EE3211
Modelling Techniques

Lecture 2

Probability Distributions

Discrete Probability Distributions

Random Variables

- **Random variable**: assigns numeric values to different events in a sample space

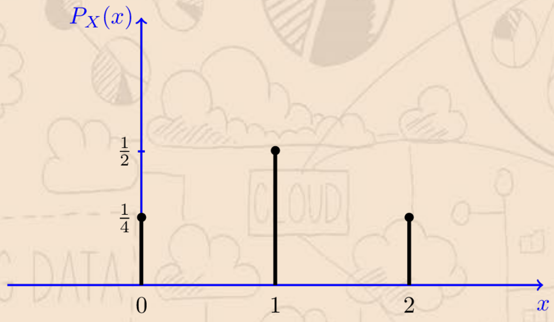
- **Discrete** and **continuous** random variables

- **Discrete random variable**: random with a discrete set of numeric values

- **Continuous random variable**: random variable without possible values enumeration

Probability-Mass Function for Discrete Random Variable

- **Probability-mass function** (pmf):



  - express values of a discrete random variable and its associated probabilities
  - Probability distribution:
    -discrete random variable X
    -Pr (X=r), all values of r have +ve probability

- Display in a table with the values and their probabilities or mathematical formula with probabilities of all values

Example (Hypertension): Suppose from previous experience with a certain drug, the drug company expects that for any clinical practice the probability that 0 patients of 4 will be brought under control is 0.008, 1 patient of 4 is 0.076, 2 patients of 4 is 0.265, 3 patients of 4 is 0.411, and all 4 patients is 0.24.

Table 4.1 Probability-mass function for the hypertension-control example

$Pr(X = r)$	.008	.076	.265	.411	.240
r	0	1	2	3	4

- $0 < Pr(X = r) \leq 1$
- $\sum Pr(X = r) = 1$

-summation is taken over all possible values that have positive probability

Relationship of Probability Distributions to Frequency Distributions

- **Frequency distribution:** list of each possible value in and a corresponding count (how frequent the value occurs)

- Divide each count by sample size:
  frequency distribution ~ probability distribution

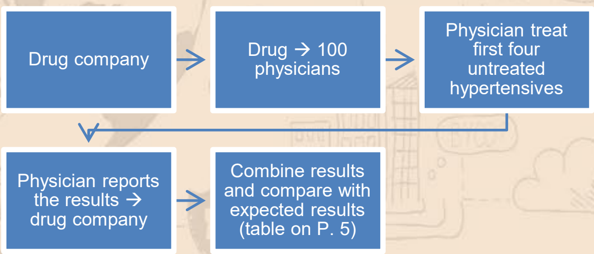
- Probability distribution: model based on very large sample
  -each value → fraction of data points in a sample

- Frequency distribution gives actual proportion of points corresponds to specific values
  -**Goodness-of-fit test**: check the appropriateness of the model
  *comparing the observed sample frequency distribution with the probability distribution

Question: How can the probability-mass function (pmf) in the table be used to judge whether the drug behaves with the same efficacy in actual practice as predicted by the drug company?

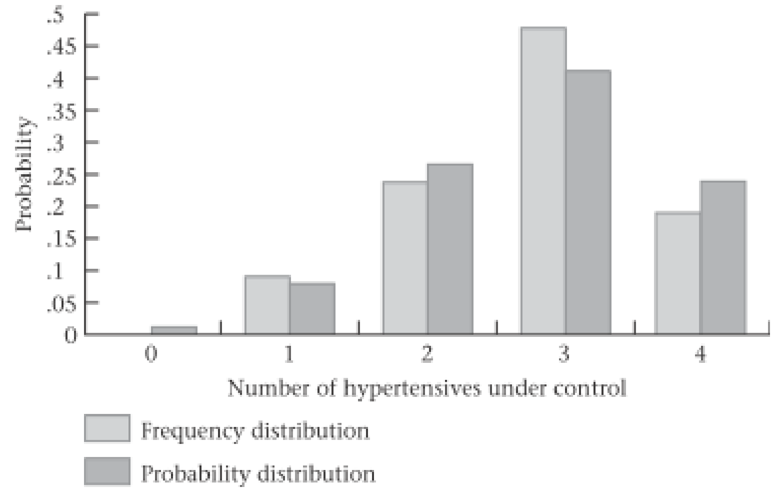
Table 4.2 Comparison of the sample-frequency distribution and the theoretical-probability distribution for the hypertension-control example

Number of hypertensives under control = r	Probability distribution $Pr(X = r)$	Frequency distribution
0	.008	.000 = 0/100
1	.076	.090 = 9/100
2	.265	.240 = 24/100
3	.411	.480 = 48/100
4	.240	.190 = 19/100



- **Statistical inference**: compare the two distributions to judge differences between the two (chance or real differences ?)
- Pmf: previous data / well-known distribution
- Pmf derived from the binomial distribution is compared with the frequency distribution to determine whether the drug behaves with the same efficacy as predicted.

Figure 4.1 Comparison of the frequency and probability distribution for the hypertension-control example



Discrete Random Variable: Expected value

- X (random variable): many values with +ve probability → pmf is not useful

  -summarize sample points by listing each data value

- Develop measure of location and spread for X

- Arithmetic mean x

  = expected value of a random variable (E(X))

  = population mean (μ)

  = the “average” value of the random variable

Expected value of a discrete random variable:

$$E(X) = \mu = \sum_{i=1}^{R} x_i Pr(X = x_i)$$

x_i's are the values the random variable assumes with positive probability

Discrete Random Variable: Expected value

Example: Number of episodes of otitis media in the first 2 years of life.
Question: What is the expected number of episodes of otitis media in the first 2 years of life?

Table 4.3 Probability-mass function for the number of episodes of otitis media in the first 2 years of life

r	0	1	2	3	4	5	6
$Pr(X = r)$	.129	.264	.271	.185	.095	.039	.017

E(X)=0(0.129)+1(.264)+2(.271)+3(.185)+4(.095)+5(.039)+6(.017)=2.038

Interpretation: on average a child would be expected to have about two episodes of otitis media in the first 2 years of life.

Discrete Random Variable: Variance + SD

- **Population variance (**Var(X) / or σ^2 **):**
  = sample variance (s^2) for a random variable

$$Var(X) = \sigma^2 = \sum_{i=1}^{R} (x_i - \mu)^2 Pr(X = x_i)$$

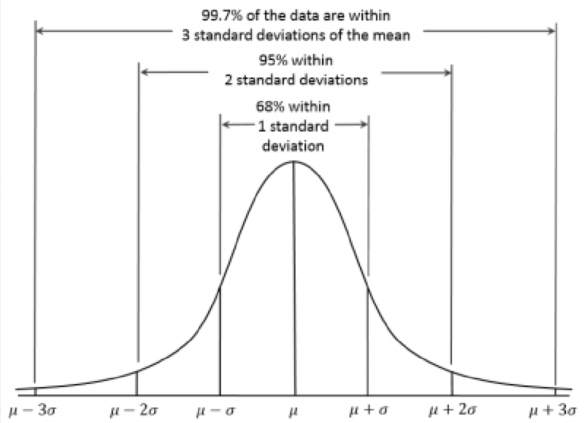
where x_i^2 are with positive probability

- spread (relative to E(X)) about values with +ve probability

- **Standard deviation (**sd(X) **/** σ)
  -square root of its variance

Discrete Random Variable: Variance + SD

- Approximately 95% of the probability mass falls within two standard deviations (2σ) of the mean of a random variable



Cumulative-Distribution Function of a Discrete Random Variable

- **Cumulative-distribution function** (cdf):
  - for a specific value of *x* of *X*: $Pr(X \leq x) = F(x)$
  - can be used to distinguish a certain variable is discrete or continuous

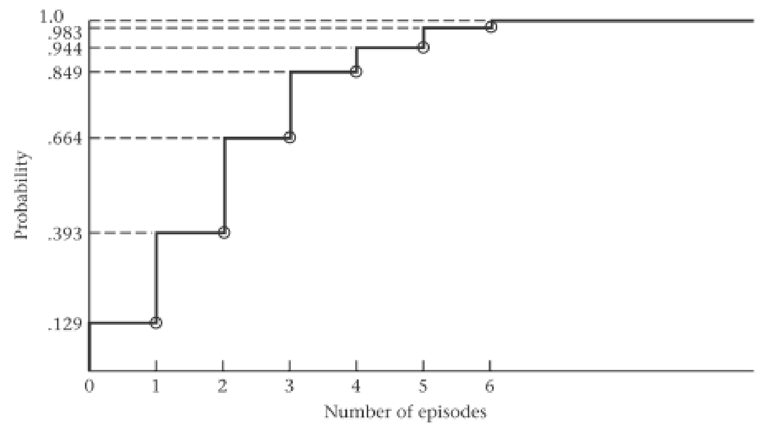


Figure 4.2 Cumulative-distribution function for the number of episodes of otitis media in the first 2 years of life

➢**Discrete** random variable : series of steps (step function)

- With the increase in number of values, the cdf approaches that of a smooth curve

➢**Continuous** random variable: smooth curve

Continuous Probability Distribution

- The normal, Gaussian or “bell-shaped” distribution: foundation of estimation and hypothesis testing

- Many random variables tend to follow a normal distribution (approximately)

  e.g. birthweights, blood pressure

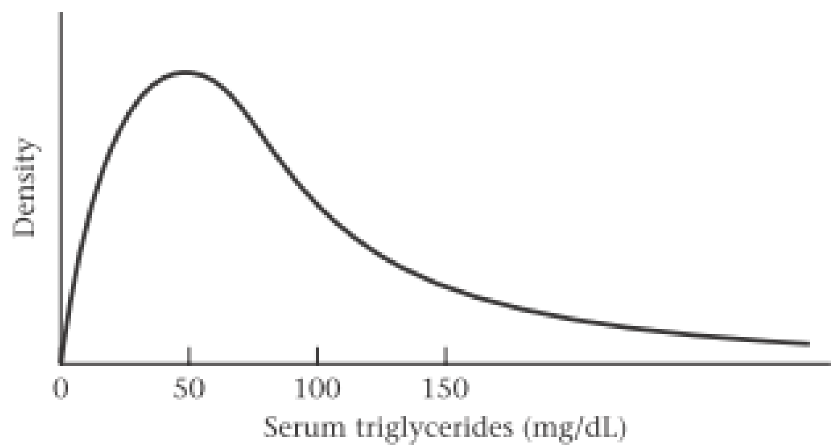
- Variables are not themselves normal: can be approximated by a normal distribution by summing many times

- Normal distribution is commonly used :

  - easy to use
  - normal distribution tables more widely available

- Continuous random variable and probability-mass function: which values are more probable than others and to what degree
- **Probability-density function** (pdf): certain ranges of values occur more frequently than others
  - large values in regions of high probability
  - small values in regions of low probability

Figure 5.2 The pdf for serum triglycerides



Example:
Serum triglyceride level: asymmetric and positively skewed

- Pdf of the continuous random variable

- **Cumulative-distribution function** (cdf): probability that X will take on values ≤ a
  - area under the pdf to the left of a
  - similar to that for discrete random variable

- Continuous random variable: $Pr(X < x) = Pr(X \leq x)$
  - $Pr(X = x) = 0$

- Expected value and variance for continuous random variables have the same meaning as for discrete random variables

- **Expected value** of a continuous random variable X:

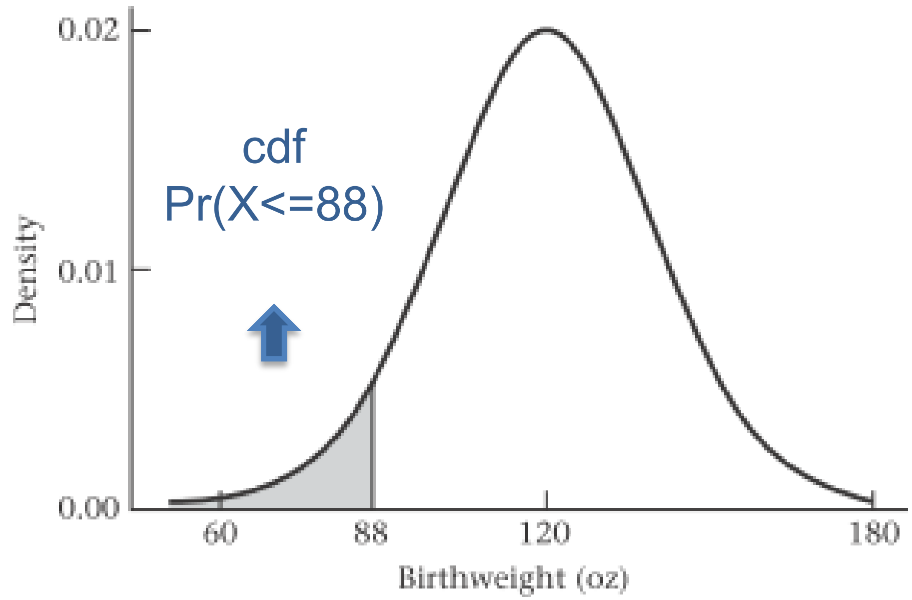
  - $E(X) = \mu$

  – average value of the random variable

88oz:
-special meaning in obstetrics
-cutoff for identifying low birthweight infants
-with higher risk for different unfavorable outcomes
e.g. mortality in the first year of life

Probability density function (pdf): birthweights in general population

Figure 5.3 The pdf for birthweight



Continuous random variable X:

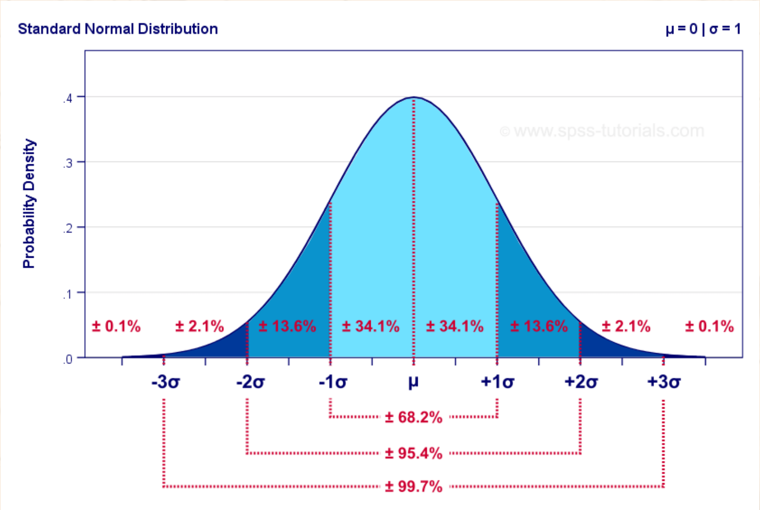
- Variance: Var(X) = σ^2
  - average squared distance of each value of the random variable from its expected value = $E(X^2) - \mu^2$
- Standard deviation = σ = squared root of the variance = $\sqrt{Var(X)}$

$$E(X) = \mu$$

$$\begin{aligned}
Var(X) &= E[(X-\mu)^2] \\
&= E[X^2 - 2\mu X + \mu^2] \\
&= E[X^2] - E[2\mu X] + E[\mu^2] \\
&= E[X^2] - 2\mu E[X] + \mu^2 \\
&= E[X^2] - 2\mu \cdot \mu + \mu^2 \\
&= E[X^2] - 2\mu^2 + \mu^2 \\
&= E[X^2] - \mu^2
\end{aligned}$$

Normal Distribution

- most widely used continuous distribution
- Also called Gaussian distribution (Karl Friedrich Gauss)
- Important in statistics
  E.g. body weights or blood pressures follows normal distribution



- Other distributions that are not themselves normal can be made normal by transformation
  e.g. positively skewed serum triglyceride concentrations
  → log transformation
  → normal distribution
- An approximating distribution to other distributions
  - convenient to work with (hypothesis testing)

Normal Distribution

- Random variables can be approximated by a normal distribution by summing

- Many physiologic measures (genetic + environmental risk factors) can be approximated by normal distribution

- Most estimation procedures and hypothesis tests assume the random variable being considered has an underlying normal distribution

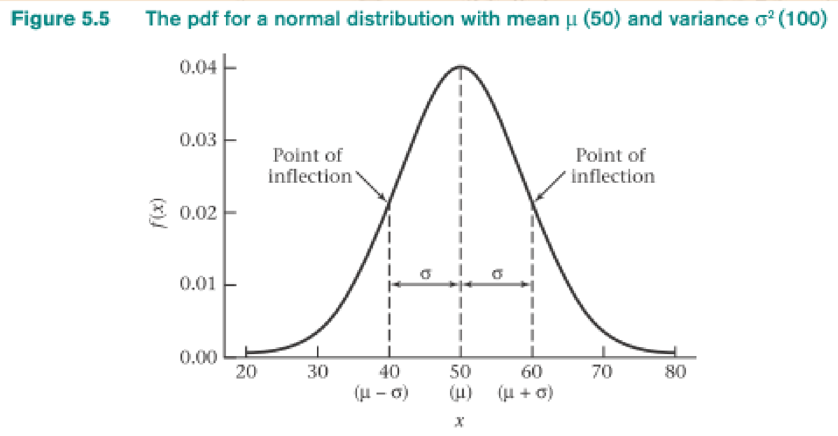
- Probability density function of the normal distribution:

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

-for some parameters μ, σ, where $\sigma > 0$

-exp is the power to which "e" ($\approx$2.17828) is raised

Normal Distribution



- Bell-shaped curve:
  -mode at μ
  -symmetric around μ
  -points of inflection on either side of μ at μ - σ and $\mu = \sigma$
  *A **point of inflection:** slope of the curve changes direction

- Slope:
  - increase to the left of $\mu - \sigma$
  - decrease to the right of $\mu - \sigma$
  - continues to decrease until $\mu + \sigma$

- Magnitude of σ: distance from μ to the points of inflection (visually)

- μ: expected value of the distribution

- σ^2 : variance of the distribution

Normal Distribution

$N(\mu,\sigma^2)$ distribution:

- a normal distribution with mean μ and variance σ^2
- height of the normal distribution = $1/(\sqrt{2\pi}\sigma)$

- Shape: mean μ and variance σ

Figure 5.6 Comparison of two normal distributions with the same variance and different means

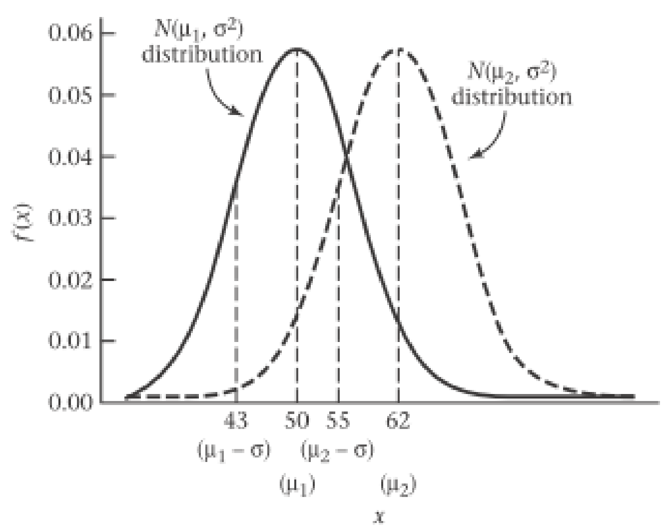
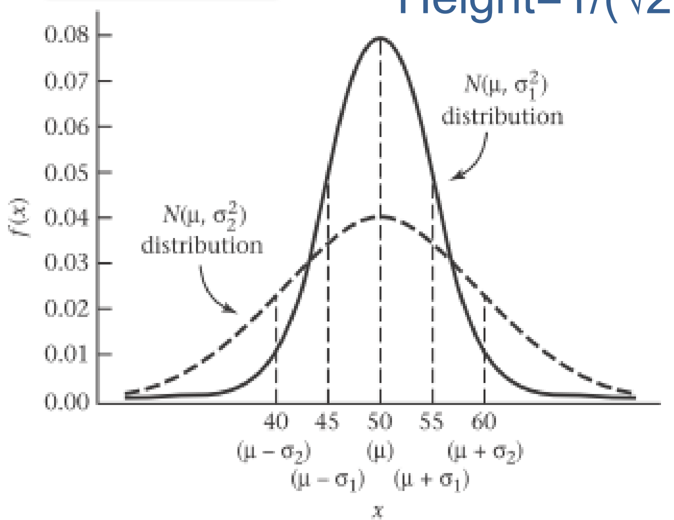


Figure 5.7 Comparison of two normal distributions with the same means and different variances

Height=$1/(\sqrt{2\pi}\sigma)$

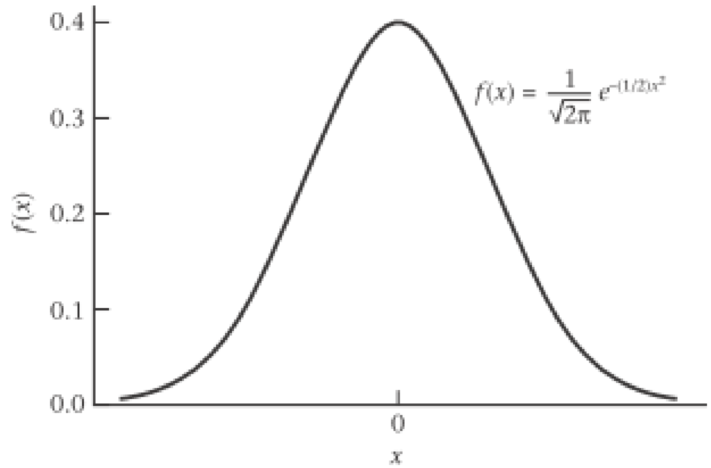


Standard / unit / normal distribution:

- mean 0 and variance 1
- $N(0,1)$ distribution.

Standard Normal Distribution

Figure 5.8 The pdf for a standard normal distribution



Pdf for N(0, 1):

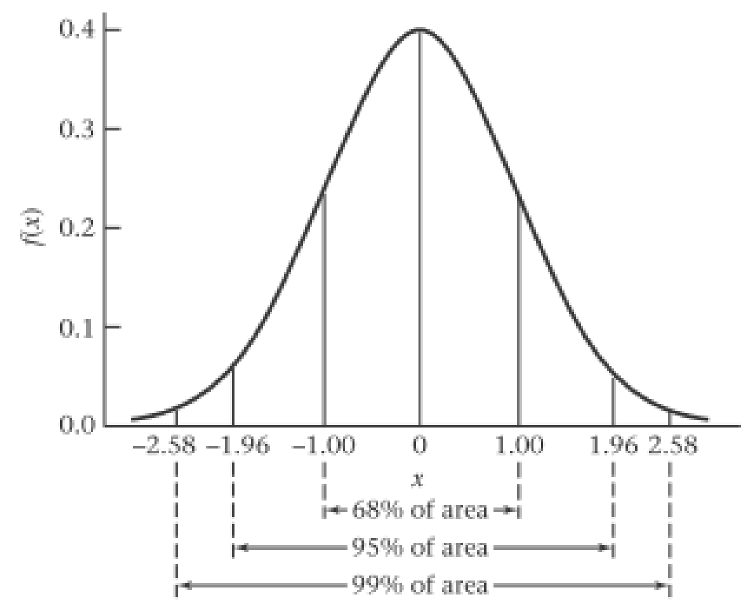
$$f(x) = \frac{1}{\sqrt{2\pi}} e^{(-1/2)x^2}, \quad -\infty < x < +\infty$$

- symmetric about 0
  - f(x) = f(-x)

- 68% of the area under the standard normal density lies between +1 and -1
- 95% of the area lies between +2 and -2
- 99% lies between +2.5 and -2.5

Pr(-1 < X < 1) = 0.6827
Pr(-1.96 < X < 1.96) = 0.95
Pr(-2.576 < X < 2.576) = 0.99

Figure 5.9 Empirical properties of the standard normal distribution



Standard Normal Distribution

- **Cumulative-distribution function** (cdf) for X~N(0,1):
  $\phi(x) = Pr(X \leq x)$
  "~": is distributed as
  X ~ N(0,1) : random variable X is distributed as an N(0,1) distribution

Figure 5.10 The cdf [$\Phi(x)$] for a standard normal distribution

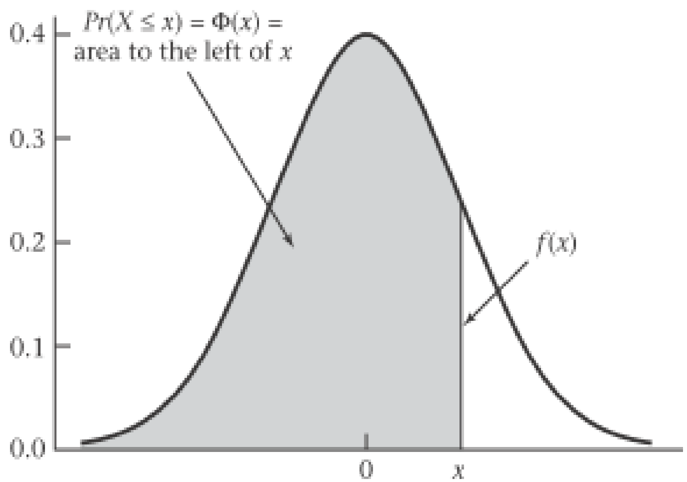
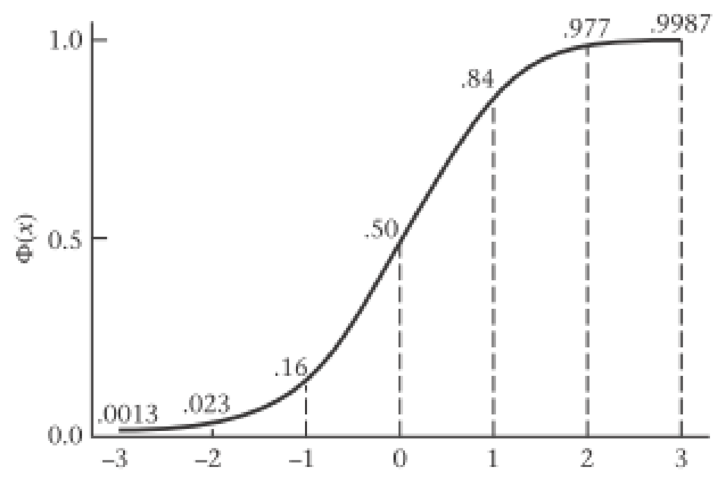


Figure 5.11 The cdf for a standard normal distribution [$\Phi(x)$]

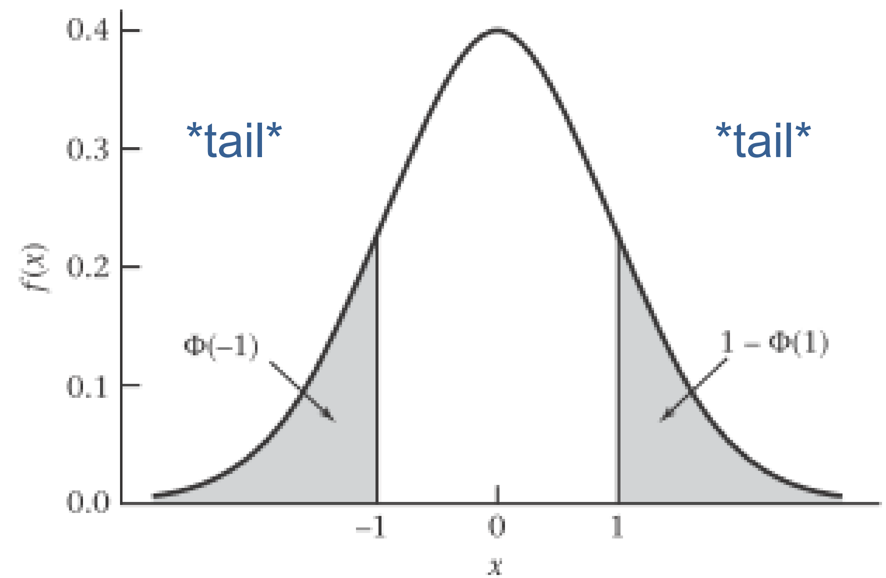


- x becomes small: area to the left of x approaches 0
- x becomes large: area approaches 1

Symmetry Properties of the Standard Normal Distribution

$\phi(-x) = Pr(X \leq -x) = Pr(X \geq x) = 1 - pr(X \leq x) = 1 - \phi(x)$

Figure 5.12 Illustration of the symmetry properties of the normal distribution



tail:

- normal range for a biological quantity: range within x standard deviations of the mean
- Probability of a value in this range = $Pr(-x \leq X \leq x)$ for N(0,1)

Using (Electronic) Tables for the Normal Distribution

- Statistical inference: percentiles of a normal distribution

  E.g. Normal range: the upper and lower fifth percentiles
- The (100 × *u*)th percentile of a standard normal distribution:

  z_u : $Pr(X < z_u) = u$, $X \sim N(0,1)$

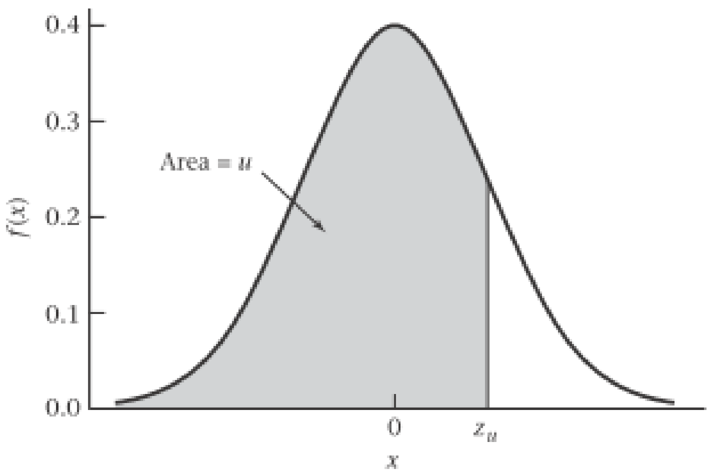
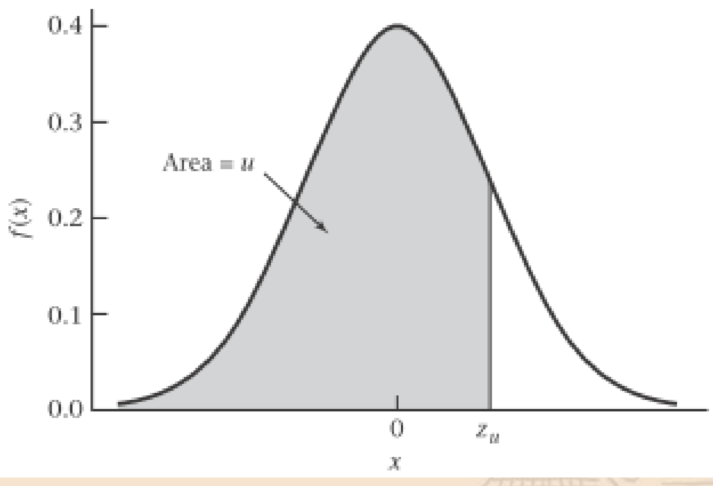


Figure 5.13 Graphic display of the (100 × *u*)th percentile of a standard normal distribution (z_u)

Using (Electronic) Tables for the Normal Distribution

- Z_u : inverse normal function
  - given value of x → normal tables → area to the left of x
  e.g. $\phi(x)$ for $X \sim N(0,1)$
- Z_u :
  - evaluate Z_u → area u in normal tables → Z_u
  - If $u < 0.5$: $z_u = -z_{1-u}$
    →obtain z_{1-u} from normal table
    * symmetry properties of the normal distribution

Figure 5.13 Graphic display of the (100 × u)th percentile of a standard normal distribution (z_u)



Conversion:
$N(\mu,\sigma^2)$ Distribution to N(0,1) distribution

$X \sim N(\mu,\sigma^2)$

- What is $Pr(a < X < b)$ for any a,b?
  - Consider the random variable $Z = (X - \mu)/\sigma$

    If $X \sim N(\mu,\sigma^2)$ and $Z = (X - \mu)/\sigma \rightarrow Z \sim n(0,1)$

Evaluation of Probabilities for Any Normal Distribution via Standardization

$X \sim N(\mu,\sigma^2)$ and $Z = (X - \mu)/\sigma$

- standardization of a normal variable

$$Pr(a < X < b) = Pr\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) = \Phi[(b-\mu)/\sigma] - \Phi[(a-\mu)/\sigma]$$

Conversion: $N(\mu,\sigma^2)$ Distribution to $N(0,1)$ distribution

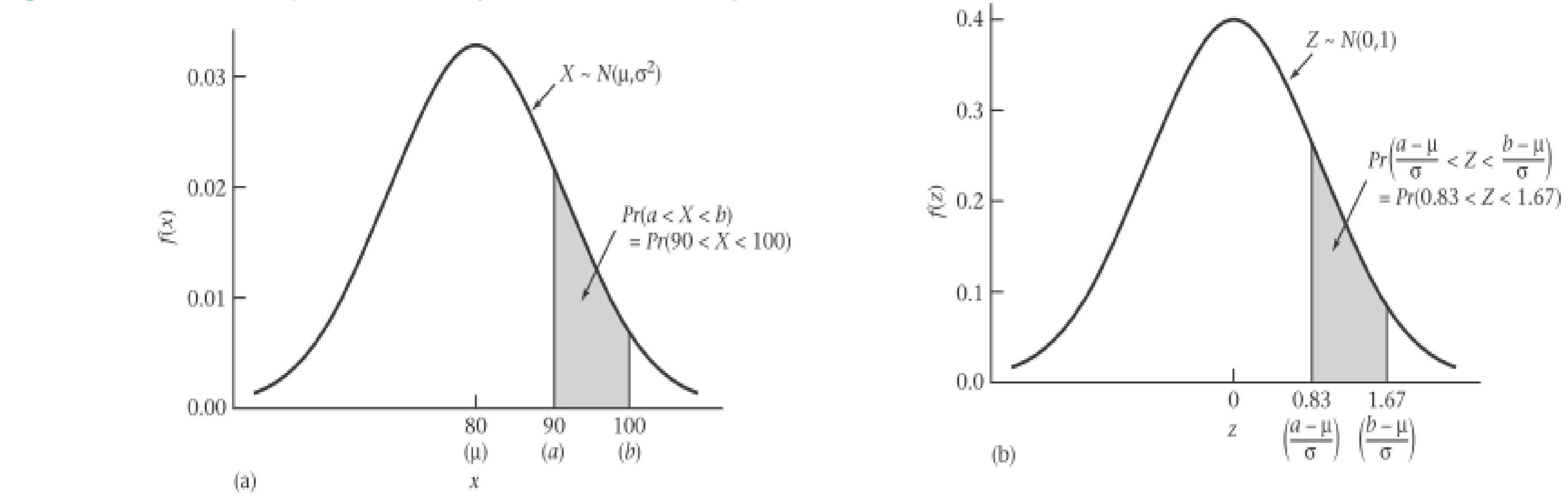
$Pr(a < X < b)$

- population mean μ is subtracted from each boundary point
- divided by the standard deviation σ

$$Pr[(a - \mu)/\sigma < Z < (b - \mu)/\sigma]$$

- standard normal tables can then be used to evaluate this probability

Figure 5.14 Evaluation of probabilities for any normal distribution using standardization



Example of Hypertension

#Suppose the distribution of DBP in 35- to 44-year old men is #normally distributed with mean=80 mm Hg and variance =144 mm #Hg. Find the upper and lower fifth percentiles of this distribution.

#We could do this using normal table or using a computer program.

#We can denote the upper and lower 5th percentiles by $X_{.05}$ and $X_{.95}$ #respectively:

$X_{.05}$ =80+$Z_{.05}$ (12) =80-1.645(12)=60.3 MM HG

$X_{.95}$ =80+$Z_{.95}$ (12) =80+1.645(12)=99.7 MM HG

$Z = (X - \mu)/\sigma$

$X= \mu + Z^* \sigma$

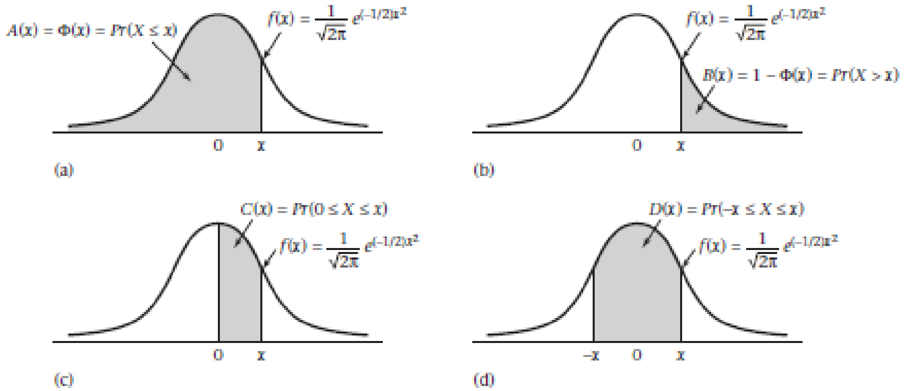
#Use the qnorm function of R, we have

$X_{.05}$ =QNORM(0.05, MEAN=80, SD=12)

$X_{.95}$ =QNORM(0.95, MEAN=80, SD=12)

```
>X=QNORM(0.05, MEAN=80, SD=12)
>X
[1] 60.26176
>Y=QNORM(0.95, MEAN=80, SD=12)
>Y
[1] 99.73824
```

TABLE 3 The normal distribution



x	A[a]	B[b]	C[c]	D[d]
1.56	.9406	.0594	.4406	.8812
1.57	.9418	.0582	.4418	.8836
1.58	.9429	.0571	.4429	.8859
1.59	.9441	.0559	.4441	.8882
1.60	.9452	.0548	.4452	.8904
1.61	.9463	.0537	.4463	.8926
1.62	.9474	.0526	.4474	.8948
1.63	.9484	.0516	.4484	.8969
1.64	.9495	.0505	.4495	.8990
1.65	.9505	.0495	.4505	.9011
1.66	.9515	.0485	.4515	.9031
1.67	.9525	.0475	.4525	.9051
1.68	.9535	.0465	.4535	.9070
1.69	.9545	.0455	.4545	.9090
1.70	.9554	.0446	.4554	.9109
1.71	.9564	.0436	.4564	.9127
1.72	.9573	.0427	.4573	.9146
1.73	.9582	.0418	.4582	.9164
1.74	.9591	.0409	.4591	.9181
1.75	.9599	.0401	.4599	.9199
1.76	.9608	.0392	.4608	.9216
1.77	.9616	.0384	.4616	.9233
1.78	.9625	.0375	.4625	.9249
1.79	.9633	.0367	.4633	.9265
1.80	.9641	.0359	.4641	.9281
1.81	.9649	.0351	.4649	.9297

TABLE 3 The normal distribution (*continued*)

x	A[a]	B[b]	C[c]	D[d]
1.82	.9656	.0344	.4656	.9312
1.83	.9664	.0336	.4664	.9327
1.84	.9671	.0329	.4671	.9342
1.85	.9678	.0322	.4678	.9357
1.86	.9686	.0314	.4686	.9371
1.87	.9693	.0307	.4693	.9385
1.88	.9699	.0301	.4699	.9399
1.89	.9706	.0294	.4706	.9412
1.90	.9713	.0287	.4713	.9426
1.91	.9719	.0281	.4719	.9439
1.92	.9726	.0274	.4726	.9451
1.93	.9732	.0268	.4732	.9464
1.94	.9738	.0262	.4738	.9476
1.95	.9744	.0256	.4744	.9488
1.96	.9750	.0250	.4750	.9500
1.97	.9756	.0244	.4756	.9512
1.98	.9761	.0239	.4761	.9523
1.99	.9767	.0233	.4767	.9534
2.00	.9772	.0228	.4772	.9545
2.01	.9778	.0222	.4778	.9556
2.02	.9783	.0217	.4783	.9566
2.03	.9788	.0212	.4788	.9576
2.04	.9793	.0207	.4793	.9586
2.05	.9798	.0202	.4798	.9596
2.06	.9803	.0197	.4803	.9606
2.07	.9808	.0192	.4808	.9615
2.08	.9812	.0188	.4812	.9625
2.09	.9817	.0183	.4817	.9634
2.10	.9821	.0179	.4821	.9643
2.11	.9826	.0174	.4826	.9651
2.12	.9830	.0170	.4830	.9660
2.13	.9834	.0166	.4834	.9668
2.14	.9838	.0162	.4838	.9676
2.15	.9842	.0158	.4842	.9684
2.16	.9846	.0154	.4846	.9692
2.17	.9850	.0150	.4850	.9700
2.18	.9854	.0146	.4854	.9707
2.19	.9857	.0143	.4857	.9715
2.20	.9861	.0139	.4861	.9722
2.21	.9864	.0136	.4864	.9729
2.22	.9868	.0132	.4868	.9736
2.23	.9871	.0129	.4871	.9743
2.24	.9875	.0125	.4875	.9749
2.25	.9878	.0122	.4878	.9756
2.26	.9881	.0119	.4881	.9762
2.27	.9884	.0116	.4884	.9768
2.28	.9887	.0113	.4887	.9774
2.29	.9890	.0110	.4890	.9780
2.30	.9893	.0107	.4893	.9786
2.31	.9896	.0104	.4896	.9791
2.32	.9898	.0102	.4898	.9797
2.33	.9901	.0099	.4901	.9802
2.34	.9904	.0096	.4904	.9807
2.35	.9906	.0094	.4906	.9812
2.36	.9909	.0091	.4909	.9817
2.37	.9911	.0089	.4911	.9822
2.38	.9913	.0087	.4913	.9827

Summary:
Discrete Probability Distribution

➢Random variables

- discrete vs. continuous variables

➢Specific attributes of random variables

- probability-mass function (probability distribution)
- cumulative density function (cdf)
- expected value and variance

➢Sample frequency distribution : sample realization of a probability distribution

- sample mean (x) and variance (s^2)
- expected value and variance (random variable)

Summary: Continuous Probability Distribution

- Continuous random variables
  - Probability-density function: analogs of probability-mass function for discrete random variables

- Expected value, variance, cumulative distribution for continuous random variables

- Normal distribution: most important continuous distribution

- The two parameters: mean μ and variance $\sigma 2$

- Normal tables (working with standard normal distribution)