Problem 1 - A heavy tailed distribution

For this problem, submit your code through the Assignments page link. Additional reading: the section on Zipf’s law from the text book.

In this problem you will be acquainted with a “difficult” distribution, which however appears in many practical applications. This is a distribution over the positive integers 1, 2, 3, ... i ..., defined by

\[ P(i) = \frac{1}{Z_i^r}, \quad r > 1 \]  

For this homework, assume that the sample space is finite, \( S = \{1, 2, \ldots, 10^4\} \).

Let \( r = 3.5 \).

a. Calculate the value of the normalization constant \( Z \).

b. Make a stem plot of the distribution in the range \( i = 1, \ldots, 20 \).

c. Calculate \( m = E[i] \) and \( v^2 = Var(i) \) the mean and variance of this distribution. Calculate the standard deviation \( v = \sqrt{Var(i)} \).

d. Now you will estimate the mean and variance from samples. For comparison, you will also generate samples from a normal distribution and estimate its mean and variance. Repeat \( N = 500 \) times:

1. take a sample of size \( n = 100 \) from \( P \)

2. estimate \( \hat{m}, \hat{v}^2 \) from this sample

3. take a sample of size \( n = 100 \) from \( Normal(m, v^2) \)

4. estimate \( \hat{m}_u, \hat{v}^2_u \) for the normal distribution

Now you have a set of \( N \) estimated means and variances from the two distributions. We will treat the means \( \hat{m} \) as independent samples (think why they should be independent) from some distribution (think what distribution this should be) and we will estimate the variance of that distribution from the \( N \) samples. We are interested in the variance of \( \hat{m} \) because, intuitively, the mean of the \( \hat{m} \) estimates should be \( m \). We are interested in how close, on average, the estimate \( \hat{m} \) is to \( m \). We shall do the same for the other quantities.
Calculate the standard deviation of $\hat{m}$, $\hat{v}^2$, $\hat{m}_u$, $\hat{v}_u^2$ from the $N$ samples, and denote them respectively $s_m$, $s_v$, $s'_m$, $s'_v$.

You are encouraged to run this experiment several times, with different values of $N$ and $n$, in order to see how stable are the results.

e. Derive the theoretical value of $s_m$ and $s'_m$. To do this, note that $\hat{m}$ is a random variable equal to $\sum_{j=1}^{n} i_j/n$ where $i_j$ is the $j$’th draw from $P$. You should obtain a formula that depends on $n$ and other known quantities. Literal answer only is required, but you may like to calculate the numerical value too.

f. Let us now compare $s_v$, $s'_v$ the standard deviations of the variance estimators. Which one is larger? What is the ratio of the largest over the smallest one?

Discuss this result (1–2 sentences) in view of your answer to e.

g. Let us now compare $s_m$, $s'_m$ the standard deviations of the mean estimators. Which one is larger? What is the ratio of the largest over the smallest one?

Discuss this result (1–2 sentences) in view of your findings in e,f,g.

h. Extra credit What should be the variance of the estimated mean of $\hat{m}$? In other words, let $\bar{m} = \sum_{j=1}^{N} \hat{m}^{(j)}/N$. What is the the variance of $\bar{m}$? (Look for a short answer).

For more extra credit: What should be the theoretical variance of $\hat{v}$?

Problem 2 - Testing the Central Limit Theorem For this problem, submit your code through the Assignments page link.

Choose two distributions from the following list. You must choose at least one distribution which is not among the first 3.

1. Uniform in $[0,1]$
2. Uniform in $\{1,2,3\}$
3. Normal$(1,4)$
4. The distribution from Problem 1
5. Continuous exponential $\lambda = 1$

For each of the two distributions $P$ of your choice, repeat:

a. Calculate $m$, $v^2$ the mean and variance of $P$ (formula and numerical values for all but the distribution in problem 1; nothing to do for distribution 4).

b.

• Repeat for $n = 5, 50, 500$.
• Repeat $N = 100$ times
  
  1. Draw a sample of size $n$ from $P$
2. Calculate the value of $Z$

\[ Z = \frac{\sum_{i=1}^{n}(X_i - m)}{\sqrt{nv^2}} \]  \hspace{1cm} (2)

- Plot the distribution of $Z$ (histogram OK) and the density of $Normal(0, 1)$ on the same graph
- Optional: plot the CDF of $Z$ and $\Phi$ the CDF of $Normal(0, 1)$ on the same graph

Optional, for nicer results use $N = 1000$ instead of $N = 100$. c.kfka