Data Modeling & Analysis Techniques

Probability Distributions

Continuous Distributions

- Uniform Distribution
- Normal Distribution
- Exponential Distribution

Continuous Distributions- Uniform

- Continuous distributions for event probability
  - Uniform distribution
    - Models the likelihood that a particular outcome will result from an experiment where every outcome value is equally likely
    - Parameterized by the range of possible outcomes, [a..b]
    - Probability density function:

$$p(x;a,b) = \frac{1}{b-a}$$

    - Mean: $\mu=(a+b)/2$ Variance: $\sigma^2=(b-a)^2/12$

Problem

- The pdf of a Uniform Distribution of X is $f(x)= 5;\ 4.9 =< x <= 5.1$
- What is the probability that a measurement of current is between 4.95 and 5.0 milliamperes.

Distributions

- Normal distribution
  - Models the likelihood of results if the results are either distributed with a "Bell curve" or, alternatively, the result of the summation of a large number of random effects. This is a good approximation for a wide range of natural processes or noise phenomena as we will see a little later
  - Parameterized by a mean, μ, and standard deviation σ
  - Probability density function:

$$p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

  - Mean: μ Variance: σ^2

Distributions

- Continuous distributions for event frequency
  - Normal distribution
    - Models the number of times an event happens in a very large (infinite) number of experiments
    - Parameterized by a mean, μ, and standard deviation σ
    - Probability density function:

$$p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Normal Distribution Example

Distributions

- Continuous distributions for inter-event timing
  - Exponential distribution
    - Models the likelihood of an event happening for the first time at time *x* in a Poisson process (i.e. a process where events occur with the same likelihood at any point in time, independent of the time since the last occurrence.
    - Parameterized by event rate, *λ*
    - Probability density function:

$$p(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & if\ x \geq 0 \\ 0 & otherwise \end{cases}$$

    - Mean: *1/μ* Variance: $1/\lambda^2$

Exponential Distribution Example

Commonly, car cooling systems are controlled by electrically driven fans. Assuming that the lifetime T in hours of a particular make of fan can be modelled by an exponential distribution with $\lambda = 0.0003$ find the proportion of fans which will give at least 10000 hours service. If the fan is redesigned so that its lifetime may be modelled by an exponential distribution with $\lambda = 0.00035$, would you expect more fans or fewer to give at least 10000 hours service?

Exponential Distribution Contd..

Answer

We know that $f(t) = 0.0003\mathrm{e}^{-0.0003t}$ so that the probability that a fan will give at least 10000 hours service is given by the expression

$$\mathrm{P}(T > 10000) = \int_{10000}^{\infty} f(t)\,dt = \int_{10000}^{\infty} 0.0003\mathrm{e}^{-0.0003t}\,dt = -\left[\mathrm{e}^{-0.0003t}\right]_{10000}^{\infty} = e^{-3} \approx 0.0498$$

Hence about 5% of the fans may be expected to give at least 10000 hours service. After the redesign, the calculation becomes

$$\mathrm{P}(T > 10000) = \int_{10000}^{\infty} f(t)\,dt = \int_{10000}^{\infty} 0.00035\mathrm{e}^{-0.00035t}\,dt = -\left[\mathrm{e}^{-0.00035t}\right]_{10000}^{\infty} = e^{-3.5} \approx 0.0302$$

and so only about 3% of the fans may be expected to give at least 10000 hours service.

Hence, after the redesign we expect *fewer* fans to give 10000 hours service.

Exponential Distribution Example

- The print jobs at UTA network follows an exponential distribution with $\lambda = 0.2$ per minute.
- What is the probability you will have print jobs arrive
  - Less than 3 min
  - More than 7 min
  - Bet 3 and 7 min

Moments

- Moments represent important aspects of the distribution and can be used to characterize mean, variance, etc.

$$E\left[(x-a)^r\right]$$

- In some cases the standard definition is difficult to compute
  - Moment generating function can sometimes help

Moment Generating Function

- The moment generating function for a random variable *X* is defined as

$$m_X(t) = E\left[e^{xt}\right]$$

  - The r^th^ moment of *X* around 0 can then be computed as:

$$\lim_{t \to 0} \frac{\partial^r}{\partial t^r} m_X(t)$$

    - Note that sometimes this can not be computed since the limit might not be defined

Moment Generating Function

- The moment generating function allows to compute, e.g., the mean and the variance
  - Mean:

$$m = \lim_{t \to 0} \frac{\partial}{\partial t} \int e^{xt} p(x)\, dx$$

  - Variance:

$$s^2 = \lim_{t \to 0} \frac{\partial^2}{\partial t^2} \int e^{(x-m)t} p(x)\, dx$$

Example: Poisson Distribution

- Probability mass function

$$P(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

- Moment generating function

$$m_X(t) = E[e^{xt}] = e^{\lambda\left(e^t - 1\right)}$$

- Mean

$$\mu = \lim_{t\to 0} \frac{\partial}{\partial t} e^{\lambda\left(e^t - 1\right)} = \lambda$$

- Variance

$$\sigma^2 = \lim_{t\to 0} \frac{\partial^2}{\partial t^2} e^{\lambda\left(e^{t-\mu} - 1\right)} = \lambda$$

Multivariate Distributions

- Multivariate distributions sometimes arise when combining the outcomes of multiple random variables
  - Sometimes we are interested of the joint effect of multiple random variables
    - Distribution of the product of two random variables
    - Distribution of the joint additive effect of multiple variables

Multivariate Distributions

- For some operations combining multiple variables we can determine the moments of the distribution relatively easily
  - Usually assumptions made about random variables
    - Independently distributed
    - Moments of the distributions of the individual variables are known
  - If variables are not independent we have to use conditional distributions and the laws of probability

Distribution of the Product

- The mean and variance of the distribution of the product of two independent random variables can be determined

$$\mu_{XY} = \sum_i \sum_j \left(x_i y_j P(x_i) P(y_j) \right) = \sum_i \left(x_i P(x_i) \sum_j \left(y_j P(y_j) \right) \right)$$

$$= \sum_i \left(x_i P(x_i) \mu_Y \right) = \mu_Y \sum_i \left(x_i P(x_i) \right) = \mu_X \mu_Y$$

Distribution of the Product

$$\sigma_{XY}^2 = \sum_i \sum_j \left(\left(x_i y_j - \mu_X \mu_Y \right)^2 P(x_i) P(y_j) \right) = \sum_i \left(P(x_i) \sum_j \left(\left(\left((x_i - \mu_X) + \mu_X \right)\left((y_j - \mu_Y) + \mu_Y \right) - \mu_X \mu_Y \right)^2 P(y_j) \right) \right)$$

$$= \sum_i \left(P(x_i) \sum_j \left(\left(\left((x_i - \mu_X)(y_j - \mu_Y) + (x_i - \mu_X)\mu_Y + (y_j - \mu_Y)\mu_X + \mu_X \mu_Y \right) - \mu_X \mu_Y \right)^2 P(y_j) \right) \right)$$

$$= \sum_i \left(P(x_i) \sum_j \left(\left((x_i - \mu_X)(y_j - \mu_Y) + (x_i - \mu_X)\mu_Y + (y_j - \mu_Y)\mu_X \right)^2 P(y_j) \right) \right)$$

$$= \sum_i \left(P(x_i) \sum_j \left(\left(\begin{array}{l} (x_i - \mu_X)^2 (y_j - \mu_Y)^2 + (x_i - \mu_X)^2 (y_j - \mu_Y)\mu_Y + (x_i - \mu_X)(y_j - \mu_Y)^2 \mu_X + (x_i - \mu_X)(y_j - \mu_Y)\mu_X \mu_Y \\ + (x_i - \mu_X)^2 \mu_Y^2 + (y_j - \mu_Y)^2 \mu_X^2 \end{array} \right) P(y_j) \right) \right)$$

$$= \sum_i \left(P(x_i) \left(\begin{array}{l} \sum_j \left((x_i - \mu_X)^2 (y_j - \mu_Y)^2 P(y_j) \right) + \sum_j \left((x_i - \mu_X)^2 (y_j - \mu_Y)\mu_Y P(y_j) \right) + \sum_j \left((x_i - \mu_X)(y_j - \mu_Y)^2 \mu_X P(y_j) \right) \\ + \sum_j \left((x_i - \mu_X)(y_j - \mu_Y)\mu_X \mu_Y P(y_j) \right) + \sum_j \left((x_i - \mu_X)^2 \mu_Y^2 P(y_j) \right) + \sum_j \left((y_j - \mu_Y)^2 \mu_X^2 P(y_j) \right) \end{array} \right) \right)$$

$$= \sum_i \left(P(x_i) \left(\begin{array}{l} (x_i - \mu_X)^2 \sum_j \left((y_j - \mu_Y)^2 P(y_j) \right) + (x_i - \mu_X)^2 \mu_Y \sum_j \left((y_j - \mu_Y) P(y_j) \right) + (x_i - \mu_X)\mu_X \sum_j \left((y_j - \mu_Y)^2 P(y_j) \right) \\ + (x_i - \mu_X)\mu_X \mu_Y \sum_j \left((y_j - \mu_Y) P(y_j) \right) + (x_i - \mu_X)^2 \mu_Y^2 \sum_j \left(P(y_j) \right) + \mu_X^2 \sum_j \left((y_j - \mu_Y)^2 P(y_j) \right) \end{array} \right) \right)$$

$$= \sum_i \left(P(x_i) \left((x_i - \mu_X)^2 \sigma_Y^2 + (x_i - \mu_X)^2 \mu_Y (\mu_Y - \mu_Y) + (x_i - \mu_X)\mu_X \sigma_Y^2 + (x_i - \mu_X)\mu_X \mu_Y (\mu_Y - \mu_Y) + (x_i - \mu_X)^2 \mu_Y^2 + \mu_X^2 \sigma_Y^2 \right) \right)$$

$$= \sum_i \left(P(x_i) \left((x_i - \mu_X)^2 \sigma_Y^2 + (x_i - \mu_X)\mu_X \sigma_Y^2 + (x_i - \mu_X)^2 \mu_Y^2 + \mu_X^2 \sigma_Y^2 \right) \right) = \sum_i \left(P(x_i) \left(\sigma_Y^2 \left((x_i - \mu_X)^2 + (x_i - \mu_X)\mu_X + \mu_X^2 \right) + (x_i - \mu_X)^2 \mu_Y^2 \right) \right)$$

$$= \sigma_Y^2 \sum_i (x_i - \mu_X)^2 P(x_i) + \sigma_Y^2 \mu_X \sum_i (x_i - \mu_X) P(x_i) + \sigma_Y^2 \mu_X^2 \sum_i P(x_i) + \mu_Y^2 \sum_i (x_i - \mu_X)^2 P(x_i)$$

$$= \sigma_Y^2 \sigma_X^2 + \sigma_Y^2 \mu_X (\mu_X - \mu_X) + \sigma_Y^2 \mu_X^2 + \mu_Y^2 \sigma_X^2 = \sigma_Y^2 \sigma_X^2 + \sigma_Y^2 \mu_X^2 + \mu_Y^2 \sigma_X^2$$

Distribution of the Sum

- The mean and variance of the distribution of the sum of two independent random variables can be determined

$$\mu_{X+Y} = \sum_i \sum_j \left((x_i + y_j)P(x_i)P(y_j)\right) = \sum_i P(x_i) \sum_j \left(x_i P(y_j) + y_j P(y_j)\right)$$

$$= \sum_i P(x_i)\left(x_i \sum_j P(y_j) + \sum_j \left(y_j P(y_j)\right)\right) = \sum_i P(x_i)\left(x_i + \mu_Y\right)$$

$$= \sum_i P(x_i)x_i + \mu_Y \sum_i P(x_i) = \mu_X + \mu_Y$$

Distribution of the Sum

$$\sigma^2_{X+Y} = \sum_i \sum_j \left(\left((x_i + y_j) - (\mu_X + \mu_Y) \right)^2 P(x_i) P(y_j) \right) = \sum_i \left(P(x_i) \sum_j \left(\left((x_i + y_j)^2 - 2(x_i + y_j)(\mu_X + \mu_Y) + (\mu_X + \mu_Y)^2 \right) P(y_j) \right) \right)$$

$$= \sum_i \left(P(x_i) \sum_j \left(\left((x_i^2 + 2x_i y_j + y_j^2) - 2(x_i \mu_X + y_j \mu_X + x_i \mu_Y + y_j \mu_Y) + (\mu_X^2 + 2\mu_Y \mu_X + \mu_Y^2) \right) P(y_j) \right) \right)$$

$$= \sum_i \left(P(x_i) \sum_j \left(\left((x_i^2 - 2x_i \mu_X + \mu_X^2) + (y_j^2 - 2y_j \mu_Y + \mu_Y^2) + 2x_i y_j - 2(y_j \mu_X + x_i \mu_Y) + 2\mu_Y \mu_X \right) P(y_j) \right) \right)$$

$$= \sum_i P(x_i) \left((x_i - \mu_X)^2 \sum_j P(y_j) + \sum_j (y_j - \mu_Y)^2 P(y_j) + 2x_i \sum_j y_j P(y_j) - 2\mu_X \sum_j y_j P(y_j) - 2x_i \mu_Y \sum_j P(y_j) + 2\mu_Y \mu_X \sum_j P(y_j) \right)$$

$$= \sum_i P(x_i) \left((x_i - \mu_X)^2 + \sigma_Y^2 + 2x_i \mu_Y - 2\mu_X \mu_Y - 2x_i \mu_Y + 2\mu_Y \mu_X \right)$$

$$= \sum_i P(x_i) \left((x_i - \mu_X)^2 + \sigma_Y^2 \right) = \sum_i (x_i - \mu_X)^2 P(x_i) + \sigma_Y^2 \sum_i P(x_i) = \sigma_X^2 + \sigma_Y^2$$

Pareto Distribution

- The Pareto distribution has two parameters, a shape parameter α and a minimum x_m
  - Models many social and physical phenomena
    - Wealth distribution (80-20 rule), hard drive failures, daily maximum rainfalls, size of fires, etc.
  - Probability density

$$p(x; a, x_m) = \begin{cases} \dfrac{a x_m^a}{x^{(a+1)}} & x \geq x_m \\ 0 & otherwise \end{cases}$$

  - Cumulative density function

$$P(y < x) = \begin{cases} 1 - \left(\dfrac{x_m}{x}\right)^a & x \geq x_m \\ 0 & otherwise \end{cases}$$

Pareto Distribution

- The Pareto distribution is heavy tailed for some parameter settings
  - Infinite mean for $\alpha \leq 1$
  - Infinite variance for $\alpha \leq 2$
- For many interesting problems the parameters fall into this region
  - E.g. 80-20 rule has $\alpha \approx 1.161$
- Heavy tailed distributions exist and model existing problems
  - Has implications on sums and products of functions and the central limit theorem