7. Sampling Distributions

Adams Wai Kin Kong
School of Computer Science and Engineering
Nanyang Technological University, Singapore
adamskong@ieee.org

7.2 Statistics and Sampling Distributions

Statistics and Sampling Distributions (1 of 4)

When you select a random sample from a population, the numerical descriptive measures you calculate from the sample are called **statistics (e.g., sample mean and sample variance)**.

These statistics vary or change for each different random sample you select; that is, they are *random variables*.

Statistics and Sampling Distributions (2 of 4)

The probability distributions for statistics are called **sampling distributions** because, in repeated sampling, they tell us:

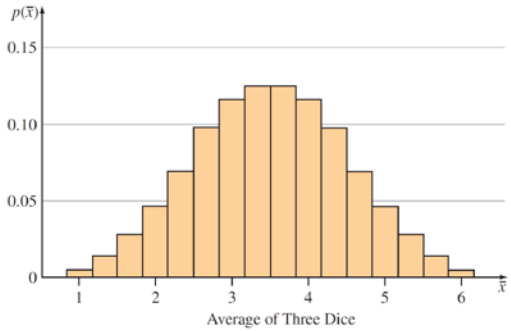
- What values of the statistic can occur.
- How often each value occurs.

Definition:

The **sampling distribution of a statistic** is the probability distribution for the possible values of the statistic that results when random samples of size *n* are repeatedly drawn from the population.

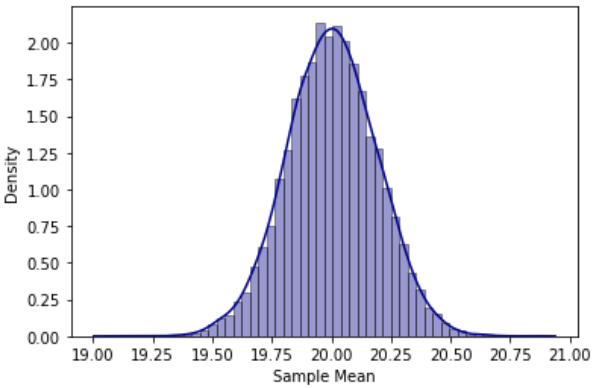
Statistics and Sampling Distributions (3 of 4)

- Three ways to find the sampling distribution of a statistic:



1. Derive the distribution *mathematically* using the laws of probability.

2. Use a *simulation* to approximate the distribution. That is, draw a large number of samples of size *n*, calculating the value of the statistic for each sample, and tabulate the results in a relative frequency histogram. When the number of samples is large, the histogram will be very close to the theoretical sampling distribution.



Statistics and Sampling Distributions (4 of 4)

3. Use *statistical theorems* to derive exact or approximate sampling distributions.

Example 7.3 – Solution (1 of 6)

We are sampling from the population shown in figure.

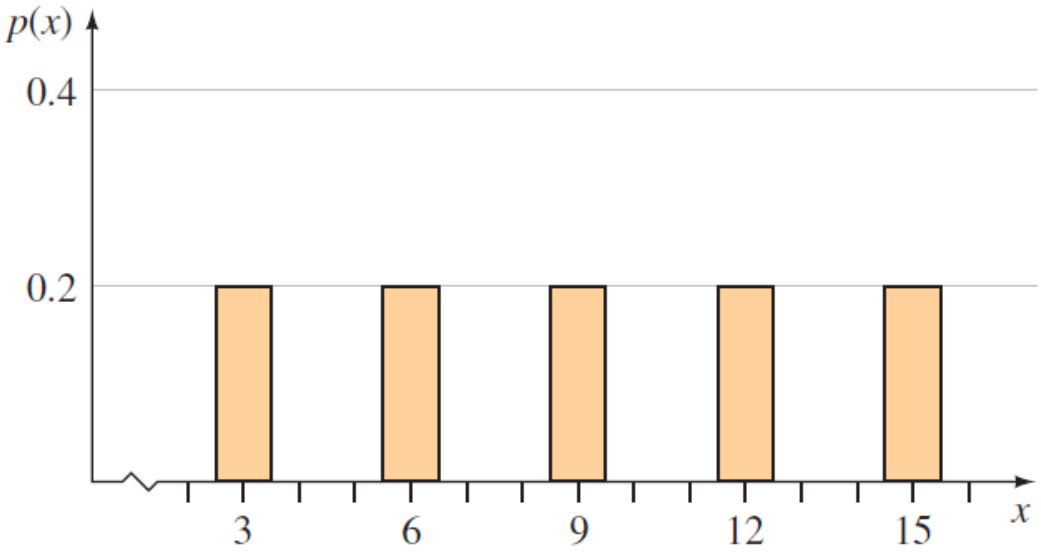


Figure 7.1

Example 7.3 – Solution (2 of 6)

It contains five distinct numbers and each is equally likely, with probability $p(x) = 1/5$. We can easily find the population mean and median as

$$\mu = \frac{3+6+9+12+15}{5} = 9 \quad \text{and} \quad M = 9$$

To find the sampling distribution, we need to know what values of $\bar{x}$ and m can occur when the sample is taken.

Example 7.3 – Solution (3 of 6)

There are $C_3^5 = 10$ possible random samples and each is equally likely, with probability 1/10. These samples, along with the calculated values of $\bar{x}$ and m for each, are listed in table.

Sample	Sample Values	$\bar{x}$	m
1	3, 6, 9	6	6
2	3, 6, 12	7	6
3	3, 6, 15	8	6
4	3, 9, 12	8	9
5	3, 9, 15	9	9
6	3, 12, 15	10	12
7	6, 9, 12	9	9
8	6, 9, 15	10	9
9	6, 12, 15	11	12
10	9, 12, 15	12	12

statistics

Values of $\bar{x}$ and m for Simple Random Sampling when n = 3 and N = 5

Table 7.3

Example 7.3 – Solution (4 of 6)

You will notice that some values of $\bar{x}$ are more likely than others because they occur in more than one sample. For example,

Number of Occurrence

$$P(\bar{x} = 8) = \frac{2}{10} = .2 \quad \text{and} \quad P(m = 6) = \frac{3}{10} = .3$$

Example 7.3 – Solution (5 of 6)

Using the values in Table 7.3, we can find the sampling distribution of $\bar{x}$ and m, shown in table and graphed in Figure 7.2 shown on the next slide.

$\bar{x}$	$p(\bar{x})$
6	.1
7	.1
8	.2
9	.2
10	.2
11	.1
12	.1

(a)

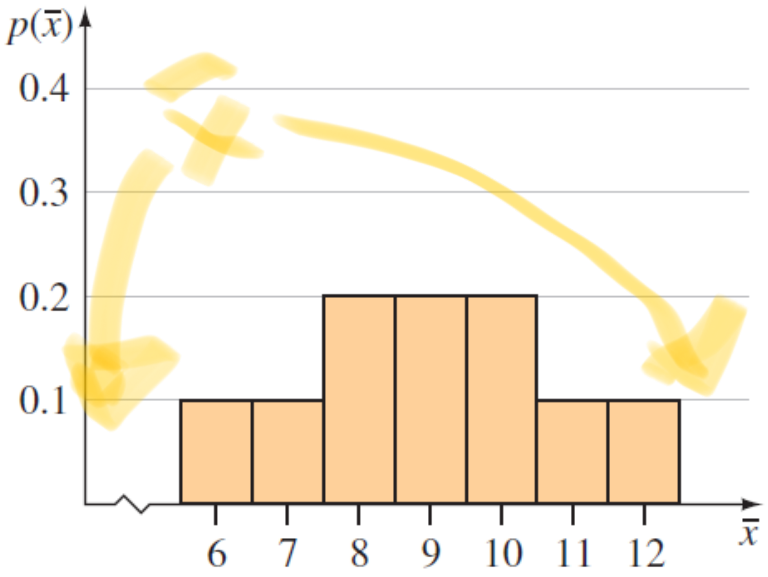
m	$p(m)$
6	.3
9	.4
12	.3

(b)

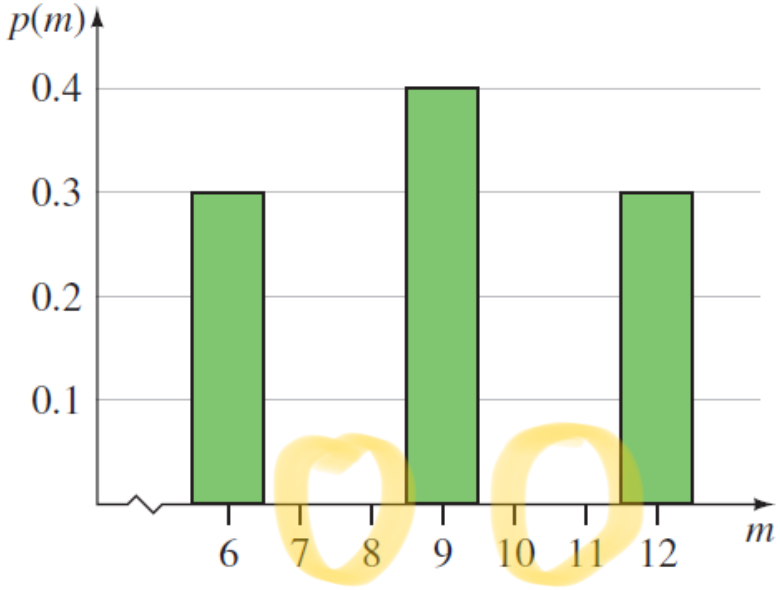
Sampling Distributions for (a) the Sample Mean and (b) the Sample Median

Table 7.4

Example 7.3 – Solution (6 of 6)



Sampling distribution of mean



Sampling distribution of median

Figure 7.2

The Central Limit Theorem and the Sample Mean

The Central Limit Theorem and the Sample Mean (1 of 1)

One important statistical theorem that describes the sampling distribution of statistics that are sums or averages is the *Central Limit Theorem*.

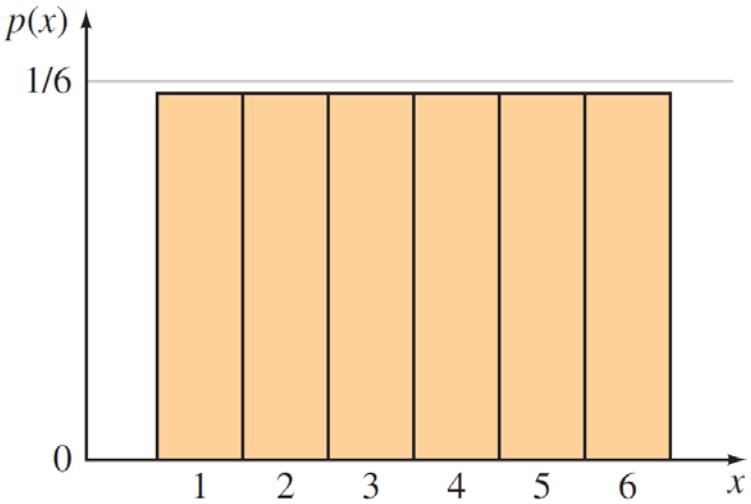
The Central Limit Theorem (1 of 12)

Under rather general conditions, this theorem states that sums and means of random samples of measurements drawn from a population tend to have an approximately normal distribution.

For example, suppose you toss a balanced die $n = 1$ time. The random variable x is the number observed on the upper face.

The Central Limit Theorem (2 of 12)

This familiar random variable can take six values, each with probability 1/6, and its probability distribution is shown in figure.



Probability distribution for x, the number appearing on a single toss of a die

The Central Limit Theorem (3 of 12)

The shape of the distribution is *flat*—generally called a *discrete uniform distribution*—and is symmetric about the mean $\mu = 3.5$, with a standard deviation $\sigma = 1.71$.

Now, take a sample of size $n = 2$ from this population; that is, toss two dice and record the sum of the numbers on the two upper faces, $\Sigma x_i = x_1 + x_2$.

The Central Limit Theorem (4 of 12)

Table shows the 36 possible outcomes, each with probability 1/36.

	First Die					
Second Die	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Sums of the Upper Faces of Two Dice
Table 7.5(a)

The sums are tabulated, and each of the possible sums is divided by $n = 2$ to obtain an average.

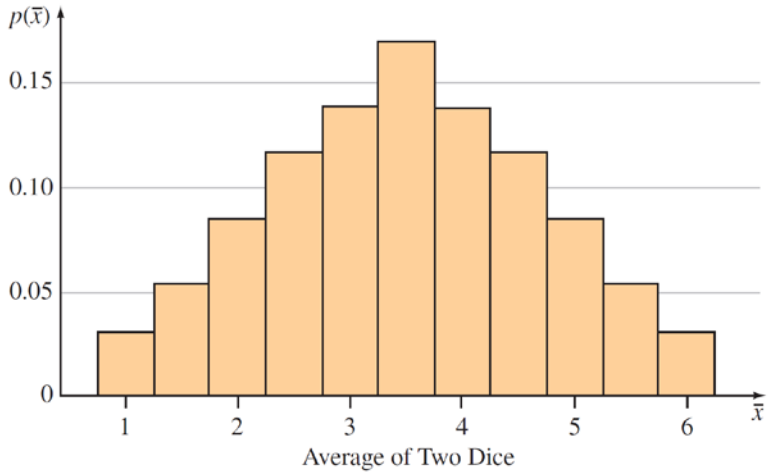
The Central Limit Theorem (5 of 12)

When all of the 36 possible averages are consolidated into a statistical table, the result is the **sampling distribution** of $\bar{x} = \sum x_i / n$ shown in table and graphed in figure.

$\bar{x}$	$p(\bar{x})$	$\bar{x}$	$p(\bar{x})$
2/2 = 1	1/36	8/2 = 4	5/36
3/2 = 1.5	2/36	9/2 = 4.5	4/36
4/2 = 2	3/36	10/2 = 5	3/36
5/2 = 2.5	4/36	11/2 = 5.5	2/36
6/2 = 3	5/36	12/2=6	1/36
7/2 = 3.5	6/36		

Sampling Distribution of $\bar{x}$

Table 7.5(b)



Sampling distribution of $\bar{x}$ for n = 2 dice

Figure 7.4

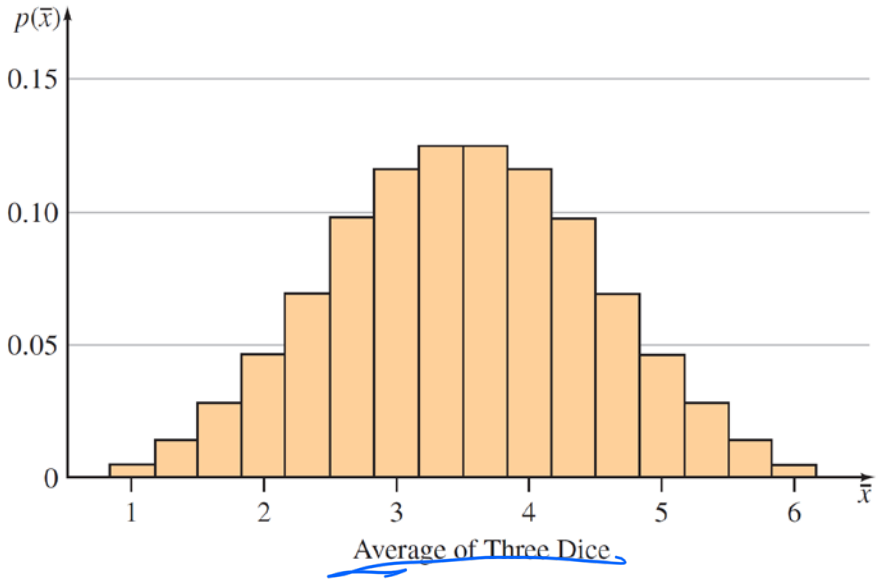
The Central Limit Theorem (6 of 12)

Notice the dramatic difference in the shape of the sampling distribution. It is now roughly mound-shaped but still symmetric about the mean $\mu = 3.5$.

Using a similar procedure, we generated the sampling distributions of $\bar{x}$ when $n = 3$ and $n = 4$.

The Central Limit Theorem (7 of 12)

For n = 3, the sampling distribution in figure clearly shows the mound shape of the normal probability distribution, still centered at μ = 3.5.



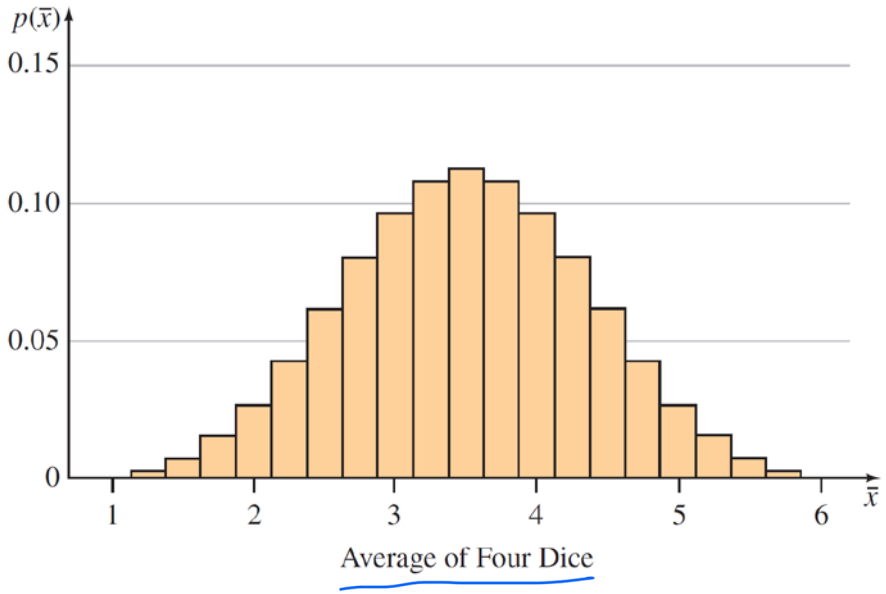
Sampling distribution of $\bar{x}$ for n=3 dice

Figure 7.5

Notice also that the spread of the distribution is slowly *decreasing* as the sample size *n increases*.

The Central Limit Theorem (8 of 12)

Figure dramatically shows that the distribution of $\bar{x}$ is approximately normally distributed based on a sample as small as $n = 4$.



Sampling distribution of $\bar{x}$ for n=4 dice

Figure 7.6

This phenomenon is the result of an important statistical theorem called the **Central Limit Theorem (CLT)**.

The Central Limit Theorem (10 of 12)

The Central Limit Theorem can be restated to apply to the **sum of the sample measurements** Σx_i, which, as n becomes large, also has an approximately normal distribution with mean $n\mu$ and standard deviation $\sigma\sqrt{n}$.

The Central Limit Theorem can be restated to apply to the **mean of the sample measurements** $\bar{x}$, which, as n becomes large, also has an approximately normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.

The Central Limit Theorem (11 of 12)

When the Sample Size Is Large Enough to Use the Central Limit Theorem?

- If the sampled population is **normal**, then the sampling distribution of $\bar{x}$ will also be normal, no matter what sample size you choose.
- When the sampled population is approximately symmetric, the sampling distribution of $\bar{x}$ becomes approximately normal for relatively small values of *n*.
- When the sampled population is **skewed**, the sample size *n* must be larger, with *n* at least 30 before the sampling distribution of $\bar{x}$ becomes approximately normal. Conservatively, we require $n \geq 30$.

The Central Limit Theorem (12 of 12)

- These guidelines suggest that, for many populations, the sampling distribution of $\bar{x}$ will be approximately normal for moderate sample sizes, but as specific applications of the Central Limit Theorem arise, we will give you the appropriate sample size n.

Standard Error of the Sample Mean

Standard Error of the Sample Mean (1 of 3)

Definition

The **standard deviation** of a statistic is also called the **standard error of the estimator** (abbreviated **SE**) because it refers to the precision of the estimator. Therefore, the standard deviation of $\bar{x}$ − given by $\sigma/\sqrt{n}$ − is referred to as the **standard error of the mean** (abbreviated as SE($\bar{x}$) SEM, or sometimes just SE

Standard Error of the Sample Mean (2 of 3)

How to Calculate Probabilities for the Sample Mean $\bar{x}$. (assuming $\bar{x}$ is *normal* or *approximately normal*)

Standard error of mean

Step 1: Find $\mu, \bar{x}, \sigma$ and calculate $SE(\bar{x}) = \sigma/\sqrt{n}$.

Step 2. Compute z-value using

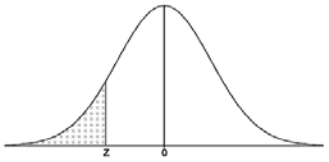
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Standard derivation of the sampling distribution

Step 3. Using the normal table or computer to compute the probability of $\bar{x}$.

Standard Error of the Sample Mean (3 of 3)

Cumulative Standard Normal Distribution Table



Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-0.00	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
-0.10	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.20	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.30	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.40	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.50	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.60	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.70	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.80	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.90	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-1.00	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-1.10	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.20	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.30	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.40	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.50	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.60	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.70	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.80	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.90	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-2.00	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-2.10	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.20	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.30	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.40	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.50	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.60	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.70	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.80	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.90	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-3.00	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-3.10	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.20	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.30	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.40	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.50	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
-3.60	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.70	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.80	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001

*Note: z-values less than -3.89 produce a probability of zero.

Example 7.4 (1 of 2)

The duration of Alzheimer's disease from the time symptoms first appear until death ranges from 3 to 20 years; the average is 8 years with a standard deviation of 4 years.

The administrator of a large medical center randomly selects the medical records of 30 deceased Alzheimer's patients from the medical center's database, and records the average duration.

Example 7.4 (2 of 2)

Find the approximate probabilities for these events:

1. The average duration is less than 7 years.
2. The average duration exceeds 7 years.
3. The average duration lies within 1 year of the population mean $\mu = 8$.

Solution:

Sampling Plan: Since the administrator has selected a random sample from the database at this medical center, he can draw conclusions about only past, present, or future patients with Alzheimer's disease at this medical center.

Example 7.4 – Solution (1 of 7)

If, on the other hand, this medical center can be considered representative of other medical centers in the country, it may be possible to draw more far-reaching conclusions.

Population of Interest: What can you say about the shape of the sampled population? It is not symmetric, because the mean $\mu = 8$ does not lie halfway between the maximum and minimum values. (3 years-20 years.)

Since the mean is closer to the minimum value, the distribution is skewed to the right, with a few patients living a long time after the onset of the disease.

Example 7.4 – Solution (2 of 7)

Sampling Distribution of $\bar{x}$

- n=30. Thus, the sampling distribution is approximately normal. (By the Central Limit Theorem)
- Mean of sampling distribution is $\mu = 8$.
- Standard derivation of the sampling distribution is $\frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{30}} = 0.73$.

Example 7.4 – Solution (3 of 7)

Ans 1: The probability that $\bar{x}$ is less than 7 is given by the shaded area in figure.

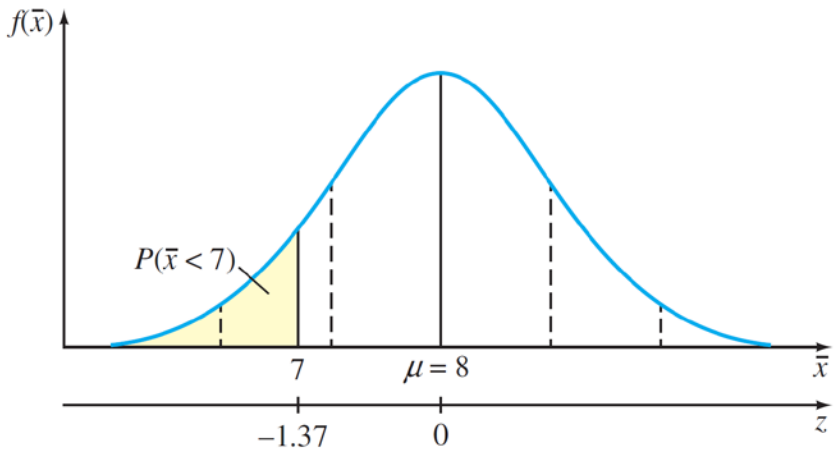


Figure 7.7

Example 7.4 – Solution (4 of 7)

Thus

$$\Pr(\bar{x} < 7) = P\left(z < \frac{7-\mu}{\sigma/\sqrt{n}}\right) = P\left(z < \frac{7-8}{0.73}\right)$$

$$= P(z < -1.37) = 0.0853. \text{ (from normal table)}$$

(Note: You must use $\frac{\sigma}{\sqrt{n}}$ (not σ) in the formula for z because you are finding an area under the sampling distribution for $\bar{x}$, not under the probability distribution for x.)

Example 7.4 – Solution (5 of 7)

Ans 2. The event that $\bar{x}$ exceeds 7 is

$$P(\bar{x} > 7) = 1 - P(\bar{x} \leq 7)$$

$$= 1 - .0853 = .9147$$

Example 7.4 – Solution (6 of 7)

Ans 3. The probability that $\bar{x}$ lies within 1 year of $\mu = 8$ is the shaded area in figure.

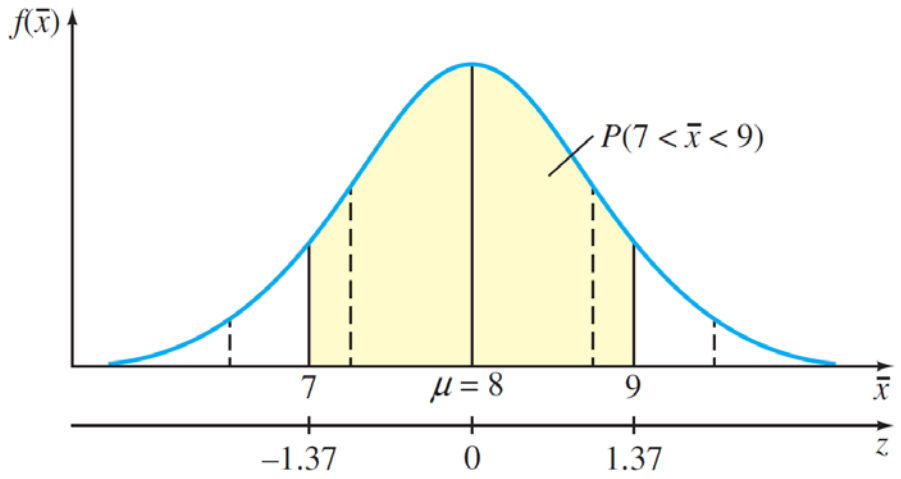


Figure 7.8

Example 7.4 – Solution (7 of 7)

The z-value corresponding to $\bar{x} = 7$ is $z = -1.37$, from part 1, and the z-value for $\bar{x} = 9$ is

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{9 - 8}{.73} = 1.37$$

The probability of interest is

$$P(7 < \bar{x} < 9) = P(-1.37 < z < 1.37)$$

$$= .9147 - .0853 = .8294$$

The Sampling Distribution of the Sample Proportion

The Sampling Distribution of the Sample Proportion (1 of 4)

Properties of the Sampling Distribution of the Sample Proportion, $\hat{p}$

If a random sample of n observations is selected from a binomial population with parameter p, then the sampling distribution of the sample proportion

$$\hat{p} = \frac{x}{n}$$

will have a mean p

and a standard deviation

$$\text{SE}(\hat{p}) = \sqrt{\frac{pq}{n}} \quad \text{where } q = 1 - p$$

The Sampling Distribution of the Sample Proportion (2 of 4)

When the sample size n is large, the sampling distribution of $\hat{p}$ can be approximated by a normal distribution with a mean p and a standard derivation $\sqrt{\frac{p(1-p)}{n}}$. The approximation will be adequate if $np > 5$ and $nq > 5$.

Can use central limit
else cannot

Since x follows a binomial distribution, $\hat{p} = \frac{x}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$, where x_i is a Bernoulli trail output. Thus, using central limiting theorem, the sampling distribution of $\hat{p}$ can be approximated by a normal distribution.

Example 7.10

In a survey, 500 parents were asked about the importance of sports for boys and girls. Of the parents interviewed, 60% agreed that boys and girls should have equal opportunities to participate in sports.

Describe the sampling distribution of the sample proportion $\hat{p}$ of parents who agree that boys and girls should have equal opportunities.

Example 7.10 – Solution (1 of 4)

You can assume that the 500 parents represent a random sample of the parents of all boys and girls in the United States and that the true proportion in the population is equal to some unknown value that you can call p.

Example 7.10 – Solution (2 of 4)

The sampling distribution of $\hat{p}$ can be approximated by a normal distribution, with mean equal to p and standard error

$$\mathrm{SE}(\hat{p}) = \sqrt{\frac{pq}{n}}$$

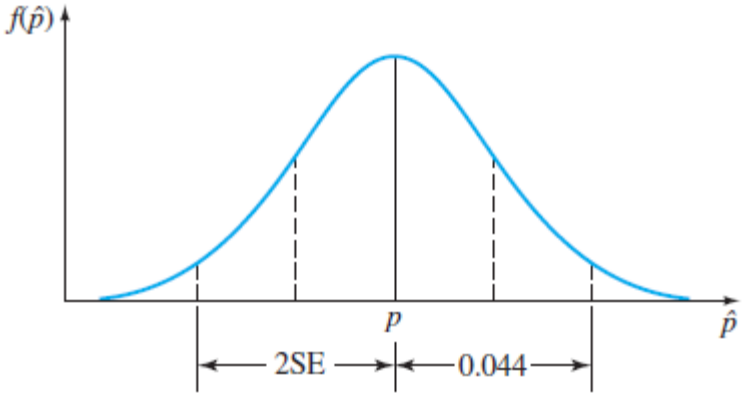


Figure 7.15

Example 7.10 – Solution (3 of 4)

You can see from Figure 7.15 that the sampling distribution of $\hat{p}$ is centered over its mean p.

Even though you do not know the exact value of p (the sample proportion $\hat{p}$ =0.6 may be larger or smaller than p), an approximate value for the standard deviation of the sampling distribution can be found using the sample proportion $\hat{p}$ =0.6 to approximate the unknown value of p.

Example 7.10 – Solution (4 of 4)

Thus,

$$\text{SE} = \sqrt{\frac{pq}{n}} \approx \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$= \sqrt{\frac{(.60)(.40)}{500}} = .022$$

Therefore, approximately 95% of the time $\hat{p}$ will fall within 2SE ≈ .044 of the (unknown) value of p.

The Sampling Distribution of the Sample Proportion (3 of 4)

How to Calculate Probabilities for the Sample Proportion $\hat{p}$

1. Find the values of n and p.
2. Check whether the normal approximation to the binomial distribution is appropriate ($np > 5$ and $nq > 5$).
3. Write down the event of interest in terms of $\hat{p}$, and locate the appropriate area on the normal curve.
4. Convert the necessary values of $\hat{p}$ to z-values using $z = \dfrac{\hat{p} - p}{\sqrt{\dfrac{pq}{n}}}$
5. Use the normal distribution table to calculate the probability.

The Sampling Distribution of the Sample Proportion (4 of 4)

step 2

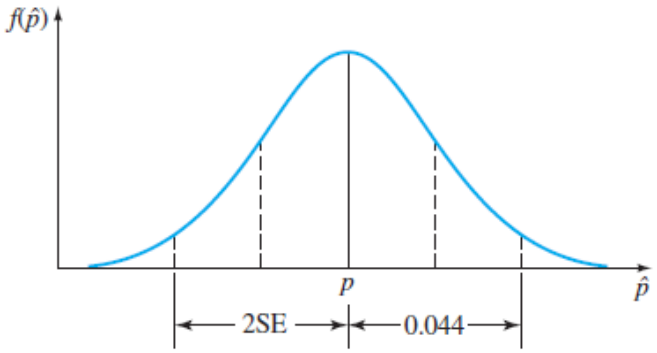
$$SE(\hat{p}) = \sqrt{\frac{pq}{n}}$$

step 3

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

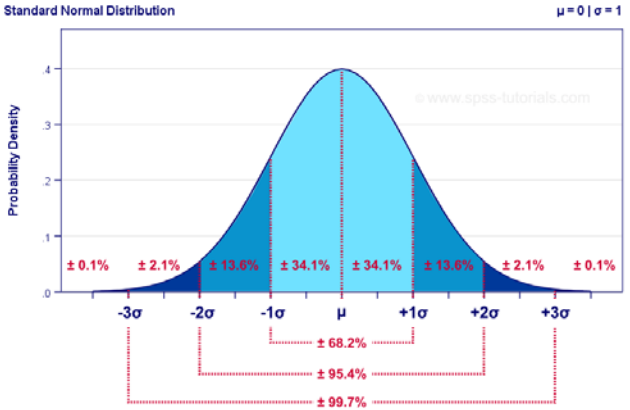
step 4

Standard normal distribution



step 1

$np > 5$ and $nq > 5$



Example 7.11 – Solution (1 of 4)

Figure shows the sampling distribution of $\hat{p}$ when $p = 0.55$, with the observed value $\hat{p}$ =0.60 from 500 samples located on the horizontal axis.

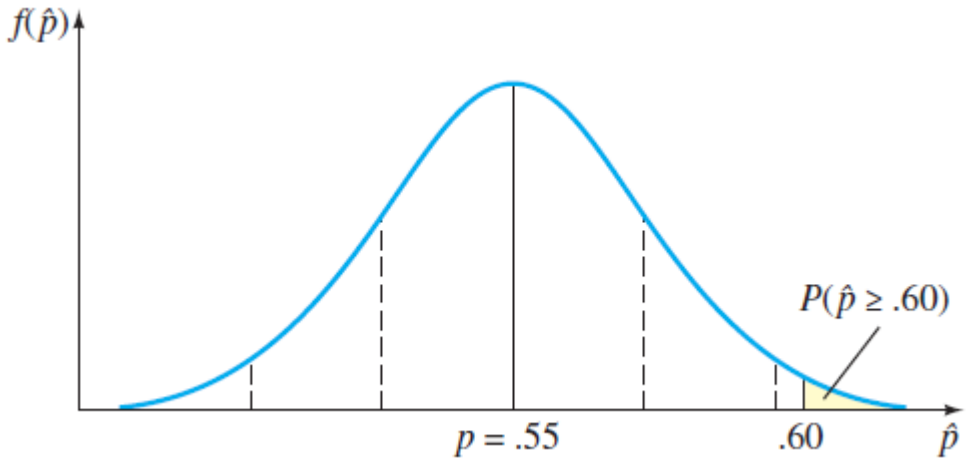


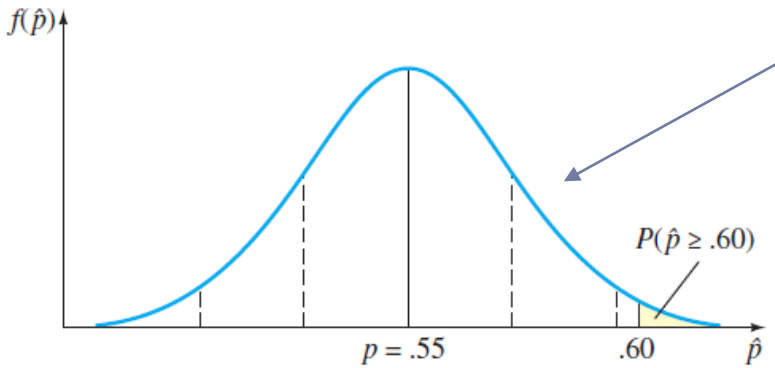
Figure 7.16

Example 7.11 – Solution (2 of 4)

The probability of observing a sample proportion $\hat{p}$ equal to or larger than 0.60 is approximated by the shaded area in the upper tail of this normal distribution with

p=0.55 and $\text{SE} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(.55)(.45)}{500}} = .0222$

$N(0.55, 0.0222^2)$



Example 7.11 – Solution (3 of 4)

To find this shaded area, first calculate the *z*-value corresponding to $\hat{p}$ =0.60:

$$z = \frac{\hat{p} - p}{\sqrt{pq/n}} = \frac{.60 - .55}{.0222} = 2.25$$

Using the normal distribution table, you find

$$P(\hat{p} > .60) \approx P(z > 2.25) = 1 - .9878 = .0122$$

Example 7.11 – Solution (4 of 4)

That is, if you were to select a random sample of $n = 500$ observations from a population with proportion p equal to 0.55, the probability that the sample proportion $\hat{p}$ would be as large as or larger than 0.60 is only 0.0122.

This probability is quite small! Either we have observed a very unlikely event, or perhaps the true value of p is not as claimed.