Obs & Stats HW 2

Ryan Hofmann

8 February 2017

A Hunt for Galaxy Clusters

1. If the probability of a single image containing a cluster is $p = 0.3$, then the probability of the image not containing a cluster is $1 - p = 0.7$. Since each image can be considered an independent event (much like a coin toss), the probability of NOT finding a cluster after k images is $(1-p)^k$. Then the probability of finding a cluster in the k^{th} image is

$$P[k] = p(1-p)^{k-1} = 0.3 \times 0.7^{k-1} \tag{1}$$

2. Rewrite the sum so that it starts at $k = 0$:

$$\sum_{k=1}^{\infty} p(1-p)^{k-1} = \sum_{k=1}^{\infty} \frac{p}{1-p}(1-p)^k = -\frac{p}{1-p} + \sum_{k=0}^{\infty} \frac{p}{1-p}(1-p)^k \tag{2}$$

The second term is a geometric series with sum $S = \frac{a}{1-r}$, with $a = \frac{p}{1-p}$ and $r = 1 - p$:

$$S = \frac{p}{1-p} \frac{1}{1-(1-p)} = \frac{1}{1-p} \tag{3}$$

Therefore,

$$\sum_{k=1}^{\infty} P[k] = \frac{1}{1-p} - \frac{p}{1-p} = \frac{1-p}{1-p} = 1 \tag{4}$$

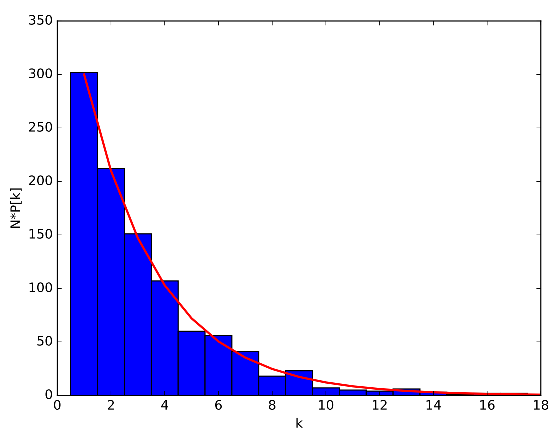
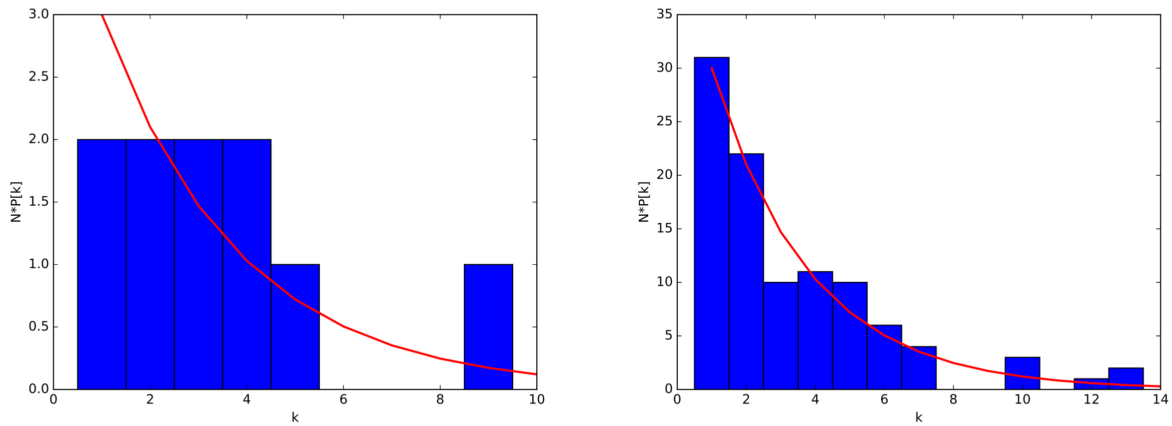


Figure 1: 1000 trials with $p = 0.3$

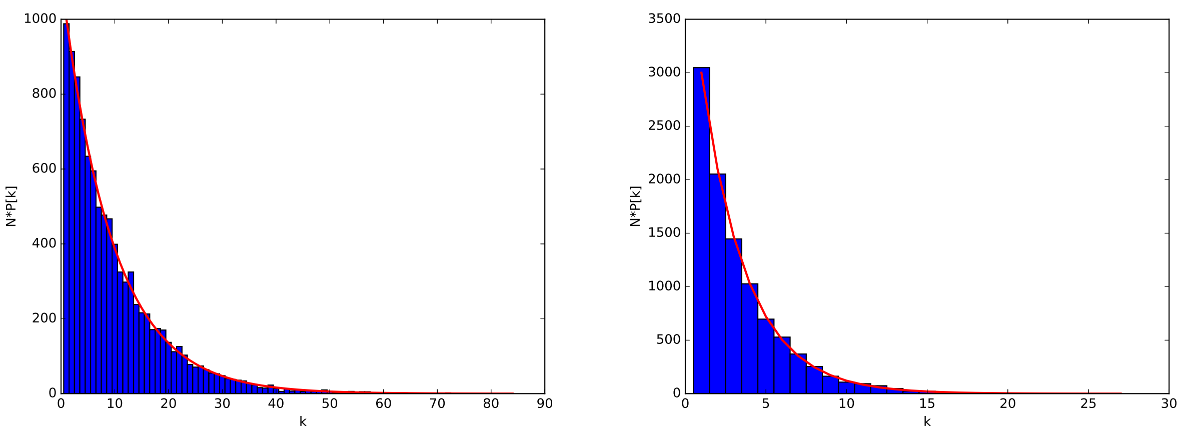
3. As Figure 1 shows, the histogram is in excellent agreement with the analytical calculation, though there is still noticeable noise in the tail of the distribution.

4. The agreement between histogram and analytic solution is much worse for smaller N, as shown in Figure 2. For $N = 1e4$, the agreement is very good for both $p = 0.1$ and $p = 0.3$ (Figure 3), but the mean is much higher for the smaller value of p; this indicates, as expected, that the more spread-out clusters are, the more observations will typically be required for a detection.



(a) $N = 10$ (b) $N = 100$

Figure 2: Distributions with $p = 0.3$ at low N



(a) $p = 0.1$ (b) $p = 0.3$

Figure 3: Distributions with $N = 1e4$

Bayesian Analysis of Tainted Beef Patties

1. Let $P(A)$ and $P(B)$ represent the probabilities of the burger coming from suppliers A and B, respectively; $P(X|A)$ and $P(X|B)$ the probabilities of an infected burger given that it came from supplier A or B; and $P(A|X)$ the probability that the burger came from supplier A, given that it was infected. Baye's Theorem states that[1]

$$P(A|X) = \frac{P(X|A)P(A)}{P(X|A)P(A) + P(X|B)P(B)} \tag{5}$$

Plugging in the values from the problem statement,

$$P(A|X) = \frac{(0.0001)(\frac{5}{6})}{(0.0001)(\frac{5}{6}) + (0.0003)(\frac{1}{6})} = \frac{5}{8} \tag{6}$$

PDF, CDF, and Moments

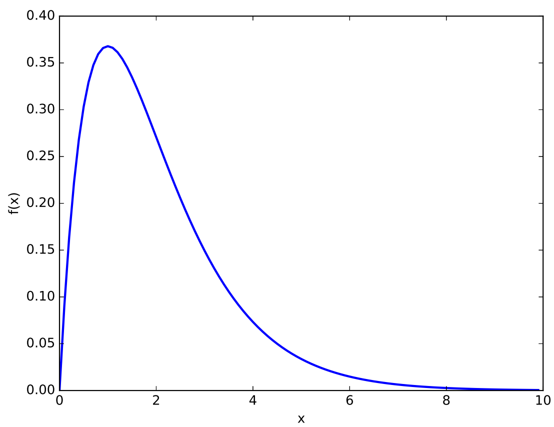


Figure 4: Probability distribution function $f(x) = xe^x$, $x \geq 0$

1. The PDF $f(x)$ is plotted in Figure 4. To verify the normalization, integrate by parts:

$$\int_{-\infty}^{\infty} f(x)dx = \int_0^{\infty} xe^{-x}dx = [xe^{-x}]_{\infty}^0 + \int_0^{\infty} e^{-x}dx = 0 + [e^{-x}]_{\infty}^0 = 1 \tag{7}$$

2. The CDF is defined as

$$F(x) = \int_0^x f(x')dx' \tag{8}$$

Integrating by parts,

$$F(x) = -xe^{-x} + 1 - e^{-x} = 1 - (1+x)e^{-x} \tag{9}$$

3. The mean is defined as

$$\mu \equiv \mathrm{E}[x] = \int_{-\infty}^{\infty} xf(x)dx = \int_0^{\infty} x^2 e^{-x}dx \tag{10}$$

Integrating by parts twice,

$$\mu = [e^{-x}(x^2 - 2x + 2)]_{\infty}^0 = 2 \tag{11}$$

The fraction of total probability below the mean can be evaluated directly using the CDF:

$$F(\mu) = F(2) = 1 - (1+2)e^{-2} = 1 - 3e^{-2} \approx 0.594 \tag{12}$$

[1] Thanks to betterexplained.com/articles/an-intuitive-and-short-explanation-of-bayes-theorem/ for a much clearer explanation than the book.

4. The variance is defined as

$$\sigma^2 \equiv \mathrm{V}[x] = \mathrm{E}[(x-\mu)^2] = \mathrm{E}[x^2] - (\mathrm{E}[x])^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2 \tag{13}$$

Plugging in $f(x)$ and μ and integrating,

$$\sigma^2 = [e^{-x}(x^3 + 3x^2 + 6x + 6)]_{\infty}^{0} - 2^2 = 6 - 4 = 2 \tag{14}$$

The fraction of the total probability contained within $\mu \pm \sigma$ is the difference of the CDF evaluated at the endpoints:

$$F(\mu+\sigma) - F(\mu-\sigma) = (1+\mu-\sigma)e^{-(\mu-\sigma)} - (1+\mu+\sigma)e^{-(\mu+\sigma)} = 0.738 \tag{15}$$

This is slightly greater than for a Gaussian distribution, where $F(\mu+\sigma) - F(\mu-\sigma) = 0.683$.