Problem 1. A Zipf \((n, \alpha = 1)\) random variable \(X\) has p.m.f.

\[ P_X(x) = c(n)/x, \text{ for } x = 1, 2, 3, \cdots, n \]

The constant \(c(n)\) is set so that \(\sum_{n=1}^{n} P_X(x) = 1\) (to make it a proper p.m.f.). Calculate \(c(n)\) for \(n = 1, 2, 3, 4, 5, 6\).

Solution 1: The requirement that

\[ \sum_{n=1}^{n} P_X(x) = 1 \]  \hspace{1cm} (1)

For \(n = 1\) : \(c(1) \cdot \left[\frac{1}{1}\right] = 1 \implies c(1) = 1; \)  \hspace{1cm} (2)

For \(n = 2\) : \(c(2) \cdot \left[\frac{1}{1} + \frac{1}{2}\right] = 1 \implies c(2) = \frac{2}{3}; \)  \hspace{1cm} (3)

For \(n = 3\) : \(c(3) \cdot \left[\frac{1}{1} + \frac{1}{2} + \frac{1}{3}\right] = 1 \implies c(3) = \frac{6}{11}; \)  \hspace{1cm} (4)

For \(n = 4\) : \(c(4) \cdot \left[\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4}\right] = 1 \implies c(4) = \frac{12}{25}; \)  \hspace{1cm} (5)

For \(n = 5\) : \(c(5) \cdot \left[\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5}\right] = 1 \implies c(5) = \frac{60}{137}; \)  \hspace{1cm} (6)

For \(n = 6\) : \(c(6) \cdot \left[\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6}\right] = 1 \implies c(6) = \frac{20}{49}; \)  \hspace{1cm} (7)

Problem 2. The Zipf \((n, \alpha = 1)\) random variable \(X\) introduced in Problem 1 is often used to model the “popularity” of a collection of \(n\) objects. For example, a Web server can deliver one of \(n\) Web pages. The pages are numbered such that the page 1 is the most requested page, page 2 is the second most requested page and so on. If page \(k\) is requested, then \(X = k\). To reduce external network traffic, an ISP gateway caches copies of the \(k\) most popular pages. Using Matlab, calculate, as a function of \(n\) for \(1 \leq n \leq 100\), how large \(k\) must be to ensure that the cache can deliver a page with probability 0.75 or more.
Solution 2: Suppose $X_n$ is a Zipf ($n, \alpha = 1$) random variable and thus has PMF

$$P_X(x) = \begin{cases} \frac{c(n)}{x} & x = 1, 2, ... n \\ 0 & \text{otherwise} \end{cases}$$

The problem asks us to find the smallest value of $k$ such that $P[X_n \leq k] \geq 0.75$. That is, if the server caches the $k$ most popular files, then with $P[X_n \leq k]$ the request is for one of the $k$ cached files. First, we might as well solve this problem for any probability $p$ rather than $p = 0.75$. Thus, in math terms, we are looking for

$$k = \min\{k'|P[X_n \leq k'] \geq p\} \quad (8)$$

What makes the Zipf distribution hard to analyze is that there is no closed form expression for

$$c(n) = \left( \sum_{x=1}^{n} \frac{1}{x} \right)^{-1} \quad (9)$$

Thus we use MATLAB to grind through calculations. The following simple program generates the Zipf distribution and returns the correct value of $k$.

The program zipfcache generalizes 0.75 to be the probability $p$. Although this program is sufficient the problem asks us to find $k$ for all values of $n$ from 1 to $10^2$. One way to do this is to call Zipfcache a hundred times to find $k$ for each value of $n$. A better way is to use the properties of the Zipf PDF. In particular

$$P[X_n \leq k'] = c(n) \sum_{K=1}^{k'} \frac{1}{x} = \frac{c(n)}{c(k')} \quad (10)$$
Thus we wish to find

\[ k = \min\{k' \mid \frac{c(n)}{c(k')} \geq p \} = \min\{k' \mid \frac{1}{c(k')} \geq \frac{p}{c(n)} \} \]  

(11)

Note that the definition of \( k \) implies that

\[ \frac{1}{c(k')} < \frac{p}{c(n)} \]  

(12)

\[ \text{for } k' = 1, \ldots, k - 1. \]  

(13)

Using the notation \(|A|\) to denote the number of elements in the set \( A \), we can write

\[ k = 1 + |k' \{ \frac{1}{c(k')} < \frac{p}{c(n)} \}| \]  

(15)

This is the basis for a very short MATLAB program:

Note that \text{zipfcacheall} uses a short Matlab program \text{countless.m} that is almost the same as \text{count.m} introduced in Example 2.47. If \( n = \text{countless}(x, y) \), then \( n(i) \) is the number of elements of \( x \) that are strictly less than \( y(i) \) while \text{count} returns the number of elements less than or equal to \( y(i) \). In any case, the commands \( k = \text{zipfcacheall}(100, 0.75); \text{plot}(1:100, k); \) is sufficient to produce this figure of \( k \) as a function of \( m \):
We see in the figure that the number of files that must be cached grows slowly with the total number of files $n$.

Finally, we make one last observation. It is generally desirable for Matlab to execute operations in parallel. The program zipfcacheall generally will run faster than $n$ calls to zipfcache. However, to do its counting all at once, countless generates and $n^2$ array. When $n$ is not too large, say $n \leq 100$, the resulting array with $n^2 = 1,0000$ elements fits in memory. For much large values of $n$, say $n = 106$ (as was proposed in the original printing of this edition of the text), countless will cause an “out of memory” error.

**Problem 3.** We measure for resistance $R$ of each resistor in a production line and we accept only the units whose resistance is between 96 and 104 ohms. Find the percentage of the accepted units if

- $R$ is uniform between 95 and 105 ohms
- $R$ is Gaussian with mean 100 and standard deviation 2 ohms.

**Solution 3:** Percentage of units between 96 and 104 ohms equals 100 p and p is calculated as follows

$$P = P(95 \leq X \leq 104) = F(104) - F(96)$$

where $F(.)$ is the cumulative density function
Case 1: \( R \) is uniform between 95 and 105 ohms

\[
F(X) = 0.1(X - 95) \quad \text{for} \quad 95 \leq X \leq 100
\]

\[
P = 0.1(104 - 95) - 0.1(96 - 95) = 0.8
\]

Case 2: \( R \) is Gaussian with mean 100 and standard deviation 2 ohms.

\[
P = \Phi\left(\frac{104 - 100}{2}\right) - \Phi\left(\frac{96 - 100}{2}\right) = 0.987
\]

**Problem 4.** The probability of heads of a random coin is a RV \( P \) uniform in the interval \((0,1)\).

- Find the probability \( P[0.3 \leq P \leq 0.7] \).
- The coin is tossed 10 times and heads shows 6 times. Given this fact, find the probability that \( P \) is between 0.3 and 0.7.

**Solution 4:**

Find the probability \( P[0.3 \leq P \leq 0.7] \)

\[
P[0.3 \leq P \leq 0.7] = \int_{0.3}^{0.7} dp = 0.4
\]

The coin is tossed 10 times and heads shows 6 times. Given this fact, find the probability that \( P \) is between 0.3 and 0.7. We have that head shows 6 times from 10 tosses and we are asked to get the conditional probability \( P[0.3 \leq P \leq 0.7 | A] \) where \( A \) is the event you get 6 times heads out of ten tosses

\[
f(p | A) = \frac{(p^6)(1-p)^4}{\int_{0}^{1} p^6 (1-p)^4 dp} = \frac{(p^6)(1-p)^4}{4329 \times 10^{-7}}
\]

\[
P[0.3 \leq P \leq 0.7 | A] = \int_{0.3}^{0.7} f(p | A) dp = \frac{10^7}{4329} \int_{0.7}^{0.3} p^6 (1-p)^4 dp = 0.768
\]