Lecture 4. 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

- Note: only this lecture and Lecture 8 involve some knowledge of calculus.

Continuous Random Variables

[Review] Continuous Random Variables

- A continuous random variable is a random variable that can assume any value in an interval.
  - thickness of an item;
  - time required to complete a task;
  - temperature of a solution;
  - height, in inches;
- These can potentially take on any value, depending only on the ability to measure accurately.

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.,

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

- 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

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$$

for $a < b$.
- 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.

Probability Density Function

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

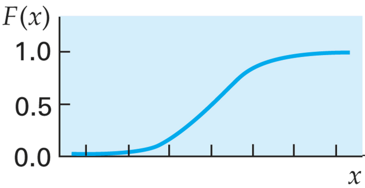
$$f(x) = \frac{d}{dx}F(x). \text{ [figure here][review here]}$$

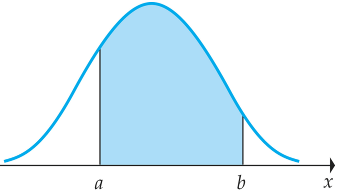
  - Notation: $\frac{d}{dx}F(x)$ is often written as $\frac{dF(x)}{dx}$ or $F'(x)$.

- Properties of PDF:

1. $f(x) \geq 0$ since $F(x)$ is nondecreasing.
   - We denote the area where $f(x) > 0$ as $\mathscr{S}$, called the support of X.[1]
2. $P(a \leq X \leq b) = \int_a^b f(x)dx$. [review here]
3. $\int_{-\infty}^{\infty} f(x)dx = \int_{\mathscr{S}} f(x)dx = 1$. [figure here]
4. $F(x_0) = \int_{-\infty}^{x_0} f(x)dx = \int_{x_m}^{x_0} f(x)dx$, where $x_m = \inf(\mathscr{S})$. [figure here]

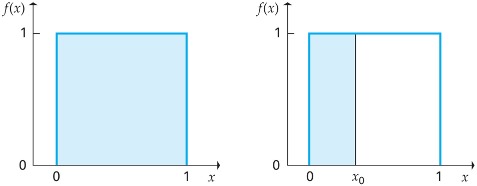
[1] (**) Usually, $\mathscr{S}$ is defined as the closure of this area, but we will not distinguish this difference in this lecture.





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

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



(a)

(b)

$$\int_{\mathscr{S}} f(x)dx = 1 \text{ and } F(x_0) = \int_{x_m}^{x_0} f(x)dx$$

[Review] Derivative and Integral

- Intuitively, $\frac{d}{dx}F(x)$ is the local slope at x, i.e.,

$$\frac{d}{dx}F(x)=\lim_{\Delta\to 0}\frac{F(x+\Delta)-F(x)}{\Delta},$$

  where Δ can be positive or negative.

- Intuitively, $\int_a^b f(x)dx$ is the area under $f(x)$ between a and b, i.e.,

$$\int_a^b f(x)dx=\lim_{\Delta\to 0}\sum_{i=0}^{n} f(x_i)\Delta, \text{ [figure here]}$$

  where we partition (a,b) into small subintervals with length Δ, $x_i=a+\left(i+\frac{1}{2}\right)\Delta$ is the middle point of each subinterval, and $n=(b-a)/\Delta-1$.[2]
  - $\sum\to\int$, $x_i\to x$, $\Delta\to dx$, $i=0\to a$, and $i=n\to b$.

- Fundamental Theorem of Calculus [figure here]: If $\frac{d}{dx}F(x)=f(x)$, then $F(b)-F(a)=\int_a^b f(x)dx$.

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

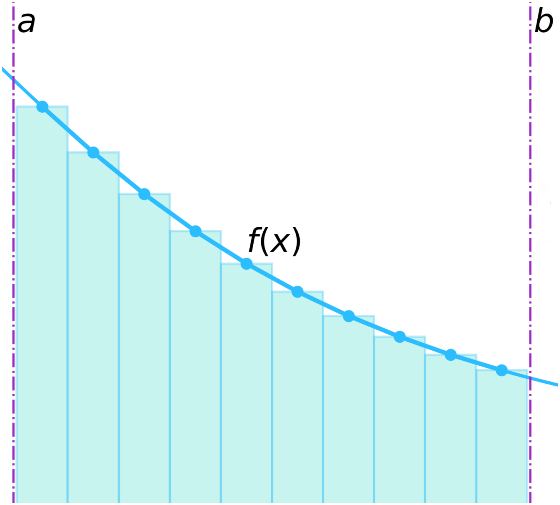


Figure: Definition of Integral of f from a to b

Co-Inventers of Calculus

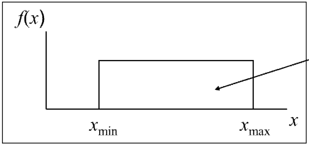
Isaac Newton (1642-1726), English

Gottfried Leibniz (1646-1716), German

- We usually say Newton and Leibniz invented calculus because they found the **fundamental theorem of calculus** which links the concept of the derivative of a function with the concept of the function's integral.

Uniform Distribution

- The uniform distribution is a probability distribution that has equal probabilities for all equal-width intervals within the range of the random variable.



Total area under the uniform probability density function is 1.0

- Assume the density between $x_{\min}$ and $x_{\max}$ is f; then

$$\left(x_{\max}-x_{\min}\right)f=1\Longrightarrow f=\frac{1}{x_{\max}-x_{\min}}.$$

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

$$f\left(x\right)=\frac{1}{b-a}\cdot\mathbf{1}(a\leq x\leq b),$$

and the cdf

$$F\left(x_0\right)=\int_a^{x_0}\frac{1}{b-a}dx=\frac{x_0-a}{b-a}\text{ for }x_0\in\left[a,b\right],$$

where $\mathbf{1}(\cdot)$ is the indicator function which equals 1 when the statement in the parentheses is true and 0 otherwise.

Example: Gasoline Sales

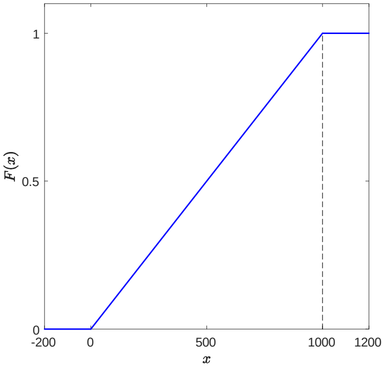
- Assume the gasoline sales at a gasoline station is equally likely from 0 to 1,000 gallons during a day; then the gasoline sales follow a uniform (probability) distribution:

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

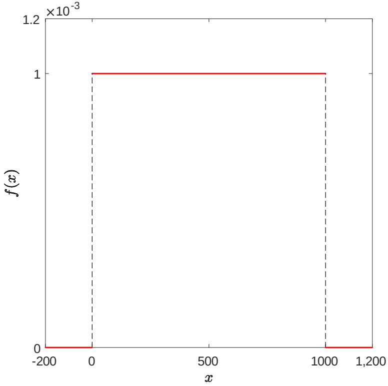
whose pdf is

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

[figure here]



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

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

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

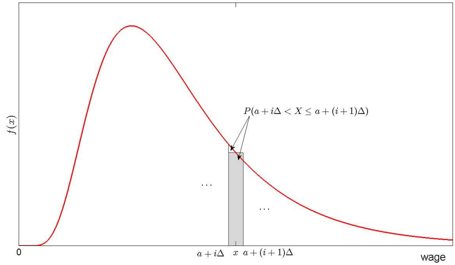


Figure: PDF of Wage: wage $\sim\exp\left(N\left(\mu,\sigma^2\right)\right)$ with $N\left(\mu,\sigma^2\right)$ defined below, $a=0,b=\infty$

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:

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

- 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.

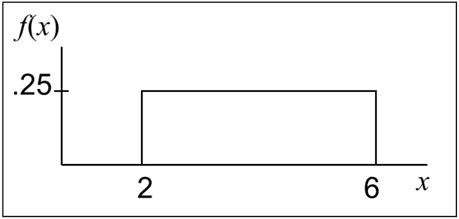
- The standard deviation of X, $\sigma_X = \sqrt{\sigma_X^2}$.
- Example: For the uniform distribution on (a,b),

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

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

[Example] Mean and Variance of the Uniform Distribution

- Suppose $X \sim U(2,6)$.
- $f(x) = \frac{1}{6-2} = 0.25$ for $2 \le x \le 6$.



- $\mu_X = \frac{a+b}{2} = \frac{2+6}{2} = 4$.
- $\sigma_X^2 = \frac{(b-a)^2}{12} = \frac{(6-2)^2}{12} = 1.333$.

Linear Functions of Random Variables

- Let $W = a + bX$, where X has mean μ_X and variance σ_X^2, and a and b are constant fixed numbers.
- Then the mean of W is
$$\mu_W = a + b\mu_X.$$
- The variance of W is
$$\sigma_W^2 = b^2\sigma_X^2.$$
- The standard deviation of W is
$$\sigma_W = |b|\,\sigma_X.$$
- An important special case of the result for the linear function of random variable is z-score of X, $Z = \frac{X-\mu_X}{\sigma_X}$: $\mu_Z = 0$ and $\sigma_Z^2 = 1$.

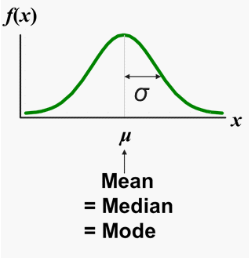
The Normal Distribution

The Normal Distribution

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

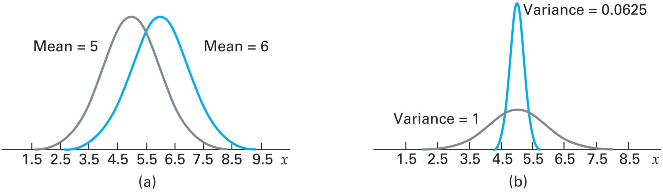
$$f\left(x|\mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ for } x\in\mathscr{S}=(-\infty,\infty)\,, \tag{1}$$

  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 $X\sim N\left(\mu,\sigma^2\right)$.
- Since $f\left(x|\mu,\sigma^2\right)$ is symmetric and bell-shaped, its mean, median and mode are equal $(=\mu)$. Its spead is determined by σ $\left[f\left(\mu|\mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\right]$.



continue

- By varying the parameters μ and σ, we obtain different normal distributions:



Copyright ©2013 Pearson Education, publishing as Prentice Hall

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

- The cdf of the normal distribution, $F\left(x_0|\mu,\sigma^2\right)=\int_{-\infty}^{x_0}f\left(x|\mu,\sigma^2\right)dx$, does not have an analytic form (i.e., a closed-form expression), but computation of probabilities based on the normal distribution is direct nowadays. [see below]
- 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.
- (**) It can be shown that $\sigma =$IQR$/1.349$, which can be treated as a robust representation of σ because IQR does not involve the two tails of X.

- Because $f\left(x|\mu, \sigma^2\right)$ is symmetric, the skewness of X is 0.
- Recall from Lecture 1 that the (population) skewness is

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

and the sample skewness is

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

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

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?]:

$$\text{Kurt}_X = \frac{E\left[\left(X-\mu_X\right)^4\right]}{\sigma_X^4} =: \frac{\mu_4}{\sigma^4},$$

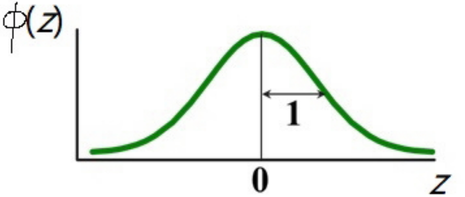
  where μ_4 is the fourth central moment.
  - It can be shown that the kurtosis of $N\left(\mu,\sigma^2\right)$ is 3, which is chosen as a benchmark, i.e., if a distribution's kurtosis is larger than 3, it is called heavy tailed, and if less than 3, called light tailed. Heavy-tailed phenomena seem more 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.
- The pdf (cdf) of $N(0,1)$ is often denoted as $\phi(\cdot)$ $(\Phi(\cdot))$[4].
- 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 $X \sim N\left(\mu, \sigma^2\right)$.[5]

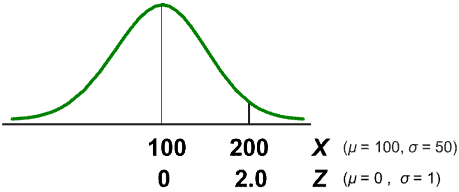


[4] The textbook uses $f(z)$ and $F(z)$, while we use $\phi(z)$ $\Phi(z)$, for the standard normal pdf and cdf; $f(\cdot)$ and $F(\cdot)$ are reserved for the pdf and cdf of $N(\mu, \sigma^2)$.

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

Relationship Between X and Z

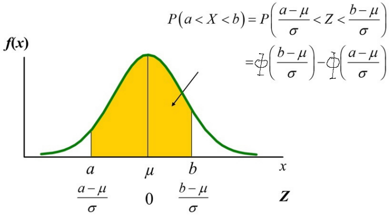
- If X is distributed normally with mean of 100 and standard deviation of 50, the Z value for $X = 200$ is
$$Z = \frac{X-\mu}{\sigma} = \frac{200-100}{50} = 2.0,$$
which says that $X = 200$ is two standard deviations (2 increments of 50 units) above the mean of 100.



- Note that the distribution is the same, only the scale has changed. We can express the problem in original units (X) in standardized units (Z).

continue

- Finding Normal Probabilities:



$$P(a<X<b)=P\left(\frac{a-\mu}{\sigma}<Z<\frac{b-\mu}{\sigma}\right)$$
$$=\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)$$

- The relationship between $F\left(x|\mu,\sigma^2\right)$ ($f\left(x|\mu,\sigma^2\right)$) and $\Phi(z)$ ($\phi(z)$):

$$F\left(x|\mu,\sigma^2\right) = P\left(X\le x\right)=P\left(\frac{X-\mu}{\sigma}\le\frac{x-\mu}{\sigma}\right)=\Phi\left(\frac{x-\mu}{\sigma}\right),$$

$$(*)\ f\left(x|\mu,\sigma^2\right) = \frac{d}{dx}F\left(x|\mu,\sigma^2\right)=\frac{d}{dx}\Phi\left(\frac{x-\mu}{\sigma}\right)=\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right),\ \text{[check!]}$$

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

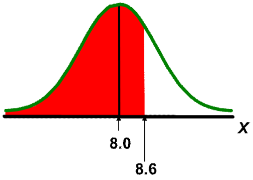
- The Standard Normal Distribution table in the textbook (Appendix Table 1) shows values of $\Phi(z)$, which imply values of $F\left(\mu+\sigma z|\mu,\sigma^2\right)(=\Phi(z))$ for any μ and $\sigma>0$.

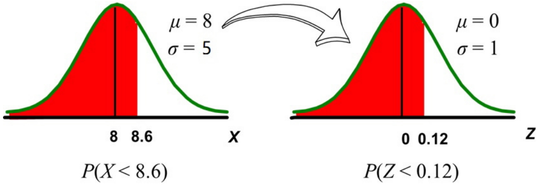
Summary: General Procedure for Finding Normal Probabilities

- To find $P\left(a < X < b\right)$ when $X \sim N\left(\mu, \sigma^2\right)$.

1. Draw the normal curve for the problem in terms of X.
2. Translate X-values to Z-values.
3. Use the Cumulative Normal Table.

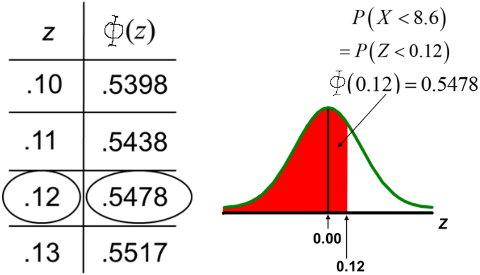
- Suppose $X \sim N\left(8, 5^2\right)$. Find $P\left(X < 8.6\right)$.





$$Z = \frac{X - \mu}{\sigma} = \frac{8.6 - 8}{5} = 0.12.$$

continue



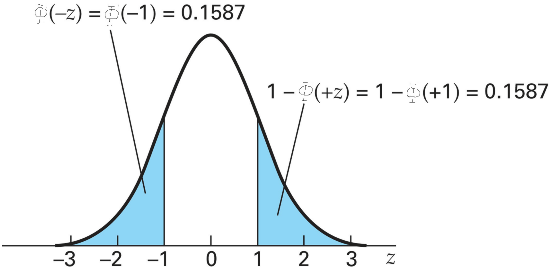
z	$\Phi(z)$
.10	.5398
.11	.5438
.12	.5478
.13	.5517

- $P(X>8.6)=1-P(X<8.6)=1-P(Z<0.12)=1-\Phi(0.12)$.
- $P(X<7.4)=P\left(Z<\frac{7.4-8}{5}\right)=P(Z<-0.12)=\Phi(-0.12)\overset{?}{=}1-\Phi(0.12)$.

  - By symmetry of $\Phi(\cdot)$, $\Phi(-z)=1-\Phi(z)$ for $z>0$ [figure here]; this is also why Appendix Table 1 only reports $\Phi(z)$ for $z>0$.

Upper αth Quantile of Standard Normal

- The upper αth quantile of $N(0,1)$, i.e., the solution to $P(Z>z)=1-\Phi(z)=\alpha$, is denoted as z_α.
  - The αth quantile of $N(0,1)$ is the solution to $\Phi(z)=\alpha$, i.e., $\Phi^{-1}(\alpha)$.
  - Because $1-\Phi(z)=\Phi(-z)$, solving $\Phi(-z)=\alpha$ we have $z_\alpha=-\Phi^{-1}(\alpha)$.
  - If $X\sim N\left(\mu,\sigma^2\right)$, then its upper αth quantile is $\mu+\sigma z_\alpha$ because $P(X>\mu+\sigma z_\alpha)=P(Z>z_\alpha)=\alpha$.



Copyright ©2013 Pearson Education, publishing as Prentice Hall

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.
- If the data are indeed from a normal distribution, then the plot will be a straight line. [figure here]
- (**) 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 $y = \mu + \sigma x$.

  - In the normal probability plot, we invert the x and y axes [so if $x_i \sim N\left(\mu, \sigma^2\right)$, then $\Phi^{-1}\left(\frac{i}{n+1}\right) = \frac{x_{(i)} - \mu}{\sigma}$] and mark the position $\Phi^{-1}\left(\frac{i}{n+1}\right)$ on the y axis as $\frac{i}{n+1}$.

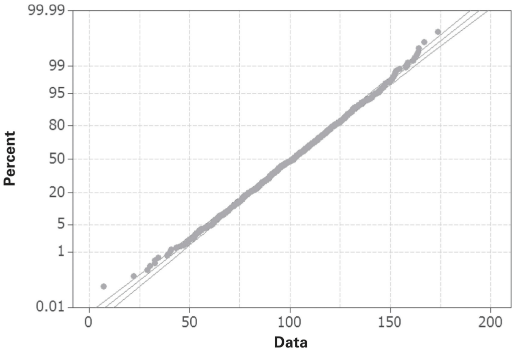


Figure: Normal Probability Plot for Data Simulated from $N\left(100, 25^2\right)$: $n = 1000$

- The horizontal axis indicates the data points ranked in order from the smallest to the largest.
- The vertical axis indicates the cumulative normal probabilities of the ranked data values if the sample data were obtained from a "normal" population; it has a transformed cumulative normal scale [What transformation? see Justification].
- Two dotted lines provide an interval within which data points from a normal distribution would occur most cases.

Nonlinear Plots Indicate Nonnormality

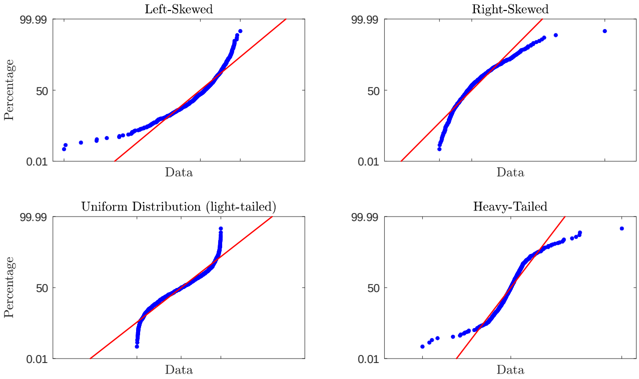


Figure: Normal Probability Plot for Different Types of Data with $n = 1000$: the red line is the plot from normal distribution with the same mean and variance

- (**) Why are the normal probability plots like those above? [Take the uniform distribution as an example]

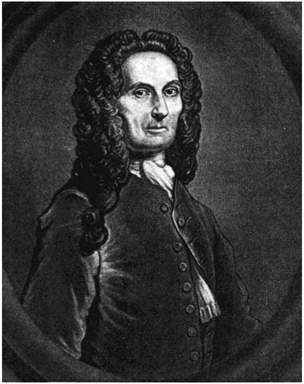
Normal Distribution Approximation for Binomial Distribution

Normal Distribution Approximation for Binomial Distribution

- Recall that a binomial r.v. $X = \sum_{i=1}^{n} X_i$, where $X_i \overset{iid}{\sim}$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]
  - Because normal distributions are easier to handle, this approximation can simplify the analysis of some problems.
  - 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 $X_i \overset{iid}{\sim}$(symmetric) Bernoulli$(0.5) \in \{-1, 1\}$

History of the De Moivre-Laplace Theorem



Abraham de Moivre (1667-1754), French[6]

Pierre-Simon Laplace (1749-1827), French[7]

- 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.

[6] 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.

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

Continuity Correction

- Because we are using a continuous distribution to approximate a discrete one, we often conduct a continuity correction to improve the approximation quality.
- Specifically,

$$P\left(X\leq x\right)\approx P\left(X\leq x+0.5\right).$$

- Since $P\left(X\leq x\right)=P\left(X<x+1\right)$, $P\left(X<x+1\right)\approx P\left(X\leq x+0.5\right)$.

- For example,

$$\begin{aligned}
P\left(a\leq X\leq b\right) &= P\left(X\leq b\right)-P\left(X<a\right)\\
&\approx P\left(X\leq b+0.5\right)-P\left(X\leq a-0.5\right)\\
&= P\left(\frac{X-np}{\sqrt{np\left(1-p\right)}}\leq\frac{b+0.5-np}{\sqrt{np\left(1-p\right)}}\right)\\
&\quad -P\left(\frac{X-np}{\sqrt{np\left(1-p\right)}}\leq\frac{a-0.5-np}{\sqrt{np\left(1-p\right)}}\right)\\
&\approx \Phi\left(\frac{b+0.5-np}{\sqrt{np\left(1-p\right)}}\right)-\Phi\left(\frac{a-0.5-np}{\sqrt{np\left(1-p\right)}}\right).
\end{aligned}$$

- (*) The continuity correction can provide a more accurate approximation particularly when n is not very large or p is not close to 0.5.

[Example] Approximating X

- If X is the number of customers after $n = 200$ people browsed a store's website, and based on past experiences, the probability of visiting the store after browsing is $p = 0.4$, the manager wants to predict the probability of the number of customers falling in an interval, say, $[76, 80]$.
- Solution: From the normal approximation,

$$\begin{aligned} P(76 \le X \le 80) &\approx \Phi\left(\frac{80.5-80}{\sqrt{200\times 0.4\times(1-0.4)}}\right)-\Phi\left(\frac{75.5-80}{\sqrt{200\times 0.4\times(1-0.4)}}\right) \\ &= \Phi(0.0722)-\Phi(-0.6495)=0.5288-0.2580=0.2708. \end{aligned}$$

- For comparison, the exact probability

$$P(76 \le X \le 80) = 0.2717.$$

- The textbook does not apply the continuity correction and approximates $P(76 \le X \le 80)$ by

$$\Phi\left(\frac{80-80}{\sqrt{200\times 0.4\times(1-0.4)}}\right)-\Phi\left(\frac{76-80}{\sqrt{200\times 0.4\times(1-0.4)}}\right)=0.219.$$

[Example] Approximating X/n

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

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

  where the last approximation is from using $\hat{p}$ to substitute p, i.e., p is estimated rather than known a priori.

- In the example above, suppose we observe 80 customers after 200 browsings, then the proportion

$$\hat{p} = \frac{X}{n} \overset{approx.}{\sim} N\left(0.4, \frac{0.4 \times (1-0.4)}{200}\right) = N(0.4, 0.0012).$$

- So for $0 < \underline{p} < \overline{p} < 1$,

$$\begin{aligned} P\left(\underline{p} \le \hat{p} \le \overline{p}\right) &= P\left(\frac{\underline{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \le \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \le \frac{\overline{p}-p}{\sqrt{\frac{p(1-p)}{n}}}\right) \\ &\approx \Phi\left(\frac{\overline{p}-\hat{p}}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}\right) - \Phi\left(\frac{\underline{p}-\hat{p}}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}\right) = \Phi\left(\frac{\overline{p}-0.4}{\sqrt{0.0012}}\right) - \Phi\left(\frac{\underline{p}-0.4}{\sqrt{0.0012}}\right). \end{aligned}$$

  - The continuity correction can also be applied here by replacing $\overline{p}$ by $\overline{p} + \frac{0.5}{n}$ and $\underline{p}$ by $\underline{p} - \frac{0.5}{n}$ [presumably, $\underline{p}$ and $\overline{p}$ take the form of m/n for some integer m; otherwise, correction is not required].

Summary of Approximations of Binomial(n, p)

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

- (**) If n large and $p \leq 0.05$ such that $np \leq 7$ and $np(1-p) > 5$, then Binomial (n, p) can be approximated by both Poisson(np) and $N(np, np(1-p))$.
  - This indicates that the Poisson(λ) distribution can be approximated by a normal distribution, e.g., when λ is large (say $\lambda > 100$), then Poisson$(\lambda) \approx N(\lambda, \lambda)$.
- When approximating a discrete distribution by a continuous one, we may conduct the continuity correction to reduce the approximation error. For example,
  $P(\text{Binomial}(n, p) \leq x) \approx P(N(np, np(1-p)) \leq x + 1/2)$, and
  $P(\text{Poisson}(\lambda) \leq x) \approx P(N(\lambda, \lambda) \leq x + 1/2)$.

The Exponential Distribution

The Exponential Distribution

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

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

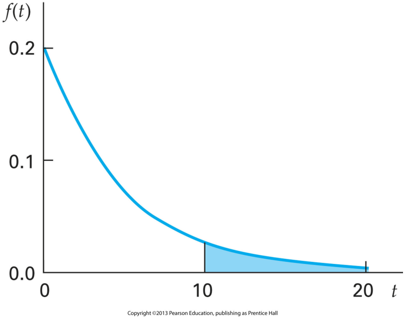
  denoted as $T \sim \text{Exponential}(\lambda)$, i.e., T can take only positive values, and the distribution is not symmetric – the right tail is heavier than that of the normal distribution, and the left tail is thinner. [figure here]

- The cdf of Exponential(λ) is

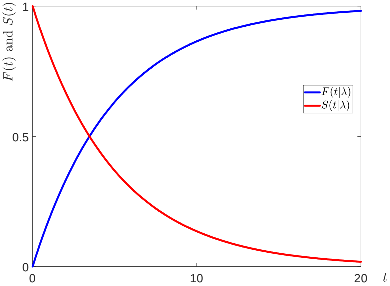
$$F(t|\lambda) = \int_0^t \lambda e^{-\lambda\tau} d\tau = \left(1 - e^{-\lambda t}\right) \cdot \mathbf{1}(t > 0).$$[figure here]

- $F(t|\lambda)$ 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.

- Examples:
  - Time between trucks arriving at an unloading dock.
  - Time between transactions at an ATM Machine.
  - Time between phone calls to the main operator.



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



$F(t|\lambda)$ and $S(t|\lambda)$ with $\lambda = 0.2$

Properties of the Exponential Distribution

- The cdf $F(t|\lambda)$ implies that the survivor function is

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

  for $t > 0$.
  - In survival analysis, $S(t|\lambda)$ is more popular than $F(t|\lambda)$ 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

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

  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.
  - λ in Exponential(λ) is the mean number of arrivals in a unit time.

[Example] Exponential CDF

- Customers arrive at the service counter at the rate of 15 per hour. What is the probability that the arrival time between consecutive customers is less than three minutes?
- Solution: The mean number of arrivals per hour is 15, so $\lambda = 15$.
- Three minutes is .05 hours.
- $P\left(\text{arrival time} < 0.05\right) = 1 - e^{-\lambda t} = 1 - e^{-15\times 0.05} = 0.5276$.
- So there is a 52.76% probability that the arrival time between successive customers is less than three minutes.

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
$$F(x_1, \cdots, x_K) = P(X_1 \leq x_1 \cap \cdots \cap X_K \leq x_K).$$
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
$$f(x_1, \cdots, x_K) = \frac{d^K}{dx_1 \cdots dx_K} F(x_1, \cdots, x_K)$$
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

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

- 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
$$\sigma_W^2 = Var(W) = \sum_{i=1}^{K} a_i^2 \sigma_i^2 + 2\sum_{i=1}^{K-1} \sum_{j>i} a_i a_j \sigma_{ij},$$
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
$$\begin{aligned} Cov(X, Y) &= \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X \mu_Y, \\ Corr(X, Y) &= \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}. \end{aligned}$$

Linear Combinations of Normal Random Variables

- If X and Y are jointly normally distributed random variables [see the next slide for its pdf], then the linear combination, $W = aX + bY$, is also normally distributed.
- Example: Two tasks must be performed by the same worker,

$$\begin{aligned}
&X = \text{minutes to complete task 1, } \mu_X = 20, \sigma_X = 5,\\
&Y = \text{minutes to complete task 2, } \mu_Y = 30, \sigma_Y = 8,\\
&X \text{ and } Y \text{ are normally distributed and independent.}^{8}
\end{aligned}$$

- What is the mean and standard deviation of the time to complete both tasks?
- $W = X + Y$, so $\mu_W = \mu_X + \mu_Y = 20 + 30 = 50$.
- Since X and Y are independent, $Cov(X,Y) = 0$, thus $\sigma_W^2 = \sigma_X^2 + \sigma_Y^2 + 2Cov(X,Y) = 5^2 + 8^2 = 89$, and $\sigma_W = \sqrt{89} = 9.434$.
- Because a linear combination of jointly normal r.v.'s is also normally distributed, and a normal distribution is determined completely by its mean and variance, we have $W \sim N(50, 89)$.

[8] This assumption may not be valid since time cannot be negative while normal distributions can take negative values, and X and Y may not be independent because the same worker finishes both tasks; anyway, we hope this assumption is approximately satisfied.

(**) Bivariate Normal Distribution

- The bivariate normal pdf is

$$f_{X_1X_2}(x_1,x_2|\mu_1,\mu_2,\sigma_1,\sigma_2,\rho)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$
$$\cdot\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2-2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\right]\right\},$$

where $\mu_1,\mu_2\in\mathbb{R}$, $\sigma_1,\sigma_2>0$, and $\rho\in(-1,1)$.

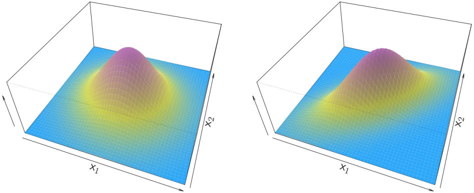


Figure: Two Bivariate Normal Density Functions: left: $\rho=0$; right: $\rho=0.7$

- *Corr* $(X_1,X_2)=\rho$; when $\rho=0$, $f_{X_1X_2}(x_1,x_2)=f_{X_1}(x_1|\mu_1,\sigma_1)f_{X_2}(x_2|\mu_2,\sigma_2)$, so no correlation implies independence for the bivariate normal distribution!