Introduction to Probability and Statistics

Review of the most common distributions

March 19, 2009

Here is a list of some of the most common probability distributions encountered in statistics, with their basic properties and uses. The first three in each category (discrete, continuous) are the most important.

1 Discrete random variables

1. **Bernoulli**

   (a) **Parameter**: $0 \leq p \leq 1$

   (b) **Mass function**: $P(X = 1) = p$, $P(X = 0) = 1 - p$

   (c) **Mean and variance**: p and $p(1 - p)$

   (d) **Use**: Models the outcome of a single trial, success probability p

2. **Binomial**

   (a) **Parameters**: Positive integer n, $0 \leq p \leq 1$

   (b) **Mass function**: $P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$, $k = 0, 1, \ldots, n$.

   (c) **Mean and variance**: np and $np(1 - p)$

   (d) **Use**: Counts the number of successes when n independent repetitions of a trial are performed, each with success probability p

3. **Poisson**

   (a) **Parameter**: $\lambda > 0$

   (b) **Mass function**: $P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$, $k = 0, 1, 2, \ldots$

   (c) **Mean and variance**: Both λ

   (d) **Uses**:

i. Approximates the binomial when number of trials n is large and probability of success p is small, $\lambda = np$

ii. Models the number of times a rare event occurs in a time period, when λ is the average number of occurrences

4. **Hypergeometric**

   (a) **Parameters**: N, M, n

   (b) **Mass function**: $P(X = k) = \frac{\binom{N}{k}\binom{M}{n-k}}{\binom{N+M}{n}}$, $k = 0, 1, \ldots \min\{N, n\}$

   (c) **Mean** : $\frac{nN}{N+M}$

   (d) **Use**: Models the number of desirable elements obtained, if n items are drawn from a pool containing N desirable and M undesirable elements

5. **Geometric**

   (a) **Parameter**: p

   (b) **Mass function**: $P(X = k) = p(1-p)^{k-1}$, $k = 1, 2, \ldots$

   (c) **Mean** : $1/p$

   (d) **Use**: Models the number of repetitions of a trial until first success, when success probability is p

   (e) **Remark**: It is the only memoryless continuous distribution: $P(X > s + t | X > s) = P(X > t)$ for all $s, t \geq 0$

2 Continuous random variables

1. **Uniform**

   (a) **Parameters**: $\alpha < \beta$

   (b) **Density function**: $f(x) = \frac{1}{\beta - \alpha}$, $\alpha \leq x \leq \beta$

   (c) **Mean and variance**: $\frac{\alpha+\beta}{2}$ and $\frac{(\beta-\alpha)^2}{12}$

   (d) **Use**: Models the selection of a random number between α and β

2. **Exponential**

   (a) **Parameter**: $\lambda > 0$

   (b) **Density function**: $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$

   (c) **Mean and variance**: $\frac{1}{\lambda}$ and $\frac{1}{\lambda^2}$

(d) **Use**: Models the time spent waiting until an event first occurs, $1/\lambda$ the average waiting time (equivalently λ is the averag number of occurrences per unit time)

(e) **Remark**: It is the only memoryless continuous distribution: $P(X > s + t | X > s) = P(X > t)$ for all $s, t \geq 0$

3. **Normal**

   (a) **Parameters**: μ, σ^2

   (b) **Density function**: $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

   (c) **Mean and variance**: μ and σ^2 (**Standard Normal** if $\mu = 0$, $\sigma^2 = 1$)

   (d) **Uses**:

      i. Models the distribution of many observable physical quantities

      ii. Models the error made when measuring device makes measurement

      iii. Approximates the limit of the sum of independent random variables with the same mean and variance

   (e) **Remark**: If X is normal with parameters μ and σ^2 and Z is the standard normal, then $X = \sigma Z + \mu$

4. **Chi-squared**

   (a) **Parameter**: n

   (b) **Description**: $X = Z_1^2 + Z_2^2 + \ldots + Z_n^2$ where the Z_i's are independent standard normals

   (c) **Uses**:

      i. Models distance of a point in n-dimensional space from origin if coordinates are independent standard normals

      ii. Suitably scaled, is the sample variance of a sample drawn from a normal population