IE208: Applied Statistics (Review of Chapters 1-7)

Chapter 1. Probability Theory

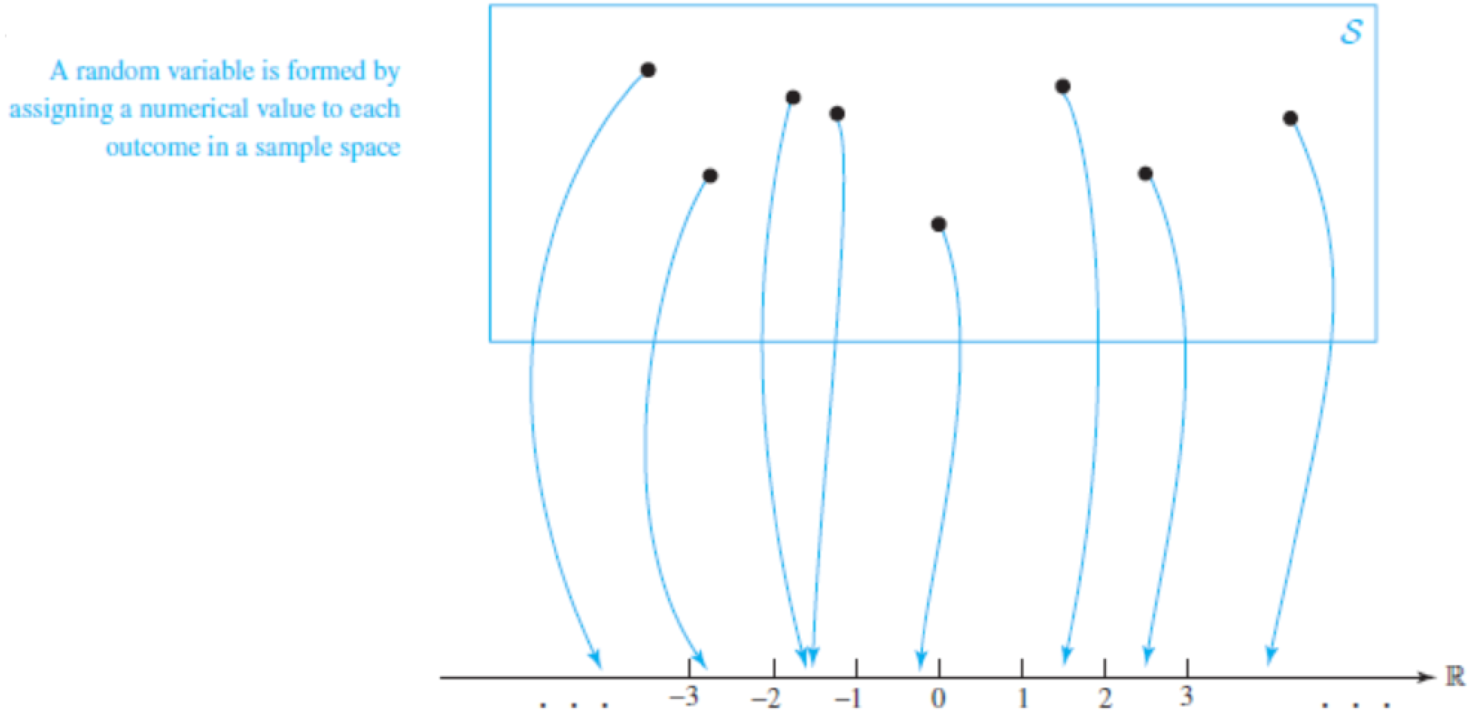
Chapter 1

- Sample space
- Event
  - Intersection of events
  - Union of events
- Conditional probability
- Independent events
- Posterior probabilities

Chapter 2. Random Variables

Random Variable

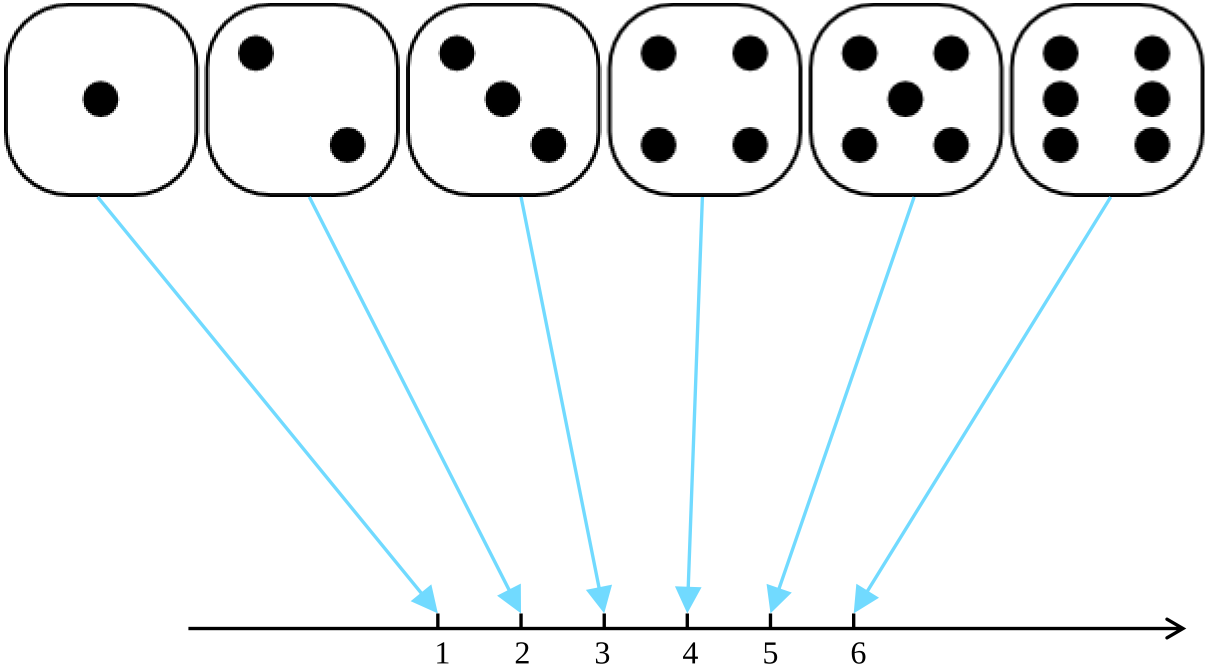
- A random variable is obtained by assigning a numerical value to each outcome of a particular experiment



Random variable: 확률변수

Random Variable

- A random variable is obtained by assigning a numerical value to each outcome of a particular experiment



Probability Mass Function

The **probability mass function (p.m.f.)** of a random variable X is a set of probability values p_i assigned to each of the values x_i taken by the *discrete* random variable. These probability values must satisfy $0 \leq p_i \leq 1$ and $\sum_i p_i = 1$. The probability that the random variable takes the value x_i is said to be p_i, and this is written $P(X = x_i) = p_i$.

- Here, the importance is **discrete** values

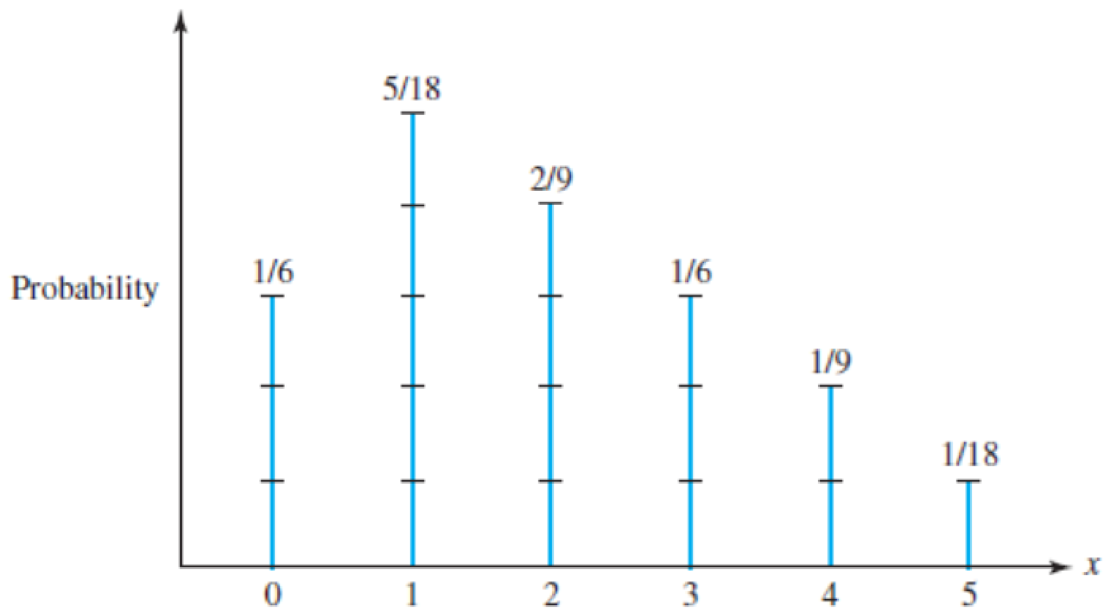
Probability mass function: 확률질량함수

Probability Mass Function

- *Example*: X = positive difference between the scores of two dice

x_i	0	1	2	3	4	5
p_i	1/6	5/18	2/9	1/6	1/9	1/18

Tabular presentation of the probability mass function for dice example



Line graph of the probability mass function for dice example

Cumulative Distribution Function

The **cumulative distribution function (c.d.f.)** of a random variable X is the function

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

For example, suppose that the probability mass function is known. The cumulative distribution function can then be calculated from the expression

$$F(x) = \sum_{y:y\leq x} P(X = y)$$

In other words, the value of $F(x)$ is constructed by simply adding together the probabilities $P(X = y)$ for values y that are no larger than x.

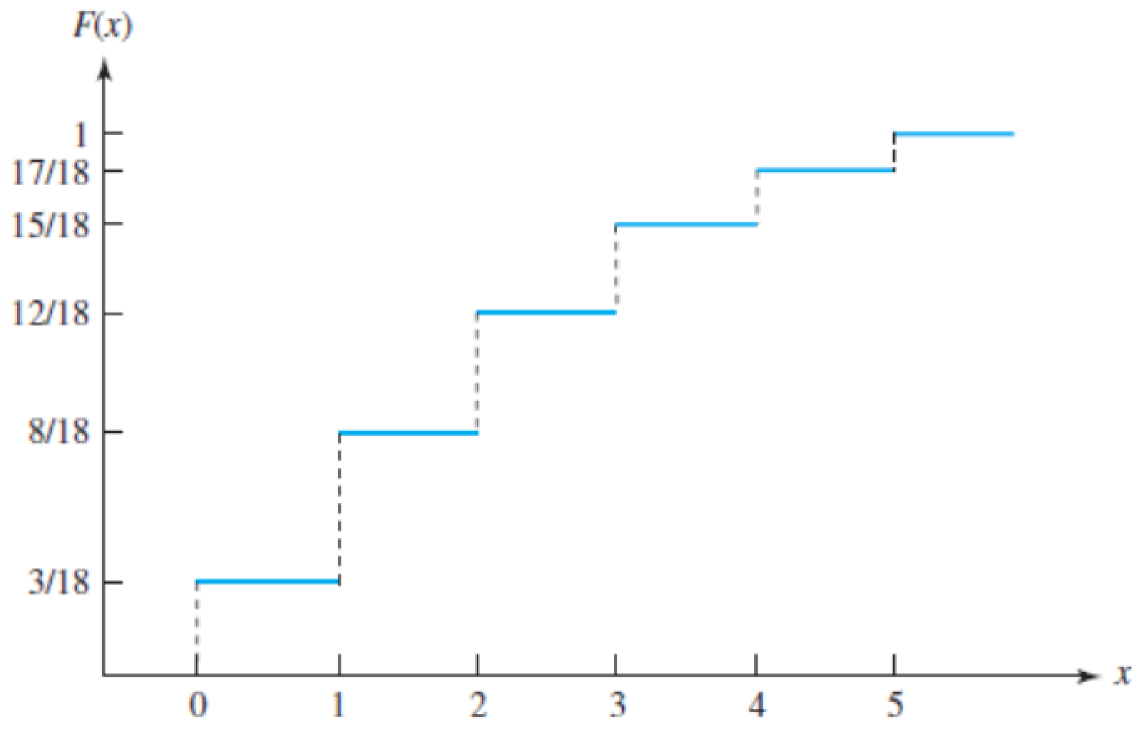
Cumulative distribution function: 누적분포함수

Cumulative Distribution Function

- *Example*: X = positive difference between the scores of two dice

x_i	0	1	2	3	4	5
p_i	1/6	5/18	2/9	1/6	1/9	1/18

Tabular presentation of the probability mass function for dice example



Probability Density Function

- But some random variables are continuous
- *Example*: Random variable X is the time to failure of a newly charged battery

Probability Density Function

A **probability density function** $f(x)$ defines the probabilistic properties of a *continuous* random variable. It must satisfy $f(x) \geq 0$ and

$$\int_{\text{state space}} f(x)\,dx = 1$$

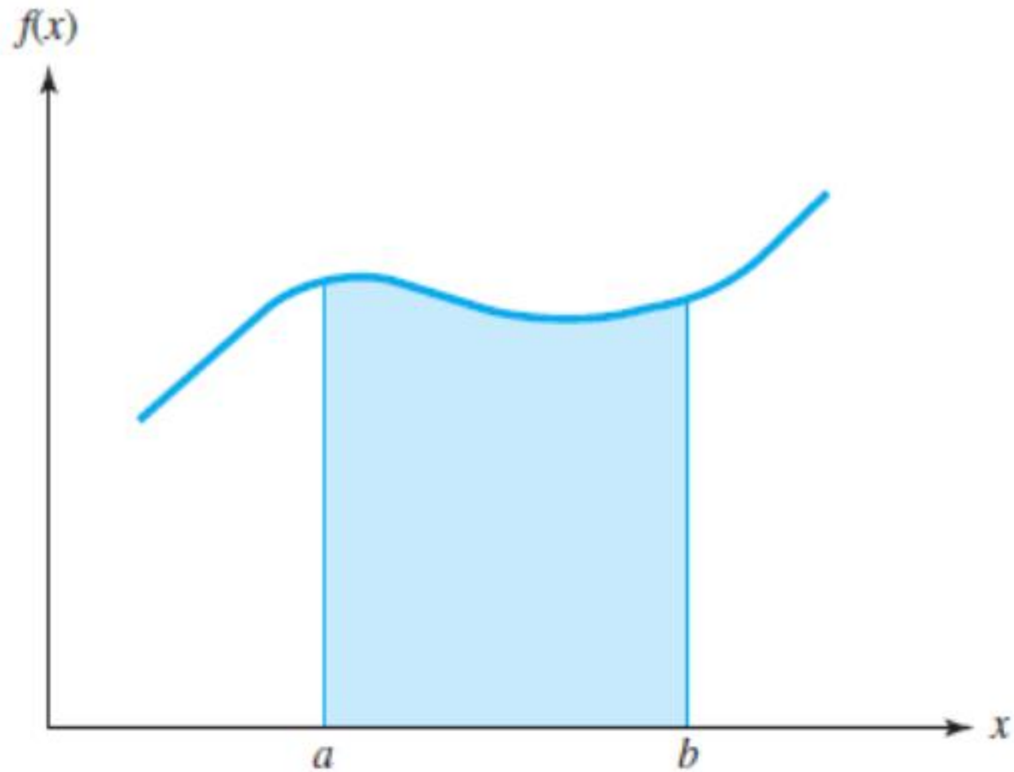
The probability that the random variable lies between two values is obtained by integrating the probability density function between the two values.

Probability density function: 확률밀도함수

Probability Density Function

- Probability density function between two points

$P(a \le x \le b)$ is the area under the probability density function $f(x)$ between the points a and b



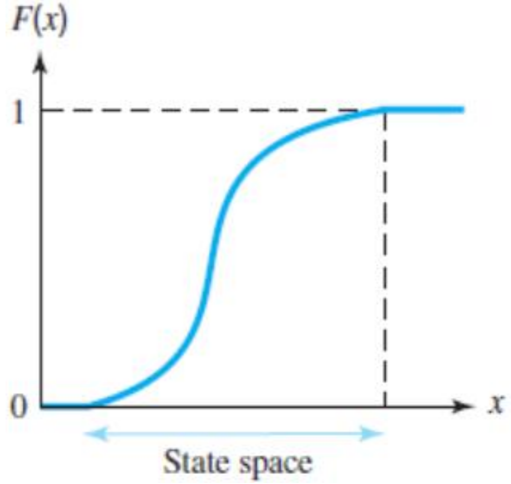
Cumulative Distribution Function

- Exactly the same as for a discrete random variable

The **cumulative distribution function** of a continuous random variable X is defined in exactly the same way as for a discrete random variable, namely

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

For a continuous random variable, the cumulative distribution function $F(x)$ is a continuous nondecreasing function that takes the value 0 prior to and at the beginning of the state space and increases to a value of 1 at the end of and after the state space.



Cumulative Distribution Function

Like the probability density function, the cumulative distribution function summarizes the probabilistic properties of a continuous random variable, and knowledge of either function allows the other function to be constructed. For example, if the probability density function $f(x)$ is known, then the cumulative distribution function can be calculated from the expression

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)\,dy$$

In practice, the lower integration limit $-\infty$ can be replaced by the lower endpoint of the state space since the probability density function is 0 outside the state space. The probability density function can be obtained by differentiating the cumulative distribution function

$$f(x) = \frac{dF(x)}{dx}$$

In addition, notice that the cumulative distribution function provides a convenient way of obtaining the probability that a random variable lies within a certain region, since

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

Random Variables

- Distributions provide full detail
- But how can we describe random variables (probability distributions) using summary measures?
  - Expectation
  - Median
  - Variance
  - Standard deviation
  - …

Expectation

- Expectation, mean, average
- $E(X)$: expectation of X

Expectation: 기대값

Expectation

- *Example*: What is the expectation of the score (value)?

If a fair die is rolled, the expected value of the outcome is

$$E(X) = \left(1 \times \frac{1}{6}\right) + \left(2 \times \frac{1}{6}\right) + \left(3 \times \frac{1}{6}\right) + \left(4 \times \frac{1}{6}\right) + \left(5 \times \frac{1}{6}\right) + \left(6 \times \frac{1}{6}\right)$$
$$= 3.5$$

Since each outcome is equally likely, this is just the normal arithmetic average of the six outcomes.

Expectation

- For discrete random variables

Expected Value of a Discrete Random Variable

The **expected value** or **expectation** of a discrete random variable with a probability mass function $P(X = x_i) = p_i$ is

$$E(X) = \sum_i p_i x_i$$

$E(X)$ provides a summary measure of the average value taken by the random variable and is also known as the **mean** of the random variable.

Expectation

- For continuous random variables

Expected Value of a Continuous Random Variable

The **expected value** or **expectation** of a continuous random variable with a probability density function $f(x)$ is

$$E(X) = \int_{\text{state space}} x f(x)\, dx$$

The expected value provides a summary measure of the average value taken by the random variable, and it is also known as the **mean** of the random variable.

Median

- Median is another measure of the “middle” value
- Random variable is equally likely to be either smaller or larger than the median

Median

The **median** of a continuous random variable X with a cumulative distribution function $F(x)$ is the value x in the state space for which

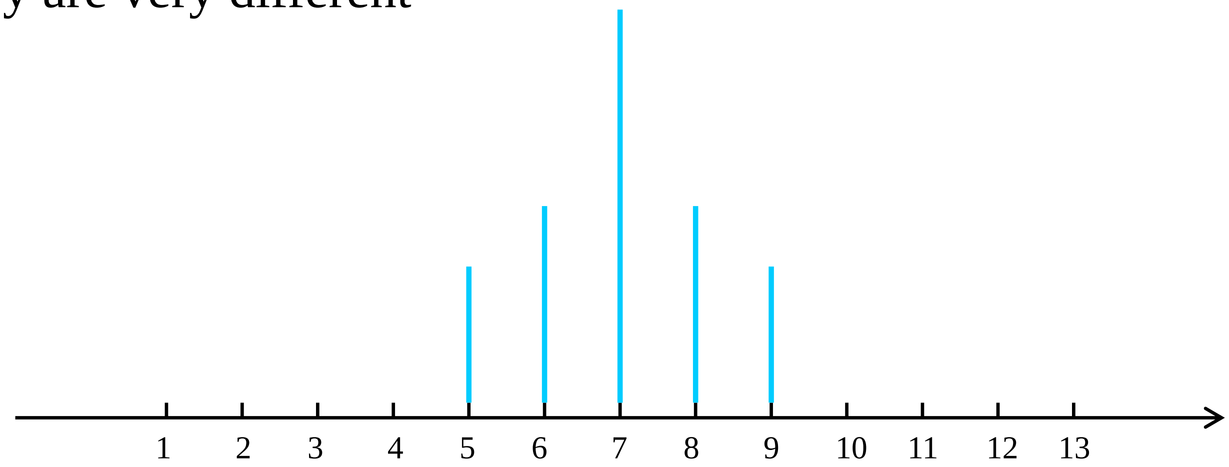
$$F(x) = 0.5$$

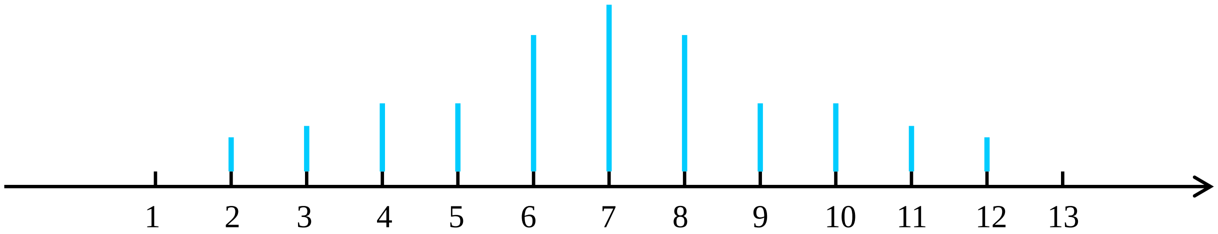
The random variable is then equally likely to fall above or below the median value.

Median: 중앙값

Expectation and Median

- Two distributions have the same expectation and median
- But they are very different





Measure of Spread

- Measure of the spread or deviation of the random variable
  - Variance
  - Standard deviation
  - Percentiles
  - Interquartile
  - …

Variance

- Variance: measures the spread or deviation of the random variable about its mean value
- Expected value of the squares of the deviations of the random variable values about the expected value
- Variance is always positive
- Larger values of the variance indicate a greater spread in the distribution of the random variable about the mean

Variance: 분산

Variance

Variance

The **variance** of a random variable X is defined to be

$$\mathrm{Var}(X) = E((X - E(X))^2)$$

or equivalently

$$\mathrm{Var}(X) = E(X^2) - (E(X))^2$$

The variance is a positive quantity that measures the spread of the distribution of the random variable about its mean value. Larger values of the variance indicate that the distribution is more spread out.

- Discrete random variable:

$$\mathrm{Var}(X) = E((X - E(X))^2) = \sum_i p_i (x_i - E(X))^2$$

Variance

- *Example*: (discrete random variable)

$$\mathrm{Var}(X) = E((X - E(X))^2) = \sum_i p_i (x_i - E(X))^2$$

$$= \frac{1}{6}(1-3.5)^2 + \frac{1}{6}(2-3.5)^2 + \frac{1}{6}(3-3.5)^2 + \frac{1}{6}(4-3.5)^2 + \frac{1}{6}(5-3.5)^2 + \frac{1}{6}(6-3.5)^2$$

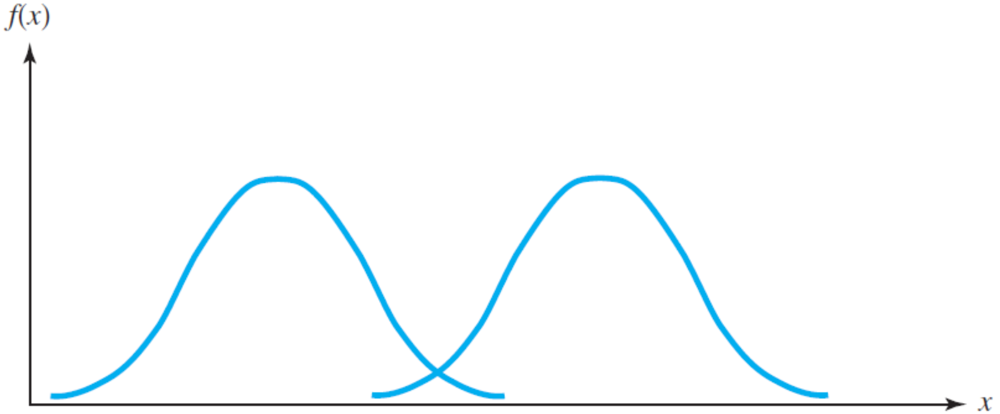
$$= \frac{1}{6}\left((1-3.5)^2 + (2-3.5)^2 + (3-3.5)^2 + (4-3.5)^2 + (5-3.5)^2 + (6-3.5)^2\right)$$

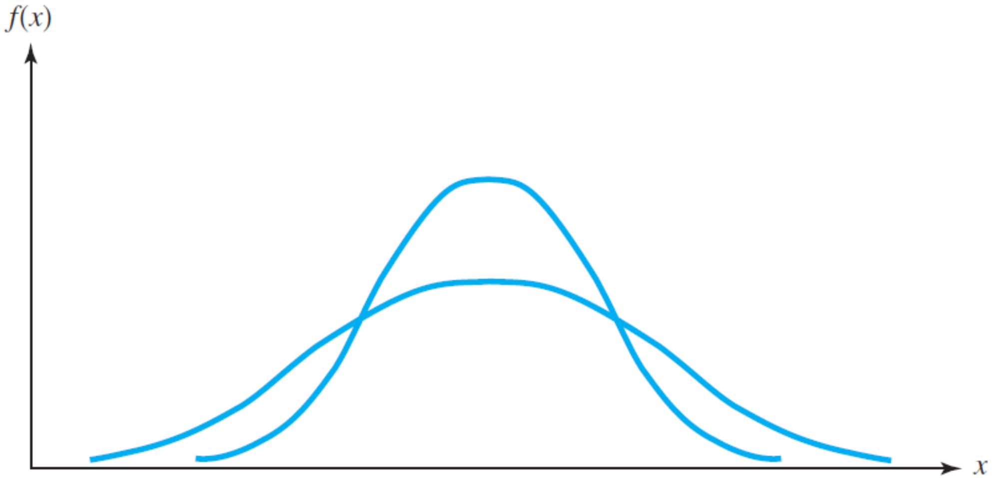
$$\approx 2.92$$

$$E[X^2] = 1^2 \times \frac{1}{6} + 2^2 \times \frac{1}{6} + 3^2 \times \frac{1}{6} + 4^2 \times \frac{1}{6} + 5^2 \times \frac{1}{6} + 6^2 \times \frac{1}{6} = 15.17$$

$$\mathrm{Var}(X) = E[X^2] - E[X]^2 = 15.17 - 3.5^2 \approx 2.9$$

Expectation and Variance





Standard Deviation

Standard Deviation

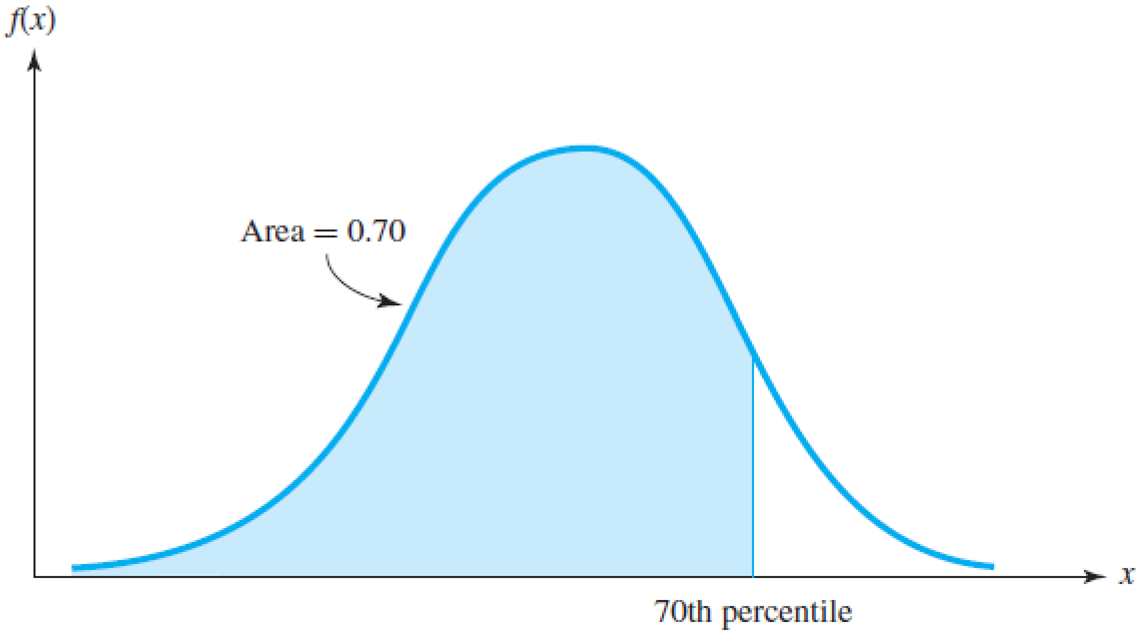
The **standard deviation** of a random variable X is defined to be the positive square root of the variance. The symbol σ^2 is often used to denote the variance of a random variable, so that σ represents the standard deviation.

Standard deviation: 표준편차

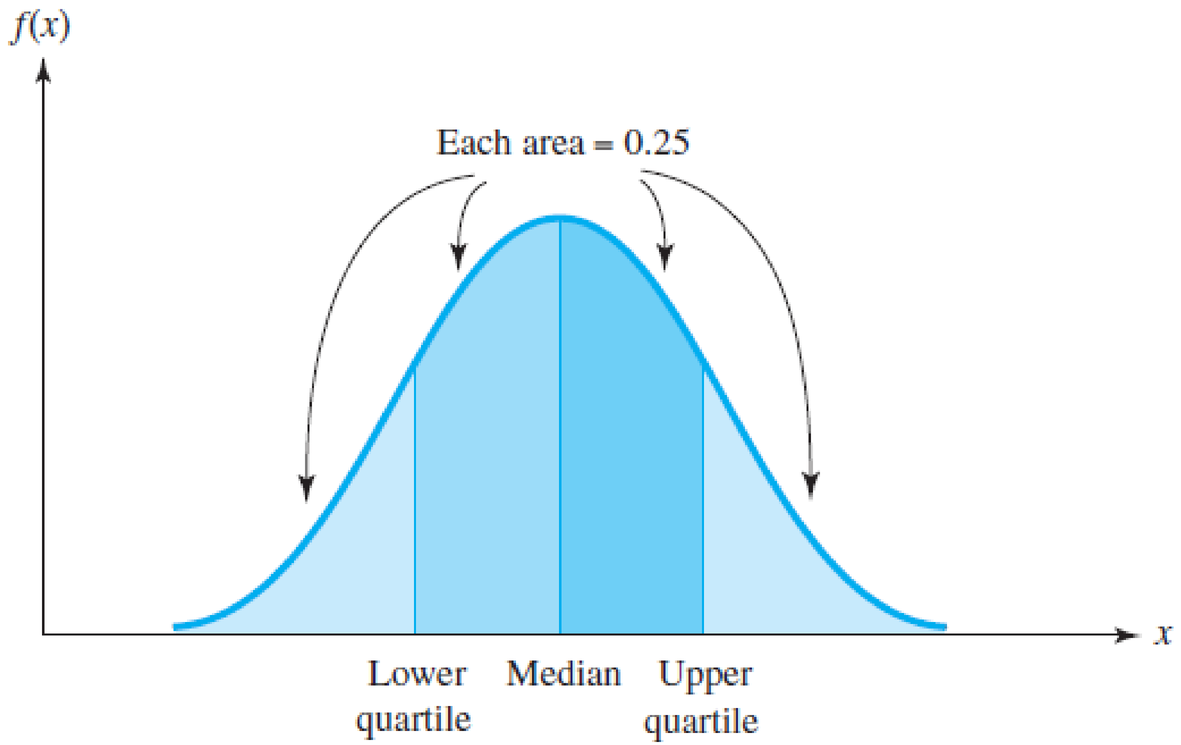
Percentiles

- The $p \times 100$th quantile of a random variable X with a cumulative distribution function $F(x)$ is defined to be the value x for which

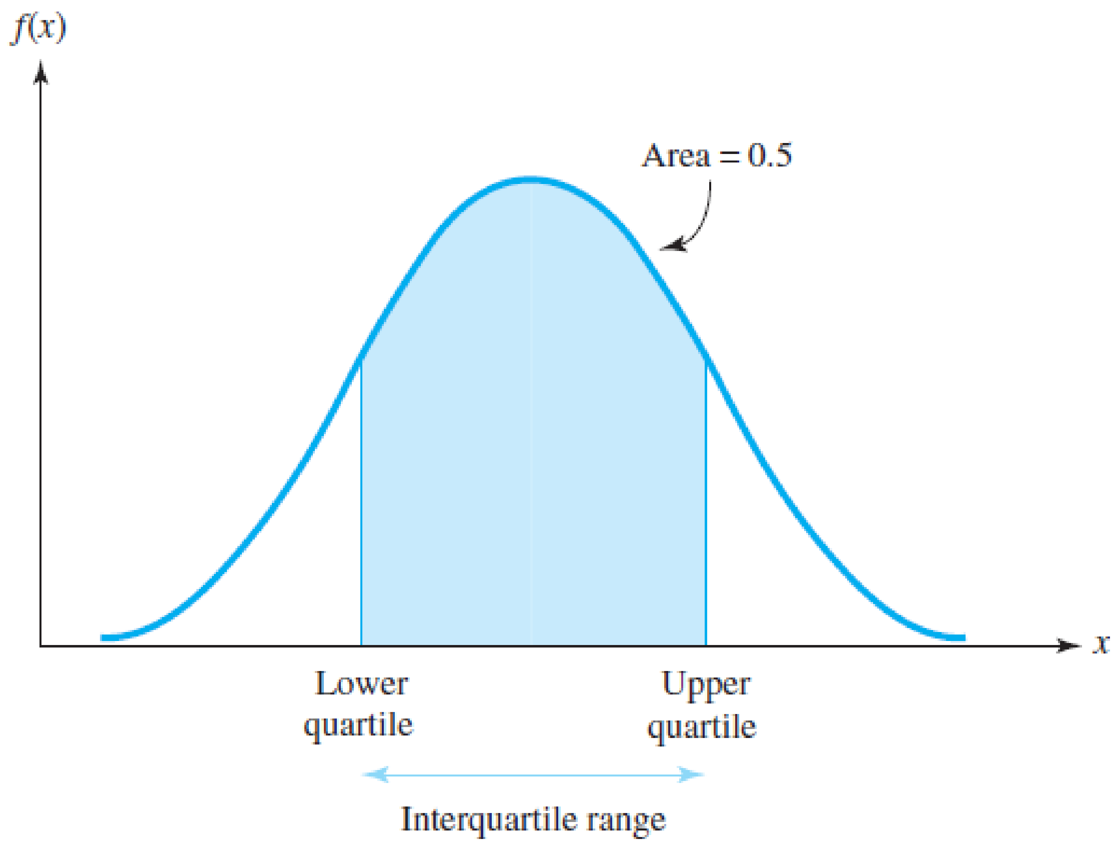
$$F(x) = p$$



Percentiles



Interquartile



Two Random Variables

- How do we measure the relationship between two random variables?
  - Independence
  - Covariance
  - Correlation
  - …

Independence

- Two random variables X and Y are independent if one does not affect the probability of occurrence of the other

Independent Random Variables

Two random variables X and Y are defined to be **independent** if their joint probability mass function or joint probability density function is the *product* of their two marginal distributions. If the random variables are discrete, then they are independent if

$$p_{ij} = p_{i+}p_{+j}$$

for all values of x_i and y_j. If the random variables are continuous, then they are independent if

$$f(x, y) = f_X(x)f_Y(y)$$

for all values of x and y. If two random variables are independent, then the probability distribution of one of the random variables does not depend upon the value taken by the other random variable.

Covariance

- If two random variables are dependent, we should measure the strength of the dependence
- Covariance: indicates the strength of the dependence of two random variables on each other

Covariance: 공분산

Covariance

Covariance

The **covariance** of two random variables X and Y is defined to be

$$\mathrm{Cov}(X, Y) = E((X - E(X))(Y - E(Y))) = E(XY) - E(X)E(Y)$$

The covariance can be any positive or negative number, and independent random variables have a covariance of 0.

Correlation

- More convenient way to assess the strength of the dependence between two random variables
- Correlation takes values between -1 and 1

Correlation: 상관관계

Correlation

Correlation

The **correlation** between two random variables X and Y is defined to be

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

The correlation takes values between -1 and 1, and independent random variables have a correlation of 0.

Functions of Random Variables

- Linear functions of random variable
- Sums of random variables
- Averaging random variables

Linear Functions of RV

- *Example*:

  X: Score of a fair die

  Y: Two times the score of a fair die

- What is the relationship between $E(X)$ and $E(Y)$?

$$X: 1, 2, 3, 4, 5, 6$$

$$Y: 2, 4, 6, 8, 10, 12$$

$$2 \times E(X) = E(Y)$$

Linear Functions of RV

- *Example*:

  X: Score of a fair die

  Z: Two times the score of a fair die plus 10

- What is the relationship between $E(X)$ and $E(Z)$?

$$X: 1, 2, 3, 4, 5, 6$$

$$Y: 12, 14, 16, 18, 20, 22$$

$$2 \times E(X) + 10 = E(Z)$$

Linear Functions of RV

Linear Functions of a Random Variable

If X is a random variable and $Y = aX + b$ for some numbers $a, b \in \mathbb{R}$, then

$$E(Y) = aE(X) + b$$

and

$$\mathrm{Var}(Y) = a^2\mathrm{Var}(X)$$

Sums of RV

Sums of Random Variables

If X_1 and X_2 are two random variables, then

$$E(X_1 + X_2) = E(X_1) + E(X_2)$$

and

$$\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2) + 2\text{Cov}(X_1, X_2)$$

If X_1 and X_2 are independent random variables so that $\text{Cov}(X_1, X_2) = 0$, then

$$\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2).$$

Multiple Random Variables

- What about more than two random variables?

Averaging Independent RVs

Averaging Independent Random Variables

Suppose that $X_1, \ldots, X_n$ is a sequence of independent random variables each with an expectation μ and a variance σ^2, and with an **average**

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}$$

Then

$$E(\bar{X}) = \mu$$

and

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

Chapter 3. Discrete Probability Distributions

Binomial Distribution

Binomial distribution: 이항분포

Binomial Distribution

Consider an experiment consisting of

- n Bernoulli trials
- that are independent and
- that each have a constant probability p of success.

Then the total number of successes X is a random variable that has a **binomial distribution** with parameters n and p, which is written

$$X \sim B(n, p)$$

The probability mass function of a $B(n, p)$ random variable is

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

for $x = 0, 1, \ldots, n$, with

$$E(X) = np \qquad \text{and} \qquad \mathrm{Var}(X) = np(1-p)$$

Binomial Distribution

- *Example*: (Tossing a fair coin n times)

Suppose that a fair coin is tossed n times. The distribution of the number of heads obtained, X, is binomial with parameters n and $p = 0.5$. The expected number of heads obtained is therefore

$$E(X) = np = \frac{n}{2}$$

and the variance is

$$\mathrm{Var}(X) = np(1 - p) = \frac{n}{4}$$

Additional Discrete Distributions

- Geometric distribution:

  The number of trials up to and including the *first* success in a sequence of independent Bernoulli trials with a constant success probability p has a geometric distribution with parameter p

- Negative binomial distribution:

  The number of trials up to and including the rth success in a sequence of independent Bernoulli trials with a constant success probability p has a negative binomial distribution with parameters p and r

- Poisson distribution:

  Often useful to model the number of times that a certain event occurs per unit of time, distance, or volume and the distribution has a single variable λ

Chapter 4. Continuous Probability Distributions

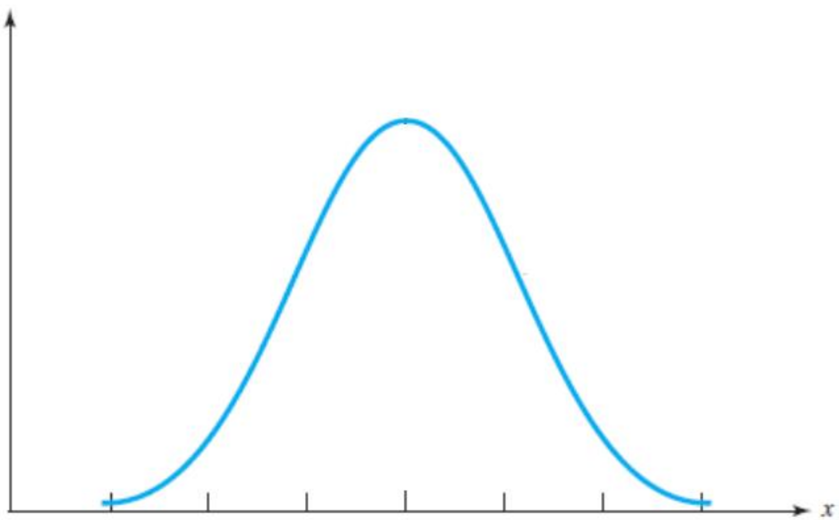
Continuous Distributions

- Uniform distribution
- Exponential distribution
- Normal distribution
- ...

Chapter 5. The Normal Distribution

Normal Distribution

- Normal or Gaussian distribution
- Most important among all continuous distributions
- "Bell-shaped" curve (symmetric)



Normal distribution: 정규분포

Normal Distribution

The Normal Distribution

The **normal** or **Gaussian distribution** has a probability density function

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

for $-\infty \leq x \leq \infty$, depending upon two parameters, the mean and the variance

$$E(X) = \mu \qquad \text{and} \qquad \text{Var}(X) = \sigma^2$$

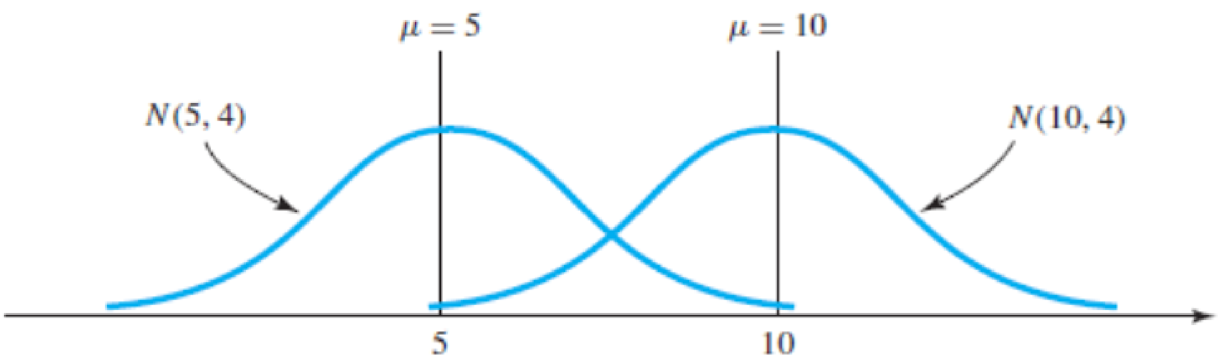
of the distribution. The probability density function is a bell-shaped curve that is symmetric about μ. The notation

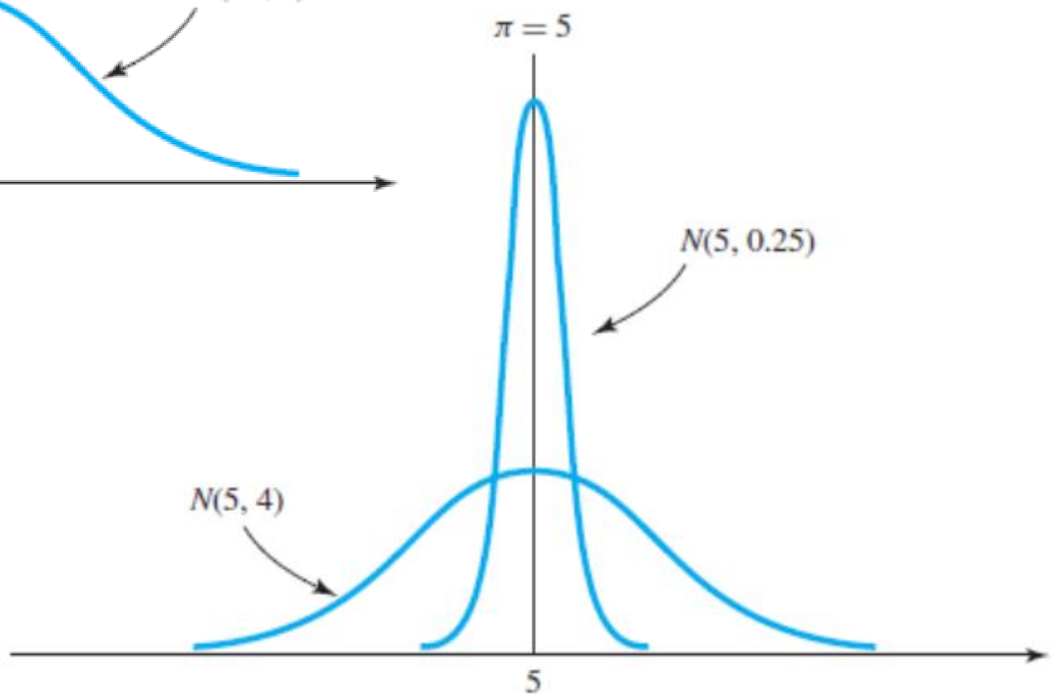
$$X \sim N(\mu, \sigma^2)$$

denotes that the random variable X has a normal distribution with mean μ and variance σ^2. In addition, the random variable X can be referred to as being "normally distributed."

Normal Distribution

- Shape depends on mean and variance



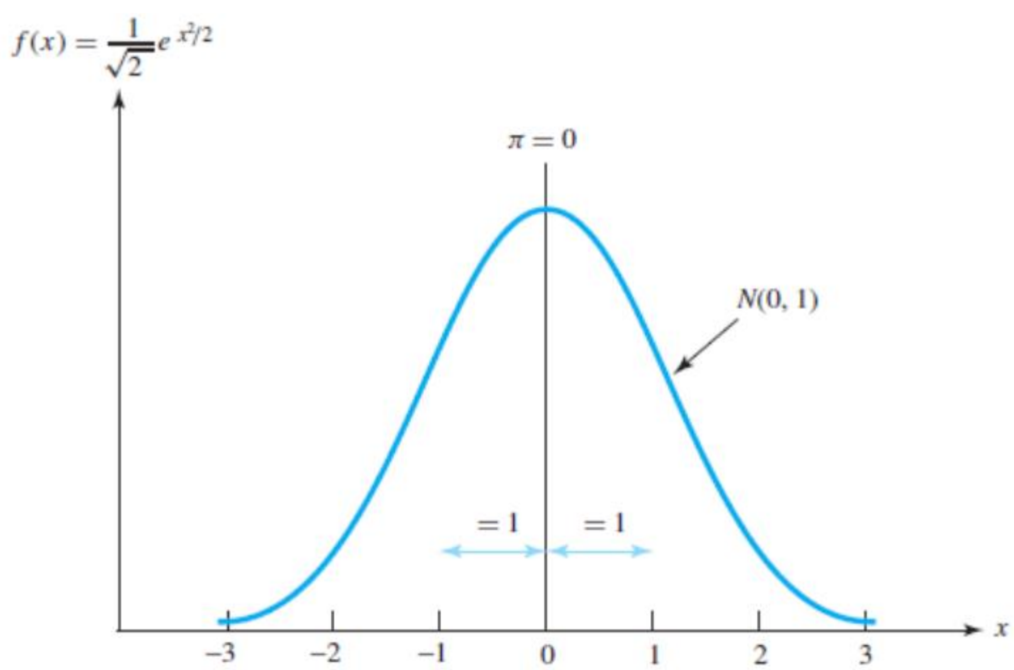


Standard Normal Distribution

A normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$ is known as the **standard normal distribution**. Its probability density function has the notation $\phi(x)$ and is given by

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

for $-\infty \leq x \leq \infty$



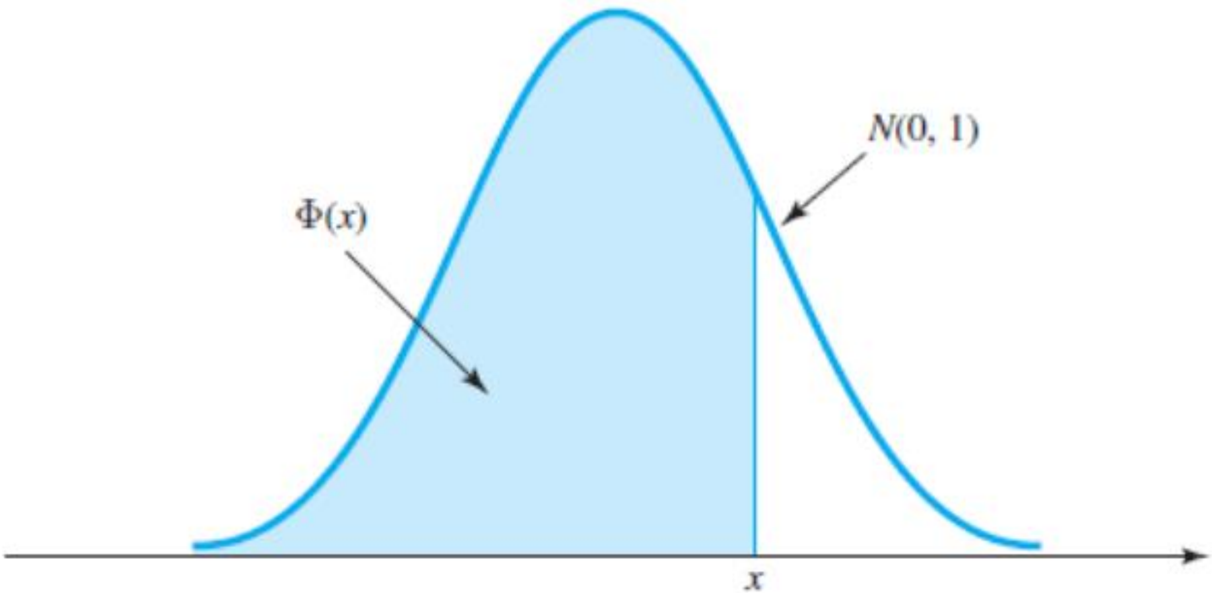
Standard normal distribution: 표준정규분포

Standard Normal Distribution

The notation $\Phi(x)$ is used for the cumulative distribution function of a standard normal distribution, which is calculated from the expression

$$\Phi(x) = \int_{-\infty}^{x} \phi(y)\, dy$$

$\Phi(x)$ is the cumulative distribution function of a standard normal distribution

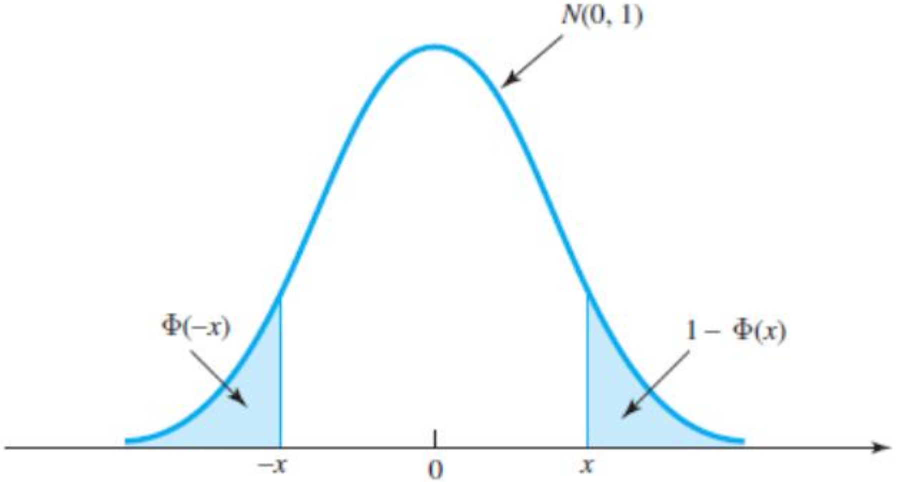


Standard Normal Distribution

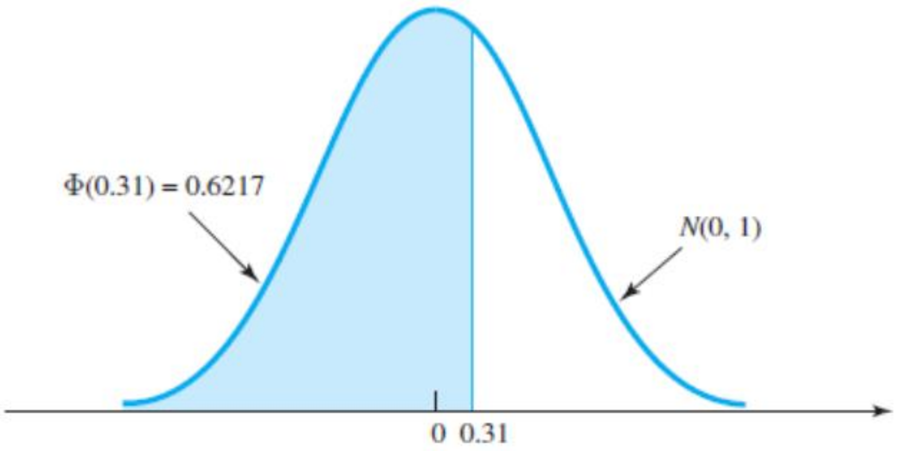
$\Phi(-x) = 1 - \Phi(x)$

$$1 - \Phi(x) = P(Z \geq x) = P(Z \leq -x) = \Phi(-x)$$

$$\Phi(x) + \Phi(-x) = 1$$



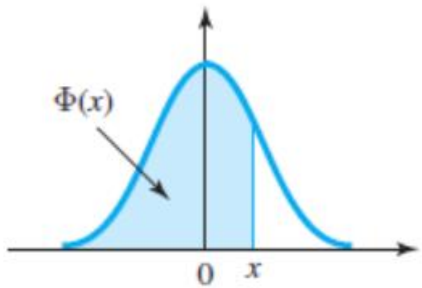
Probability calculations for a standard normal distribution



Standard Normal Distribution

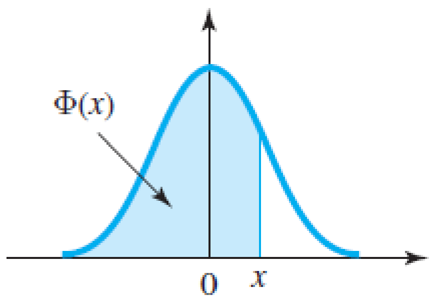
- (From the back of the textbook)

Table I: Cumulative Distribution Function of the Standard Normal Distribution



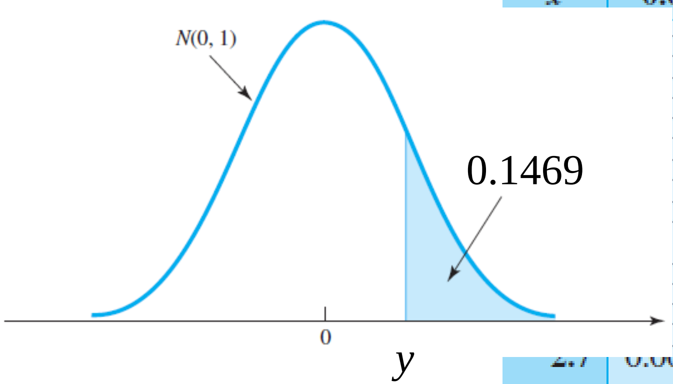
x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0061
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110

Find x where $\Phi(x) = 0.1$ ≈ -1.28



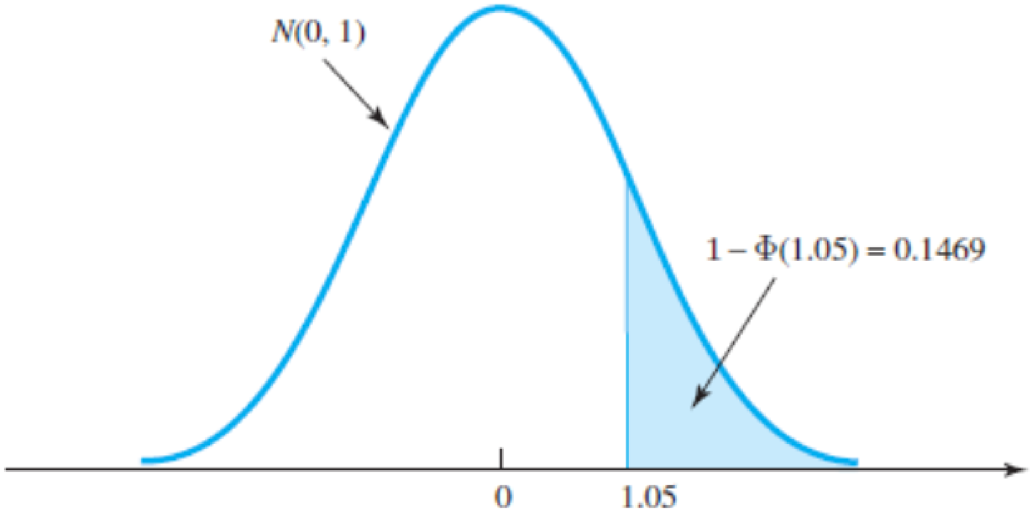
x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0061
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0352	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0722	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379

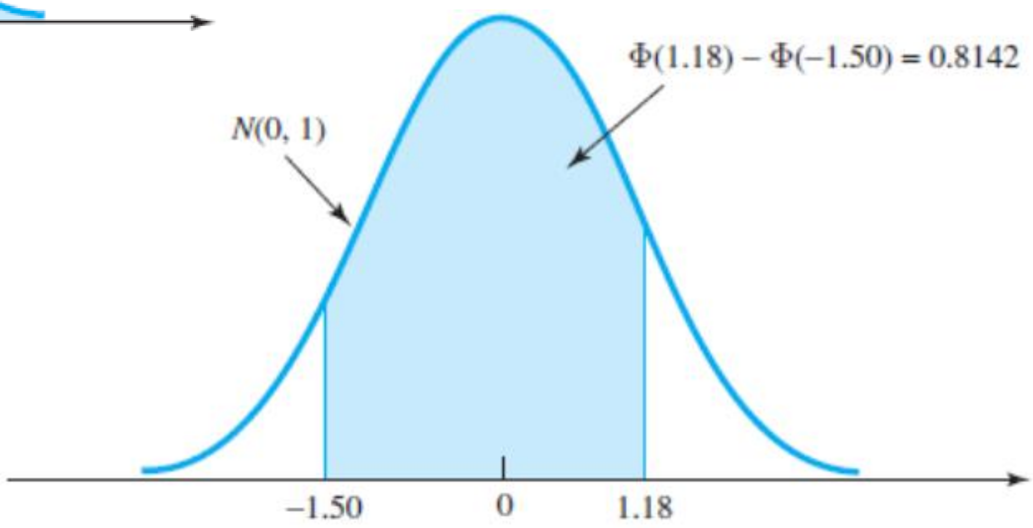
Find y shown below



x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
[illegible]	003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
[illegible]	005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
[illegible]	007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
[illegible]	010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
[illegible]	013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
[illegible]	019	0.0018	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
[illegible]	026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
[illegible]	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0061
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0352	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0722	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	$\boxed{0.1469}$	0.1446	0.1423	0.1401	0.1379

Standard Normal Distribution

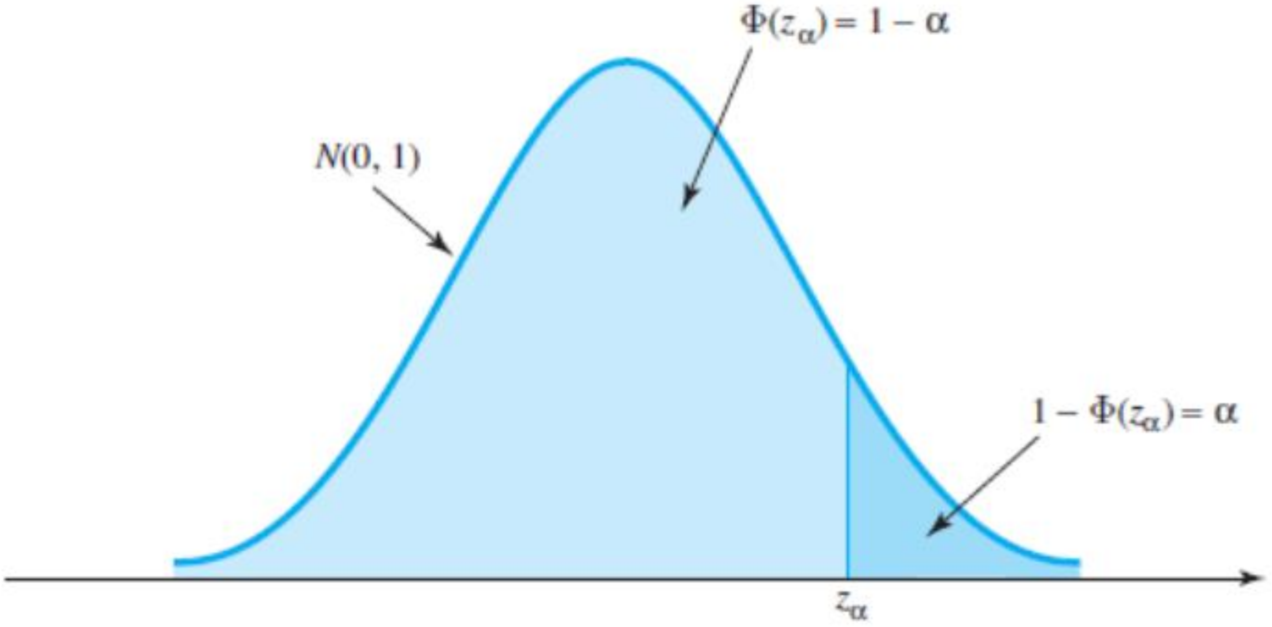


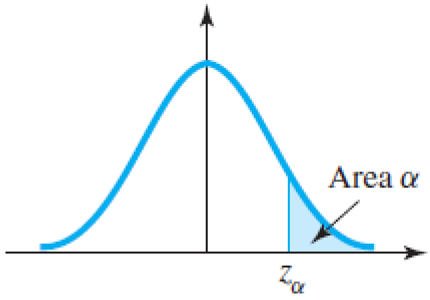


Standard Normal Distribution

- Critical points

$$\Phi(z_\alpha) = 1 - \alpha$$



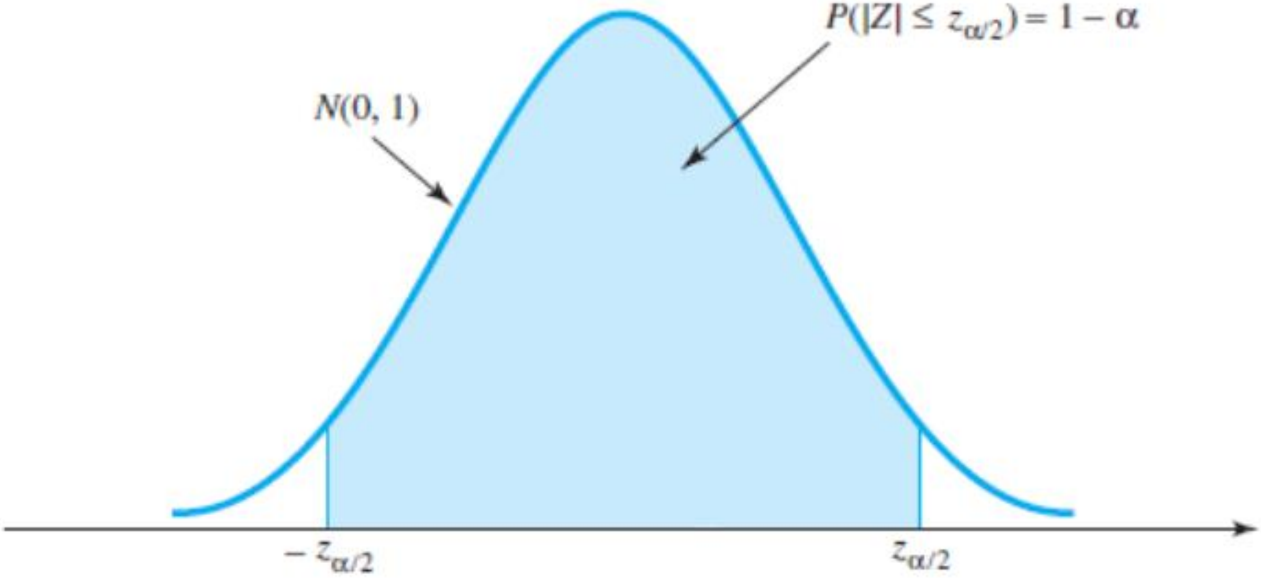


α	z_α
0.10	1.282
0.05	1.645
0.025	1.960
0.01	2.326
0.005	2.576

Standard Normal Distribution

- Critical points

$$\Phi(z_\alpha) = 1 - \alpha$$



Normal Distribution

Probability Calculations for Normal Distributions

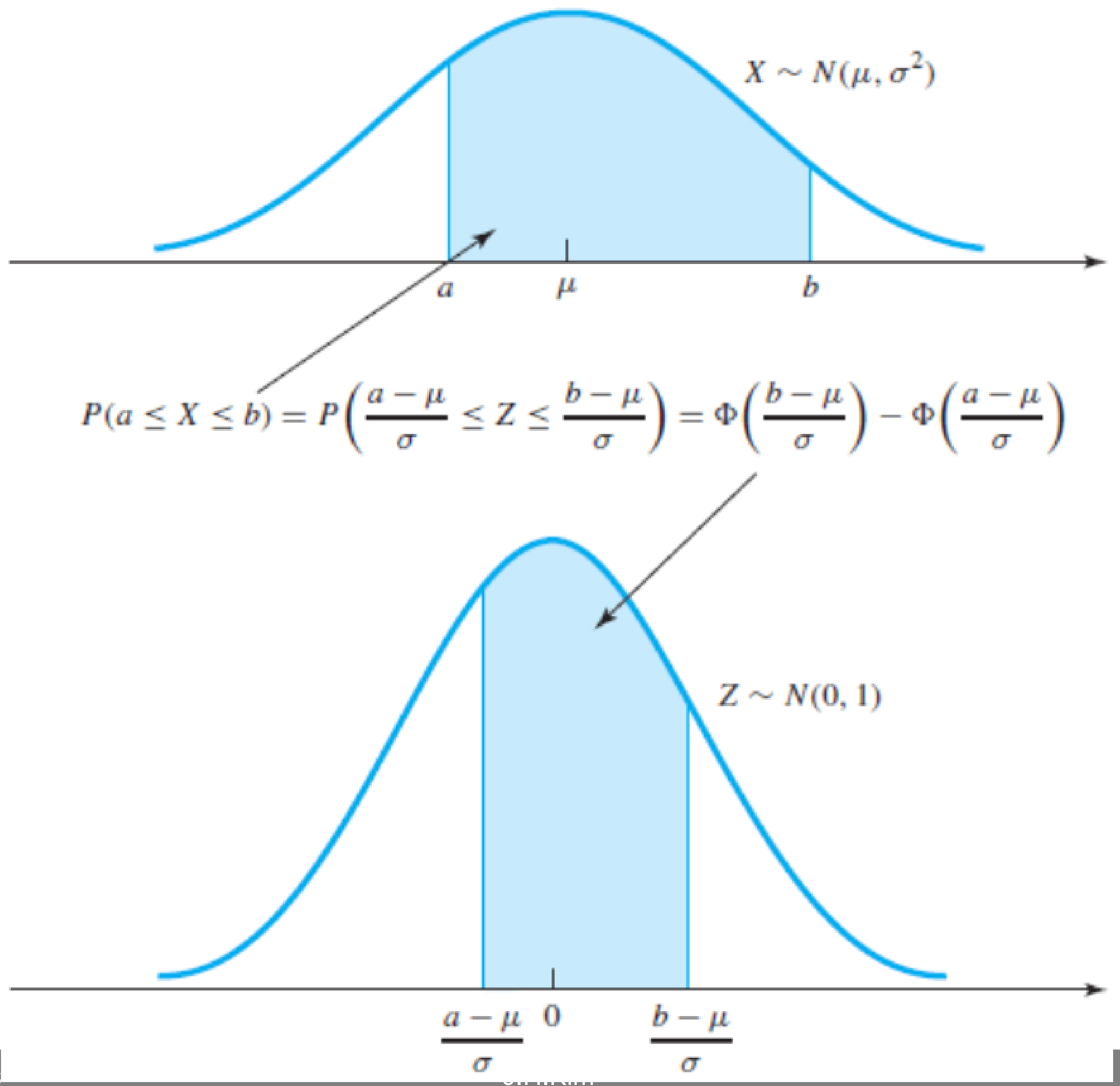
If $X \sim N(\mu, \sigma^2)$, then

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

The random variable Z is known as the "standardized" version of the random variable X. This result implies that the probability values of a general normal distribution can be related to the cumulative distribution function of the standard normal distribution $\Phi(x)$ through the relationship

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

Normal Distribution



Normal Distribution

Normal Random Variables

- There is a probability of about 68% that a normal random variable takes a value within *one* standard deviation of its mean.
- There is a probability of about 95% that a normal random variable takes a value within *two* standard deviations of its mean.
- There is a probability of about 99.7% that a normal random variable takes a value within *three* standard deviations of its mean.

α	z_α
0.10	1.282
0.05	1.645
0.025	1.960
0.01	2.326
0.005	2.576

Normal Random Variables

Averaging Independent Normal Random Variables

If $X_i \sim N(\mu, \sigma^2)$, $1 \leq i \leq n$, are independent random variables, then their average $\bar{X}$ is distributed

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Notice that averaging reduces the variance to σ^2/n, so that the average $\bar{X}$ has a tendency to be closer to the mean value μ than do the individual random variables X_i. This tendency increases as n increases and the average of more and more random variables X_i is taken.

Central Limit Theorem

The Central Limit Theorem

If $X_1, \ldots, X_n$ is a sequence of independent identically distributed random variables with a mean μ and a variance σ^2, then the distribution of their average $\bar{X}$ can be approximated by a

$$N\left(\mu, \frac{\sigma^2}{n}\right)$$

distribution. Similarly, the distribution of the sum $X_1 + \cdots + X_n$ can be approximated by a

$$N(n\mu, n\sigma^2)$$

distribution.

Central limit theorem: 중심극한정리

Distributions Related to Normal

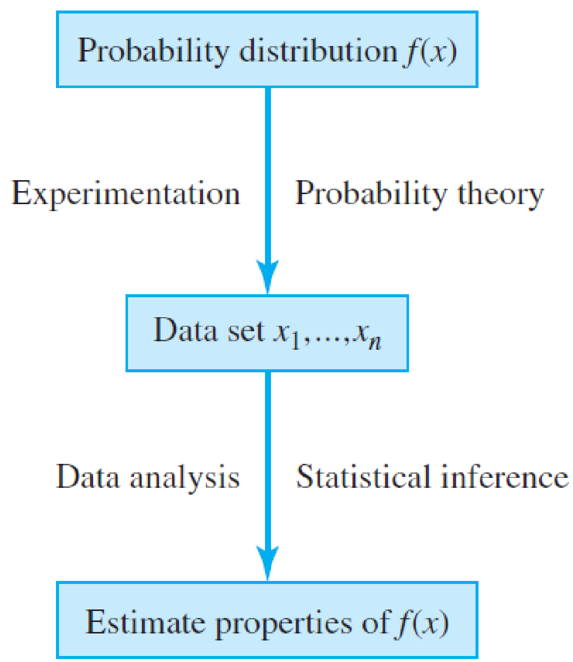
- Lognormal distribution
- Chi-square distribution
- *t*-distribution
- *F*-distribution

(These distributions are used in this course. We will talk more about them later.)

Chapter 6. Descriptive Statistics

Statistical Inference

Of course, in most applications the probability mass function or probability density function of a random variable is not known by an experimenter, and one of the first tasks of the experimenter is to find out as much as possible about the probability distribution of the random variable under consideration. This is done through *experimentation* and the collection of a **data set** relating to the random variable. The science of deducing properties of an underlying probability distribution from such a data set is known as the science of **statistical inference**.



Population and Sample

Populations and Samples

A **population** consists of all possible observations available from a particular probability distribution. A **sample** is a particular subset of the population that an experimenter measures and uses to investigate the unknown probability distribution. A **random sample** is one in which the elements of the sample are chosen at random from the population, and this procedure is often used to ensure that the sample is *representative* of the population.

Population: 모집단, Sample: 표본

Parameter

Parameters

In statistical inference, the term **parameter** is used to denote a quantity θ, say, that is a property of an unknown probability distribution. For example, it may be the mean, variance, or a particular quantile of the probability distribution. Parameters are unknown, and one of the goals of statistical inference is to estimate them.

Parameter: 모수

Statistic

Statistics

In statistical inference, the term **statistic** is used to denote a quantity that is a property of a sample. For example, it may be a sample mean, a sample variance, or a particular sample quantile. Statistics are random variables whose observed values can be calculated from a set of observed data observations. Statistics can be used to estimate unknown parameters.

Statistic: 통계량

Sample Statistics

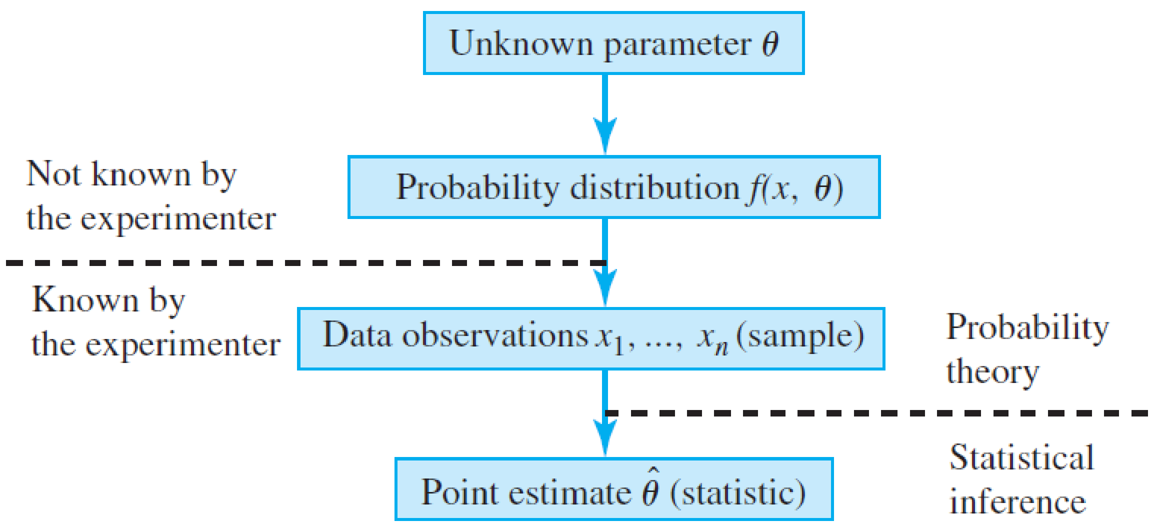
- **Sample mean**: $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$
- Sample median: Value of the *middle* of the ordered data points
- Sample trimmed mean: Obtained by deleting some of the largest and some of the smallest data (e.g., top 10% and bottom 10%)
- **Sample variance**: $s^2 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$

Chapter 7. Statistical Estimation and Sampling Distributions

Point Estimates

Point Estimates of Parameters

A **point estimate** of an unknown parameter θ is a statistic $\hat{\theta}$ that represents a "best guess" at the value of θ. There may be more than one sensible point estimate of a parameter.



Point estimate: 점추정치

The statistic $\hat{\theta}$ is the "best guess" of the parameter θ.

Unbiased Estimates

- *Example*: If we want a point estimate for mean, which one is better?

Sample mean

or

Sample median

*Need to consider **unbiasedness** and **variance** of the estimate*

Unbiased Estimates

Unbiased and Biased Point Estimates

A point estimate $\hat{\theta}$ for a parameter θ is said to be **unbiased** if

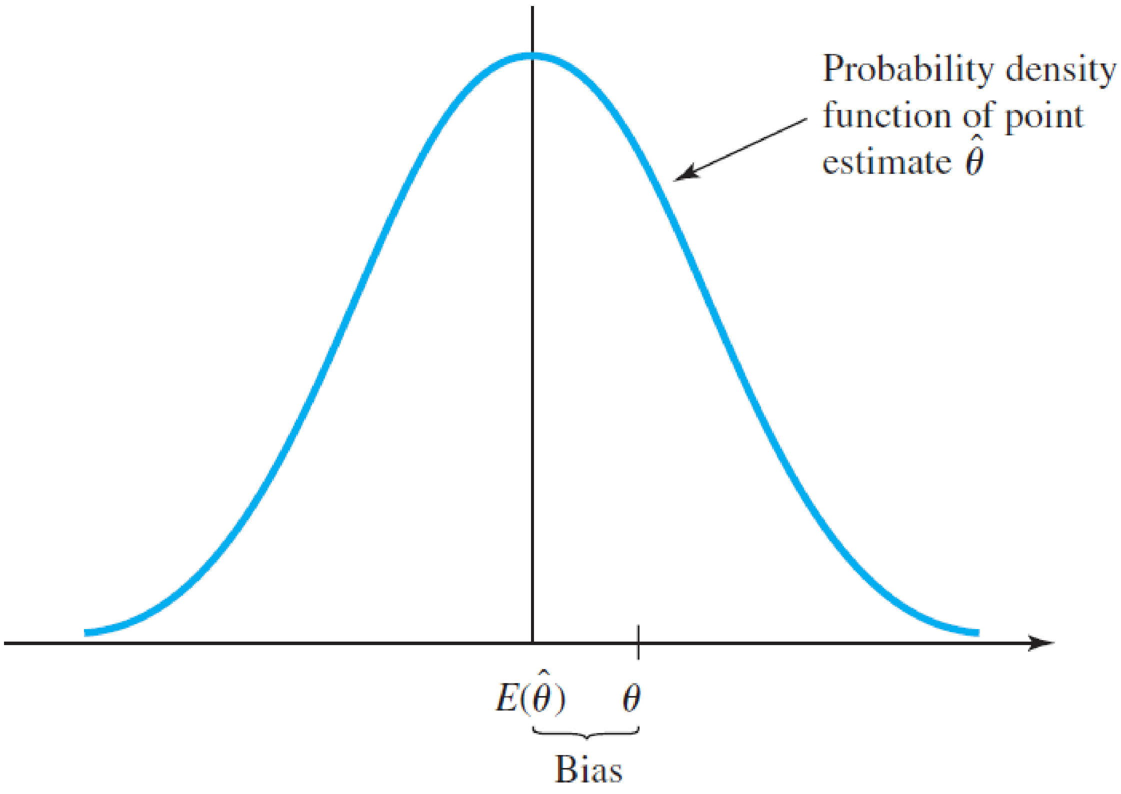
$$E(\hat{\theta}) = \theta$$

Unbiasedness is a good property for a point estimate to possess. If a point estimate is not unbiased, then its **bias** can be defined to be

$$\text{bias} = E(\hat{\theta}) - \theta$$

All other things being equal, the smaller the absolute value of the bias of a point estimate, the better.

Unbiased Estimates



Unbiased Estimates

- Is sample mean unbiased?

Now suppose that $X_1, \ldots, X_n$ is a sample of observations from a probability distribution with a mean μ and a variance σ^2. Is the sample mean

$$\hat{\mu} = \bar{X}$$

an unbiased point estimate of the population mean μ? Clearly it is since

$$E(X_i) = \mu, \quad 1 \leq i \leq n$$

so that

$$E(\hat{\mu}) = E(\bar{X}) = \frac{1}{n} E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu$$

Unbiased Estimates

- Is sample variance unbiased?

The sample variance S^2 is an unbiased estimate of the population variance σ^2. This is because

$$
\begin{aligned}
E(S^2) &= \frac{1}{n-1} E\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right) \\
&= \frac{1}{n-1} E\left(\sum_{i=1}^{n}((X_i - \mu) - (\bar{X} - \mu))^2\right) \\
&= \frac{1}{n-1} E\left(\sum_{i=1}^{n}(X_i - \mu)^2 - 2(\bar{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) + n(\bar{X} - \mu)^2\right)
\end{aligned}
$$

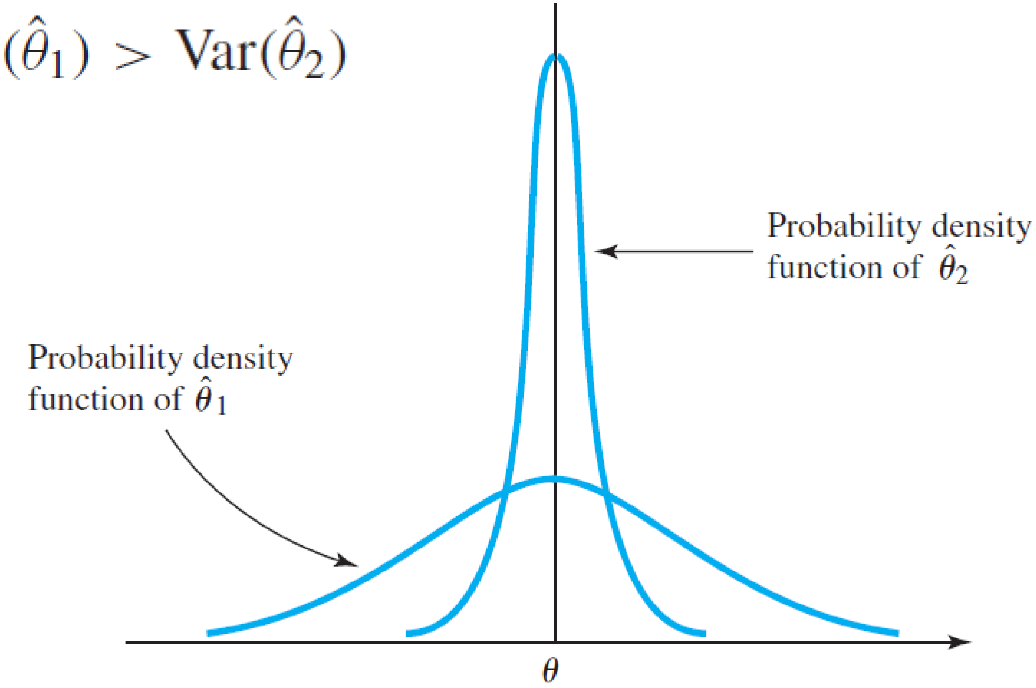
$$\vdots \quad \Rightarrow \quad E(S^2) = \frac{1}{n-1}\left(\sum_{i=1}^{n}\sigma^2 - n\left(\frac{\sigma^2}{n}\right)\right) = \sigma^2$$

so that S^2 is indeed an unbiased estimate of σ^2.

Minimum Variance Estimates

As well as looking at the expectation $E(\hat{\theta})$ of a point estimate $\hat{\theta}$, it is important to consider the variance $\mathrm{Var}(\hat{\theta})$ of the point estimate. It is generally desirable to have unbiased point estimates with as small a variance as possible.

$$\mathrm{Var}(\hat{\theta}_1) > \mathrm{Var}(\hat{\theta}_2)$$



Minimum Variance Estimates

- Minimum variance unbiased estimate

The best possible situation is to be able to construct a point estimate that is unbiased and that also has the smallest possible variance. An unbiased point estimate that has a smaller variance than any other unbiased point estimate is called a **minimum variance unbiased estimate** (MVUE).

Sampling Distributions

- **Sampling distribution** of a statistic is the probability distribution of that statistic
- Standard deviation of the statistic is called the **standard error**
  - Smaller values indicate that the point estimate is likely to be more accurate because its variability is smaller

Sampling distribution: 표본분포, Standard error: 표준오차

Sampling Distributions

- Sampling distribution of the sample mean:

Sample Mean

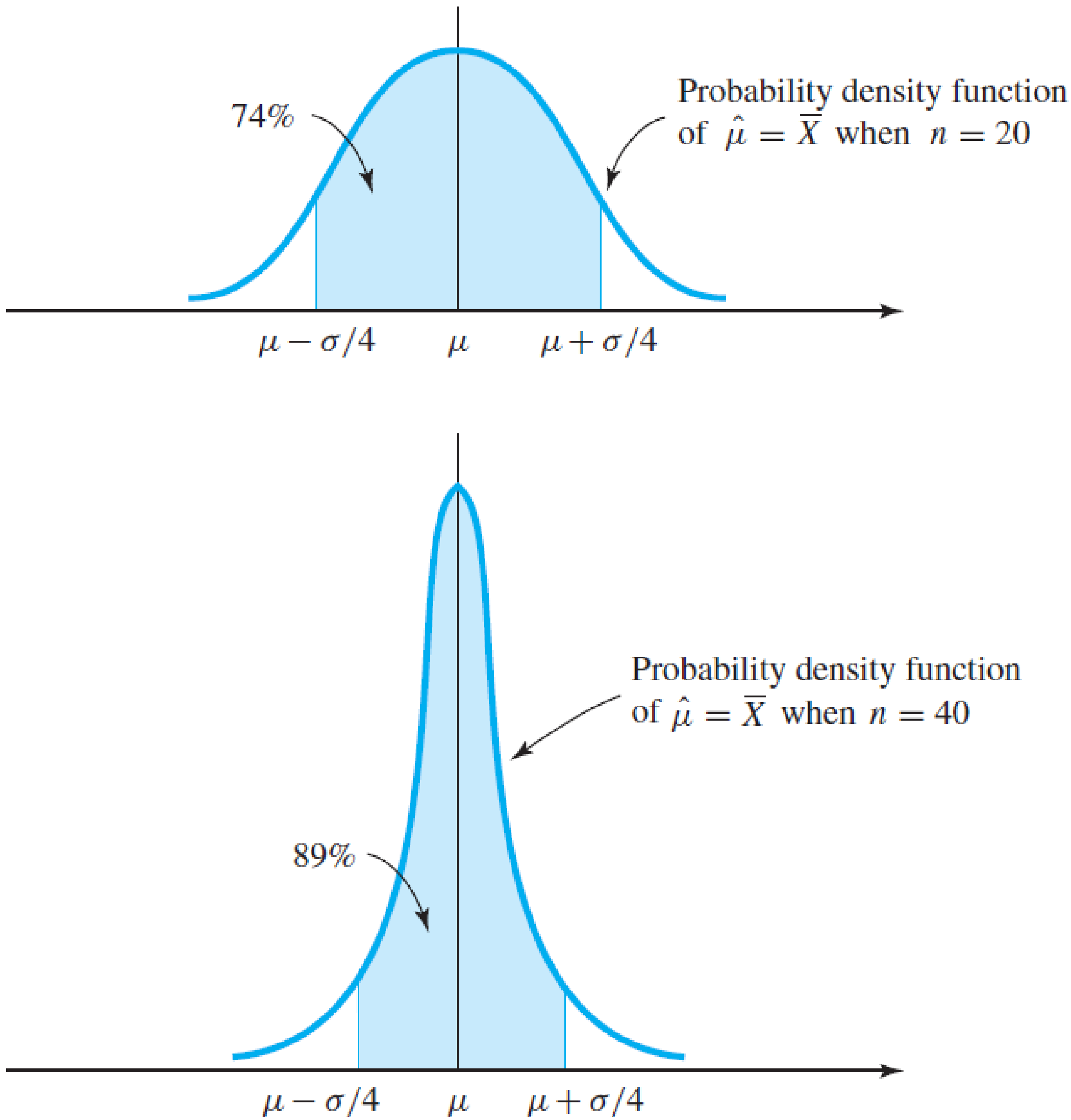
If $X_1, \ldots, X_n$ are observations from a population with a mean μ and a variance σ^2, then the central limit theorem indicates that the sample mean $\hat{\mu} = \bar{X}$ has the approximate distribution

$$\hat{\mu} = \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

The standard error of the sample mean is

$$\text{s.e.}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

which again is inversely proportional to the square root of the sample size. Thus, if the sample size is doubled, the standard error is reduced by a factor of $1/\sqrt{2} = 0.71$. Similarly, in order to *halve* the standard error, the sample size needs to be multiplied by *four*.


74%
Probability density function
of $\hat{\mu} = \overline{X}$ when $n = 20$
$\mu - \sigma/4$
μ
$\mu + \sigma/4$
Probability density function
of $\hat{\mu} = \overline{X}$ when $n = 40$
89%
$\mu - \sigma/4$
μ
$\mu + \sigma/4$

Sampling Distributions

- Sampling distribution of the sample variance:

For a sample $X_1, \ldots, X_n$ obtained from a population with a mean μ and a variance σ^2, consider the variance estimate

$$\hat{\sigma}^2 = S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

Sample Variance

If $X_1, \ldots, X_n$ are normally distributed with a mean μ and a variance σ^2, then the sample variance S^2 has the distribution

$$S^2 \sim \sigma^2 \frac{\chi^2_{n-1}}{(n-1)}$$

Sampling Distributions

- Distribution of the sample variance is very important for the problem of estimating a normal population *mean*
- Standard error of the sample mean requires the unknown variance σ^2, but the sample variance S^2 can be used instead

Sampling Distributions

- How?

The elimination of the unknown variance σ^2 is accomplished as follows using the t-distribution. The distribution of the sample mean

$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ can be rearranged as $\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu) \sim N(0, 1)$

Also, notice that

$\frac{S}{\sigma} \sim \sqrt{\frac{\chi^2_{n-1}}{(n-1)}}$ so that $\frac{\sqrt{n}(\bar{X} - \mu)}{S} = \frac{\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu)}{\left(\frac{S}{\sigma}\right)} \sim \frac{N(0, 1)}{\sqrt{\frac{\chi^2_{n-1}}{(n-1)}}} \sim t_{n-1}$

t-statistic

If $X_1, \ldots, X_n$ are normally distributed with a mean μ, then

$$\frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t_{n-1}$$

Sampling Distributions

t-statistic

If $X_1, \ldots, X_n$ are normally distributed with a mean μ, then

$$\frac{\sqrt{n}(\bar{X}-\mu)}{S} \sim t_{n-1}$$

This result is very important since in practice an experimenter knows the values of n and the observed sample mean $\bar{x}$ and sample variance s^2, and so knows everything in the quantity

$$\frac{\sqrt{n}(\bar{x}-\mu)}{s}$$

except for μ. This allows the experimenter to make useful inferences about μ.

The Chi-Square Distribution

A **chi-square** random variable with ν *degrees of freedom*, X, can be generated as

$$X = X_1^2 + \cdots + X_\nu^2$$

where the X_i are independent standard normal random variables. A chi-square distribution with ν degrees of freedom is a gamma distribution with parameter values $\lambda = 1/2$ and $k = \nu/2$, and it has an expectation of ν and a variance of 2ν.

The t-distribution

A ***t*-distribution** with ν degrees of freedom is defined to be

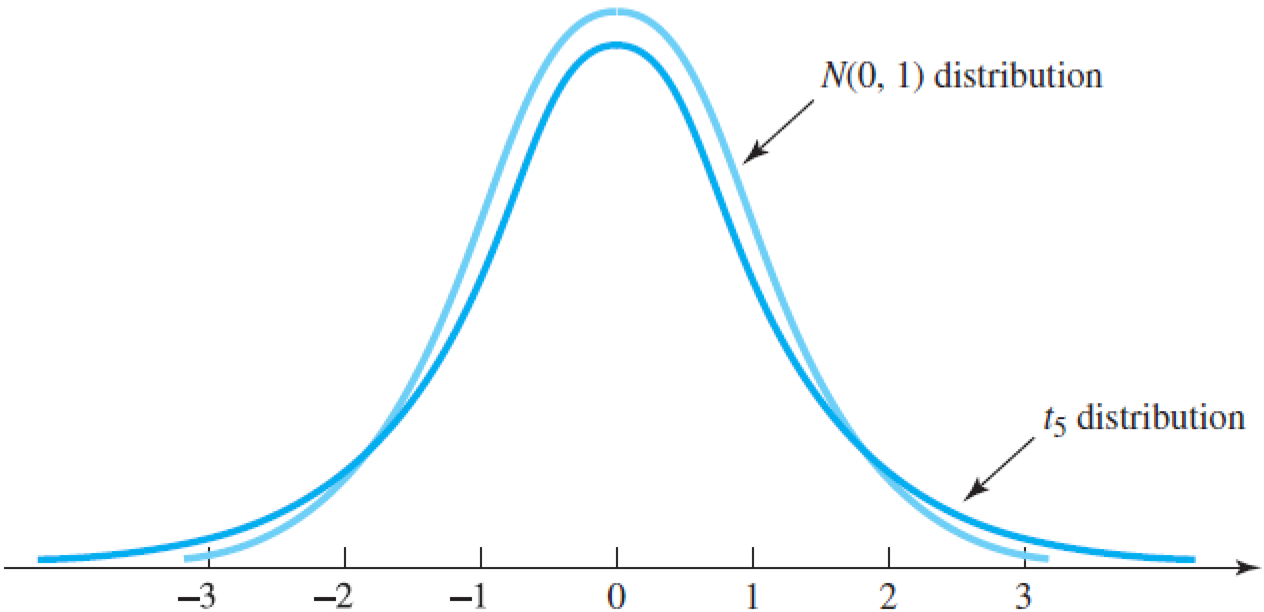
$$t_\nu \sim \frac{N(0,1)}{\sqrt{\chi_\nu^2/\nu}}$$

where the $N(0, 1)$ and χ_ν^2 random variables are independently distributed. The t-distribution has a shape similar to a standard normal distribution but is a little flatter. As $\nu \to \infty$, the t-distribution tends to a standard normal distribution.

t-Distributions

FIGURE 5.38

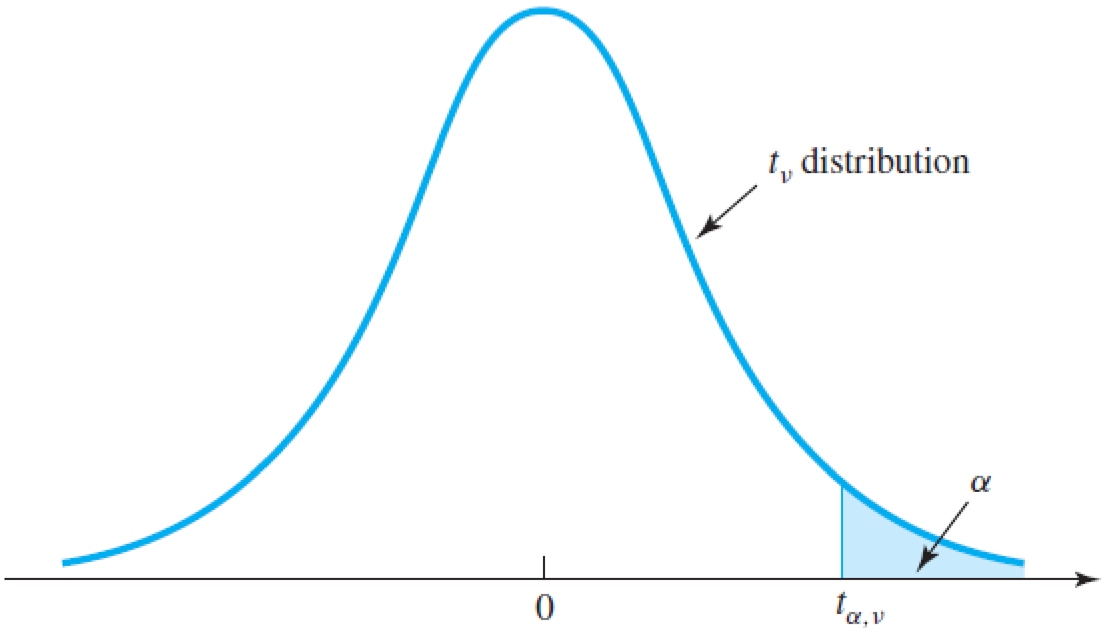
Comparison of a t-distribution and the standard normal distribution



t-Distributions

FIGURE 5.39

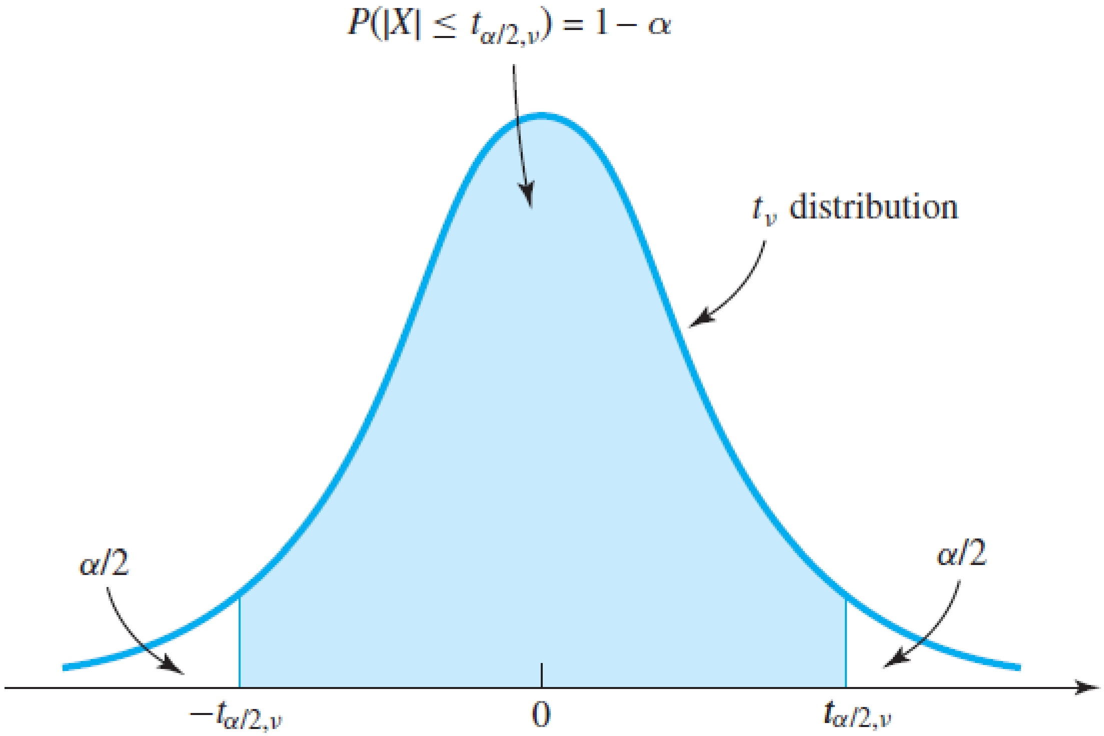
The critical points $t_{\alpha,\nu}$ of the t-distribution



t-Distributions

FIGURE 5.40

The critical points $t_{\alpha/2,\nu}$ of the t-distribution



$$P(|X| \leq t_{\alpha/2,\nu}) = P(-t_{\alpha/2,\nu} \leq X \leq t_{\alpha/2,\nu}) = 1-\alpha$$

Microsoft Excel
Google Spreadsheet

Excel

- Expectation (mean)

  =AVERAGE( )

- Median

  =MEDIAN( )

- Variance

  =VAR.P( )

  =VAR.S( )

- Standard deviation

  =STDEV.P( )

  =STDEV.S( )

- Percentiles

  =PERCENTILE.INC( )

Excel

- Covariance

  =COVARIANCE.P( )

  =COVARIANCE.S( )

- Correlation

  =CORREL( )

Excel

- Normal distribution – Critical points

  =NORM.S.INV( )

  =NORM.INV( )

Excel

- Cell reference
- Cell reference using $
- Drag
- Copy-paste (special paste)
- Ctrl+Arrow
- Format cells (number)
- Format cells (color, outline, etc.)
- Excel formulas
- Sheets