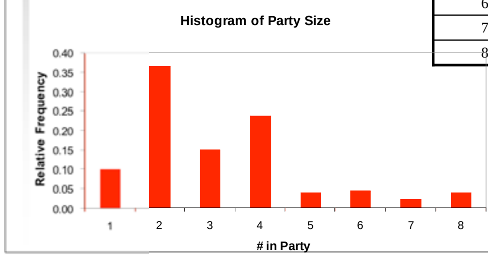
Statistical Tutorial
(Chapter 5)

Compiled and Presented by
Alana Cordick

# Histograms

- Example:

| Arrivals / Party | Frequency | Relative Frequency |
|---|---|---|
| 1 | 30 | 0.10 |
| 2 | 110 | 0.37 |
| 3 | 45 | 0.15 |
| 4 | 71 | 0.24 |
| 5 | 12 | 0.04 |
| 6 | 13 | 0.04 |
| 7 | 7 | 0.02 |
| 8 | 12 | 0.04 |




# Statistical Concepts Part 4

Continuous Distributions

# Continuous Distributions

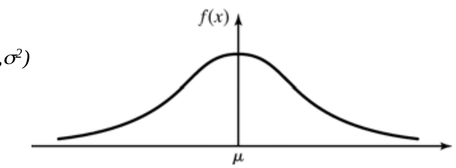
- Continuous random variables can be used to describe random phenomena in which the variable can take on any value in some interval
- In this section, the distributions studied are:
  - Normal
  - Uniform
  - Exponential
  - Gamma
  - Erlang
  - Weibull
  - Triangular
  - Lognormal
  - Beta distribution

# Normal (Gaussian) Distribution

- Continuous
  - AKA Bell Curve
  - Most common continuous distribution found in nature (as seen through the Central Limit Theorem – to come)
  - Distribution:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \quad -\infty \prec x \prec \infty$$

  - Mean: $-\infty \prec \mu \prec \infty$
  - Variance: $\sigma^2 \succ 0$
  - Denoted by: $X \sim N(\mu,\sigma^2)$




# Normal (Gaussian) Distribution

- Evaluating the distribution:
  - Use numerical methods (no closed form)
  - Independent of μ and σ, using the standard normal distribution:

$$Z \sim N(0,1)$$

  - Transformation of variables: let $Z = (X - \mu) / \sigma$,

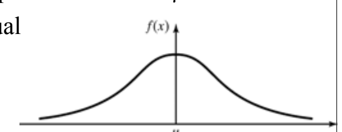
$$F(x) = P(X \le x) = P\left(Z \le \frac{x-\mu}{\sigma}\right)$$

$$= \int_{-\infty}^{(x-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \quad \text{, where } \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

$$= \int_{-\infty}^{(x-\mu)/\sigma} \phi(z) dz = \Phi\left(\tfrac{x-\mu}{\sigma}\right)$$

# Normal (Gaussian) Distribution

- Standard Normal Distribution
  - Has a μ = 0 and σ = 1 i.e. $X \sim N(0, 1)$
- Special properties:
  - $\lim_{x\to-\infty} f(x) = 0$, and $\lim_{x\to\infty} f(x) = 0$
  - $f(\mu-x)=f(\mu+x)$; the pdf is symmetric about $\mu$
  - The maximum value of the pdf occurs at $x = \mu$
  - The mean and mode are equal
  - Linearity




# Normal Distribution

- **Linearity**
  - Any linear combination of normally distributed random variable is normally distributed

If $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ then

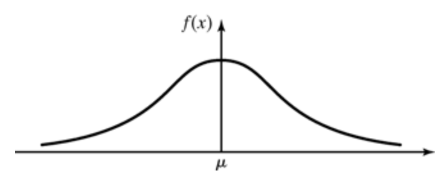
$$X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

$$X - Y \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

**In General**

$$Y = \sum_{i=1}^{n} a_i X_i$$

Normally distributed



## Normal Distribution

- Evaluating the distribution:
  - Use numerical methods (no closed form)
  - Independent of μ and σ, using the standard normal distribution:

    *Z ~ N(0,1)*

  - Transformation of variables: let Z = (X - *μ) / σ*,

$$F(x) = P(X \le x) = P\left(Z \le \frac{x-\mu}{\sigma}\right)$$
$$= \int_{-\infty}^{(x-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$
$$= \int_{-\infty}^{(x-\mu)/\sigma} \phi(z) dz = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

, where $\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$

## Uniform Distribution

- A random variable X is uniformly distributed on the interval (a,b) if its pdf is given by:

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & \text{otherwise} \end{cases}$$

- And a cdf of:

$$F(x) = \begin{cases} 0, & x \prec a \\ \frac{x-a}{b-a}, & a \le x \prec b \\ 1, & x \ge b \end{cases}$$

## Uniform Distribution

- $P(x_1 < X < x_2)$ is proportional to the length of the interval for all $x_1$ and $x_2$ such that:
  - a <= $x_1$ < $x_2$ <=b

$$P(x_1 < X < x_2) = F(x_2) - F(x_1) = \frac{x_2 - x_1}{b-a}$$

## Uniform Distribution

- Mean:

$$E(X) = \frac{a+b}{2}$$

- Variance:

$$V(X) = \frac{(b-a)^2}{12}$$

# Uniform Distribution

- Applications:
  - Random numbers, uniformly distributed between zero and 1, provide the means to generate a **truly** random event
  - Used to generate samples of random variates from all other distributions

# Uniform Distribution

- Example
  - A bus arrives every 20 minutes at a specified stop beginning at 6:40 a.m. and continuing until 8:40 a.m. A certain passenger does not know the schedule, but arrives randomly (uniformly distributed) between 7:00 a.m. and 7:30 a.m. every morning. What is the probability that the passenger wait more than 5 minutes for a bus?

# Uniform Distribution

- Example
  - First concentrate on the bus arrival times at the stop:
    - 6:40
    - 7:00
    - 7:20
    - 7:40
    - …
  - We are only interested in it's arrival between 7:00 and 7:30

# Uniform Distribution

- Example
  - So we know the bus will arrive at 7:00 and 7:20
  - What ranges of time will the passenger not have to wait more than 5 minutes?
    - 7:00
    - 7:15 – 7:20
  - Therefore, we want to find the probability that the passenger arrives during the other times
    - $P(0 < X < 15) + P(20 < X < 30)$

## Uniform Distribution

$$F(x)=\frac{x-a}{b-a}$$

- Example
  - Since X is a uniform random variable (0, 30), we have the following:

$$= F(15) + F(30) - F(20)$$

$$= \frac{15-0}{30-0} + \frac{30-0}{30-0} - \frac{20-0}{30-0}$$

$$= \frac{15}{30} + 1 - \frac{20}{30}$$

$$= \frac{5}{6}$$

*There is an 5/6 = 83% chance the passenger will have to wait more than 5 minutes*

## Exponential Distribution

- A random variable X is said to be exponentially distributed with parameter λ > 0 if its pdf is given by:

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

- With cdf:

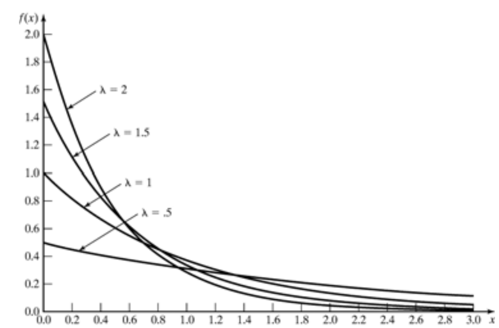
$$F(x)=\begin{cases}0, & x \prec 0\\ \int_0^x \lambda e^{-\lambda t}dt = 1-e^{-\lambda x}, & x \geq 0\end{cases}$$

## Exponential Distribution

- Application
  - Model interarrival times (time between arrivals) when arrivals are completely random
    - λ = arrivals / hour
  - Model service times
    - λ = services / minute
  - Model the lifetime of a component that fails catastrophically (i.e. light bulb)
    - λ = failure rate

## Exponential Distribution

- λ is the value where the pdf intersects the y axis



## Exponential Distribution

- "Memoryless" property
  - For all s>= 0 and t >=0
    - P(X > s + t | X > s) = P(X > t)
  - In other words, if you know a component has survived s hours so far, the same distribution of the remaining amount of time that it survives is the same as the original distribution
    - It does not remember the it already has been used for s amount of time

## Exponential Distribution

- Example:
  - Suppose the life of an industrial lamp is exponentially distributed with failure rate $\lambda$ = 1/3 (one failure every 3000 hours on the avg.)
  - Determine the probability the lamp will last longer than its mean life

$$P(X>3) = 1 - P(X \le 3) = 1 - F(3) = 1 - (1 - e^{-3/3}) = e^{-1} = 0.368$$

$$F(x) = \begin{cases} 0, & x \prec 0 \\ 1 - e^{-\lambda x}, & x \ge 0 \end{cases}$$

## Exponential Distribution

- Example:
  - The probability that an exponential random variable is greater than its mean is always 0.368
    - Regardless of the value of $\lambda$
  - The probability the industrial lamp will last between 2000 and 3000 hours is:

$$P(2 \le X \le 3) = F(3) - F(2) = (1 - e^{-3/3}) - (1 - e^{-2/3})$$
$$= (1 - e^{-1}) - (1 - e^{-2/3}) = 0.632 - 0.487 = 0.145$$

## Exponential Distribution

- Example:
  - The probability that the lamp will last for another 1000 hours given that it is operating after 2500 hours

$$P(X > 3.5 \mid X > 2.5)$$
$$= P(X > 2.5 + 1 \mid X > 2.5)$$
$$= P(X > 1)$$
$$= 1 - P(X \le 1)$$
$$= 1 - (1 - e^{-1/3})$$
$$= e^{-1/3} = 0.717$$

## Gamma Distribution

- A function used in defining the gamma distribution is the gamma function, which is defined for all $\beta > 0$ as:

$$\Gamma(\beta) = (\beta - 1)!$$

- Gamma function is a generalization of the factorial notion to all positive #'s (not just integers)

## Gamma Distribution

- A random variable X is gamma distributed with parameters $\beta$ and $\theta$ if its pdf is given by:

$$f(x) = \begin{cases} \frac{\beta\theta}{\Gamma(\beta)}(\beta\theta x)^{\beta-1} e^{-\beta\theta x}, x > 0 \\ 0, otherwise \end{cases}$$

  - Where $\beta$ is a shape parameter and $\theta$ is a scale parameter

## Gamma Distribution

- Mean: $E(X) = \frac{1}{\theta}$
- Variance: $V(X) = \frac{1}{\beta\theta^2}$
- Cdf of X:

$$F(x) = \begin{cases} 1 - \int_x^\infty \frac{\beta\theta}{\Gamma(\beta)}(\beta\theta t)^{\beta-1} e^{-\beta\theta t} dt, x > 0 \\ 0, x \le 0 \end{cases}$$

## Gamma Distribution

- Gamma and exponential distributions are related when $\beta$ is an integer
  - If the random variable, X, is the sum of $\beta$ independent, *exponentially distributed* random variables, each with parameter $\beta\theta$, then X has a *gamma distribution* with parameters $\beta$ and $\theta$
    - Independent if: $X = X_1 + X_2 + \ldots + X_\beta$

# Weibull Distribution

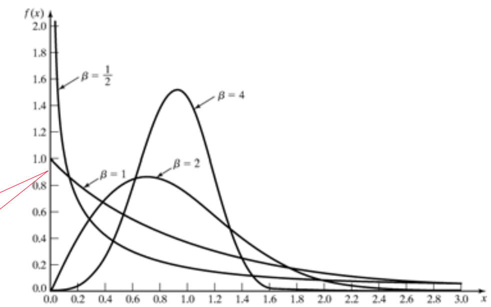
- A random variable $X$ has a Weibull distribution if its pdf has the form

$$f(x)=\begin{cases}\frac{\beta}{\alpha}\left(\frac{x-v}{\alpha}\right)^{\beta-1}\exp\left[-\left(\frac{x-v}{\alpha}\right)^{\beta}\right], & x\geq v\\ 0, & \text{otherwise}\end{cases}$$

- 3 parameters:
  - Location parameter: $\upsilon$,
  - Scale parameter: $\beta$ , $(\beta > 0)$
  - Shape parameter. $\alpha$, $(> 0)$

# Weibull Distribution

- Example: $\upsilon = 0$ and $\alpha = 1$




When $\beta = 1$,
$X \sim exp(\lambda = 1/\alpha)$

# Weibull Distribution

- Mean: $E(X)=v+\alpha\Gamma\left(\frac{1}{\beta}+1\right)$
- Variance: $V(X)=\alpha^2\left[\Gamma\left(\frac{2}{\beta}+1\right)-\left[\Gamma\left(\frac{1}{\beta}+1\right)\right]^2\right]$
- Cdf:

$$F(x)=\begin{cases}0, x<v\\ 1-e^{\left[-\left(\frac{x-v}{\alpha}\right)^{\beta}\right]}, x\geq v\end{cases}$$

# Weibull Distribution

- Applications
  - Can model a wide range of failure rates
  - When multiple components are in a process, this distribution can be used to govern the time of the first failure
    - i.e. ball bearing, capacitator, relay, and strength material failures

*Reference: http://www.itl.nist.gov/div898/handbook/apr/section1/apr162.htm*

## Erlang Distribution

- The pdf from the Gamma distribution is often referred to as the Erlang distribution of order (or number of phases) $k$ when $\beta = k$, an integer

$$f(x) = \begin{cases} \dfrac{\beta\theta}{\Gamma(\beta)}(\beta\theta x)^{\beta-1}e^{-\beta\theta x}, x > 0 \\ 0, otherwise \end{cases}$$

## Erlang Distribution

- Mean: $E(X) = \frac{1}{k\theta} + \frac{1}{k\theta} + \cdots \frac{1}{k\theta} = \frac{1}{\theta}$
- Variance: $V(X) = \frac{1}{(k\theta)^2} + \frac{1}{(k\theta)^2} + \cdots \frac{1}{(k\theta)^2} = \frac{1}{k\theta^2}$
- Cdf:

$$F(x) = \begin{cases} 1 - \sum_{i=0}^{k-1} \dfrac{e^{-k\theta x}(k\theta x)^i}{i!}, x > 0 \\ 0, x \le 0 \end{cases}$$

## Erlang Distribution

- Application
  - When a customer must complete a series of $k$ stations
    - The next customer cannot enter start the first station until the previous customer has completed all $k$ stations

## Erlang Distribution

- Example
  - A college professor of electrical engineering is leaving home for the summer, but would like to have a light burning at all times to discourage burglars. The professor rigs up a device that will hold two light bulbs. The device will switch the current to the second bulb if the first bulb fails. The box in which the light bulbs are packages says, "Average life 1000 hours, exponentially distributed". The professor will be gone 90 days (2160 hours).
    - What is the probability that a light will be burning when the summer is over and the professor returns?

## Erlang Distribution

- Example
  - Reliability function – the probability that the system will operate at least x hours
    - R(*x*) = 1 – F(*x*)
  - $\beta = k = 2$
  - $k\theta = 1 / 1000$, so $\theta = 1/2000$ per hour
  - Determine F(2160):

$$F(2160) = 1 - \sum_{i=0}^{1} \frac{e^{-(2)(1/2000)(2160)}\left[(2)(1/2000)(2160)\right]^i}{i!}$$

$$= 1 - e^{-2.16}\sum_{i=0}^{1}\frac{(2.16)^i}{i!}$$

$$= 0.636$$

## Erlang Distribution

- Example
  - Now:
    - R(x) = 1 – F(x)
    - R(x) = 1 – 0.636 = 0.364
- Therefore, there is a 36% chance that a light will be burning when the professor returns

## Summary

- Normal $f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \quad -\infty \prec x \prec \infty$
- Exponential $f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$

$$F(x) = \begin{cases} 0, & x \prec 0 \\ \int_0^x \lambda e^{-\lambda t}dt = 1 - e^{-\lambda x}, & x \geq 0 \end{cases}$$

- Weibull $f(x) = \begin{cases} \frac{\beta}{\alpha}\left(\frac{x-v}{\alpha}\right)^{\beta-1}\exp\left[-\left(\frac{x-v}{\alpha}\right)^{\beta}\right], & x \geq v \\ 0, & \text{otherwise} \end{cases}$

$$F(x) = \begin{cases} 0, x < v \\ 1 - e^{\left[-\left(\frac{x-v}{\alpha}\right)^{\beta}\right]}, x \geq v \end{cases}$$

## Review

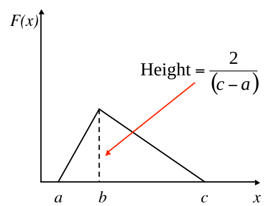
- The time to failure of a nickel-cadmium battery is Weibull distributed with parameters v=0, $\beta$=1/4, and $\alpha$=1/2 years
  - What fraction of batteries are expected to last longer than the mean life?

$$P\left(X > \frac{1}{2}\right) = 1 - P\left(X \leq \frac{1}{2}\right) = 1 - F\left(\frac{1}{2}\right) = 1 - \left[1 - e^{-\left(\frac{1/2}{1/2}\right)^{1/4}}\right] = e^{-1^{1/4}} = e^{-1} = 0.368$$

# Triangular Distribution

- A random variable X has a triangular distribution if its pdf is given by

$$f(x)=\begin{cases}\dfrac{2(x-a)}{(b-a)(c-a)}, a \le x \le b\\ 1-\dfrac{2(c-x)}{(c-b)(c-a)}, b < x \le c\\ 0, \text{otherwise}\end{cases}$$




where a ≤ b ≤ c

# Triangular Distribution

- Mean: $E(X)=\dfrac{a+b+c}{3}$
- Mode: occurs at x = b $\quad Mode = b = 3E(X)-(a+c)$
- Variance – exercise in text
- This distribution uses the mode more than the mean

# Triangular Distribution

- Cumulative Density Function

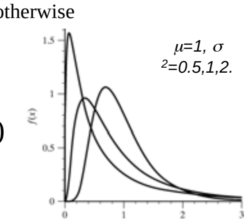
$$F(x)=\begin{cases}0, x \le a\\ \dfrac{(x-a)^2}{(b-a)(c-a)}, a < x \le b\\ 1-\dfrac{(c-x)^2}{(c-b)(c-a)}, b < x \le c\\ 1, \text{x > c}\end{cases}$$

# Lognormal Distribution

- A random variable $X$ has a lognormal distribution if its pdf has the form

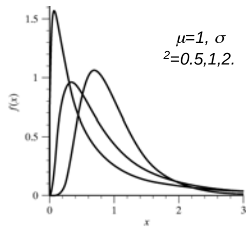
$$f(x)=\begin{cases}\dfrac{1}{\sqrt{2\pi}\sigma x}\exp\left[-\dfrac{(\ln x-\mu)^2}{2\sigma^2}\right], & x > 0\\ 0, & \text{otherwise}\end{cases}$$

- Mean E(X) = $e^{\mu+\sigma^2/2}$
- Variance V(X) = $e^{2\mu+\sigma^2/2}(e^{\sigma^2} - 1)$



# Lognormal Distribution

- Relationship with normal distribution
  - When $Y \sim N(\mu, \sigma^2)$, then $X = e^Y \sim$ lognormal($\mu$, $\sigma^2$)
  - Parameters $\mu$ and $\sigma^2$ are not the mean and variance of the lognormal




# Beta Distribution

- A random variable X is beta-distributed with parameters $\beta_1 > 0$ and $\beta_2 > 0$ if its pdf is given by:

$$f(x) = \begin{cases} \frac{x^{\beta_1 - 1}(1-x)^{\beta_2 - 1}}{B(\beta_1, \beta_2)}, & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

$$B(\beta_1, \beta_2) = \frac{\Gamma(\beta_1)\Gamma(\beta_2)}{\Gamma(\beta_1 + \beta_2)}$$

- Range: (0, 1)

# Beta Distribution

- Range (0, 1) is restrictive
  - May want a range (a, b), where a < b
    - To do this let Y be the random variable such that
      - Y = a + (b – a) X
- Mean (of Y): $E(Y) = a + (b-a)\left(\frac{\beta_1}{\beta_1 + \beta_2}\right)$
- Variance (of Y): $\sigma^2(Y) = (b-a)^2\left(\frac{\beta_1\beta_2}{(\beta_1 + \beta_2)^2(\beta_1 + \beta_2 + 1)}\right)$

# Poisson Distribution

- Definition: *N(t)* is a counting function that represents the number of events occurred in *[0,t]*
- A counting process {*N(t), t>=0*} is a Poisson process with mean rate $\lambda$ if:
  1. Arrivals occur one at a time
  2. {*N(t), t>=0*} has stationary increments (completely random, without rush or slack periods)
  3. {*N(t), t>=0*} has independent increments

# Poisson Distribution

- Properties

$$P[N(t)=n]=\frac{e^{-\lambda t}(\lambda t)^n}{n!}, \qquad \text{for } t \geq 0 \text{ and } n = 0,1,2,...$$

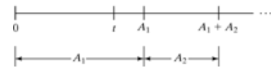
  - Equal mean and variance: $E[N(t)] = V[N(t)] = \lambda t$ *(*$= \alpha$ *as before)*
  - Since stationary increments, the number of arrivals in time $s$ to $t$, such that $s < t$, is also Poisson-distributed with mean $\lambda(t\text{-}s)$

$$P[N(t)-N(s)=n]=\frac{e^{-\lambda(t-s)}[\lambda(t-s)]^n}{n!}, \qquad n = 0,1,2,...$$

$$E[N(t)-N(s)]=\lambda(t-s)=V[N(t)-N(s)]$$

# Interarrival Times

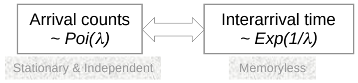
- Consider the interarrival times of a Poisson process ($A_1, A_2, \ldots$), where $A_i$ is the elapsed time between arrival $i$ and arrival $i+1$




  - The 1st arrival occurs after time t iff there are no arrivals in the interval [0,t], hence:

    $P\{A_1 > t\} = P\{N(t) = 0\} = e^{-\lambda t}$

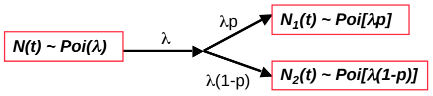
    *Since:* $P\{A_1 <= t\} = 1 - e^{-\lambda t}$ [cdf of exp($\lambda$)]
  - Interarrival times, $A_1, A_2, \ldots$, are **exponentially distributed** and independent with mean $1/\lambda$




# Splitting

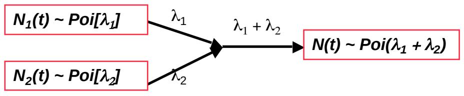
- Suppose each event of a Poisson process can be classified as Type I (high priority arrival), with probability $p$ and Type II (low priority arrival), with probability $1\text{-}p$

- $N(t) = N_1(t) + N_2(t)$, where $N_1(t)$ and $N_2(t)$ are both Poisson processes with rates $\lambda p$ and $\lambda(1\text{-}p)$




# Pooling

- Pooling:
  - Suppose two Poisson processes are pooled together
  - $N_1(t) + N_2(t) = N(t)$, where $N(t)$ is a Poisson processes with rates $\lambda_1 + \lambda_2$



# Nonstationary Poisson Process (NSPP)

- Poisson Process **without** the stationary increments, characterized by $\lambda(t)$, the arrival rate at time $t$
- Useful when arrival rates vary i.e. restaurant
- The expected number of arrivals by time $t$: $\ddot{E}(t) = \int_0^t \ddot{e}(s)ds$
- Relating stationary Poisson process $n(t)$ *with rate* $\lambda=1$ and NSPP $N(t)$ with rate $\lambda(t)$:
  - Let arrival times of a stationary process with rate $\lambda = 1$ be $t_1, t_2, \ldots$, and arrival times of a NSPP with rate $\lambda(t)$ be $T_1, T_2, \ldots$, we know: $t_i = \Lambda(T_i)$ & $T_i = \Lambda^{-1}(t_i)$

# Nonstationary Poisson Process (NSPP)

- Example: Suppose arrivals to a Post Office have rates 2 per hour from 8 am until 12 pm, and then 0.5 per hour until 4 pm
- Let t = 0 correspond to 8 am, NSPP $N(t)$ has rate function:
$$\lambda(t) = \begin{cases} 2, & 0 \le t \prec 4 \\ 0.5, & 4 \le t \prec 8 \end{cases}$$
- Expected number of arrivals by time t:

$$\Lambda(t) = \begin{cases} 2t, & 0 \le t \prec 4 \\ \int_0^4 2ds + \int_4^t 0.5ds = 2*4 + (t*0.5 - 4*0.5) = \frac{t}{2} + 6, & 4 \le t \prec 8 \end{cases}$$

# Nonstationary Poisson Process (NSPP)

- Hence, the probability distribution of the number of arrivals between 11 am and 2 pm.
  - 11 am = 3rd hour
  - 2 pm = 6th hour

$$P[N(6) - N(3) = k]$$
$$= P[N(\Lambda(6)) - N(\Lambda(3)) = k]$$
$$= P[N(9) - N(6) = k]$$
$$= e^{(9-6)}(9-6)^k/k!$$
$$= e^3(3)^k/k!$$

# Empirical Distributions

- A distribution whose parameters are the observed values in a sample of data.
  - May be discrete or continuous
  - May be used when it is impossible or unnecessary to establish that a random variable has any particular parametric distribution.
  - Advantage: no assumption beyond the observed values in the sample.
  - Disadvantage: sample might not cover the entire range of possible values.

# References

Discrete-Event System Simulation, 4th Ed.
Banks, Carson, Nelson, Nicol
Published by Prentice-Hall, 2005