Data Modeling & Analysis Techniques

Probability Distributions

Experiment and Sample Space

- A (random) experiment is a procedure that has a number of possible outcomes and it is not certain which one will occur
- The sample space is the set of all possible outcomes of an experiment (often denoted by S).
  - Examples:
    - Coin : S={H, T}
    - Two coins: S={HH, HT, TH, TT}
    - Lifetime of a system: S={0..∞}

Probability Distributions

- Probability distributions represent the likelihood of certain events
  - Probability “mass” (or density for continuous variables) represents the amount of likelihood attributed to a particular point
  - Cumulative distribution represents the accumulated probability “mass” at a particular point
    - Distributions in probability are usually given and their results are computed
    - Distributions (or their parameters) are usually the items to be estimated in statistics

Probability Distributions

- Distributions can be characterized by their moments
  - *r^th^* moment: $E\left[(x-a)^r\right]$
  - Important moments:
    - Mean: $E\left[(x-0)^1\right]$
    - Variance: $E\left[(x-\mu)^2\right]$
    - Skewness: $E\left[(x-\mu)^3\right]$

Distributions

- There are families of important distributions that are useful to model or analyze events
  - Families of distributions are parameterized
  - Different distributions are used to answer different questions about events
    - What is the probability of an individual event
    - How many times would an event happen in a repeated experiment
    - How long will it take until an event happens

Distributions

- Discrete distributions for event probability
  - Uniform distribution
    - Models the likelihood of a set of events assuming they are all equally likely
    - Parameterized by the number of discrete events, *N*
    - Probability function:

$$P(x;N) = P(X = x) = \frac{1}{N}$$

    - If the events are integers in the interval [*a..b*] (with *N=b-a+1*) we can compute a mean and variance
    - Mean: $\mu=(b+a)/2$ Variance: $\sigma^2=(N^2-1)/12$

Distributions

- Bernoulli distribution
  - Models the likelihood of one of two possible events happening
  - Parameterized by the likelihood, *p*, of event *1*
  - Probability function:

$$P(x;p) = P(X=x) = \begin{cases} p & \text{if } x=1 \\ 1-p & \text{otherwise} \end{cases}$$

  - Can be easily extended to represent more than two possible events
  - Mean: $\mu = p$ Variance: $\sigma^2 = p*(1-p)$

Distributions

- Discrete distributions for event frequency
  - Binomial distribution
    - Models the likelihood that an event will occur a certain number of times in *n* Bernoulli experiments
    - Parameterized by the likelihood, *p*, of event *1* in the Bernoulli experiment and the number of experiments, *n*
    - Probability function:

$$P(x;n,p) = \binom{n}{x} p^x (1-p)^{(n-x)}$$

    - Mean: $\mu = np$ Variance: $\sigma^2 = np(1-p)$

Distributions

- Poisson distribution
  - Models the likelihood that an even will occur a given number of times in a continuous experiment with constant likelihood that does not depend on the time since the last occurrence
  - Parameterized by the expected number of occurrences, λ, of the event within one time period
  - Probability function:

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

  - Mean: $E[x]=\mu=\lambda$ Variance: $\sigma^2=\lambda$

Distributions

- Multinomial distribution
  - Models the likelihood that each event, *i*, will occur a certain number of times in *n* independent experiments with l different events
  - Parameterized by the likelihoods, p_i, of the *l* events in the experiment and the number of experiments, *n*
  - Probability function:

$$P(x_1..x_l;n,p_1..p_l) = \frac{n!}{\prod_{i\in[1..l]} x_i!} \prod_{i\in[1..l]} p^{x_i}$$

  - Mean: $\mu_i = np_i$ Variance: $\sigma_i^2 = np_i(1-p_i)$

Distributions

- Hypergeometric distribution
  - Models the likelihood that an event type will occur a certain number of times in *n* experiments if no specific event can occur twice and they are all equally likely
  - Parameterized by the total number of events, *N*, the number of events of the event type, *M*, and the number of experiments, *n*
  - Probability function:

$$P(x;M,N,n)=\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

  - Mean: $\mu=nM/N$ Variance: $\sigma^2=n(M(n-M)(N-n)/(N^2(N-1))$

Distributions

- Discrete distributions for inter-event timing
  - Geometric distribution
    - Models the likelihood that an event will occur for the first time in the x^{th} Bernoulli experiment
    - Parameterized by the probability, *p*, of the event in each Bernoulli experiment
    - Probability function:

$$P(x;p) = \left(1-p\right)^{x-1} p$$

    - Mean: $\mu=1/p$ Variance: $\sigma^2=(1-p)/p^2$

Distributions

- 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: $\sigma^2=(b-a)^2/12$

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: σ2

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:

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

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:

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

    - Mean: $1/\mu$ Variance: $1/\lambda^2$

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:

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

  - Variance:

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

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

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

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

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

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

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

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

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

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

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

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

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

$$
\begin{aligned}
\sigma_{X+Y}^2 &= \sum_i \sum_j \Big(\big((x_i + y_j) - (\mu_X + \mu_Y) \big)^2 P(x_i) P(y_j) \Big) = \sum_i \Big(P(x_i) \sum_j \Big(\big((x_i + y_j)^2 - 2(x_i + y_j)(\mu_X + \mu_Y) + (\mu_X + \mu_Y)^2 \big) P(y_j) \Big) \Big) \\
&= \sum_i \Big(P(x_i) \sum_j \Big(\big((x_i^2 + 2x_i y_j + y_j^2) - 2(x_i \mu_X + y_j \mu_X + x_i \mu_Y + y_j \mu_Y) + (\mu_X^2 + 2\mu_Y \mu_X + \mu_Y^2) \big) P(y_j) \Big) \Big) \\
&= \sum_i \Big(P(x_i) \sum_j \Big(\big((x_i^2 - 2x_i \mu_X + \mu_X^2) + (y_j^2 - 2y_j \mu_Y + \mu_Y^2) + 2x_i y_j - 2(y_j \mu_X + x_i \mu_Y) + 2\mu_Y \mu_X \big) P(y_j) \Big) \Big) \\
&= \sum_i P(x_i) \Big((x_i - \mu_X)^2 \sum_j P(y_j) + \sum_j (y_j - \mu_Y)^2 P(y_j) + 2x_i \sum_j y_j P(y_j) - 2\mu_X \sum_j y_j P(y_j) - 2x_i \mu_Y \sum_j P(y_j) + 2\mu_Y \mu_X \sum_j P(y_j) \Big) \\
&= \sum_i P(x_i) \Big((x_i - \mu_X)^2 + \sigma_Y^2 + 2x_i \mu_Y - 2\mu_X \mu_Y - 2x_i \mu_Y + 2\mu_Y \mu_X \Big) \\
&= \sum_i P(x_i) \Big((x_i - \mu_X)^2 + \sigma_Y^2 \Big) = \sum_i (x_i - \mu_X)^2 P(x_i) + \sigma_Y^2 \sum_i P(x_i) = \sigma_X^2 + \sigma_Y^2
\end{aligned}
$$

Hypergeometric to Binomial

- If the population is large and the number of samples drawn is small, then the Hypergeometric distribution can be approximated by the Binomial distribution.
  - $p=M/N$

Normal to Standard Normal

- We usually denote Normal as: $N(m, \sigma^2)$
- The standard normal as: $N(0,1)=Z$
- If random variable X is normally distributed, i.e., $X= N(m, \sigma^2)$ then $Z=(X-m)/\sigma$

Binomial to Poisson

- Binomial pdf:

$$P(x;n,p) = \binom{n}{x} p^x (1-p)^{(n-x)}$$

- Binomial is hard to calculate for large *n*
- Poisson asks a similar question but in continuous time (no discrete time steps)

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

- If n is lage and p is small, then the binomial can be approximated by a Poisson distribution with rate $\lambda = np$

Binomial to Normal

- We cannot use Poisson to approximate binomial if *p* is not very small (as *np* goes towards infinity).
- However, we can use the Normal distribution: $N(np, np(1-p))$
- Thus we can also approximate the Poisson as $N(\lambda, \lambda)$ for large λ-s

Distributions

- If a function is always positive and converges to 0, $\lim_{x-\infty} f(x) = 0$, can we make it into a pdf ?
- E.g. $f(x) = a/x$ between 1 and ∞ and 0 otherwise.
- No
  - There are functions of this type that do not represent probability distributions

Heavy Tailed Distributions

- How about “quicker” convergence:
  - $f(x)=a/x^2$ between 1 and ∞ and 0 otherwise.
- Can this be made into a pdf
  - Yes
- What is its mean
  - Infinite – the tail is too heavy
  - i.e. there are distributions that do not have numeric mean
- What is its variance
  - Infinite

Lower Polynomial Powers

- How about even "quicker" convergence:
  - $f(x)=a/x^3$ between 1 and ∞ and 0 otherwise.
- Can this be made into a pdf
  - Yes
- What is its mean
  - 2
- What is its variance
  - Infinite
  - i.e. there are distributions that have a numeric mean but do no numeric variance

Lower Polynomial Powers

- How about even "quicker" convergence:
  - $f(x)=a/x^4$ between 1 and ∞ and 0 otherwise.
- Can this be made into a pdf
  - Yes
- Does it have a finite mean
  - Yes
- Does it have a finite variance
  - Yes

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), heard drive failures, daily maximum rainfalls, size of fires, etc.
  - Probability density

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

  - Cumulative density function

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

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