours have passed. If the probability of not catching a fish is 0.6, what is the probability that he has to wait five days to catch three fishes?

► Let $X$ be the number of days required for catching three fishes. This variable has negative binomial distribution with parameters 3 and 0.4. Therefore

$$P(X = 5) = \binom{4}{2} (0.4)^3 (0.6)^2 = 0.138$$

▲

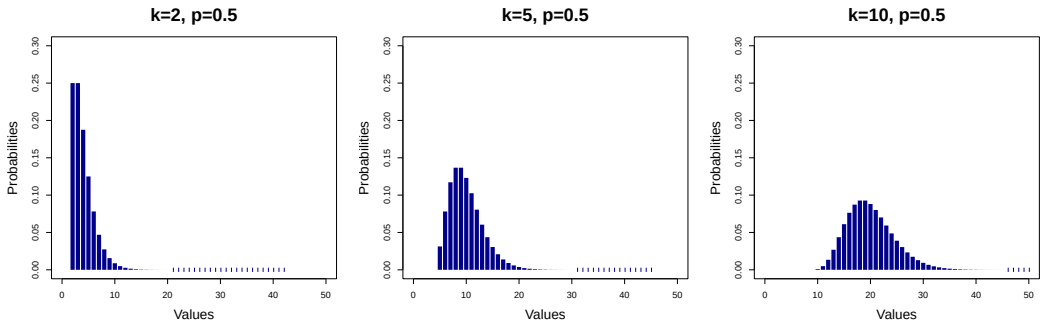


Figure 2.5: Negative Binomial distribution with $p = 0.5$ for three values of $k$.

Observe that a negative binomial random variable with parameters $k$ and $p$ is the sum of $k$ random variables with geometric distribution with parameter $p$. This observation is useful to generate random variables with negative binomial distribution and also to calculate the moments of these variables.

**Moments**

By the remark above, if $X$ has a negative binomial distribution with parameters $k$ and $p$ then $X = Y_1 + \cdots + Y_k$ where the $Y_i$ are independent geometric r. v. with parameter $p$. Hence

$$\mathrm{E}(X) = \mathrm{E}(Y_1 + \cdots + Y_k) = k\,\mathrm{E}(Y_1) = \frac{k}{p}$$

and

$$\mathrm{Var}(X) = \mathrm{Var}(Y_1 + \cdots + Y_k) = k\,\mathrm{Var}(Y_1) = \frac{k(1-p)}{p}.$$

## 2.3 The Weak Law of Large Numbers.

**Lemma 2.1 (Markov's Inequality)** *Let $X \geq 0$ be a random variable and let $a$ be a positive number, then*

$$P(X \geq a) \leq \frac{1}{a}\,\mathrm{E}(X). \tag{2.12}$$

*Proof.* If $\mathrm{E}(X) = +\infty$, there is nothing to prove. Otherwise, let $A$ be the event $A = \{X \geq a\}$, we have the following inequalities,

$$X(\omega) \geq X(\omega)\mathbf{1}_A(\omega) \geq a\mathbf{1}_A(\omega).$$

The first one is due to the fact that $X(\omega) \geq 0$ and $\mathbf{1}_A(\omega)$ is equal to 0 or 1. As for the second, if $\omega \notin A$ then $\mathbf{1}_A(\omega) = 0$ and both sides are equal. If $\omega \in A$, on the one hand $\mathbf{1}_A(\omega) = 1$ and on the other, given the definition of $A$, $X(\omega) \geq a$. We get that

$$\mathrm{E}(X) \geq \mathrm{E}(a\mathbf{1}_A) = a\,\mathrm{E}(\mathbf{1}_A) = aP(A).$$

Dividing by $a$ we get (2.12). ■

**Lemma 2.2 (Tchebychefff's Inequality)** *Let $X$ be a random variable with finite variance and let $\varepsilon > 0$. Then*

$$P(|X - \mathrm{E}(X)| \geq \varepsilon) \leq \frac{1}{\varepsilon^2}\mathrm{Var}(X). \tag{2.13}$$

*Proof.* The events $\{|X - \mathrm{E}(X)| \geq \varepsilon\}$ and $\{(X - \mathrm{E}(X))^2 \geq \varepsilon^2\}$ are equal. Using Markov's inequality with $\varepsilon^2$ in place of $a$ and $(X - \mathrm{E}(X))^2$ instead of $X$, we get

$$P(|X - \mathrm{E}(X)| \geq \varepsilon) = P((X - \mathrm{E}(X))^2 \geq \varepsilon^2) \leq \frac{1}{\varepsilon^2}\,\mathrm{E}((X - \mathrm{E}(X))^2) = \frac{1}{\varepsilon^2}\,\mathrm{Var}(X)$$

which is (2.13). ■

**Theorem 2.2 (Weak Law of Large Numbers)** *Let $\{X_n\}_{n\geq 1}$ be a sequence of independent random variables with*

$$\mathrm{E}(X_n) = \mu, \qquad \mathrm{Var}(X_n) = \sigma^2 \quad n = 1, 2, \ldots$$

*Then, for any $\varepsilon > 0$, we have*

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| \geq \varepsilon\right) = 0. \tag{2.14}$$

*Proof.* Let $S_n = X_1 + X_2 + \cdots + X_n$ and let's use Tchebychefff's inequality for $S_n/n$. We have

$$\mathrm{E}\left(\frac{S_n}{n}\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) = \frac{1}{n}n\mu = \mu,$$

$$\mathrm{Var}\left(\frac{S_n}{n}\right) = \frac{1}{n^2}\,\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Using (2.13) with $\varepsilon > 0$

$$p\left(\left|\frac{S_n}{n} - \mu\right| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2}\frac{\sigma^2}{n} \to 0$$

as $n \to \infty$ (for $\varepsilon > 0$ fixed). This completes the proof. ■

### 2.3.1 Examples and Comments.

1. Let's consider again the case of the binomial distribution in which the variable $S_n$ represents the number of times a certain event $A$ with probability $p = P(A)$ happens in $n$ independent observations. **As we have seen** it is usual to estimate the parameter $p$ – which is often unknown– using the relative frequency

   $$\hat{p}_n = \frac{S_n}{n}$$

   of the number of times the event $A$ occurs in the $n$ observations of the experiment. By the law of large numbers we have that

   $$P(|\hat{p}_n - p| \geq \varepsilon) \to 0 \quad \text{for any } \varepsilon > 0 \text{ as } n \to \infty. \tag{2.15}$$

   Property (2.15), relative to the distance between $\hat{p}_n$ (which is a function of the empirical observations that we make), and the number $p$ (which is a parameter of the problem), says that if the number $n$ of observations is large enough, then the probability that the distance between $\hat{p}_n$ and $p$ is bigger than a given number $\varepsilon$ is small.

   Tchebychefff's inequality also gives a bound on the speed with which the first term of (2.15) goes to zero when $n \to \infty$. In the example we are considering, using the results we have seen in section 2.2.1,

   $$\mathrm{E}(\hat{p}_n) = p \qquad \mathrm{Var}(\hat{p}_n) = \frac{p(1-p)}{n}.$$

   and then

   $$P(|\hat{p}_n - p| \geq \varepsilon) \leq \frac{1}{\varepsilon^2}\frac{p(1-p)}{n}. \tag{2.16}$$

   This inequality is useful to solve problems like this: Assume $p$ is the unknown probability that a certain product manufactured in a plant is defective and $S_n$ is the number of defective objects in a sample of $n$ objects inspected using random sampling with replacement. Then

   $$\hat{p}_n = \frac{S_n}{n}.$$

   We are frequently interested in the problem of finding the size $n$ of the sample so that certain tolerance margins are satisfied. For instance, assume that we do not know the true value of $p$ and we want to estimate it by $\hat{p}_n$ (see figure 2.6), but we want the probability that $p$ differs from $\hat{p}_n$ by at most 0.01, to be is bigger or equal to 0.95. That is, we want to choose $n$ so that

   $$P(|\hat{p}_n - p| \leq 0.01) \geq 0.95$$

   or equivalently

   $$P(|\hat{p}_n - p| \geq 0.01) \leq 0.05. \tag{2.17}$$

   Using (2.16) we have

   $$P(|\hat{p}_n - p| \geq 0.01) \leq \frac{1}{(0.01)^2}\frac{p(1-p)}{n}. \tag{2.18}$$

If we have no additional information about $p$, we may use the fact that $p(1-p) \leq 1/4$ for $0 \leq p \leq 1$. Choosing $n$ so that

$$\frac{1}{(0.01)^2}\frac{1}{4n} \leq 0.05, \tag{2.19}$$

we get

$$P(|\hat{p}_n - p| \geq 0.01) \leq 0.05.$$

It is easy to see that (2.19) is satisfied as long as $n \geq 50.000$.

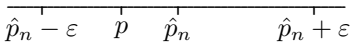


Figure 2.6

In general, Tchebycheff's inequality is not sharp. In this example one could get sharper bounds for the left-hand side of (2.18), which would reduce significantly the value of $n$ needed to verify (2.17). In particular, if we use the normal approximation –which we will study later on– one gets that if, for instance, the true value of $p$ is $1/2$ and $n = 50,000$, then the left-hand side of (2.17) is bounded by 0.00001, i.e. it is significantly smaller than the bound obtained with Tchebycheff's inequality. We will revisit this point in the next section.

2. In the proof of the weak law of large numbers, we have assumed that the variables $X_1, X_2, \ldots$ are independent and have equal variance. It is not hard to show that a similar proof is valid if the variables are uncorrelated and

$$\frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2 \to 0$$

as $n \to +\infty$, where $\sigma_i = \mathrm{Var}(X_i)$. In fact, there are many generalizations of the weak law of large numbers, some of which have technically complicated proofs and are not within the scope of this text. The interested reader may look up the classic text of W. Feller, included in the references.

3. It is interesting to observe that the weak law expresses, in some sense, the commonly accepted idea that empirical averages may replace the (theoretical) expected value. In this case, the expected value is $\mu$ and the empirical average is

$$\frac{X_1 + X_2 + \cdots + X_n}{n},$$

which is the arithmetic mean of the observations. The weak law says that if we take repeated observations, independently and under similar conditions, the empirical average is close to the expected values in the sense made precise by expression (2.14). In other words, if the number of observations is large it is unlikely that they will be far apart.

### 2.3.2 Weierstrass Approximation Theorem

This theorem says that any continuous function $f$ defined on a bounded interval $[a, b]$ can be uniformly approximated by polynomials: Given $\varepsilon > 0$ there exists a polynomial $g$ such that

$$\sup_{x\in[a,b]} |f(x) - g(x)| < \varepsilon.$$

Serge Bernstein gave a proof of this result using the WLLN, which is presented below for functions defined in $[0, 1]$. The extension to the general case is easy.

**Proposition 2.1 (Bernstein Polynomials)** *Let $f : [0,1] \to \mathbb{R}$ be a continuous function, then*

$$\sup_{x\in[0,1]} \Big| f(x) - \sum_{k=0}^{n} \binom{n}{k} f\big(\frac{k}{n}\big) x^k (1-x)^{n-k} \Big| \to 0$$

*as $n \to \infty$.*

*Proof.* Let $\varepsilon > 0$. Since $f$ is continuous in a bounded closed interval, it is uniformly continuous and there exists $\delta > 0$ such that if $0 \le x, y \le 1$ satisfy $|x - y| < \delta$ then $|f(x) - f(y)| < \varepsilon$. Consider now binomial random variables $S_n$ with parameters $n$ and $x$. The random variables $f(S_n/n)$ have expected value

$$\begin{aligned} \mathrm{E}\Big[f\Big(\frac{S_n}{n}\Big)\Big] &= \sum_{k=0}^{n} f\big(\frac{k}{n}\big) P(S_n = k) \\ &= \sum_{k=0}^{n} \binom{n}{k} f\big(\frac{k}{n}\big) x^k (1-x)^{n-k} \end{aligned}$$

By the WLLN, there exists an integer $n_0$ independent of $x$, such that, if $n \ge n_0$

$$P\Big(\Big|\frac{S_n}{n} - x\Big| > \delta\Big) < \varepsilon.$$

Using the triangular inequality we have

$$\begin{aligned} \Big|\mathrm{E}\Big[f\Big(\frac{S_n}{n}\Big)\Big] - f(x)\Big| &= \Big|\sum_{k=0}^{n} \Big(f\Big(\frac{k}{n}\Big) - f(x)\Big) P(S_n = k)\Big| \\ &\le \sum_{k=0}^{n} \Big|f\Big(\frac{k}{n}\Big) - f(x)\Big| P(S_n = k) \\ &\le \sum_{|\frac{k}{n}-x|\le\delta} \Big|f\Big(\frac{k}{n}\Big) - f(x)\Big| P(S_n = k) + \sum_{|\frac{k}{n}-x|>\delta} \Big(\Big|f\Big(\frac{k}{n}\Big)\Big| + |f(x)|\Big) P(S_n = k) \\ &\le \sum_{|\frac{k}{n}-x|\le\delta} \varepsilon P(S_n = k) + \sum_{|\frac{k}{n}-x|>\delta} 2 \sup_{0\le x\le 1} |f(x)| P(S_n = k) \\ &= \varepsilon P\Big(\Big|\frac{S_n}{n} - x\Big| \le \delta\Big) + 2 \sup_{0\le x\le 1} |f(x)| P\Big(\Big|\frac{S_n}{n} - x\Big| > \delta\Big). \end{aligned}$$

Thus, for every $n \ge n_0$,

$$\Big|\mathrm{E}\Big[f\Big(\frac{S_n}{n}\Big)\Big] - f(x)\Big| \le \varepsilon + 2\varepsilon \sup_{0\le x\le 1} |f(x)|$$

and this concludes the proof. ∎

## 2.4 Chernoff's Inequality

Tchebycheff's inequality gives a general bound for the probability $P(|X - \mathrm{E}(X)| \ge \varepsilon)$ which in many cases, as we have previously mentioned, can be improved. In this section we will do this for binomial random variables. However, it is important to observe that if we want an inequality that holds for any r. v. with finite second moment, one cannot do better than Tchebycheff's inequality.

Let $S_n = \sum_1^n X_i$ where the $X_i$ are independent and have Bernoulli distribution with success probability $p$.

**Lemma 2.3** *For any $u \in \mathbb{R}$,*

$$\mathrm{E}\left(e^{uS_n}\right) = (1 - p + pe^u)^n$$

*Proof.* Since the $X_i, 1 \leq i \leq n$ are independent, the variables $e^{uX_i}, 1 \leq i \leq n$ are also independent. On the other hand, for any $i$,

$$\mathrm{E}(e^{uX_i}) = pe^u + (1-p)e^0 = 1 - p + pe^u$$

Using independence

$$\begin{aligned}\mathrm{E}\left(e^{uS_n}\right) &= \mathrm{E}\left(\prod_{i=1}^{n} e^{uX_i}\right) = \prod_{i=1}^{n} \mathrm{E}(e^{uX_i}) \\ &= (1 - p + pe^u)^n\end{aligned}$$

■

If $X$ is any random variable, the function $\psi(u) = \mathrm{E}(e^{uX})$, for those values of $u$ for which this expected value is well-defined, is known as the *moment generating function* of $X$. In the previous lemma we calculated this function for variables with binomial distribution.

**Theorem 2.3 (Chernoff's Inequality)** *Let $S_n$ be a random variable having binomial distribution with parameters $n$ and $p$ and let $\lambda = np = \mathrm{E}(S_n)$. For $0 < \lambda + t < n$*

$$P(S_n \geq \lambda + t) \leq \Big(\frac{\lambda}{\lambda + t}\Big)^{\lambda + t} \Big(\frac{n - \lambda}{n - \lambda - t}\Big)^{n - \lambda - t} \tag{2.20}$$

*Proof.* Let $u > 0$. Since $f(x) = e^{ux}$ is an increasing function, we have

$$P(S_n \geq \lambda + t) = P\big(e^{uS_n} \geq e^{u(\lambda + t)}\big) \leq e^{-u(\lambda + t)}\, \mathrm{E}\left(e^{uS_n}\right)$$

where the inequality is a consequence of Markov's inequality. Using lemma 2.3 we get

$$P(S_n \geq \lambda + t) \leq (1 - p + pe^u)^n e^{-u(\lambda + t)} \tag{2.21}$$

Set $g(u) = (1 - p + pe^u)^n e^{-u(\lambda + t)}$ and let us find the value of $u$ minimizing this function. Taking derivatives we get

$$\begin{aligned}g'(u) &= -e^{-u(\lambda + t)}(\lambda + t)(1 - p + pe^u)^n + e^{-u(\lambda + t)} n (1 - p + pe^u)^{n-1} pe^u \\ &= e^{-u(\lambda + t)}(1 - p + pe^u)^{n-1}\big[pne^u - (\lambda + t)(1 - p + pe^u)\big] \\ &= e^{-u(\lambda + t)}(1 - p + pe^u)^{n-1}\big[-(\lambda + t)(1 - p) + pe^u(n - \lambda - t)\big]\end{aligned}$$

For this expression to vanish, it is necessary that the third factor vanishes, since the other two are always positive. This happens if

$$e^u = \frac{(\lambda + t)(1 - p)}{p(n - \lambda - t)}. \tag{2.22}$$

To see that this values corresponds to a minimum for $g$, observe that $g(0) = 1$, $g(u) \to \infty$ $(u \to \infty)$ since $\lambda + t < n$ and it is easy to verify that $g'(0) < 0$, so that the function initially decreases. Since $g$ is continuous, it must have a minimum at the value given by (2.22).

Let us go back to equation (2.21) and replace $e^u$ by the expression in (2.22)

$$\begin{aligned}P(S_n \geq \lambda + t) &\leq \Big(1 - p + \frac{(\lambda + t)(1 - p)}{(n - \lambda - t)}\Big)^n \Big(\frac{p(n - \lambda - t)}{(\lambda + t)(1 - p)}\Big)^{\lambda + t} \\ &= \Big(\frac{n(1 - p)}{n - \lambda - t}\Big)^n \Big(\frac{np(n - \lambda - t)}{n(\lambda + t)(1 - p)}\Big)^{\lambda + t}\end{aligned}$$

recalling that $\lambda = np$ and therefore $n(1-p) = n - \lambda$ we get

$$= \Big(\frac{n-\lambda}{n-\lambda-t}\Big)^n \Big(\frac{n-\lambda-t}{n-\lambda}\Big)^{\lambda+t} \Big(\frac{\lambda}{\lambda+t}\Big)^{\lambda+t}$$

which proves the inequality. ■

For $0 < x < 1$ define

$$\varphi(x) = x \log \frac{x}{p} + (1-x) \log \frac{1-x}{1-p}. \tag{2.23}$$

**Corolary 2.1** *If $n \geq 1$ and $\varepsilon \in (0, 1-p)$ then*

$$P\Big(\frac{S_n}{n} \geq p + \varepsilon\Big) \leq e^{-n\varphi(p+\varepsilon)} \tag{2.24}$$

*Proof.* Recall that $\lambda = np$, using (2.20) we get

$$P\Big(\frac{S_n}{n} \geq p + \frac{t}{n}\Big) \leq \Big(\frac{p}{p+\frac{t}{n}}\Big)^{np+t} \Big(\frac{1-p}{1-p-\frac{t}{n}}\Big)^{n-np-t}$$

as long as $0 < \lambda + t < n$. Putting $\varepsilon = t/n$ this condition becomes $0 < p + \varepsilon < 1$ and we get

$$\begin{aligned} P\Big(\frac{S_n}{n} \geq p + \varepsilon\Big) &\leq \exp\Big\{ n(p+\varepsilon) \log \frac{p}{p+\varepsilon} + n(1-p-\varepsilon) \log \frac{1-p}{1-p-\varepsilon} \Big\} \\ &= \exp\Big\{ -n\Big[ (p+\varepsilon) \log \frac{p+\varepsilon}{p} + (1-p-\varepsilon) \log \frac{1-p-\varepsilon}{1-p} \Big] \Big\} \\ &= \exp\{-n\varphi(p+\varepsilon)\}. \end{aligned}$$

■

**Corolary 2.2** *If $0 < \varepsilon < p$ then $\varphi(p-\varepsilon)$ is well defined, positive and*

$$P\Big(\frac{S_n}{n} \leq p - \varepsilon\Big) \leq e^{-n\varphi(p-\varepsilon)}. \tag{2.25}$$

*Proof.* This follows by interchanging successes and failures, $S_n$ and $n - S_n$ and $p$ and $1-p$. ■

**Corolary 2.3** *If $0 < \varepsilon < \min(p, 1-p)$ then*

$$P\Big(\Big|\frac{S_n}{n} - p\Big| \geq \varepsilon\Big) \leq e^{-n\varphi(p+\varepsilon)} + e^{-n\varphi(p-\varepsilon)}. \tag{2.26}$$

*Proof.* This is a consequence of the two previous corolaries. ■

### 2.4.1 Comparison with Tchebycheff's Inequality

In the binomial case, Tchebycheff's inequality says that

$$P(|X - \mathrm{E}(X)| \geq \varepsilon) \leq \frac{p(1-p)}{n\varepsilon^2}.$$

Observe that for $n$ and $p$ fixed, this bound goes to $\infty$ as $\varepsilon \to 0$, so that the inequality ceases to be useful for small values of $\varepsilon$. Let us look at some examples to compare this inequality with Chernoff's bound.

**Examples 2.1**
Consider the symmetric case, which corresponds to $p = 0.5$.

1. For $n = 100$ and $\varepsilon = 0.1$ Tchebycheff's inequality gives

$$P\Big(\Big|\frac{1}{100}S_{100} - \frac{1}{2}\Big| > 0.1\Big) \leq \frac{1}{4} = 0.25$$

while using Chernoff's inequality we have $h_+(0.1) = h_-(0.1) = 0.02014$ and

$$e^{-100h_+(0.1)} + e^{-100h_-(0.1)} = 0.267,$$

So that Chernoff's bound is not better than Tchebycheff's in this case. However, if we set $\varepsilon = 0.05$, Tchebycheff's bound is 1, which is useless, while Chernoff's is 0.005

2. For $n = 1000$ and $\varepsilon = 0.1$, Tchebycheff's bound is 0.025 while Chernoff's is less than $3.6 \times 10^{-9}$. Setting $\varepsilon = 0.05$ with Tchebycheff we get 0.1 while for Chernoff the bound is less than 0.0067.

3. Finally, for $n = 10,000$ and $\varepsilon = 0.1$ we get 0.0025 and $7.14 \times 10^{-88}$ while for $\varepsilon = 0.05$ the bounds are 0.01 and $1.8 \times 10^{-22}$.

**Notation for the Asymptotic Behavior of Functions**

Let us introduce the usual notation for comparison of functions, proposed by E. Landau. Let $f$ and $g$, be two real functions.

- We say that $f = \mathcal{O}(g)$ or more precisely that $f(x) = \mathcal{O}(g(x))$ as $x \to \infty$, if there exists a constant $k > 0$ and a number $x_o \in \mathbb{R}$ such that

$$|f(x)| \leq kg(x) \quad \text{for all } x \geq x_0.$$

In this case we say that $f$ is, at most, of the same order than $g$ as $x \to \infty$.

- We say that $f(x) = o(g(x))$ as $x \to \infty$ if for every $\varepsilon > 0$ there exists $x_0 \in \mathbb{R}$ such that

$$|f(x)| \leq \varepsilon g(x) \quad \text{for every } x \geq x_0$$

or equivalently, if

$$\frac{|f(x)|}{g(x)} \to 0 \quad \text{as } x \to \infty.$$

In particular, if $f(x) \to 0$ as $x \to \infty$ we write $f(x) = o(1)$ as $x \to \infty$.

- We write $f \sim g$ or $f(x) \sim g(x)$ as $x \to \infty$ if $g(x) \neq 0$ for every large $x$ and

$$\frac{f(x)}{g(x)} \to 1 \quad \text{as } x \to \infty.$$

Similar notations are used when $x \to 0$ or $x \to -\infty$, with the obvious changes, and also for sequences.

**Example 2.2**
The function $f(x) = 2x^3 - x^2 + 5$ satisfies the following relations: $f(x) = \mathcal{O}(x^3)$, $f(x) = o(x^4)$ and $f(x) \sim 2x^3$ as $x \to \infty$.

For our next result we need the second order expansion of the function $\log(1 + x)$ for $|x| < 1$, about $x = 0$:

$$\log(1 + x) = x - \frac{x^2}{2} + \mathcal{O}(x^3) \tag{2.27}$$

as $x \to 0$, which can be obtained form a Taylor expansion.

**Proposition 2.2** *As $\varepsilon \to 0$ we have that*

$$\varphi(p+\varepsilon) = \frac{\varepsilon^2}{2p(1-p)} + \mathcal{O}(\varepsilon^3). \tag{2.28}$$

*Proof.* Using (2.27) we have

$$\begin{aligned}
\varphi(p+\varepsilon) &= (p+\varepsilon)\log\Big(1+\frac{\varepsilon}{p}\Big) + (1-p-\varepsilon)\log\Big(1-\frac{\varepsilon}{1-p}\Big) \\
&= (p+\varepsilon)\Big(\frac{\varepsilon}{p} - \frac{1}{2}\Big(\frac{\varepsilon}{p}\Big)^2 + \mathcal{O}(\varepsilon^3)\Big) + (1-p-\varepsilon)\Big(\frac{-\varepsilon}{1-p} - \frac{1}{2}\Big(\frac{-\varepsilon}{1-p}\Big)^2 + \mathcal{O}(\varepsilon^3)\Big) \\
&= \frac{\varepsilon^2}{2p} + \frac{\varepsilon^2}{2(1-p)} + \mathcal{O}(\varepsilon^3) \\
&= \frac{\varepsilon^2}{2p(1-p)} + \mathcal{O}(\varepsilon^3).
\end{aligned}$$

■

## 2.5 Chernoff's Inequality (2)

In this section we shall prove that Chernoff's inequality (Corollary 2.1) is asymptotically the best possible bound. For this we need Stirling's formula, which we shall prove in the appendix to this chapter.

**Lemma 2.4 (Stirling's formula)** *As $n \to \infty$,*

$$n! \sim \sqrt{2\pi n}\, n^n\, e^{-n}.$$

**Proposition 2.3** *For every $\varepsilon \in (0, 1-p)$*

$$\lim_{n\to\infty} \frac{1}{n} \log\Big(P\Big(\frac{S_n}{n} \geq p+\varepsilon\Big)\Big) = -\varphi(p+\varepsilon) \tag{2.29}$$

*Proof.* Corollary 2.1 says that for $n \geq 1$

$$\frac{1}{n}\log P\big(S_n \geq n(p+\varepsilon)\big) \leq -\varphi(p+\varepsilon).$$

To find a lower bound, let $\alpha_n = 1 + [n(p+\varepsilon)]$ which is the smallest integer strictly bigger than $n(p+\varepsilon)$. Then

$$P\big(S_n \geq n(p+\varepsilon)\big) \geq P(S_n = \alpha_n) = \frac{n!}{\alpha_n!(n-\alpha_n)!} p^{\alpha_n}(1-p)^{n-\alpha_n}$$

Using Stirling's formula,

$$P(S_n = \alpha_n) \sim \frac{1}{\sqrt{2\pi}} \Big(\frac{n}{\alpha_n(n-\alpha_n)}\Big)^{1/2} \Big(\frac{np}{\alpha_n}\Big)^{\alpha_n} \Big(\frac{n(1-p)}{n-\alpha_n}\Big)^{n-\alpha_n} \tag{2.30}$$

and therefore the logarithms of both sides are also asymptotically equivalent. From the relations $\alpha_n \sim n(p+\varepsilon)$ and $n - \alpha_n \sim n(1-p-\varepsilon)$ we get

$$\Big(\frac{n}{\alpha_n(n-\alpha_n)}\Big)^{1/2} \sim \Big(\frac{n}{n^2(p+\varepsilon)(1-p-\varepsilon)}\Big)^{1/2} = \big(n(p+\varepsilon)(1-p-\varepsilon)\big)^{-1/2}$$

and therefore

$$\lim_{n\to\infty} \frac{1}{n} \log\Big(\Big(\frac{n}{2\pi\alpha_n(n-\alpha_n)}\Big)^{1/2}\Big) = 0. \tag{2.31}$$

On the other hand,

$$\begin{aligned}\alpha_n \log\Big(\frac{np}{\alpha_n}\Big) &= n(p+\varepsilon)\log\Big(\frac{np(p+\varepsilon)}{\alpha_n(p+\varepsilon)}\Big) + \big(\alpha_n - n(p+\varepsilon)\big)\log\Big(\frac{np}{\alpha_n}\Big)\\ &= n(p+\varepsilon)\log\Big(\frac{p}{p+\varepsilon}\Big) + n(p+\varepsilon)\log\Big(\frac{n(p+\varepsilon)}{\alpha_n}\Big) + \big(\alpha_n - n(p+\varepsilon)\big)\log\Big(\frac{np}{\alpha_n}\Big) \qquad (2.32)\end{aligned}$$

Using again that $\alpha_n \sim n(p+\varepsilon)$ we get

$$\lim_{n\to\infty}\frac{1}{n}\Big(n(p+\varepsilon)\log\Big(\frac{n(p+\varepsilon)}{\alpha_n}\Big)\Big) = 0. \qquad (2.33)$$

Since the difference $\alpha_n - n(p+\varepsilon)$ is bounded, we have

$$\lim_{n\to\infty}\frac{1}{n}\Big(\big(\alpha_n - n(p+\varepsilon)\big)\log\Big(\frac{np}{\alpha_n}\Big)\Big) = 0. \qquad (2.34)$$

Using (2.33) and (2.34) in (2.32) we get

$$\lim_{n\to\infty}\frac{1}{n}\log\Big(\frac{np}{\alpha_n}\Big)^{\alpha_n} = (p+\varepsilon)\log\Big(\frac{p}{p+\varepsilon}\Big).$$

Similarly, it can be shown that

$$\lim_{n\to\infty}\frac{1}{n}\log\Big(\frac{n(1-p)}{n-\alpha_n}\Big)^{n-\alpha_n} = (1-p-\varepsilon)\log\Big(\frac{1-p}{1-p-\varepsilon}\Big)$$

and this concludes the proof. ■

## 2.6 The de Moivre - Laplace Central Limit Theorem.

We have now the tools to prove the de Moivre - Laplace theorem, that concerns the approximation of the binomial distribution by the normal distribution. Recall that the standard normal distribution has density

$$\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2} \qquad (2.35)$$

and we denote the corresponding distribution function by

$$\Phi(x) = \int_{-\infty}^{x}\phi(t)\,dt = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-t^2/2}\,dt. \qquad (2.36)$$

As before, $S_n$ denotes a random variable with binomial distribution, which represents the number of successes in $n$ independent Bernoulli trials with success probability $p$, $0 < p < 1$ and $q = 1-p$. The distribution of $S_n$ is given by

$$P(S_n = k) = p_{n,k} = \binom{n}{k}p^k q^{n-k} \quad k = 0, 1, \ldots, n,$$

with $\mathrm{E}(S_n) = np$, $\mathrm{Var}(S_n) = npq$.

In the next two theorems we shall prove that as the number of trials increases to infinity, the distribution of

$$S_n' = \frac{S_n - np}{\sqrt{npq}}$$

converges to the normal distribution, given by (2.36).

**Theorem 2.4** *Let $a < b$ be real numbers, then*

$$P(a < S_n' \leq b) \longrightarrow \Phi(b) - \Phi(a) = \int_a^b \phi(t)\, dt, \quad \text{as } n \to \infty. \tag{2.37}$$

*Proof.* We start by defining the set

$$H_n = \{k \in \{0, 1, \ldots, n\} : a < \frac{k - np}{\sqrt{npq}} \leq b\}.$$

Observe that $H_n$ also depends on $a, b$ and $p$, which will remain fixed for the rest of the proof.

We have to look at the behavior of the following expression as $n \to \infty$

$$\begin{aligned} P(a < S_n' \leq b) &= P(a < \frac{S_n - np}{\sqrt{npq}} \leq b) \\ &= \sum_{k \in H_n} P(S_n = k) \\ &= \sum_{k \in H_n} p_{n,k} \end{aligned} \tag{2.38}$$

We will now show that, as $n \to \infty$ and for $k \in H_n$

$$p_{n,k} = \frac{1}{\sqrt{npq}} \phi(x_{k,n}) e^{\alpha_{k,n}}$$

where $\alpha_{n,k} \to 0$ as $n \to \infty$ and

$$x_{k,n} = \frac{k - np}{\sqrt{npq}}, \tag{2.39}$$

Observe that

$$x_{k+1,n} - x_{k,n} = \frac{1}{\sqrt{npq}}$$

So that the sum in (2.39) looks like a Riemann sum for the integral in (2.37), except for the exponential factor.

We start by giving an approximation for each term $p_{n,k}$, where $k \in H_n$. To simplify the notation we set $\delta = k - np$. Observe that with this notation, $k = np + \delta$ and $n - k = nq - \delta$. If $k \in H_n$, then

$$\frac{\delta}{n} \to 0 \quad \text{as } n \to \infty \tag{2.40}$$

since $\frac{|\delta|}{\sqrt{n}} = \frac{|k - np|}{\sqrt{n}}$ is a bounded sequence, because

$$a\ \sqrt{pq} < \frac{k - np}{\sqrt{n}} \leq b\ \sqrt{pq}$$

and since $1/\sqrt{n} \to 0$ we have

$$\frac{\delta}{n} = \frac{\delta}{\sqrt{n}}\ \frac{1}{\sqrt{n}} \to 0.$$

On the other hand,

$$p_{n,k} = \binom{n}{k} p^k q^{n-k} = \frac{n!}{k!\,(n-k)!} p^k q^{n-k}.$$

We now use Stirling's formula with the following notation:

$$n! = \sqrt{2\pi n}\, n^n e^{-n} e^{\gamma_n}$$

where $\gamma_n \to 0$ (and in consequence $e^{\gamma_n} \to 1$) as $n \to \infty$.

Observe now that under condition (2.40) –which is verified since we are only interested in values of $k \in H_n$– $k$ goes to $+\infty$ as $n \to \infty$. To see this, assume it is false so that $k$ is bounded and $k/n \to 0$, then

$$0 = \lim_{n\to\infty} \frac{\delta}{n} = \lim_{n\to\infty} \frac{k-np}{n} = \lim_{n\to\infty} \frac{k}{n} + p = p \neq 0.$$

This contradiction proves that $k \to \infty$ and one can similarly show that $n-k$ also goes to $+\infty$. Therefore, we can use Stirling's formula for all factorials appearing in $p_{n,k}$:

$$\begin{aligned} p_{n,k} &= \frac{\sqrt{2\pi n} n^n e^{-n} p^k q^{n-k} e^{\gamma_n - \gamma_k - \gamma_{n-k}}}{\sqrt{2\pi k} k^k e^{-k} \sqrt{2\pi(n-k)} (n-k)^{(n-k)} e^{-(n-k)}} \\ &= \frac{1}{\sqrt{2\pi}} \quad \left(\frac{n}{k(n-k)}\right)^{1/2} \left(\frac{np}{k}\right)^k \left(\frac{nq}{n-k}\right)^{n-k} e^{\gamma_n - \gamma_k - \gamma_{n-k}}. \end{aligned} \tag{2.41}$$

The first factor is a constant while the second can be written as

$$\begin{aligned} \left(\frac{n}{k(n-k)}\right)^{1/2} &= \left(\frac{n}{(np+\delta)(nq-\delta)}\right)^{1/2} \\ &= \left(\frac{n}{np\left(1+\frac{\delta}{np}\right) nq\left(1-\frac{\delta}{qn}\right)}\right)^{1/2} \\ &= \frac{1}{\sqrt{npq}} \left(1+\frac{\delta}{np}\right)^{-1/2} \left(1-\frac{\delta}{nq}\right)^{-1/2} \\ &= \frac{1}{\sqrt{npq}} e^{\gamma'_n} \end{aligned} \tag{2.42}$$

with $\gamma'_n \to 0$ as $n \to \infty$.

Taking logarithm of the third and fourth factors in (2.41) we get

$$\begin{aligned} k \log\left(\frac{np}{k}\right) + (n-k)\log\left(\frac{nq}{n-k}\right) &= -(np+\delta)\log\left(\frac{np+\delta}{np}\right) - (nq-\delta)\log\left(\frac{nq-\delta}{nq}\right) \\ &= -(np+\delta)\log\left(1+\frac{\delta}{np}\right) - (nq-\delta)\log\left(1-\frac{\delta}{nq}\right). \end{aligned}$$

Now we use the MacLaurin expansion for the function $\log(1+x)$ with $|x| < 1$:

$$\log(1+x) = x - \frac{x^2}{2} + \frac{1}{3}\frac{1}{(1+\theta x)^3} x^3, \qquad (0 < \theta < 1).$$

If $|x| < 1/2$ and $A = \frac{1}{3}\frac{1}{(1+\theta x)^3}$ then $|A| < 3$, i.e. in this case we have

$$\log(1+x) = x - \frac{x^2}{2} + Ax^3 \qquad \text{with } |A| < 3. \tag{2.43}$$

Since $\frac{\delta}{n} \to 0$, we have that both $\frac{\delta}{np}$ and $\frac{\delta}{nq}$ are less than $1/2$ for large values of $n$. Using now (2.43) for $\log(1+\frac{\delta}{np})$ and $\log(1-\frac{\delta}{nq})$, we get

$$\begin{aligned} & k \log\left(\frac{np}{k}\right) + (n-k)\log\left(\frac{nq}{n-k}\right) \\ &= -(np+\delta)\left(\frac{\delta}{np} - \frac{\delta^2}{2n^2p^2} + A\frac{\delta^3}{n^3p^3}\right) - (nq-\delta)\left(-\frac{\delta}{nq} - \frac{\delta^2}{2n^2q^2} - A'\frac{\delta^3}{n^3q^3}\right) \\ &= -\frac{\delta^2}{2n}\left(\frac{1}{p}+\frac{1}{q}\right) + B\frac{\delta^3}{n^2} = -\frac{\delta^2}{2npq} + B\frac{\delta^3}{n^2}, \end{aligned}$$

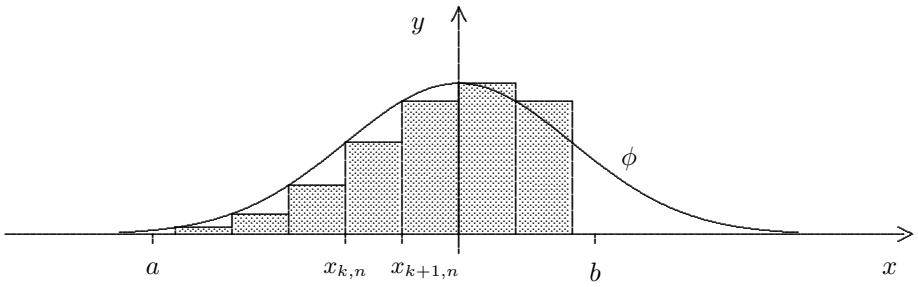


Figure 2.7: Approximation of the gaussian density

where $|A| < 3$, $|A'| < 3$ and $|B|$ is bounded by a certain fixed constant, say $|B| \leq M$. Taking now exponentials and replacing in (2.41) with $\theta_{n,k} = \gamma_n - \gamma_k - \gamma_{n-k} - \gamma'_n$,

$$p_{n,k} = \frac{1}{\sqrt{2\pi npq}} \exp\{-\frac{\delta^2}{2npq} + B\frac{\delta^3}{n^2} + \theta_{n,k}\} \tag{2.44}$$

putting $x_{n,k} = \frac{\delta}{\sqrt{npq}} = \frac{k-np}{\sqrt{npq}}$ this is (see figure 2.7)

$$\begin{aligned} &= \frac{1}{\sqrt{npq}}\phi(x_{n,k}) \exp\{B\frac{\delta^3}{n^2} + \theta_{n,k}\} \\ &= \frac{1}{\sqrt{npq}}\phi(x_{n,k}) e^{\alpha_{n,k}} \end{aligned} \tag{2.45}$$

with $\alpha_{n,k} = B\frac{\delta^3}{n^2} + \theta_{n,k}$, where $\frac{\delta^3}{n^2} \to 0$ as $n \to \infty$ since

$$\frac{\delta^3}{n^2} = \left(\frac{\delta}{\sqrt{n}}\right)^3 \frac{1}{\sqrt{n}}$$

and the first factor is bounded since $k \in H_n$. Observe that this implies that, as $n \to \infty$,

$$\sup_{k \in H_{a,b}} \alpha_{n,k} \to 0$$

Let's go back to the initial sum. Substituting $p_{n,k}$ by the expression in (2.45), we get

$$\begin{aligned} P(a < S'_n \leq b) &= \sum_{k \in H_{a,b}} \frac{1}{\sqrt{npq}}\phi(x_{k,n}) e^{\alpha_{n,k}} \\ &= \sum_{k \in H_{a,b}} \frac{1}{\sqrt{npq}}\phi(x_{k,n}) + \sum_{k \in H_{a,b}} \frac{1}{\sqrt{npq}}\phi(x_{k,n}) \left[e^{\alpha_{n,k}} - 1\right]. \end{aligned} \tag{2.46}$$

Since $\frac{1}{\sqrt{npq}}$ is equal to $x_{k+1,n} - x_{k,n}$, and

$$\frac{1}{\sqrt{npq}} \to 0$$

the first term goes to

$$\int_a^b \phi(t)\, dt \tag{2.47}$$

as $n \to \infty$.

As regards the second term in (2.46), since

$$|e^x - 1| \leq e^{|x|} - 1,$$

it is bounded by

$$\sup_{k \in H_{a,b}} \big(\exp(|\alpha_{n,k}|) - 1\big) \sum_{k \in H_{a,b}} \frac{1}{\sqrt{npq}} \phi(x_{k,n}),$$

and now the first factor vanishes as $n \to \infty$ while the second converges to the integral (2.47). In consequence, the second term in (2.46) goes to zero. This ends the proof of the theorem. ■

**Theorem 2.5**

$$P(S'_n \leq x) \to \Phi(x) = \int_{-\infty}^{x} \phi(t)\, dt \quad \textit{as } n \to \infty$$

*Proof.* We rely on theorem 2.4. Let $\varepsilon > 0$ and choose $a$ and $b$ so that $a < x < b$ (see figure 2.8) and

$$\Phi(a) < \frac{\varepsilon}{2}, \qquad 1 - \Phi(b) < \frac{\varepsilon}{2}. \tag{2.48}$$

Then

$$P(a < S'_n \leq x) \leq P(S'_n \leq x) = 1 - P(S'_n > x) \leq 1 - P(b \geq S'_n > x).$$

The first term goes to

$$\Phi(x) - \Phi(a)$$

while the last term goes to

$$1 - (\Phi(b) - \Phi(x)) = \Phi(x) + (1 - \Phi(b))$$

as $n \to \infty$.

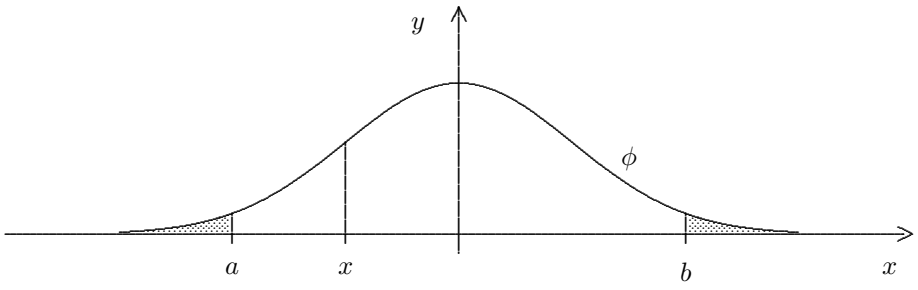


Figure 2.8: Choosing the numbers $a$ and $b$.

Hence, there exits $N$ so that for $n \geq N$ we have

$$\Phi(x) - \Phi(a) - \frac{\varepsilon}{2} \leq P(S'_n \leq x) \leq \Phi(x) + (1 - \Phi(b)) + \frac{\varepsilon}{2}$$

and taking into account the way $a$ and $b$ were chossen, for $n \geq N$

$$\Phi(x) - \varepsilon \leq P(S'_n \leq x) \leq \Phi(x) + \varepsilon$$

i.e.

$$|P(S'_n \leq x) - \Phi(x)| \leq \varepsilon.$$

and this proves the result. ■

## 2.7 Examples.

1. A die is thrown 6,000 times. Use the normal distribution to approximate the probability that the number 6 appears between 990 and 1,010 times.

▶ Let $X$ be the number of times we get the number 6 in 6,000 throws of the die. We know that $X \sim b(6,000, 1/6)$, therefore

$$P(990 \le X \le 1010) = P\left(\frac{990-1,000}{\sqrt{6,000\frac{1}{6}\frac{5}{6}}} \le \frac{X-1,000}{\sqrt{6,000\frac{1}{6}\frac{5}{6}}} \le \frac{1010-1,000}{\sqrt{6,000\frac{1}{6}\frac{5}{6}}}\right)$$
$$\simeq \frac{1}{\sqrt{2\pi}}\int_{-0.346}^{0.346} e^{-x^2}dx \simeq 0.27.$$

◀

2. In 5,000 throws of a coin there are 2,800 heads. Is it reasonable to assume that the coin is not biased?

▶ The question may be rephrased in this way: If the coin is not loaded, how exceptional is it that the number of heads in 5,000 throws exceeds its mean value of 2,500 by at least 300? If the probability of this event is very small, instead of thinking that a very extraordinary event has occurred, one will tend to attribute it to the fact that, in reality, the coin is loaded, and consequently the probability has been calculated on the wrong basis.

First we us Tchebycheff's inequality to bound this probability:

$$n = 5,000, \qquad p = 0.5 \qquad np = 2,500$$

$$P(S_n \ge 2,800) = P(S_n - np \ge 300) = \frac{1}{2}P(|S_n - np| \ge 300)$$
$$\le \frac{1}{2}\frac{1}{(300)^2}\operatorname{Var}(S_n)$$
$$= \frac{1}{2}\frac{5,000}{4(300)^2} \simeq 0.0068.$$

Thus, the probability of the event in question is bounded above by 0,0068.

If, instead, we resort to the approximation by the normal distribution, we have

$$P(S_n - np \ge 300) = P\left(\frac{S_n - np}{\sqrt{np(1-p)}} \ge \frac{300}{\sqrt{np(1-p)}}\right) \simeq \frac{1}{\sqrt{2\pi}}\int_a^{+\infty} e^{-x^2/2}dx$$

with $a = 300/\sqrt{5,000/4} \simeq 8.48$.

The last integral is bounded by

$$\frac{1}{\sqrt{2\pi}\,a}e^{-a^2/2}$$

and replacing the value of $a$ we get a value of $0.12\times10^{-17}$ for this probability, which is astronomically small. ◀

3. Between 2000 and 2009 inclusive, 13,249,775 girls and 13,230,179 boys were born in Mexico. We want to see if this result is compatible with the hypothesis that the sex of newborns is randomly distributed with a 0.5 probability.

► To see this we consider the sex of each newborn to be a Bernoulli variable with a 0.5 probability of success, and we will say that a success occurs if a girl is born. The total number of trials is 26,479,954 and the expected value of the number of girls is $np = 13,239,977$. The difference between the number of girls and their expected value is 9,798 and we want to find the probability of getting a difference greater than or equal to this value:

$$\begin{aligned} P(S_n \geq 13,239,977) &= P\Big(\frac{S_n - np}{\sqrt{npq}} \geq \frac{13,239,977 - np}{\sqrt{npq}}\Big) \\ &\simeq P(N \geq 3,808) \\ &\simeq 7 \times 10^{-5} \end{aligned}$$

This number is small enough to suggest that the probability of boys and girls is not the same. ◄

4. Let us revisit the problem considered in section 2.3.1 regarding the empirical estimation of an unknown proportion $p$ using the relative frequency $\hat{p}_n = S_n/n$. We want to determine the sample size so that with probability at least 0.95, the error in the estimation will be less than or equal to 0.01, i.e. we want to determine the value of $n$ so that

$$P(|\hat{p}_n - p| \leq 0.01) \geq 0.95$$

Using Tchebycheff's inequality, we saw then that a sample of size 50,000 would suffice but, as we remarked, this number is, in fact, too large. Let us see how we can use the normal approximation furnished by the CLT to get a more reasonable value for $n$. First, let us consider this problem in a more general framework. Suppose that we want the error to be less than $\varepsilon > 0$ with probability at least $1 - \alpha$. $\alpha$ is known as the confidence level. Now we want

$$P(|\hat{p}_n - p| \leq \varepsilon) \geq 1 - \alpha.$$

This is equivalent to

$$P(|\hat{p}_n - p| > \varepsilon) \leq \alpha.$$

We have

$$\begin{aligned} \alpha \geq P(|\hat{p}_n - p| > \varepsilon) &= P\Big(\Big|\frac{S_n}{n} - p\Big| > \varepsilon\Big) \\ &= P\Big(\Big|\frac{S_n - np}{n}\Big| > \varepsilon\Big) \\ &= P\Big(\Big|\frac{S_n - np}{\sqrt{npq}}\Big| > \frac{\varepsilon\sqrt{n}}{\sqrt{pq}}\Big) \\ &\approx P\Big(|N| > \frac{\varepsilon\sqrt{n}}{\sqrt{pq}}\Big) \end{aligned}$$

where $N$ is a standard normal random variable. Since the standard Gaussian distribution is symmetric with respect to the origin

$$P\Big(|N| > \frac{\varepsilon\sqrt{n}}{\sqrt{pq}}\Big) = 2P\Big(N > \frac{\varepsilon\sqrt{n}}{\sqrt{pq}}\Big)$$

and in consequence we want $n$ so that

$$P\Big(N > \frac{\varepsilon\sqrt{n}}{\sqrt{pq}}\Big) = \frac{\alpha}{2}$$

The number $z_{1-\alpha}$ that satisfies

$$P(N > z_{1-\alpha}) = \alpha \qquad \text{or equivalently} \qquad P(N \leq z_{1-\alpha}) = 1 - \alpha$$

is known as the quantile of order $1-\alpha$ of the standard normal distribution. This value can be obtained from a table of the normal distribution or using the function `qnorm(1-α)` in `R`.

To determine $n$ in our problem, we need the $1-\alpha/2$ quantile, which for $\alpha = 0.05$ is 1.96. Thus we want

$$\frac{\varepsilon\sqrt{n}}{\sqrt{pq}} = 1.96 \qquad \text{or} \qquad n = \frac{(1.96)^2 pq}{\varepsilon^2}$$

and we see that the sample size depends on $\varepsilon^2$ and the product $pq = p(1-p) = p - p^2$. If we now nothing a priori about $p$, we can bound this last expression by its maximum, which is $1/4$. Thus we get the expression

$$n = \frac{(1.96)^2}{4\varepsilon^2}$$

As examples, for $\varepsilon = 0.01$ we get $n = 9,604$, for $\varepsilon = 0.02$, $n = 2,401$ and for $\varepsilon = 0.03$, $n = 1,067$.

If we change the value of $\alpha$ to 0.01, then $z_{0.995} = 2.576$ and the values of $n$ for $\varepsilon = 0.01, 0.02$ and 0.03 are 16,590, 4,148 and 1,844, respectively. ◀

## 2.8 Extensions

1. To extend in several directions the de Moivre - Laplace theorem, a similar technique can be used. Looking back at the proof, we see that it depends only on two things

$$\frac{\delta}{n} \to 0 \tag{2.49}$$

and

$$\frac{\delta^3}{n^2} \to 0. \tag{2.50}$$

But (2.50) implies (2.49) since

$$\frac{\delta}{n} = \frac{1}{n^{1/3}} \left(\frac{\delta^3}{n^2}\right)^{1/3}.$$

therefore, as long as (2.50) holds for every integer $k$ such that

$$a \le \frac{k-np}{\sqrt{npq}} \le b,$$

a similar conclusion will hold. This observation permits the generalization of the previous theorems to the case where $a$ and $b$ vary with $n$. For instance, the proof we have given for theorems 2.4 and 2.5 also show that if

$$\frac{a_n^3}{\sqrt{n}} \to 0 \quad \text{as } n \to \infty$$

i.e., if $a_n$ goes to $+\infty$, but at a slower rate than $n^{1/6}$, then

$$P\left(\frac{S_n - np}{\sqrt{npq}} > a_n\right) \sim \int_{a_n}^{+\infty} \phi(t)dt \quad \text{as } n \to \infty \tag{2.51}$$

where, as before, $\phi$ is the standard normal density and the symbol "$\sim$" says that both terms are asymptotically equivalent.

These results allow the study of "large deviations" of the number of successes $S_n$ (with binomial distribution), from the expected value $np$. To be able to obtain sharp inequalities, it is useful to study the speed of convergence to zero as $n \to \infty$, of

$$1 - \Phi(x) = \int_x^{+\infty} \frac{1}{\sqrt{2\pi}} e^{t^2/2} dt.$$

Let us prove first that if $x > 0$ then

$$1 - \Phi(x) \leq \frac{1}{\sqrt{2\pi}\, x} e^{-x^2/2}. \tag{2.52}$$

Indeed,

$$1 - \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2} dt \leq \frac{1}{\sqrt{2\pi}} \int_x^\infty \frac{t}{x} e^{-t^2/2} dt,$$

since, in the integration range, $t/x \geq 1$. Then

$$\begin{aligned} 1 - \Phi(x) &\leq \frac{1}{\sqrt{2\pi}\, x} \int_x^\infty t e^{-t^2/2} dt \\ &= -\frac{1}{\sqrt{2\pi}\, x} e^{-t^2/2} \Big|_x^{+\infty} = \frac{1}{\sqrt{2\pi}\, x} e^{-x^2/2}. \end{aligned}$$

Second, we will show that

$$1 - \Phi(x) \sim \frac{1}{\sqrt{2\pi}\, x} e^{-x^2/2} \quad \text{as } x \to \infty \tag{2.53}$$

We want to show now that

$$\lim_{x\to\infty} \frac{1 - \Phi(x)}{\frac{1}{\sqrt{2\pi}\, x} e^{-x^2/2}} = 1,$$

Using L'Hôpital's rule,

$$\frac{-\frac{1}{\sqrt{2\pi}} e^{-x^2/2}}{\frac{1}{\sqrt{2\pi}} \left(-\frac{1}{x^2} e^{-x^2/2} + \frac{1}{x}(-x e^{-x^2/2})\right)} = \frac{1}{\frac{1}{x^2} + 1} \overset{x\to\infty}{\to} \longrightarrow 1,$$

and this proves (2.53).

2. In applications, it is generally interesting to know not only that as the number of observations increases indefinitely, the binomial distribution tends to the normal distribution, but also, what is the speed of convergence. In other words, it is interesting to have an estimate of the error made when, for a given n, we replace the binomial distribution with the normal distribution. This error depends on the value of n, but it also depends on the value of p[1]; the closer it is to 1/2, the faster the convergence is.

   Following the procedure in the proof of the de Moivre - Laplace theorem, to give a bound for the error we are considering, the fundamental step is to refine formula (2.45), that approximates the probability function for the binomial distribution by the density of the normal distribution. In order to do this, on the one hand we need an estimation of the error when using Stirling's formula –that may be obtained using the same method we used for the proof– and on the other, we have to use more terms in the expansion of $\log(1+x)$. For instance, we may use the following improvement on Stirling's formula

   $$n! = \sqrt{2\pi n} n^n e^{-n} e^{\alpha_n}$$

   where $0 < \alpha_n < \frac{1}{12n}$, and for the expansion of the logarithm we can take, for example

   $$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{1}{5}\frac{x^5}{(1+\theta x)^5} \quad \text{for } 0 < \theta < 1, |x| < 1.$$

[1] When p is close to 0 or 1 and n is not very large, the Poisson approximation to the binomial distribution may be more accurate. For this and other issues on approximation, as well as for the presentation of various examples, we recommend that the reader consult Vol. 1 of W. Feller's book, included in the bibliography.

Finally, in (2.42), we also use a Mac-Laurin expansion to approximate

$$\left(1+\frac{\delta}{np}\right)^{-1/2}\left(1-\frac{\delta}{nq}\right)^{-1/2}.$$

Replacing in (2.41), if we add, for instance, the condition

$$|k-np| = |\delta| \leq C\,\sqrt{npq} \tag{2.54}$$

where $C$ is a constant, and we make $n$ large enough so that

$$\frac{C}{\sqrt{npq}} < \frac{1}{3}, \tag{2.55}$$

instead of (2.45) we get

$$p_{n,k} = \frac{1}{\sqrt{npq}}\phi(x_{k,n})e^{\varepsilon_n} \tag{2.56}$$

where

$$|\varepsilon_n| \leq \frac{(1+C^3)(q-p)}{\sqrt{npq}} + \frac{(1+C^4)}{\sqrt{npq}}. \tag{2.57}$$

Observe that the constant $C$ appearing in (2.54), also appears in the error bound (2.57), so that for $n$ fixed, the greater the C, the less accurate the approximation. $C$ measures how far are the values of $k$ from the mean value $np$ for the random variable $S_n$ under consideration. The further away from the average np these k-values are, the greater the n necessary to obtain the same approximation.

Condition (2.55) can be modified, since

$$\frac{C}{\sqrt{npq}} \to 0 \quad \text{as } n \to \infty.$$

The smaller the right-hand side –which we chose to be 1/3– the sharper the error bound that can obtained in place of (2.57). This implies choosing a larger value of $n$.

As a final remark, observe that (2.57) suggests that the error we make when using the normal distribution as an approximation of the binomial, may be a function of $p$. If $p = q = 1/2$, the first term in the right hand side of (2.57) is 0, and the convergence to as $n \to \infty$ is faster. On the other hand, given $n$, the bound depends on $npq$ and is less accurate the more distant $p$ is from 1/2 (i.e. the closer $p$ is to 0 or 1).

3. The theorem we have proved concerns sums of Bernoulli random variables appropriately normalized: If $S_n = \sum_1^n X_i$ then

$$S_n' = \frac{S_n - np}{\sqrt{npq}}$$

Observe that $\mathrm{E}(S_n) = np$ and $\mathrm{Var}(S_n) = npq$ so that we can write

$$S_n' = \frac{S_n - \mathrm{E}(S_n)}{\sqrt{\mathrm{Var}(S_n)}}. \tag{2.58}$$

**Theorem 2.6** *Let $X_n n \geq 1$ be a sequence of independent identically distributed random variables with finite positive variance and let $S_n'$ be defined as in (2.58). then, for any real $x$, as $n \to \infty$,*

$$P(S_n' \leq x) \to \int_{-\infty}^{x} \phi(t)\,dt = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{t^2/2}\,dt. \tag{2.59}$$

In consequence, a CLT is valid for iid variables as long as they have finite variance, which is a very general condition.

Recall that we defined the empirical or sample average for the variables $X_1, \ldots, X_n$ as $\bar{X}_n = \frac{1}{n}\sum_1^n X_i$. Observe that if divide numerator and denominator in (2.58) by $n$, we get the equivalent expression

$$S'_n = \sqrt{n}\,\frac{\bar{X}_n - \mathrm{E}(\bar{X}_n)}{\sqrt{\mathrm{Var}(\bar{X}_n)}} \tag{2.60}$$

If the variables $X_i$ have mean $\mu$ and variance $\sigma^2$ then $\mathrm{E}(\bar{X}_n) = \mu$, $\mathrm{Var}(\bar{X}_n) = \sigma^2$ and the previous expression becomes

$$S'_n = \sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma}. \tag{2.61}$$

In this context, the proof requires tools that are beyond the scope of this course. The same questions we raised in remark **??** are valid here. In particular, the speed of convergence to the limit will depend on the shape of the common distribution of the $X'_i s$. symmetric distributions with bounded or light tails will have faster speeds of convergence than asymmetric or heavy tailed distributions. In the next section we present a simulation example with different distributions that exemplify these facts.

## 2.9 The R Section

In this section we review some aspects of the simulation of Bernoulli trials and their associated random variables and also consider the Law of Large Numbers and the Central Limit Theorem. We start by considering simulation of variables in R.

### 2.9.1 Random Variable Generation in R

The R language has a series of built-in routines to generate random variables. The precise syntax of the corresponding instruction depends on the distribution, but they all have the common format r*dist*, where *dist* designates the distribution; for example, to generate values from the normal distribution we use `rnorm`. Depending on the distribution, it may be necessary to specify one or more parameters. Table 2.1 presents the most common distributions, the required parameters and their default values. `n` always represents the sample size.

In addition, R has the function `sample` that produces samples with or without replacement from finite sets of values. The syntax is

```
sample(x, size, replace = FALSE, prob = NULL)
```

where

- `x` is the set from which we want to sample, written as a vector,
- `size` is the sample size,
- `replace` indicates if sampling is with replacement (`replace = TRUE`) or not and
- `prob` is a vector of probabilities, if sampling is not uniform.

Additionally, associated with each distribution are three other functions that correspond to the density, distribution and quantile functions. These have a syntax similar to the one we have reviewed for the generation of random variables, but changing the first letter to `d, p` or `q`. Thus, for the normal distribution, `dnorm(x)` corresponds to the density values at the points of the vector `x`, `pnorm(q)` corresponds to the distribution function values at the points of the vector `q` and `qnorm(p)` gives the quantiles corresponding to the probability vector `p`.

Table 2.1: R Functions for generating values from several common distributions.

| **Distribution** | **R Function** |
|---|---|
| Binomial | `rbinom(n, size, prob)` |
| Poisson | `rpois(n, lambda)` |
| Geometric | `rgeom(n, prob)` |
| Hypergeometric | `rhyper(nn, m, n, k)` |
| Negative Binomial | `rnbinom(n, size, prob)` |
| Multinomial | `rmultinom(n, size, prob)` |
| Uniform | `runif(n, min=0, max=1)` |
| Exponential | `rexp(n, rate=1)` |
| Gaussian | `rnorm(n, mean=0, sd=1)` |
| Gamma | `rgamma(n, shape, scale=1)` |
| Weibull | `rweibull(n, shape, scale=1)` |
| Cauchy | `rcauchy(n, location=0, scale=1)` |
| Beta | `rbeta(n, shape1, shape2)` |
| t | `rt(n, df)` |
| Fisher | `rf(n, df1, df2)` |
| $\chi^2$ | `rchisq(n, df)` |
| Logístic | `rlogis(n, location=0, scale=1)` |
| Lognormal | `rlnorm(n, meanlog=0, sdlog=1)` |

### 2.9.2 Bernoulli Trials

To simulate Bernoulli variables in R we use the commands to generate binomial variables with parameter $n = 1$. The following instructions generate 100 Bernoulli variables with parameter $p = 0.1$ and graph them. The first instruction sets the seed of the random number generator so that the results can be reproduced.

```
set.seed(192837)
index <- 1:100
sampl <- rbinom(100,1,0.1)
plot(index, sampl, type='h',lwd=2)
```

The following instructions are for graphing samples of size 100 for Bernoulli variables with success probabilities $p = 0.1; 0.3; 0.5$ and 0.7, on four panels. The results are shown on figure 2.9

```
par(mfrow=c(4,1))
plot(index,rbinom(100,1,0.1), type='h',lwd=2, main='Bernoulli trials, p=0.1',
        ylab='',bty='n',yaxt='n')
plot(index,rbinom(100,1,0.3), type='h',lwd=2, main='Bernoulli trials, p=0.3',
        ylab='',bty='n',yaxt='n')
plot(index,rbinom(100,1,0.5), type='h',lwd=2, main='Bernoulli trials, p=0.5',
        ylab='',bty='n',yaxt='n')
plot(index,rbinom(100,1,0.7), type='h',lwd=2, main='Bernoulli trials, p=0.7',
        ylab='',bty='n',yaxt='n')
par(mfrow=c(1,1))
```

We have previously studied the geometric distribution, which is associated to sequences of Bernoulli trials and corresponds to counting the number of trials until the first success occurs. The following instruction indicates in which trial is the first success in the vector `sampl`.

```
index[sampl==1][1]
[1]  7
```

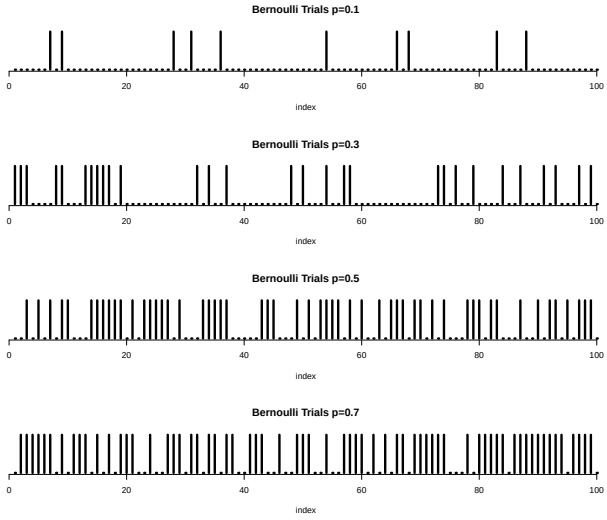


Figure 2.9: Samples of size 100 of Bernoulli trials with success probabilities 0.1, 0.3, 0.5 and 0.7. Vertical bars correspond to successes.

Let us explain in more detail this command. The condition that appears within brackets, `sampl==1`, creates a logical vector of the same length as `sampl` whose entries are either `TRUE` or `FALSE` depending on whether the condition is satisfied by the corresponding entry in the vector `sampl`. Thus, if the first entry is a success, the vector produced by the instruction `sampl == 1` will have `TRUE` as its first entry and otherwise it will have `FALSE`. Hence, `index[sampl==1]` is the vector of the indices that correspond to successes in the sequence, and we are asking for the first component of this vector, i.e. the index that corresponds to the first success in the sequence.
On the other hand, if we wanted the position in the sequence of the third success, we would write

```
index[sampl==1][3]
[1]  28
```

To facilitate the repetition of this procedure we will define a function that will allow us to find the first success:

```
geo1 <- function(n=1000,p=0.1) (1:n)[rbinom(n,1,p)==1][1]
```

This function has two parameters, the sample size `n`, which by default has a value of 1,000, and the probability of success `p`, by has a default value of 0.1. Both can be set when the function is called, but since they have default values, the function is executed even if no values are entered for these parameters:

```
geo1()
[1]  7
```

and the result is the number of trials until the firs success. We have included the number of trials as a parameter because if we want to use very small values of the probability of success, it may be necessary to raise its value. If we want to repeat this function we may us the command `replicate`:

```
replicate(10,geo1())
[1] 1 3 5 9 19 7 8 19 1 28
```

which gives a sample of size 10 from the geometric distribution with success probability 0.1. Next we use the function `tabulate` to create a table of values for a sample of size 10,000 from the geometric distribution with $p = 0.1$. We truncate the values in the table at 50 and graph them. Results are shown in figure 2.10 (left).

The commands in `R` to produce this graph are:

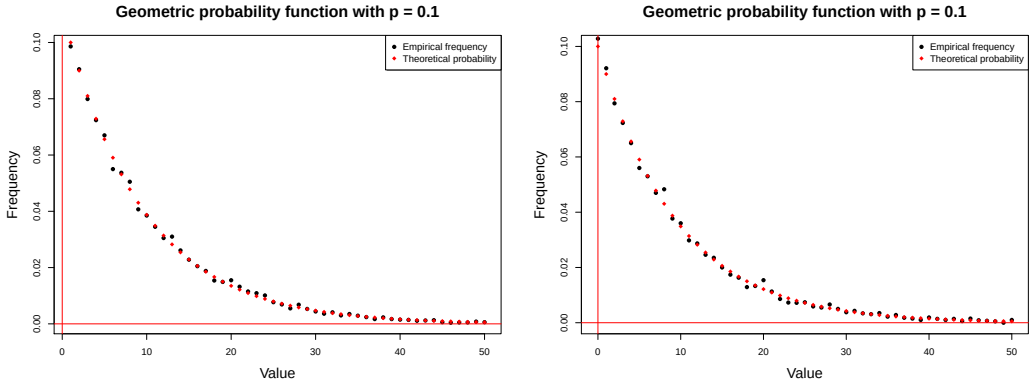


Figure 2.10: Frecuencia empírica de los valores for una distribución geométrica con parámetro $p = 0.1$ basado en una muestra simulada de tamaño 10,000: (izq) using la función `geo1` que construimos, (der) using la función `rgeom`.

```
values <- tabulate(replicate(10000,geo1()),50)
plot(1:50,values/10000,pch=16, cex.main = 1.7,
    main='Geometric Distribution with p=0.1',
    xlab='Values', ylab='Frequency', cex.lab=1.5)
abline(h=0, col='red')
abline(v=0, col='red')
points(1:50,dgeom(0:49,0.1),col='red',pch=18)
legend('topright',c('Empirical frequency','Theoretical probability'),
    pch=c(16,18),col=c('black','red'))
```

In `R` we have the function `rgeom` to generate samples from the geometric distribution, but this is not exactly the same distribution we have simulated. The function `rgeom` counts the number of *failures* before the first success, and does not include the trial in which the first success appears. Thus, the values differ in 1 and the distribution included in `R` has 0 as a potential value, which corresponds to a sequence starting with a success. We draw a similar graph using this function, in figure 2.10 (right).

However, we need to do a little trick to use the function `tabulate` in this context, since this function disregards values of the vector to be tabulated that are smaller or equal to 0 and in the vectors we will consider there will be values equal to 0. An easy way to deal with this inconvenience is to add a unit to the values before tabulation, and then draw the results starting from the value 0. This is done in the next set of commands.

```
plot(0:49,tabulate(rgeom(10000,0.1)+1,51),pch=16,xlab='k',ylab='Frequency',
        main='Geometric Distribution with p=0.1'), cex.main= 1.7, cex.lab= 1.5
abline(h=0, col='red')
abline(v=0, col='red')
points(0:50,dgeom(0:50,0.1),col='red',pch=18)
legend('topright',c('Empirical frequency','Theoretical probability'),
    pch=c(16,18),col=c('black','red'))
```

It is important to remember that there two different definitions of the geometric distribution in the literature, and the distribution included in `R` does not coincide with the one we have included on this notes. Therefore, in order to obtain values corresponding to the distribution defined in this notes, we must add a unit to the values generated with the `rgeom` function or, equivalently, move the values to the right one unit when making the graphs.

Next, we want to make a graphical representation of several instances of the geometric distribution, using vertical bars for the values of the probability function. The first command in the list below saves in `op` the default values of the graphical parameters, so if we modify them for a particular graph, we can restore them at the end with the command `par(op)`. In the process of drawing these figures, we will change the margins twice, since the default values tend to be too big for a multipanel representations. The resulting graph is presented in figure 2.11.

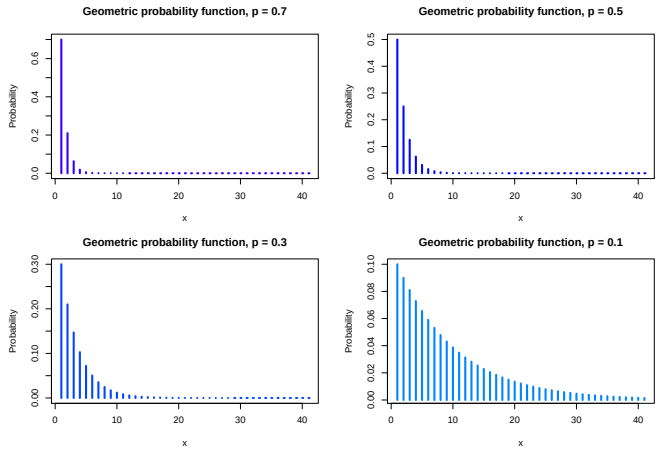


Figure 2.11: Probability functions for the geometric distribution with parámeters $0.1, 0.3, 0.5$ and $0.7$.

```
op <- par(no.readonly = TRUE)
xval <- 0:40; clrs <- rainbow(50)
par(mfrow=c(2,2))
par(mar=c(4,4,4,2)+0.1)
plot(xval+1, dgeom(xval,0.7),type='h',lwd=3, xlab='x', col = clrs[36],
     main='Geometric probability function, p = 0.7',ylab='Probability')
plot(xval+1, dgeom(xval,0.5),type='h',lwd=3, xlab='x',col=clrs[34],
     main='Geometric probability function, p = 0.5',ylab='Probability',)
par(mar=c(5,4,3,2)+0.1)
plot(xval+1, dgeom(xval,0.3),type='h',lwd=3, xlab='x',col=clrs[32],
     main='Geometric probability function, p = 0.3',ylab='Probability')
plot(xval+1, dgeom(xval,0.1),,type='h',lwd=3, xlab='x',col=clrs[30],
     main='Geometric probability function, p = 0.1',ylab='Probability')
par(mfrow=c(1,1))
par(op)
```

The colors for the previous figures were chosen using the `rainbow` palette. Figure 2.12 shows the colors included in a `rainbow(50)` palette.

Another distribution associated to sequences of Bernoulli trials is the negative binomial, that counts the number of trials needed to get the $k$-th success. One way to simulate this distribution is adding $k$ variables with geometric distribution, as in the next example in which $k = 3$.

```
set.seed(1234)
sum(replicate(3,geo1()))
[1] 35
```

An alternative way is to modify the function `geo1` to find the $k$-th success instead of the first one:

```
bineg <- function(k=2,nn=1000,pp=0.1) (1:nn)[rbinom(nn,1,pp)==1][k]
```

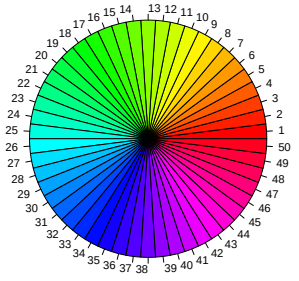

Figure 2.12: Choice of colors using the `rainbow(50)` palette.

```
bineg()
[1] 10
bineg(10)
[1] 95
```

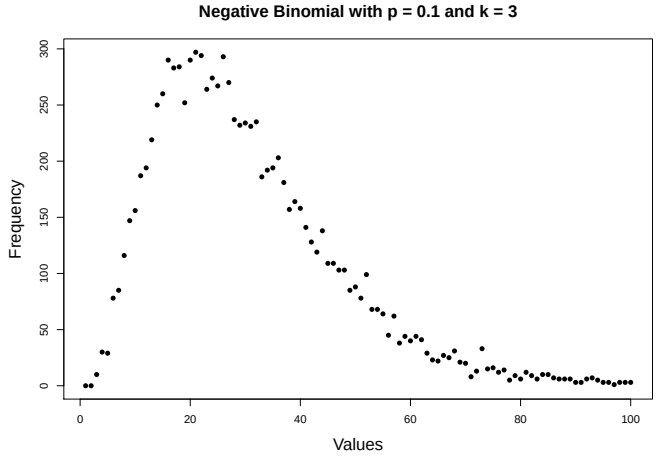


Figure 2.13: Empirical frequency for 10,000 values simulated from a negative binomial distribution with parameters $p = 0.1$ and $k = 3$.

Next, we simulate a sample of size 10,000 and graph the relative frequency of the results up to the value 100 for $k = 3$. The result is shown in figure 2.13.

```
valuesnb <- tabulate(replicate(10000,bineg(3)),100)
plot(1:100,valuesnb,pch=16,main='Negative Binomial with p=0.1 and k=3',
     xlab='Values',ylab='Frequency', cex.main=1.5, cex.lab=1.5)
```

The function `rnbinom`, which generates samples from a negative binomial distribution, as well as the functions `dnbinom` for the density, `pnbinom` for the distribution functions and `qnbinom` for the quantile function, are also available in `R`, but as in the case of the geometric distribution, the values of this

variables in R do not include the successes and only count the number of failures previous to the $k$-th success. Therefore, for the values given by R to coincide with those defined in this course, we have to add $k$ to the values of this variable.

Next, we give some graphics for the probability function of variables with negative binomial distributions, with different values for the parameters.

```
op <- par(no.readonly = TRUE)
par(mfrow=c(2,2))
par(mar=c(4,4,4,2)+0.1)
plot(xval+3, dnbinom(xval,3,0.7), type='h',main='Negative Binomial Distribution,
     k=3,p=0.7',lwd=2, ylab='Densidad', xlab='x',xlim=c(0,50))
plot(xval+3, dnbinom(xval,3,0.5),col=clrs[4],lwd=2,type='h',ylab='Density',
     main='Negative Binomial Distribution, k=3, p=0.5', xlab='x',xlim=c(0,50))
par(mar=c(5,4,3,2)+0.1)
plot(xval+3, dnbinom(xval,3,0.3),col=clrs[28],lwd=2,type='h',ylab='Density',
     main='Negative Binomial Distribution, k=3, p=0.3', xlab='x',xlim=c(0,50))
plot(xval+3, dnbinom(xval,3,0.1),col=clrs[34],lwd=2,type='h',ylab='Density',
     main='Negative Binomial Distribution, k=3, p=0.1', xlab='x',xlim=c(0,50))

par(mar=c(4,4,4,2)+0.1)
plot(xval+3, dnbinom(xval,3,0.3), type='h',main='Negative Binomial Distribution,
     k=3,p=0.3',lwd=2,ylab='Density', xlab='x',xlim=c(0,50))
plot(xval+5, dnbinom(xval,5,0.3),col=clrs[4],lwd=2,type='h',ylab='Density'
     ,main='Negative Binomial Distribution, k=5, p=0.3', xlab='x',xlim=c(0,50))
par(mar=c(5,4,3,2)+0.1)
plot(xval+7, dnbinom(xval,7,0.3),col=clrs[28],lwd=2,type='h',ylab='Density',
     main='Negative Binomial Distribution, k=7, p=0.3', xlab='x',xlim=c(0,50))
plot(xval+9, dnbinom(xval,9,0.3),col=clrs[34],lwd=2,type='h',ylab='Density',
     main='Negative Binomial Distribution, k=9, p=0.3', xlab='x',xlim=c(0,50))
par(mfrow=c(1,1))
par(op)
```

Results are shown in the following figures

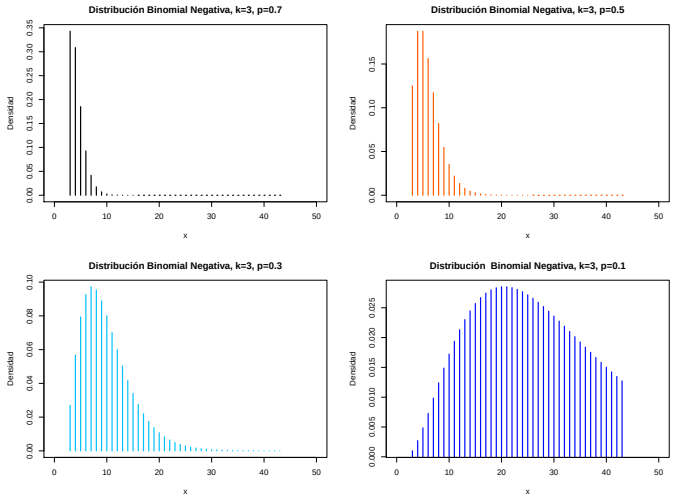


Figure 2.14: Negative binomial probability function with $k = 3$ and $p = 0.7, 0.5, 0.3$ and $0.1$.

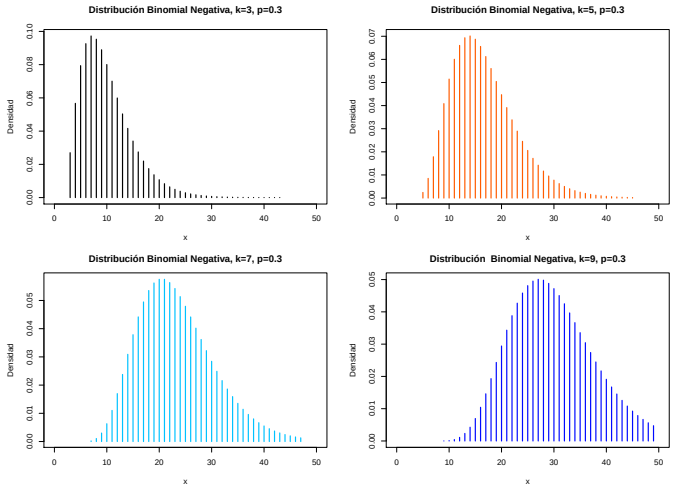


Figure 2.15: Negative binomial probability function with $p = 0.3$ and $pk = 3, 5, 7$ and 9.

Finally, we consider the binomial distribution, which is naturally linked to Bernoulli trials, since a binomial variable with parameters $n$ and $p$ counts the number os successes in a sequence of $n$ independent Bernoulli trials with success probability $p$. We present several animations in R. For a start, we set $n = 10$ and allow $p$ to vary between 0 and 1 in steps of 0.01.

```
binom <- function(p){
    plot(0:10,dbinom(0:10,10,p),type='h',lwd=5,ylim=c(0,0.5),
    xlab='Values',ylab='Probability')
    Sys.sleep(0.1)}
ignorar <- sapply((0:100)/100,binom)
```

Next we set $p = 0.5$ and vary $n$ between 1 and 100:

```
binom2<- function(n){
    plot(0:n,dbinom(0:n,n,0.5),type='h',lwd=5,ylim=c(0,0.5),
    xlab='Values',ylab='Probability')
    Sys.sleep(0.1)}
ignorar <- sapply((0:100),binom2)
```

In the following figures we change the x-axis scale and graph for $n$ up to 200.

```
binom3<- function(n){
    plot(0:n,dbinom(0:n,n,0.5),type='h',lwd=5,ylim=c(0,0.5),
    xlab='Values',ylab='Probability',
    xlim=c((n/2)-2*sqrt(n),(n/2)+2*sqrt(n)))
    Sys.sleep(0.08) }
ignorar <- sapply((0:200),binom3)
```

Finally, we set the value of $p$ to 0.25 and let $n$ vary from 0 to 200. The first set of figures is without modifying the scale.

```
binom4<- function(n){
    plot(0:n,dbinom(0:n,n,0.25),type='h',lwd=5,ylim=c(0,0.5),
    xlab='Values',ylab='Probability')
    Sys.sleep(0.08)}
ignorar <- sapply((0:200),binom4)
binom5<- function(n){
```

```
    plot(0:n,dbinom(0:n,n,0.25),type='h',lwd=5,ylim=c(0,0.5),
    xlab='Values',ylab='Probability',
    xlim=c((n/4)-2*sqrt(n),(n/4)+2*sqrt(n)))
    Sys.sleep(0.08)}
ignorar <- sapply((0:200),binom4)
```

## 2.10 Appendix: Proof of Stirling's Formula.

Stirling's formula gives an asymptotically equivalent expression for factorials that is very useful when one is interested in studying the behavior of expressions involving factorials as the argument goes to infinity. We have seen an example in the previous section and we shall use it again to prove that the binomial distribution may be approximated by the normal distribution, in the next section.

**Lemma 2.5 (Stirling's formula)** *As $n \to \infty$,*

$$n! \sim \sqrt{2\pi n}\, n^n\, e^{-n}.$$

*Proof.* We want to show that

$$\lim_{n\to\infty} \frac{n!}{\sqrt{n}\, n^n\, e^{-n}} = \sqrt{2\pi}.$$

The proof is divided in two steps. In the first one we prove that the limit in the left-hand side exists and is finite, and in the second we show that it is equal to $\sqrt{2\pi}$.

**Step 1.** Consider the function $f(x) = \log x$ for $x \geq 1$ and let us bound below the area $A_n$ limited by its graph, the $x$ axis and the vertical line going through $x = n$, replacing each piece of the curve $P_k\, P_{k+1}$ by the segment $\overline{P_k\, P_{k+1}}$ (see figure 2.16). Since the area of the trapezoid $k\, P_k\, P_{k+1}\, (k+1)$ is

$$\frac{1}{2}(P_k + P_{k+1}) = \frac{1}{2}(\log k + \log(k+1))$$

we get

$$\begin{aligned} A_n = \int_1^n \log x\, dx &> \sum_{k=1}^{n-1} \frac{1}{2}(\log k + \log(k+1)) \\ &= \frac{1}{2}(\log 1 + \log 2 + \log 2 + \log 3 + \cdots + \log n) \\ &= \log 2 + \log 3 + \cdots + \log(n-1) + \frac{1}{2}\log n. \end{aligned} \tag{2.62}$$

Let $a_n = A_n - (\log 2 + \log 3 + \cdots + \log(n-1) + \frac{1}{2}\log n)$, be the difference between the area $A_n$ and the sum of the areas of the trapezoids. It is clear from the graphs that the sequence $\{a_n\}$ is monotonically increasing.

Let now $b_k$ be the shaded area in figure 2.17, where the segment $\overline{P'_k\, P'_{k+1}}$ is tangent to the graph of $f(x) = \log x$ at the midpoint of the interval, $x = k + 1/2$. It is clear that

$$a_n < \sum_{k=1}^{n-1} b_k, \qquad (n > 1), \tag{2.63}$$

where $b_k$ is the difference of the areas of the trapezoids $k\, P'_k\, P'_{k+1}\, (k+1)$ and $k\, P_k\, P_{k+1}\, (k+1)$. Therefore

$$\begin{aligned} b_k &= \log(k + \frac{1}{2}) - \frac{1}{2}[\log k + \log(k+1)] = \frac{1}{2}\log\frac{(k+\frac{1}{2})^2}{k(k+1)} \\ &= \frac{1}{2}\log(1 + \frac{1/4}{k(k+1)}) \leq \frac{1}{8k(k+1)} \end{aligned}$$

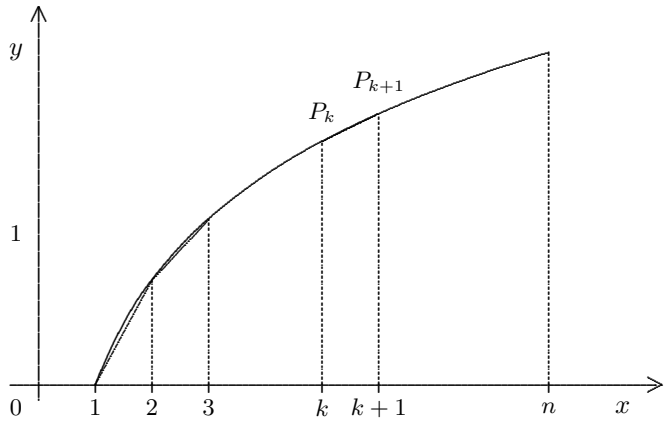


Figure 2.16: Approximating the function $\log x$ by linear segments.

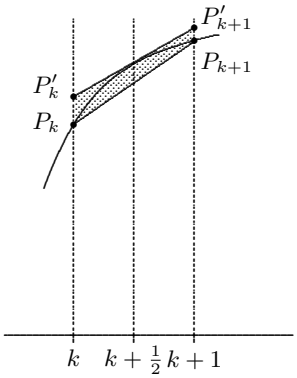


Figure 2.17: Bound for the approximation to the function $\log x$

since $\log(1+x) \le x$. Replacing this in (2.63), we get

$$a_n < \frac{1}{8}\sum_{k=1}^{n-1}\frac{1}{k(k+1)} < \frac{1}{8}\sum_{k=1}^{\infty}\frac{1}{k^2} = C < \infty$$

since the series $\sum_k 1/k^2$ is convergent.

Thus, the sequence $\{a_n\}$ is monotonically increasing and bounded above, and therefore has a finite limit that we shall call $a$. On the other hand, integrating by parts

$$A_n = \int_1^n \log x\, dx = x\, \log x\Big|_1^n - \int_1^n dx = n\log n - n + 1$$

and therefore

$$a_n = n\, \log n - n + 1 - \Big(\log 2 + \cdots + \log(n-1) + \frac{1}{2}\log n\Big).$$

Exponentiating both terms, we get

$$e^{a_n - 1} = \frac{n^n\, e^{-n}\, n^{1/2}}{n!},$$

whence

$$\frac{n!}{n^n e^{-n} n^{1/2}} \to \frac{1}{e^{a-1}} = \alpha \text{ as } n \to \infty.$$

**Step 2.** Let us show now that $\alpha = \sqrt{2\pi}$. For this we consider

$$\begin{aligned} I_n &= \int_0^{\pi/2} (\sin x)^n\, dx = \int_0^{\pi/2} (\sin x)^{n-2}(\sin x)^2\, dx \\ &= \int_0^{\pi/2} (\sin x)^{n-2}\, (1 - (\cos x)^2)\, dx \\ &= I_{n-2} - \int_0^{\pi/2} (\sin x)^{n-2} \cos x \cos x\, dx \\ &= I_{n-2} - \Big[\frac{1}{n-1}(\sin x)^{n-1}\cos x\Big|_0^{\pi/2} + \int_0^{\pi/2} \frac{1}{n-1}(\sin x)^{n-1}\sin x\, dx \\ &= I_{n-2} - \frac{1}{n-1} I_n \end{aligned}$$

whence

$$I_n = \frac{n-1}{n} I_{n-2}. \tag{2.64}$$

Hence, considering separately the cases in which $n$ is even or odd, and using (2.64) iteratively, we get

$$I_{2p} = \frac{2p-1}{2p}\frac{2p-3}{2p-2}\cdots\frac{3}{4}\frac{1}{2}\, I_0; \qquad I_{2p+1} = \frac{2p}{2p+1}\frac{2p-2}{2p-1}\cdots\frac{4}{5}\frac{2}{3}\, I_1.$$

Also

$$I_0 = \int_0^{\pi/2} dx = \frac{\pi}{2}, \qquad I_1 = \int_0^{\pi/2} \sin x\, dx = 1,$$

which means that

$$I_{2p} = \frac{2p-1}{2p}\frac{2p-3}{2p-2}\cdots\frac{3}{4}\frac{1}{2}\frac{\pi}{2}; \qquad I_{2p+1} = \frac{2p}{2p+1}\frac{2p-2}{2p-1}\cdots\frac{4}{5}\frac{2}{3}. \tag{2.65}$$

Observe now that $\{I_n\}$ is a decreasing sequence, since for $0 \le x \le \pi/2$ we have $0 \le \sin x \le 1$, $\sin^n x$ decreases with $n$, and therefore, so does the integral $I_n$. From this we get that

$$\frac{I_{2p+2}}{I_{2p}} < \frac{I_{2p+1}}{I_{2p}} < 1.$$

Using (2.64) we have

$$\frac{I_{2p+2}}{I_{2p}} = \frac{2p+1}{2p+2} \to 1 \quad \text{as } p \to \infty$$

and therefore the sequence in the middle

$$\frac{I_{2p+1}}{I_{2p}}$$

also goes to 1 as $p \to \infty$. Using now (2.65) we have that

$$\frac{I_{2p+1}}{I_{2p}} = \frac{[2p(2p-2)\cdots 4\cdot 2]^2}{(2p+1)[(2p-1)(2p-3)\cdots 5\cdot 3]^2}\frac{2}{\pi} \to 1 \quad \text{as } p \to \infty,$$

hence

$$\frac{2p(2p-2)\cdots 4\cdot 2}{(2p-1)(2p-3)\cdots 5\cdot 3}\ \sqrt{\frac{2}{\pi(2p+1)}} \to 1 \quad \text{as } p \to \infty.$$

Multiplying numerator and denominator by

$$2p(2p-2)\cdots 4\cdot 2 = 2^p p(p-1)(p-2)\cdots 1 = 2^p p!$$

we get

$$\frac{(2^p p!)^2}{(2p)!}\ \sqrt{\frac{2}{\pi(2p+1)}} \to 1 \quad \text{as } p \to \infty.$$

We now use the step 1 result. We know that

$$n! \sim \alpha\, n^n\, e^{-n}\sqrt{n}.$$

Substituting

$$\frac{(2^p p!)^2}{(2p)!}\ \sqrt{\frac{2}{\pi(2p+1)}} \sim \frac{(2^p \alpha\, p^p\, e^{-p}\sqrt{p})^2}{\alpha(2p)^{2p}\, e^{-2p}\sqrt{2p}}\ \sqrt{\frac{2}{\pi(2p+1)}} \sim \frac{\alpha}{\sqrt{2\pi}}.$$

Since the limit is 1, then necessarily $\alpha = \sqrt{2\pi}$. This finishes the proof of Stirling's formula. ■