Chapter 4

Sums and Means of Random Variables

4.1 Exact results for sums of random variables

We are often interested in the properties of the sum or mean of a sequence of random variables. In this section, we will introduce methods we can sometimes use to obtain exact probability information about these derived random variables. In later sections, we will discuss methods that, although more generally applicable, give only approximations to probabilities of interest.

We will need the following, general results (proved in Chapter 3) for independent random variables $X_1, X_2, \ldots, X_p$ and real constants $a_0, a_1, \ldots, a_p$:

$$E(X_1 + \cdots + X_p) = \sum_{i=1}^{p} E(X_i) \tag{4.1}$$

$$E(a_0 + a_1X_1 + \cdots + a_pX_p) = a_0 + \sum_{i=1}^{p} a_iE(X_i) \tag{4.2}$$

$$\text{Var}(X_1 + \cdots + X_p) = \sum_{i=1}^{p} \text{Var}(X_i) \tag{4.3}$$

$$\text{Var}(a_0 + a_1X_1 + \cdots + a_pX_p) = \sum_{i=1}^{p} a_i^2\text{Var}(X_i) \tag{4.4}$$

Example 1

A CD player has a shuffle feature. The next track to be played is always selected randomly from all the tracks on the CD, so that each track is equally-likely to be played. We will call the playing of a randomly-selected track a trial. Suppose there are ten tracks on a CD. What can be said about the random variable, S, the number of trials required until every track on the CD is played at least once?

Let the discrete random variable X_1 be the number of trials required until the first new track is played. Clearly X_1 takes the value 1 with probability 1.

Let the discrete random variable X_2 be the number of further trials required, after the first track is played, until another new track is played. On each trial, there is probability 0.1 of playing the track that has already been played and probability 0.9 of playing a new track. Trials are independent. So, $X_2 \sim \text{Geo}(0.1)$.

Let the random variable X_i $(i = 3, \ldots, 10)$ be the number of further trials required, after the $(i-1)$th different track is played for the first time, until the ith different track is played. Then $X_i \sim \text{Geo}(\theta_i)$ where $\theta_i = (i-1)/10$.

$$\begin{aligned} E(X_i) &= \frac{1}{1-\theta_i} = \\ \text{Var}(X_i) &= \frac{\theta_i}{(1-\theta_i)^2} = \end{aligned}$$

Now, $S = X_1 + X_2 + \cdots + X_{10}$, where $X_1, \ldots, X_{10}$ are independent random variables. So, using (4.1) and (4.3):

$$\begin{aligned} E(S) &= \\ &= \\ \text{Var}(S) &= \\ &= \end{aligned}$$

On average, then, 29.29 tracks have to be played in order to hear all ten different tracks on the CD at least once. This might not seem a lot. However, the variance is very large relative to the expected value so there is a good chance that many more than this average number will be required on a particular occasion. We are not in a position to obtain probabilities related to S algebraically, but the results from 10,000 simulations are plotted in the Figure 4.1. The distribution of S is highly skewed towards larger values.

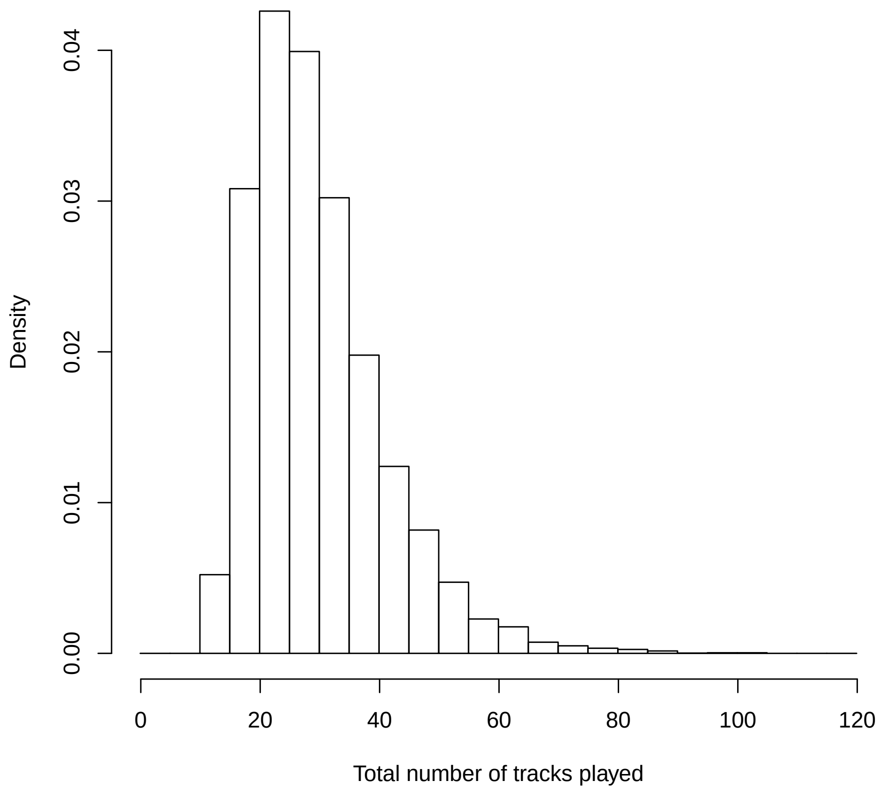

Figure 4.1: Histogram of 10,000 simulations for the number of tracks need to be played in order to play all the 10 tracks on the CD that is being played in shuffle mode.

There are many important cases where we can use moment-generating functions to determine the entire probability distribution of the sum, S, of a sequence of random variables. The following result is fundamental.

Proposition 4.1

Suppose that $X_1, \ldots, X_n$ are independent random variables, each with a finite moment generating function $M_i(t)$. Then the moment-generating function of

$$S = X_1 + \cdots + X_n$$

is

$$M_S(t) = \prod_{i=1}^{n} M_i(t).$$

Proof

$$\begin{aligned} M_S(t) = E(e^{St}) &= E\big(e^{(X_1+\cdots+X_n)t}\big) \\ &= E\big(e^{X_1 t} e^{X_2 t} \cdots e^{X_n t}\big) \\ &= E\big(e^{X_1 t}\big) E\big(e^{X_2 t}\big) \cdots E\big(e^{X_n t}\big) \qquad \text{independence} \\ &= \prod_{i=1}^{n} M_i(t). \end{aligned}$$

Proposition 4.1 allows us to obtain a variety of important results, known as **reproductive properties**, that arise when the sum of independent random variables drawn from a given family of distributions has a distribution that also belongs to that family.

Example 2

Suppose that $X_1, \ldots, X_n$ are independent random variables and that $X_i \sim \text{Poi}(\lambda_i)$ for $\lambda_i > 0$ $(i = 1, \ldots, n)$. Let $S = X_1 + \cdots + X_n$. Then:

Example 3

Suppose that $X_1, \ldots, X_n$ are independent random variables and that $X_i \sim N(\mu_i, \sigma_i^2)$. The moment-generating function of X_i is

$$M_i(t) = \exp\left(\mu_i t + \frac{1}{2}\sigma_i^2 t^2\right).$$

Let $S = a_0 + a_1X_1 + a_2X_2 + \cdots + a_nX_n$, where $a_0, a_1, a_2, \ldots, a_n$ are real constants (not all of which are zero). Then, since the random variables are independent:

$$\begin{aligned}
M_S(t) &= e^{a_0 t}\prod_{i=1}^{n} M_i(a_i t) \\
&= e^{a_0 t}\prod_{i=1}^{n} \exp\left(\mu_i a_i t + \frac{1}{2}\sigma_i^2 a_i^2 t^2\right) \\
&= \exp\left[(a_0 + a_1\mu_1 + \cdots + a_n\mu_n)t + \frac{1}{2}(a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2)t^2\right].
\end{aligned}$$

This is the moment-generating function of the Normal distribution with expected value $a_0 + a_1\mu_1 + \cdots + a_n\mu_n$ and variance $a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2$. Using the Uniqueness Property of moment-generating functions, $S \sim N\big(a_0 + a_1\mu_1 + \cdots + a_n\mu_n, a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2\big)$.

Note that, from (4.2) and (4.4), we already knew these expressions for $E(S)$ and $\mathrm{Var}(S)$; it is the fact that S is normally distributed that is the new result.

Now suppose that $X_1, \ldots, X_n$ are identically distributed, as well as independent. This means that $\mu_1 = \cdots = \mu_n \equiv \mu$ and $\sigma_1 = \cdots = \sigma_n \equiv \sigma$, for some real value μ and $\sigma > 0$. Consider the sum of the random variables

$$S = X_1 + \cdots + X_n.$$

Setting $a_0 = 0$ and $a_1 = \cdots = a_n = 1$, the general result we have just proved shows that

$$S \sim N(n\mu, n\sigma^2).$$

Consider next the sample mean of the random variables,

$$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n).$$

Setting $a_0 = 0$ and $a_1 = \cdots = a_n = \frac{1}{n}$, then the same general result shows that

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

4.2 The Central Limit Theorem

In the last section, we saw how to determine the entire probability distribution of the sum of some sequences of random variables. We were then able to calculate exact probabilities associated with the sum or average. Limit theorems allow us to find approximations to probabilities of interest, even when we cannot write down the probability distribution of the sum explicitly.

Proposition 4.2: Chebyshev's Inequality

Let X be a random variable with *finite* expected value μ. If c is a real constant such that $E[(X-c)^2]$ is finite, then for any value $\epsilon > 0$

$$P(|X-c| < \epsilon) \geqslant 1 - \frac{1}{\epsilon^2} E[(X-c)^2].$$

In particular, if X has *finite* variance, $\sigma^2 = E[(X-\mu)^2]$, then for any value $\epsilon > 0$

$$P(|X-\mu| < \epsilon) \geqslant 1 - \frac{\sigma^2}{\epsilon^2}.$$

Proof

This result is easily proved when X is a continuous random variable with probability density function $f_X(x)$. In this case,

$$\begin{aligned}
E[(X-c)^2] &= \int_{-\infty}^{\infty} (x-c)^2 f_X(x)\,dx \\
&\geqslant \int_{-\infty}^{c-\epsilon} (x-c)^2 f_X(x)\,dx + \int_{c+\epsilon}^{\infty} (x-c)^2 f_X(x)\,dx \\
&\geqslant \int_{-\infty}^{c-\epsilon} \epsilon^2 f_X(x)\,dx + \int_{c+\epsilon}^{\infty} \epsilon^2 f_X(x)\,dx \\
&= \epsilon^2 \left[\int_{-\infty}^{c-\epsilon} f_X(x)\,dx + \int_{c+\epsilon}^{\infty} f_X(x)\,dx\right] \\
&= \epsilon^2 \left[1 - P(|X-c| < \epsilon)\right] \\
\implies P(|X-c| < \epsilon) &\geqslant 1 - \frac{1}{\epsilon^2} E[(X-c)^2].
\end{aligned}$$

The second part of the theorem follows immediately from the first by putting $c = \mu$.

Proposition 4.3: The Weak Law of Large Numbers

Let $X_1, X_2, \ldots$ be a sequence of independent and identically-distributed random variables, each with finite expected value μ. For $n = 1, 2, \ldots,$ let

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then, for any $\epsilon > 0$,

$$P(|\overline{X}_n - \mu| < \epsilon) \to 1 \qquad \text{as } n \to \infty.$$

Proof

We will prove this result for the simple case where the random variables have *finite* variance σ^2; however this assumption is not required for the Weak Law to hold.

For all $n \geqslant 1$, $\overline{X}_n$ has expected value μ and variance σ^2/n (from (4.2)) and (4.4)). By Chebyshev's Inequality, for any $\epsilon > 0$,

$$P(|\overline{X}_n - \mu| < \epsilon) \geqslant 1 - \frac{\sigma^2/n}{\epsilon^2} \to 1 \quad \text{as } n \to \infty.$$

Definition

Let $X_1, X_2, \ldots$ be a sequence of random variables defined on a sample space S. The sequence X_n is said to **converge in probability** to the random variable X if, for every $\epsilon > 0$,

$$P(|X_n - X| < \epsilon) \to 1 \quad \text{as } n \to \infty.$$

The Weak Law of Large Numbers shows that $\overline{X}_n$ converges in probability to μ as $n \to \infty$.

Proposition 4.4: The Strong Law of Large Numbers

Let X_1, $X_2, \ldots$ be a sequence of independent and identically-distributed random variables, each with *finite* expected value μ. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then, for all real constants $\epsilon > 0$ and $\delta > 0$, there is an integer N such that

$$P(\text{all } |\overline{X}_n - \mu| < \epsilon; n = N, N+1, \ldots) > 1 - \delta.$$

A proof can be found in the volume by Feller[1].

Definition

Let X and the sequence $X_1, X_2, \ldots$ be random variables defined on a sample space S. Let E be the set of all elements of S for which $X_n(s)$ converges to $X(s)$ as $n \to \infty$, i.e.

$$E = \{s \in S : X_n(s) \to X(s) \text{ as } n \to \infty\}.$$

If $P(E) = 1$, then the sequence X_n is said to **converge almost surely** to X.

The Strong Law of Large Numbers shows that $\overline{X}_n$ converges almost surely to μ as $n \to \infty$. Almost sure convergence is a stronger property than convergence in probability. If a sequence of random variables converges almost surely then it also converges in probability (but the reverse is not necessarily true).

Informally, the Laws of Large Numbers tell us that the probability distribution of $\overline{X}_n$ becomes more and more concentrated at its expected value μ as $n \to \infty$. Although interesting and important, this does not help us to calculate probabilities of interest associated with $\overline{X}_n$ since it does not tell us how close $\overline{X}_n$ is to μ for a given value of n. The Central Limit Theorem provides a means of doing this, at least approximately.

---

[1] Feller, W. (1968). *An Introduction to Probability Theory and its Applications, Vol. 1*, 3rd edn, Wiley, New York.

Proposition 4.5: The Central Limit Theorem

Suppose that $X_1, \ldots, X_n$ is a sequence of independent and identically-distributed random variables, each with finite expected value μ and finite variance σ^2. For sufficiently large values of n,

$$Z_n \equiv \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} \sim N(0,1)$$

approximately, in the sense that $\lim_{n\to\infty} P(Z_n \leqslant z) \to \Phi(z)$ for all real values z.

Proof

Since $X_1, \ldots, X_n$ all follow the same probability distribution, then they have a common moment-generating function $M_X(t)$. Since their common expected value and variance are finite, it follows that $M_X(t)$ is finite and (at least) twice differentiable in some interval around $t = 0$. Expanding $M_X(t)$ in a Taylor series around the value $t = 0$ gives:

$$\begin{aligned} M_X(t) &= M_X(0) + \frac{M_X'(0)}{1!}t + \frac{M_X''(0)}{2!}t^2 + \cdots \\ &= 1 + \frac{E(X)}{1!}t + \frac{E(X^2)}{2!}t^2 + \cdots \\ &= 1 + \mu t + \tfrac{1}{2}(\mu^2 + \sigma^2)t^2 + \cdots . \end{aligned}$$

Now

$$Z_n = \sum_{i=1}^{n} \frac{1}{\sqrt{n}\sigma} X_i - \frac{\sqrt{n}\mu}{\sigma}.$$

So Z_n has moment-generating function

$$M_Z(t) = \left[M_X\left(\frac{t}{\sqrt{n}\sigma}\right)\right]^n \exp\left(-\frac{\sqrt{n}\mu}{\sigma}t\right).$$

Taking the (natural) logarithm of both sides gives

$$\begin{aligned} \log M_Z(t) &= n \log\left[M_X\left(\frac{t}{\sqrt{n}\sigma}\right)\right] - \frac{\sqrt{n}\mu}{\sigma}t \\ &= n \log\left[1 + \mu\frac{t}{\sqrt{n}\sigma} + (\mu^2 + \sigma^2)\frac{t^2}{2n\sigma^2} + \cdots\right] - \frac{\sqrt{n}\mu}{\sigma}t. \end{aligned}$$

In general, when $-1 < u < 1$,

$$\log(1+u) = u - \frac{u^2}{2} + \frac{u^3}{3} - \cdots .$$

For sufficiently large values of n, $u \equiv \mu\frac{t}{\sqrt{n}\sigma} + (\mu^2+\sigma^2)\frac{t^2}{2n\sigma^2} + \cdots < 1$, so

$$\begin{aligned} \log M_Z(t) = n &\left[\mu\frac{t}{\sqrt{n}\sigma} + (\mu^2 + \sigma^2)\frac{t^2}{2n\sigma^2} + \cdots\right] \\ &- \frac{n}{2}\left[\mu\frac{t}{\sqrt{n}\sigma} + (\mu^2 + \sigma^2)\frac{t^2}{2n\sigma^2} + \cdots\right]^2 + \cdots - \frac{\sqrt{n}\mu}{\sigma}t \\ = n\mu&\frac{t}{\sqrt{n}\sigma} + n(\mu^2+\sigma^2)\frac{t^2}{2n\sigma^2} - \frac{n}{2}\left[\mu\frac{t}{\sqrt{n}\sigma}\right]^2 + \cdots - \frac{\sqrt{n}\mu}{\sigma}t \\ = \frac{1}{2}&t^2 + \cdots \to \frac{1}{2}t^2 \text{ as } n \to \infty. \end{aligned}$$

That is, $\lim_{n\to\infty} \log M_Z(t) = \frac{1}{2}t^2$. Therefore,

$$\lim_{n\to\infty} M_Z(t) = \exp(\tfrac{1}{2}t^2).$$

This is the moment-generating function of the $N(0,1)$ distribution. Using the Uniqueness Property of moment-generating functions, this means that Z_n converges in distribution to the $N(0,1)$ distribution as $n \to \infty$.

The Central Limit Theorem is often used in one of the following two equivalent forms:

(a) $\sum_{i=1}^{n} X_i$ approximately follows the $N(n\mu, n\sigma^2)$ distribution for 'sufficiently large' n.

(b) $\overline{X}$ approximately follows the $N\big(\mu, \frac{\sigma^2}{n}\big)$ distribution for 'sufficiently large' n.

We have already derived these expected values and variances for the sum and average of independent and identically-distributed random variables; they are exact values that we do not need the Central Limit Theorem to justify. The new information from the Central Limit Theorem is the approximate Normality of the distribution of the sum and average for sufficiently large values of n. (We have already shown that, in the special case when $X_1, \ldots, X_n$ are normally distributed, then these are exact not approximate results.)

Example 4

It is usual for even very large financial transactions, to the value of hundreds of thousands of pounds, to be settled to fractions of pence. Suppose, instead, that financial institutions agreed to round all settlements of transactions between them to the nearest whole £1. How much would an individual institution stand to lose?

Suppose that the institution deals in n separate transactions in a certain period (e.g., one week). Let £X_i denote the amount it gains on the ith transaction as a result of rounding; a positive value of X_i is a true gain while a negative value of X_i is a loss. It seems reasonable to suppose that $X_i \sim U\big(-\frac{1}{2}, \frac{1}{2}\big)$, so $E(X_i) = 0$ and $\text{Var}(X_i) = \frac{1}{12}$.

Let S_n denote the total gain (or loss) on the n transactions. Then, using the Central Limit Theorem:

4.3 The Normal approximation to discrete distributions

The Normal distribution is commonly used to evaluate approximate tail probabilities for the Binomial distribution when the sample size, n, is large. The Central Limit Theorem justifies this approximation, using the following argument.

Suppose $X \sim \text{Bi}(n, \theta)$. Then, X is the number of 'successes' in n independent trials, where each trial has success probability θ. Let X_i be the number of successes in the ith trial ($i = 1, \ldots, n$). Then X_i can only take the value 0 or 1, with $P(X_i = 0) = 1 - \theta$ and $P(X_i = 1) = \theta$. This means that $X_1, \ldots, X_n$ are independent and identically distributed random variables, each with expected value θ and variance $\theta(1-\theta)$.

By definition, $X = \sum_{i=1}^{n} X_i$, so the Central Limit Theorem tells us that, approximately, for sufficiently large values of n:

$$Z_n \equiv \frac{X - n\theta}{\sqrt{n\theta(1-\theta)}} \sim N(0, 1).$$

In order to give meaning to the phrase 'sufficiently large' n, it is usual to adopt the following approach to the Normal approximation to the Binomial distribution.

- If $X \sim \text{Bi}(n, \theta)$, where $n \geqslant 20$, $n\theta \geqslant 5$ and $n(1-\theta) \geqslant 5$, then approximately

$$X \sim N\big(n\theta, n\theta(1-\theta)\big).$$

A similar argument justifies the following approximation to the Poisson distribution.

- If $X \sim \text{Poi}(\theta)$, where $\theta \geqslant 30$, then approximately

$$X \sim N(\theta, \theta).$$

Both these results require a discrete distribution to be approximated by a continuous one. This introduces some inconsistencies, since the probability that a Normal random variable equals any particular value is 0. Using a **continuity correction** improves the approximation.

For integer values, m, correcting for continuity means evaluating:

$$\begin{aligned} P(m - \tfrac{1}{2} < X < m + \tfrac{1}{2}) &\quad \text{rather than} \quad P(X = m), \\ P(X < m + \tfrac{1}{2}) &\quad \text{rather than} \quad P(X \leqslant m) \text{ or } P(X < m + 1), \\ P(X > m - \tfrac{1}{2}) &\quad \text{rather than} \quad P(X \geqslant m) \text{ or } P(X > m - 1). \end{aligned}$$

Example 5

A random digit generator is to be used to obtain $100k$ random digits, for some positive integer k. Let the random variable X be the number of zeros that is generated. On average, one-tenth of all the digits generated should be zeros, so X has the $\text{Bi}(100k, 0.1)$ distribution. X should take a value close to $10k$ on average. Let us find the probability that $9k < X < 11k$, i.e., the probability that the proportion of zeros lies between 0.09 and 0.11.

Solution

$X \sim \mathrm{Bi}(100k, 0.1) \implies \mathrm{E}(X) = 100k \times 0.1 = 10k, \mathrm{Var}(X) = 100k \times 0.1 \times (1 - 0.1) = 9k.$

For a sufficiently large value of k,

$$\frac{X - 10k}{\sqrt{9k}} = \frac{X - 10k}{3\sqrt{k}} \sim N(0, 1), \text{ approximately.}$$

So, using a continuity correction,

$$P\left(9k + \tfrac{1}{2} < X < 11k - \tfrac{1}{2}\right) = P\left(\frac{1 - 2k}{6\sqrt{k}} < \frac{X - 10k}{3\sqrt{k}} < \frac{2k - 1}{6\sqrt{k}}\right) = 2\Phi\left(\frac{2k - 1}{6\sqrt{k}}\right) - 1.$$

This probability is strictly increasing with k. As the sample size increases (i.e., as $k \to \infty$), the probability that the proportion of zeros lies between 0.09 and 0.11 tends to 1. This is what we would expect from the laws of large numbers. The Central Limit Theorem provides further insight into the rate of convergence. For example, the probability exceeds 0.9 for sample sizes of at least 2600 ($k = 26$) and exceeds 0.99 for sample sizes of at least 6100: see Figure 4.2.

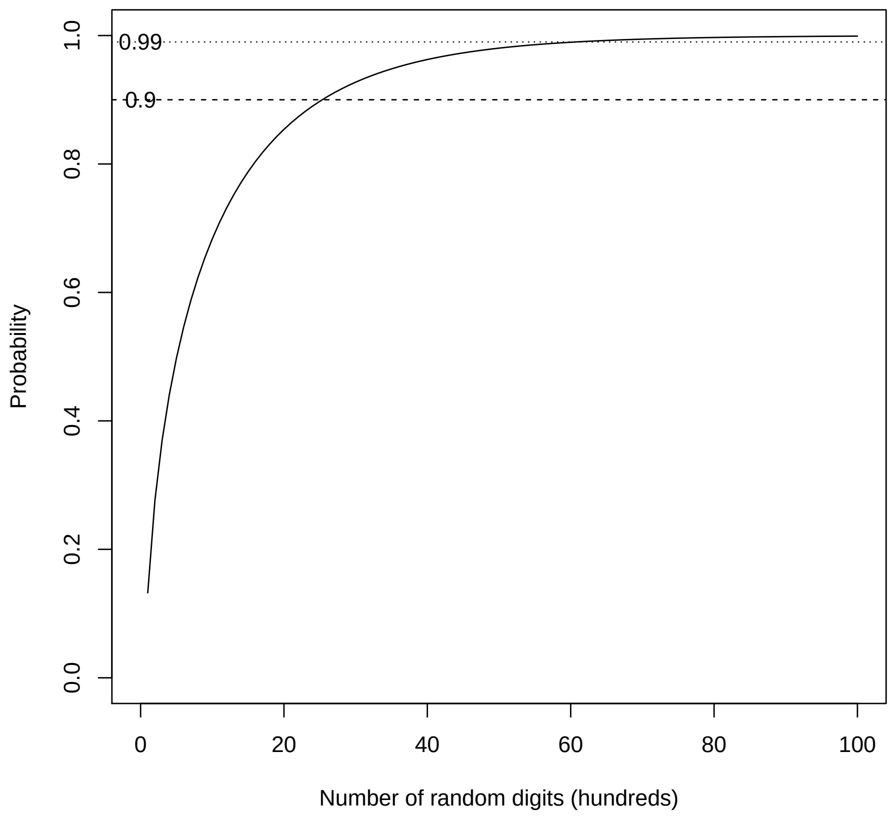

Figure 4.2: The probability that the proportion of zeros is between 0.09 and 0.11.

THE END