XX50215 Statistics for Data Science

Problems 4 - Solutions

1. Given a random variable X with the general discrete uniform distribution (N_0, N_1), that puts equal probability on each of the values N_0, N_0+1,…N_1 where $N_0 < N_1$ and both are integers. Find expressions for EX and Var X.

The pmf of X is $f(x) = \frac{1}{N_1 - N_0 + 1}$, $x = N_0, N_0 + 1, \ldots, N_1$. Then

$$\begin{aligned} EX &= \sum_{x=N_0}^{N_1} x \frac{1}{N_1 - N_0 + 1} = \frac{1}{N_1 - N_0 + 1} \left(\sum_{x=1}^{N_1} x - \sum_{x=1}^{N_0 - 1} x \right) \\ &= \frac{1}{N_1 - N_0 + 1} \left(\frac{N_1(N_1+1)}{2} - \frac{(N_0 - 1)(N_0 - 1 + 1)}{2} \right) \\ &= \frac{N_1 + N_0}{2}. \end{aligned}$$

Similarly, using the formula for $\sum_1^N x^2$, we obtain

$$\begin{aligned} Ex^2 &= \frac{1}{N_1 - N_0 + 1} \left(\frac{N_1(N_1+1)(2N_1+1) - N_0(N_0-1)(2N_0-1)}{6} \right) \\ \mathrm{Var}X &= EX^2 - EX = \frac{(N_1 - N_0)(N_1 - N_0 + 2)}{12}. \end{aligned}$$

2. A manufacturer receives a lot of 100 parts from a vendor. The lot will be unacceptable if more than five parts are defective. The manufacturer decides to select a subset of the parts randomly from the lot for inspection. The lot will be accepted if no defective parts are found in the subset. What is the smallest number of parts that should be selected to ensure that the probability of accepting an unacceptable lot is less than 0.1?

Let X = number of defective parts in the sample. Then $X \sim$ hypergeometric$(N = 100, M, K)$ where M = number of defectives in the lot and K = sample size.

If there are 6 or more defectives in the lot, then the probability that the lot is accepted $(X = 0)$ is at most

$$P(X = 0 \mid M = 100, N = 6, K) = \frac{\binom{6}{0}\binom{94}{K}}{\binom{100}{K}} = \frac{(100 - K) \cdot \cdots \cdot (100 - K - 5)}{100 \cdot \cdots \cdot 95}.$$

By trial and error we find $P(X = 0) = .10056$ for $K = 31$ and $P(X = 0) = .09182$ for $K = 32$. So the sample size must be at least 32.

3. Assuming the number of chocolate chips in a cookie have a Poisson distribution, if we want the probability that a randomly chosen cookie has at least two chocolate chips to be greater than 0.99, what is the smallest value of the mean of the distribution that ensures this probability?

Let $X \sim \text{Poisson}(\lambda)$. We want $P(X \geq 2) \geq .99$, that is,

$$P(X \leq 1) = e^{-\lambda} + \lambda e^{-\lambda} \leq .01.$$

Solving $e^{-\lambda} + \lambda e^{-\lambda} = .01$ by trial and error (numerical bisection method) yields $\lambda = 6.6384$.

4. Show that the Poisson family is an exponential family.

The probability mass function (i.e., the density respect to counting measure) of a Poisson random variable is given as follows:

$$p(x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}.$$

Rewriting this expression we obtain:

$$p(x \mid \lambda) = \frac{1}{x!} \exp\{x \log \lambda - \lambda\}.$$

Thus the Poisson distribution is an exponential family distribution, with:

$$\begin{aligned} \eta &= \log \lambda \\ T(x) &= x \\ A(\eta) &= \lambda = e^{\eta} \\ h(x) &= \frac{1}{x!}. \end{aligned}$$

Moreover, we can obviously invert the relationship between η and λ:

$$\lambda = e^{\eta}.$$

5. Let *f(x)* be any pdf with mean μ and variance σ^2. Show how to create a location-scale family based on f(x) such that the standard pdf of the family say f**(x)* has mean 0 and variance 1.

Let $X \sim f(x)$ have mean μ and variance σ^2. Let $Z = \frac{X-\mu}{\sigma}$. Then

$$\text{E}Z = \left(\frac{1}{\sigma}\right) \text{E}(X - \mu) = 0$$

and

$$\text{Var}Z = \text{Var}\left(\frac{X-\mu}{\sigma}\right) = \left(\frac{1}{\sigma^2}\right) \text{Var}(X - \mu) = \left(\frac{1}{\sigma^2}\right) \text{Var}X = \frac{\sigma^2}{\sigma^2} = 1.$$

Then compute the pdf of Z, $f_Z(z) = f_x(\sigma z + \mu) \cdot \sigma = \sigma f_x(\sigma z + \mu)$ and use $f_Z(z)$ as the standard pdf.

6. Using the identities given on Lecture6/Slide14 calculate the variance of a binomial random variable.

Using the form [1] from the lecture slides.

$h(x) = \binom{n}{x} I(x = 0, 1, \cdots)$, $c(p) = (1-p)^n I(0 < p < 1)$, $w_1(p) = \ln \frac{p}{1-p}$, and $t_1(x) = x$.

When differentiating $w_1(p)$ and $\ln c(p)$ with respect to p,

$$\frac{d}{dp} w_1(p) = \frac{1}{p(1-p)} \quad \Rightarrow \quad \frac{d^2}{dp^2} w_1(p) = \frac{2p-1}{p^2(1-p)^2}$$
$$\frac{d}{dp} \ln c(p) = -\frac{n}{1-p} \quad \Rightarrow \quad \frac{d^2}{dp^2} \ln c(p) = -\frac{n}{(1-p)^2}$$

Therefore,

$$\mathrm{E}\left(\frac{1}{p(1-p)} X\right) = \frac{n}{1-p} \quad \Leftrightarrow \quad \mathrm{E}(X) = np$$
$$\mathrm{Var}\left(\frac{1}{p(1-p)} X\right) = \frac{n}{(1-p)^2} - \mathrm{E}\left(\frac{2p-1}{p^2(1-p)^2} X\right) \quad \Leftrightarrow \quad \mathrm{Var}(X) = np(1-p)$$