Lecture 04. Continuous Random Variables (Chapter 5)

Ping Yu

HKU Business School
The University of Hong Kong

Plan of This Lecture

- Continuous Random Variables
- Expectations for Continuous Random Variables
- The Normal Distribution
- Normal Distribution Approximation for Binomial Distribution
- The Exponential Distribution
- Jointly Distributed Continuous Random Variables

Continuous Random Variables

Cumulative Distribution Function

- The cumulative distribution function (cdf), $F(x)$, for a continuous r.v. expresses the probability that X does not exceed the value x, as a function of x, i.e.,

$$F(x) = P(X \le x).$$

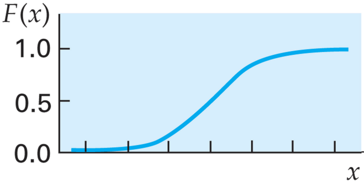
  - This definition is the same as in the discrete r.v. case, but there $F(x)$ is a step function so is not differentiable.
  - This definition of cdf implies $P(a < X \le b) = F(b) - F(a)$; recall that the probability of a single value is zero for a continuous r.v., so whether a and b are included in the interval or not does not affect the result.

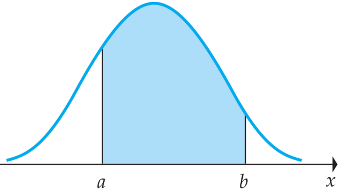
- The counterpart of pmf for a continuous r.v. is the probability density function (pdf), which is defined as

$$f(x) = \frac{d}{dx}F(x).$$ [figure here]

  - Since $F(x)$ is nondecreasing, $f(x) \ge 0$. We denote the area where $f(x) > 0$ as $\mathscr{S}$, called the support of X.[1]
  - $P(a \le X \le b) = \int_a^b f(x)dx$, $\int_{-\infty}^{\infty} f(x)dx = \int_{\mathscr{S}} f(x)dx = 1$ and $F(x) = \int_{-\infty}^{x} f(x)dx = \int_{x_m}^{x} f(x)dx$, where $x_m = \inf(\mathscr{S})$. [figure here]

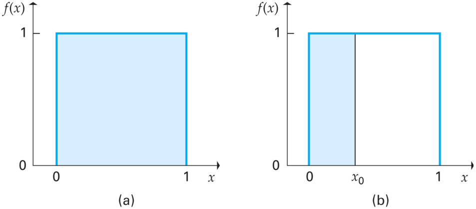
[1](**) Usually, $\mathscr{S}$ is defined as the closure of this area, but we will not distinguish this difference in this lecture.





S-Shaped CDF Implies the Bell-Shaped PDF at Right

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$



$$\int_{\mathscr{S}} f(x)dx = 1 \text{ and } F(x) = \int_{x_m}^{x} f(x)dx$$

Example: Gasoline Sales

- Assume the gasoline sales at a gasoline station is equally likely from 0 to 1,000 gallons during a day; then the gasoline sales follow a uniform (probability) distribution:

$$F(x) = \begin{cases} 0, & \text{if } x < 0, \\ 0.001x, & \text{if } 0 \le x \le 1000 \\ 1, & \text{if } x > 1000, \end{cases}$$

whose pdf is

$$\begin{aligned} f(x) &= \begin{cases} 0.001, & \text{if } 0 \le x \le 1000 \\ 0, & \text{otherwise.} \end{cases} \\ &= 0.001 \cdot 1(0 \le x \le 1000), \end{aligned}$$

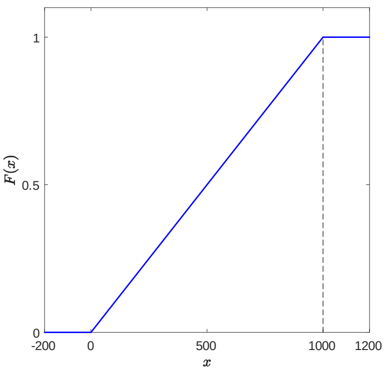
where $1(\cdot)$ is the indicator function which equals 1 when the statement in the parentheses is true and 0 otherwise. [figure here]

- In general, the uniform distribution on (a,b) has the pdf

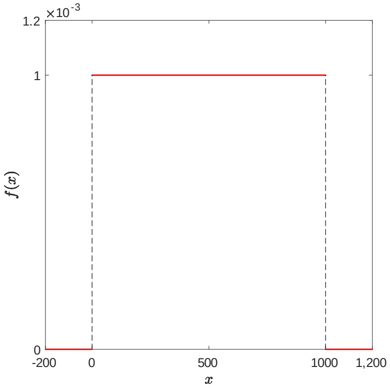
$$f(x) = \frac{1}{b-a} \cdot 1(a \le x \le b),$$

and the cdf

$$F(x) = \frac{x-a}{b-a} \text{ for } x \in [a,b].$$



CDF



PDF

- Degenerate S-shaped cdf and bell-shaped pdf?
- We denote a r.v. X with a uniform distribution on (a, b) as $X \sim U(a, b)$.

Approximate PDF by PMF

- Suppose $\mathscr{S} = (a, b)$, where a can be $-\infty$ and b can be ∞. We can partition $\mathscr{S}$ into small subintervals with length Δ, and then approximate the pdf $f(x)$ by the pmf

$$p\left(a + \left(i + \frac{1}{2}\right)\Delta\right) = P(a + i\Delta < X \leq a + (i+1)\Delta) = \int_{a+i\Delta}^{a+(i+1)\Delta} f(x)\,dx,$$

where $i = 0, 1, \cdots, \frac{b-a}{\Delta} - 1$.[2]

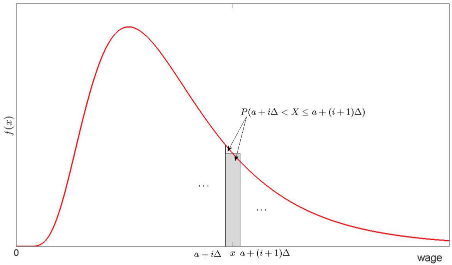


Figure: PDF of Wage: wage $\sim \exp(N(\mu, \sigma^2))$ with $N(\mu, \sigma^2)$ defined below, $a = 0, b = \infty$

[2] $i = 0$, $a + i\Delta = a$, and $i = \frac{b-a}{\Delta} - 1$, $a + (i+1)\Delta = a + \frac{b-a}{\Delta}\Delta = b$.

Expectations for Continuous Random Variables

Mean

- The mean (or expected value, or expectation) of a continuous r.v. can be defined through an approximation of a discrete r.v. in the previous slide:

$$\begin{aligned}\mu_X := E[X] &\approx \sum_{i=0}^{(b-a)/\Delta-1}\left(a+\left(i+\tfrac{1}{2}\right)\Delta\right)P\left(a+i\Delta < X \le a+(i+1)\Delta\right)\\ &\approx \sum_{i=0}^{(b-a)/\Delta-1}\left(a+\left(i+\tfrac{1}{2}\right)\Delta\right)f\left(a+\left(i+\tfrac{1}{2}\right)\Delta\right)\Delta\\ &\xrightarrow{\Delta\to 0} \int_a^b xf(x)dx.\end{aligned}$$

- $\sum \to \int$, $a+\left(i+\frac{1}{2}\right)\Delta \to x$, and $\Delta \to dx$.
- The mean is the center of gravity of a pole (a,b) with density at x being $f(x)$.
- In general, the mean of any function of X, $g(X)$, is

$$E[g(X)] = \int_{\mathscr{S}} g(x)f(x)dx.$$

- Recall that $E[g(X)] \neq g(E[X])$ unless $g(X)$ is linear in X.

Variance

- The variance of X is defined as

$$\sigma_X^2 = E\left[(X-\mu_X)^2\right] = E\left[X^2\right] - \mu_X^2.$$

  - μ_X measures the center of the distribution, while σ_X^2 measures the dispersion or spread of the distribution.

- The standard deviation of X, $\sigma_X = \sqrt{\sigma_X^2}$.
- Example: For the uniform distribution on (a,b),

$$\begin{aligned} \mu_X &= \int_a^b x \frac{1}{b-a} dx = \frac{a+b}{2}, \\ \sigma_X^2 &= \int_a^b x^2 \frac{1}{b-a} dx - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}, \end{aligned}$$

  i.e., the mean is the center of the range, and when the range (a,b) is wider, the variance is larger.

- For $W = a + bX$ with a and b being constant fixed numbers,

$$\mu_W = a + b\mu_X, \sigma_W^2 = b^2\sigma_X^2 \text{ and } \sigma_W = |b|\,\sigma_X.$$

  - For the z-score of X, $Z = \frac{X-\mu_X}{\sigma_X}$, $\mu_Z = 0$ and $\sigma_Z^2 = 1$.

The Normal Distribution

The Normal Distribution

- The pdf for a normally distributed r.v. X is

$$f\left(x|\mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ for } -\infty<x<\infty, \quad (1)$$

where $\mu\in\mathbb{R}$, and $\sigma^2\in(0,\infty)$, e is Euler's number, and $\pi=3.14159\cdots$ is Archimedes' constant (the ratio of a circle's circumference to its diameter).
- Since the normal distribution depends only on μ and σ^2, we denote a r.v. X with pdf (1) as $X\sim N\left(\mu,\sigma^2\right)$.

- The cdf of the normal distribution, $F\left(x|\mu,\sigma^2\right)=\int_{-\infty}^{x}f\left(x|\mu,\sigma^2\right)dx$, does not have an analytic form (i.e., a closed-form expression), but computation of probabilities based on the normal distribution is direct nowadays.
- This distribution has many applications in business and economics, e.g., the dimensions of parts, the heights and weights of human beings, the test scores, the stock prices, etc. all roughly follow normal distributions.
- As will be discussed in Lecture 5, the distribution of sample mean will converge to a normal distribution when the sample size gets large.

History of the Normal Distribution

- Normal Distribution is also called Gaussian Distribution.

Carl F. Gauss (1777-1855), Göttingen[3]

[3]He is also referred to as the "prince of mathematics", known for many things, e.g., the least squares.

Properties of the Normal Distribution: Mean, Variance and Skewness

- For $X \sim N\left(\mu, \sigma^2\right)$,

$$\mu_X = \mu \text{ [easy] and } \sigma_X^2 = \sigma^2 \text{ [proof not required]},$$

so a normal distribution is determined completely by its mean and variance.

- The shape of $f\left(x|\mu, \sigma^2\right)$ is a symmetric bell-shaped curve centered on the mean μ, which implies the skewness of X is 0. [figure here]
  - Recall from Lecture 1 that the (population) skewness is

$$\text{Skew}_X = \frac{E\left[(X - \mu_X)^3\right]}{\sigma_X^3} =: \frac{\mu_3}{\sigma^3},$$

and the sample skewness is

$$\text{skewness} = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{ns^3},$$

where μ_3 is the third central moment, and s is the sample standard deviation.

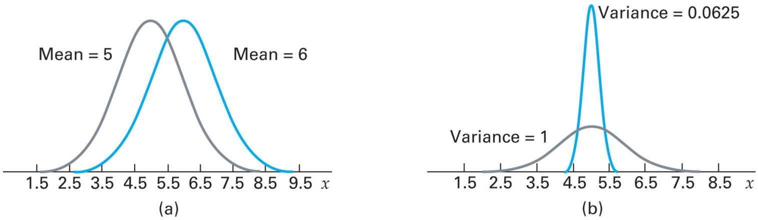


Figure: $f\left(x|\mu,\sigma^2\right)$ with Different μ and σ^2's: (a) same σ^2, different μ's; (b) same $\mu = 5$, different σ^2's.

- The normal distribution is a symmetric, bell-shaped distribution with a single peak. Its peak corresponds to the mean, median, and mode of the distribution.

Properties of the Normal Distribution: Kurtosis

- The tail of $f\left(x|\mu,\sigma^2\right)$ is approximately e^{-x^2} which shrinks to zero very quickly. The (population) kurtosis is often used to measure the heaviness of a distribution's tail [why?]:
$$\text{Kurt}_X = \frac{E\left[\left(X-\mu_X\right)^4\right]}{\sigma_X^4} =: \frac{\mu_4}{\sigma^4},$$
where μ_4 is the fourth central moment.
- It can be shown that the kurtosis of $N\left(\mu,\sigma^2\right)$ is 3, which is chosen as a benchmark, i.e., if a distribution's kurtosis is larger than 3, it is called heavy tailed, and if less than 3, called light tailed. Heavy-tailed phenomena seem more frequent than light-tailed ones since the tail of a normal distribution is already very thin.
- The sample kurtosis is
$$\text{kurtosis} = \frac{\sum_{i=1}^n \left(x_i-\bar{x}\right)^4}{ns^4}.$$

The Standard Normal Distribution

- If $X \sim N(0,1)$, then we call X follows the standard normal distribution.
- Because $Z = \frac{X-\mu}{\sigma}$ has mean 0 and variance 1 for $X \sim N\left(\mu, \sigma^2\right)$, and the normal distribution is completely determined by its mean and variance, we conclude that $Z \sim N(0,1)$ if $X \sim N\left(\mu, \sigma^2\right)$.[4]
- The pdf of $N(0,1)$ is often denoted as $\phi(\cdot)$, and the cdf is denoted as $\Phi(\cdot)$.
- The upper αth quantile of $N(0,1)$, i.e., the solution to $1-\Phi(z)=\alpha$, is denoted as z_α. [The αth quantile of $N(0,1)$ is the solution to $\Phi(z)=\alpha$, i.e., $\Phi^{-1}(\alpha)$]
  - By symmetry of $\Phi(\cdot)$, $1-\Phi(z)=\Phi(-z)$, so $z_\alpha = -\Phi^{-1}(\alpha)$. [figure here]
- The relationship between $F\left(x|\mu,\sigma^2\right)$ and $\Phi(z)$ and between $f\left(x|\mu,\sigma^2\right)$ and $\phi(z)$:

$$
\begin{aligned}
F\left(x|\mu,\sigma^2\right) &= P(X \le x) = P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right), \\
f\left(x|\mu,\sigma^2\right) &= \frac{d}{dx}\Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right),
\end{aligned}
$$

where the second result is from the chain rule and $\Phi'(\cdot) = \phi(\cdot)$.

[4] (**) Linear transformations maintain normality, but nonlinear ones need not.

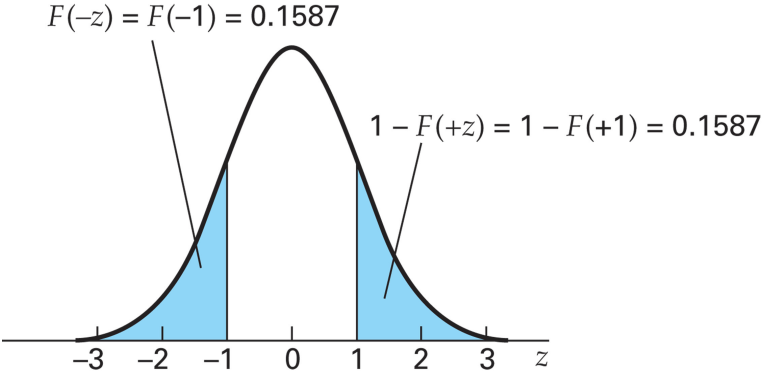


Figure: Normal Density Function with Symmetric Upper and Lower Values

Normal Probability Plot

- Since the normal distribution is most-used, we often need a way to check whether the data in hand are approximately normally distributed.
- Normal probability plots (or QQ-plots with Q for "quantile") provide an easy way to achieve this goal; Lecture 8 will provide a more rigorous test.
- If the data are indeed from a normal distribution, then the plot will be a straight line. [figure here]
  - The vertical axis has a transformed cumulative normal scale.
  - Two dotted lines provide an interval within which data points from a normal distribution would occur most cases.
- (**) Justification for QQ-plots: Suppose we order the data $\{x_i\}_{i=1}^n$ from the smallest to the largest, and denote the order statistics as $\left\{x_{(i)}\right\}_{i=1}^n$. If x_i is from the standard normal distribution, we expect the points $\left\{\left(\frac{i}{n+1}, \Phi\left(x_{(i)}\right)\right)\right\}_{i=1}^n$ in the xy-plane to lie approximately on the line $y = x$. The same must then hold for the points $\left\{\left(\Phi^{-1}\left(\frac{i}{n+1}\right), x_{(i)}\right)\right\}_{i=1}^n$. More generally, if $x_i \sim N\left(\mu, \sigma^2\right)$, then $x_{(i)} \approx \mu + \sigma\Phi^{-1}\left(\frac{i}{n+1}\right)$. So we expect the points meantioned above to lie on the line $y = \mu + \sigma x$.
- If the data are not from a normal distribution, then the plot will deviate from a straight line. [figure here]

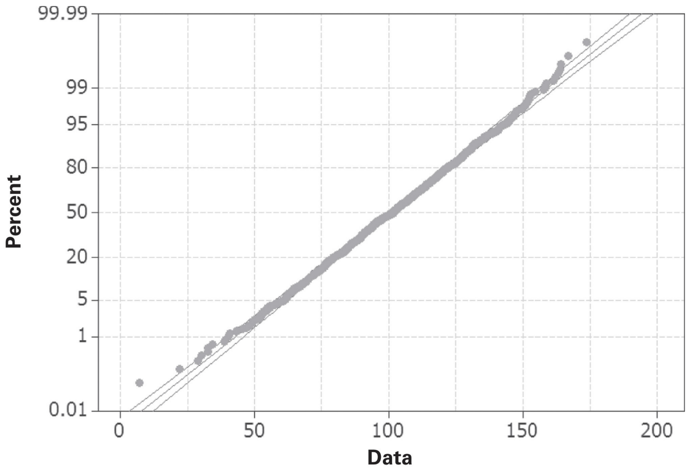


Figure: Normal Probability Plot for Data Simulated from a Normal Distribution

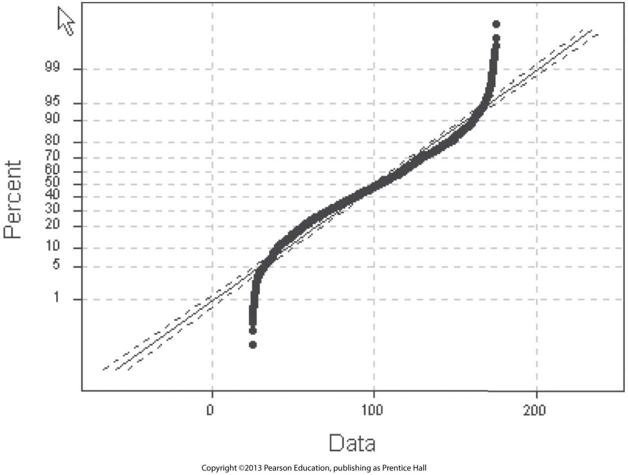


Figure: Normal Probability Plot for Data Simulated from a Uniform Distribution

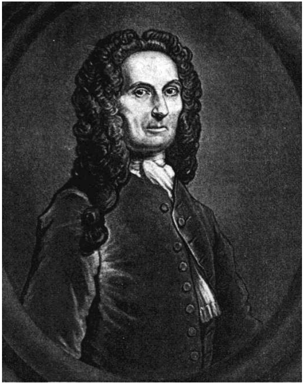
Normal Distribution Approximation for Binomial Distribution

Normal Distribution Approximation for Binomial Distribution

- Recall that a binomial r.v. $X = \sum_{i=1}^{n} X_i$, where $X_i \stackrel{iid}{\sim}$Bernoulli(p) with iid meaning "independent and identically distributed".
- When $\underline{np(1-p) > 5}$, $N(np, np(1-p))$ provides a good approximation of Binomial(n,p), which is known as the De Moivre-Laplace theorem. [figure here]
  - A more rigorous justification when p is fixed and $n \to \infty$ is provided in Lecture 5.
  - Interestingly, we are using a continuous r.v. to approximate a discrete r.v..

Figure: Galton Board (or quincunx, or bean machine): $X = \sum_{i=1}^{n} X_i$, where $X_i \stackrel{iid}{\sim}$Bernoulli(0.5) $\in \{-1, 1\}$

History of the De Moivre-Laplace Theorem



Abraham de Moivre (1667-1754), French[5]

Pierre-Simon Laplace (1749-1827), French[6]

- The phonomenon of de Moivre–Laplace theorem was first observed by de Moivre in a private manuscript circulated in 1733 and published in 1738 with the title "The Doctrine of Chances". Later, Laplace formally proved the theorem in 1810.

[5] He was exiled to England due to religious persecution, and became a friend of Newton there. To make a living, he became a private tutor of mathematics.

[6] He was referred to as the French Newton. His students include Poisson and Napoleon.

continue

- Because normal distributions are easier to handle, this approximation can simplify the analysis of some problems. For example, if X is the number of customers after n people browsed a store's website, and based on past experiences, the probability of visiting the store after browsing is p, the manager wants to predict the probability of the number of customers falling in an interval, say, $[a,b]$.
  - From the normal approximation,

$$P(a \leq X \leq b) = P\left(\frac{a-np}{\sqrt{np(1-p)}} \leq \frac{X-np}{\sqrt{np(1-p)}} \leq \frac{b-np}{\sqrt{np(1-p)}}\right)$$
$$\approx \Phi\left(\frac{b-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a-np}{\sqrt{np(1-p)}}\right).$$

- The normal approximation can also be applied to the proportion (or percentage) r.v., $P = X/n$:

$$P \overset{approx.}{\sim} N\left(\frac{np}{n}, \frac{np(1-p)}{n^2}\right) \approx N\left(P, \frac{P(1-P)}{n}\right),$$

  where the last approximation is from using P to substitute p, i.e., p is estimated rather than known a priori.

Summary of Approximations of Binomial(n, p)

Conditions	Approximating Distributions
n large and $p \le 0.05$ such that $np \le 7$	Poisson(np)
N large and $\frac{n}{N}$ small	Hypergeometric(n, N, S) with $\frac{S}{N} \approx p$
n large such that $np(1-p) > 5$	$N(np, np(1-p))$

The Exponential Distribution

The Exponential Distribution

- The pdf for an exponentially distributed r.v. T is

$$f(t|\lambda) = \lambda e^{-\lambda t} \cdot 1(t > 0),$$

denoted as $T \sim$Exponential(λ), i.e., T can take only positive values, and the distribution is not symmetric – the right tail is heavier than that of the normal distribution, and the left tail is thinner. [figure here]

- The cdf of exponential(λ) is

$$F(t|\lambda) = \left(1 - e^{-\lambda t}\right) \cdot 1(t > 0),$$

which implies that the survivor function is

$$S(t) := P(T > t) = 1 - F(t|\lambda) = e^{-\lambda t}$$

for $t > 0$.

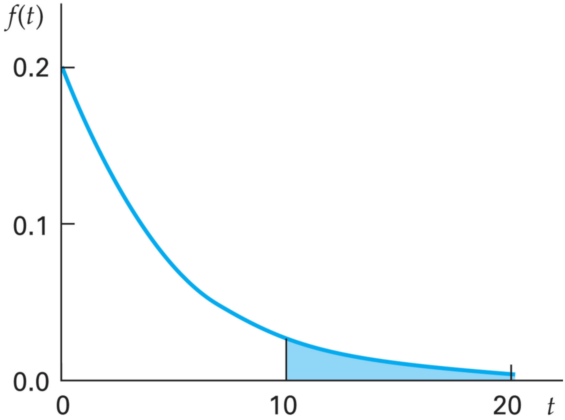


Figure: $f(t|\lambda)$ with $\lambda = 0.2$

Properties of the Exponential Distribution

- $F(t|\lambda)$ can be used to model the waiting time, i.e., the probability that an arrival will occur during an interval of time t (T is the waiting time before the first arrival), so is particularly useful for waiting-line, or queuing, problems.
  - In survival analysis, $S(t)$ is more popular since it can be used to model the survival time, i.e., the probability that a patient can survive for time t.
- The exponential distribution is closely related to the Poisson distribution: If $T_1, \cdots, T_n \overset{iid}{\sim}$ Exponential(λ), then

$$\max\left\{n \middle| \sum_{i=1}^n T_i \leq 1\right\} \sim \text{Poisson}(\lambda); \text{ [proof not required]}$$

  i.e., Poisson(λ) is the number of arrivals in a unit time.
- For $T \sim$Exponential(λ),

$$\mu_T = \frac{1}{\lambda} \text{ and } \sigma_T^2 = \frac{1}{\lambda^2},$$

  so an exponential distribution is determined completely by its mean.
  - $\mu_T = \frac{1}{\lambda}$ is expected from the mean of Poisson(λ) which is λ from Lecture 4.

Constant Harzard Function

- Constant Hazard Function: The harzard function (or hazard rate) is defined as

$$\lambda(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}, t > 0,$$

so for the exponential distribution,

$$\lambda(t) = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda$$

does not depend on t.
- $\lambda(t)$ is the probability of "instantaneous" death given than an item has survived for time t:

$$\lambda(t)\,dt \approx \frac{P(t < T \le t + dt)}{P(T > t)} = \frac{P(t < T \le t + dt \cap T > t)}{P(T > t)} = P(t < T \le t + dt | T > t).$$

- (**) $\lambda(t)$ is defined by $f(t)$ and $F(t)$. Actually, we can also recover $f(t)$ and $F(t)$ from $\lambda(t)$:

$$F(t) = 1 - e^{-\Lambda(t)}, \text{ and } f(t) = \lambda(t)\, e^{-\Lambda(t)},$$

where $\Lambda(t) = \int_0^t \lambda(s)\,ds$ if $F(0) = 0$ is the cumulative hazard function. This is because $\lambda(t) = -\frac{S'(t)}{S(t)} = -\log(S(t))'$.

Memoryless

- The constancy of the hazard function for the exponential distribution leads to the memoryless property.
- Memoryless:

$$\begin{aligned} P(T>s+t|T>s) &= \frac{P(T>s+t\cap T>s)}{P(T>s)} \\ &= \frac{P(T>s+t)}{P(T>s)} \\ &= \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} \\ &= e^{-\lambda t} \\ &= P(T>t), \end{aligned}$$

where $P(T>t)=1-F(t)=e^{-\Lambda(t)}=e^{-\int_0^t \lambda dt}=e^{-\lambda t}$ as already known.
- If T is conditioned on a failure to observe the event over some initial period of time s, the distribution of the remaining waiting time is the same as the original unconditional distribution.

Jointly Distributed Continuous Random Variables

Jointly Distributed Continuous R.V.'s

- This section is parallel to the section on multivariate discrete r.v.'s.
- Let $X_1, \cdots, X_K$ be continuous r.v.'s.

1. Their joint cdf
$$F(x_1, \cdots, x_K) = P(X_1 \le x_1 \cap \cdots \cap X_K \le x_k).$$
2. The cdf, $F(x_1), \cdots, F(x_K)$, of individual r.v.'s are called their marginal distributions.
3. The r.v.'s are independent iff for all $x_1, \cdots, x_K$,
$$F(x_1, \cdots, x_K) = F(x_1) \cdots F(x_K).$$

- The counterparts of the joint and marginal probability distributions for multivariate discrete r.v.'s are the joint pdf
$$f(x_1, \cdots, x_K) = \frac{d^n}{dx_1 \cdots dx_K} F(x_1, \cdots, x_K)$$
and the marginal pdf
$$f(x_i) = \frac{d}{dx_i} F(x_i) = \int \cdots \int f(x_1, \cdots, x_K) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_K.$$
- The independence of $X_1, \cdots, X_K$ can be equivalently defined as $f(x_1, \cdots, x_K) = f(x_1) \cdots f(x_K)$ for all $x_1, \cdots, x_K$.

Conditional Mean and Variance

- For two continuous r.v.'s (X,Y), the conditional pdf of Y given $X=x$ is

$$f(y|x)=\frac{f(x,y)}{f(x)}.$$

- The conditional mean of Y given $X=x$ is

$$\mu_{Y|X=x}=E[Y|X=x]=\int yf(y|x)\,dy.$$

- The conditional variance of Y given $X=x$ is

$$\sigma^2_{Y|X=x}=Var(Y|X=x)=\int\left(y-\mu_{Y|X=x}\right)^2 f(y|x)\,dy.$$

- These concepts can be extended to multivariate continuous r.v.'s in an obvious way.
- The results such as $E[a+bY|X=x]=a+bE[Y|X=x]$ and $Var(a+bY|X=x)=b^2Var(Y|X=x)$ for constants a and b also hold.

Mean and Variance of (Linear) Functions

- For a function of $X_1,\cdots,X_K$, $g(X_1,\cdots,X_K)$, its mean, $E[g(X_1,\cdots,X_K)]$, is defined as
$$E[g(X_1,\cdots,X_K)] = \int\cdots\int g(x_1,\cdots,x_K) f(x_1,\cdots,x_K) dx_1\cdots dx_K.$$
- If $W = \sum_{i=1}^K a_i X_i$, then
$$\mu_W = E[W] = \sum_{i=1}^K a_i \mu_i,$$
and
$$\sigma_W^2 = Var(W) = \sum_{i=1}^K a_i^2 \sigma_i^2 + 2\sum_{i=1}^{K-1}\sum_{j>i}^K a_i a_j \sigma_{ij},$$
which reduces to $\sum_{i=1}^K \sigma_i^2$ if $\sigma_{ij} = 0$ for all $i \neq j$ and $a_i = 1$ for all i.
- Actually, all the results on mean and variance for discrete r.v.'s apply to continuous r.v.'s.
- The covariance and correlation between two continuous r.v.'s (X,Y) are similarly defined as

$$\begin{aligned} Cov(X,Y) &= \sigma_{XY} = E[(X-\mu_X)(Y-\mu_Y)] = E[XY] - \mu_X\mu_Y, \\ Corr(X,Y) &= \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}. \end{aligned}$$