Sampling Distributions

Random Sample, Central Limit Theorem

Dr.Mamatha.H.R

Professor

Department of Computer Science and Engineering

PES University

Bangalore

Course material created using various Internet resources and text book

Introduction

- **Parameters** are numerical descriptive measures for populations.
  - Two parameters for a normal distribution: mean μ and standard deviation σ.
  - One parameter for a binomial distribution: the success probability of each trial p.
- Often the values of parameters that specify the exact form of a distribution are unknown.
- You must rely on the **sample** to learn about these parameters.

Sampling

Examples:

- A pollster is sure that the responses to his "agree/disagree" question will follow a binomial distribution, but p, the proportion of those who "agree" in the population, is unknown.
- An agronomist believes that the yield per acre of a variety of wheat is approximately normally distributed, but the mean μ and the standard deviation σ of the yields are unknown.

✓ If you want the sample to provide reliable information about the population, you must select your sample in a certain way!

Simple Random Sampling

- The **sampling plan** or **experimental design** determines the amount of information you can extract, and often allows you to measure the **reliability of your inference**.
- **Simple random sampling** is a method of sampling that allows each possible sample of size n an equal probability of being selected.

Sampling Distributions

- Any numerical descriptive measures calculated from the sample are called **statistics**.
- Statistics vary from sample to sample and hence are random variables. This variability is called sampling variability.
- The probability distributions for statistics are called **sampling distributions**.
- In repeated sampling, they tell us what values of the statistics can occur and how often each value occurs.

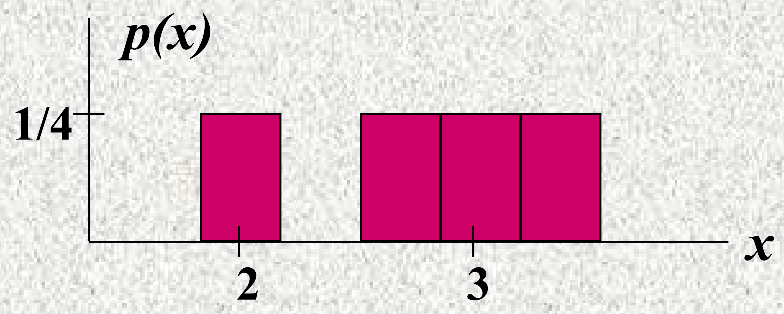
Example

<u>Population:</u> 3, 5, 2, 1

Draw samples of size $n = 3$ without replacement

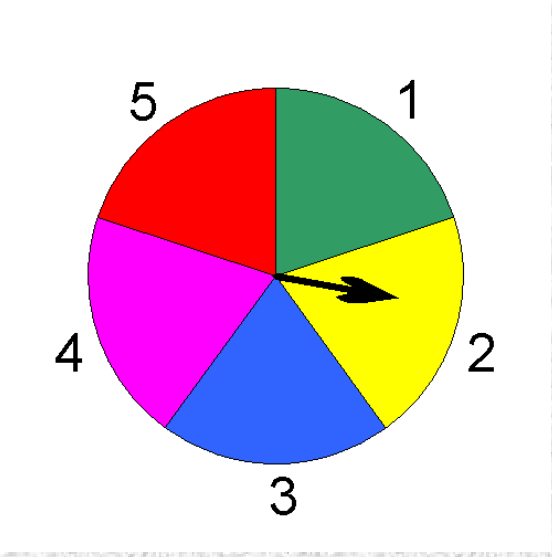
Possible samples	$\bar{x}$
3, 5, 2	$10/3 = 3.33$
3, 5, 1	$9/3 = 3$
3, 2, 1	$6/3 = 2$
5, 2, 1	$8/3 = 2.67$

Each value of x-bar is equally likely, with probability 1/4



Example

Consider a population that consists of the numbers 1, 2, 3, 4 and 5 generated in a manner that the probability of each of those values is 0.2 no matter what the previous selections were. This population could be described as the outcome associated with a spinner such as given below with the distribution next to it.



x	p(x)
1	0.2
2	0.2
3	0.2
4	0.2
5	0.2

Example

If the sampling distribution for the means of samples of size two is analyzed, it looks like

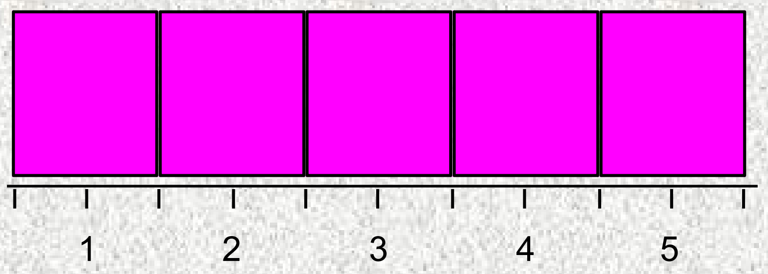
Sample	
1, 1	1
1, 2	1.5
1, 3	2
1, 4	2.5
1, 5	3
2, 1	1.5
2, 2	2
2, 3	2.5
2, 4	3
2, 5	3.5
3, 1	2
3, 2	2.5
3, 3	3

Sample	
3, 4	3.5
3, 5	4
4, 1	2.5
4, 2	3
4, 3	3.5
4, 4	4
4, 5	4.5
5, 1	3
5, 2	3.5
5, 3	4
5, 4	4.5
5, 5	5

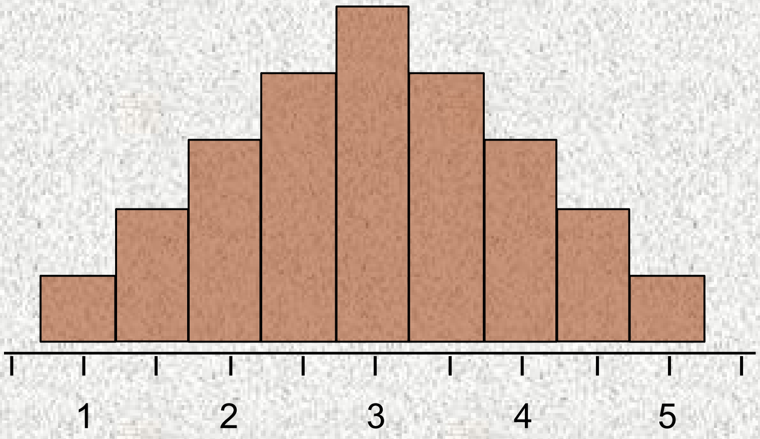
	frequency	p(x)
1	1	0.04
1.5	2	0.08
2	3	0.12
2.5	4	0.16
3	5	0.20
3.5	4	0.16
4	3	0.12
4.5	2	0.08
5	1	0.04
	25	

Example

The original distribution and the sampling distribution of means of samples with n=2 are given below.



Original distribution

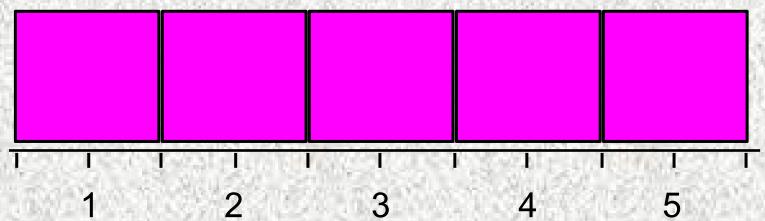


Sampling distribution

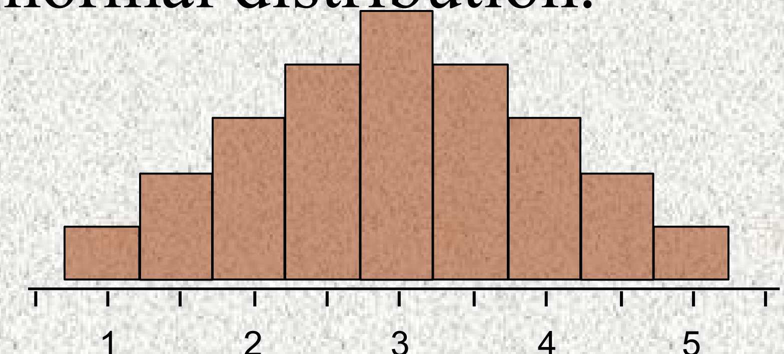
n = 2

Example

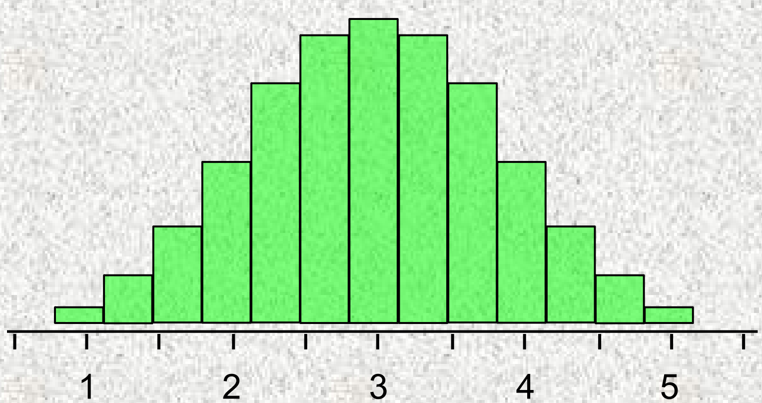
Sampling distributions for n=3 and n=4 were calculated and are illustrated below. The shape is getting closer and closer to the normal distribution.



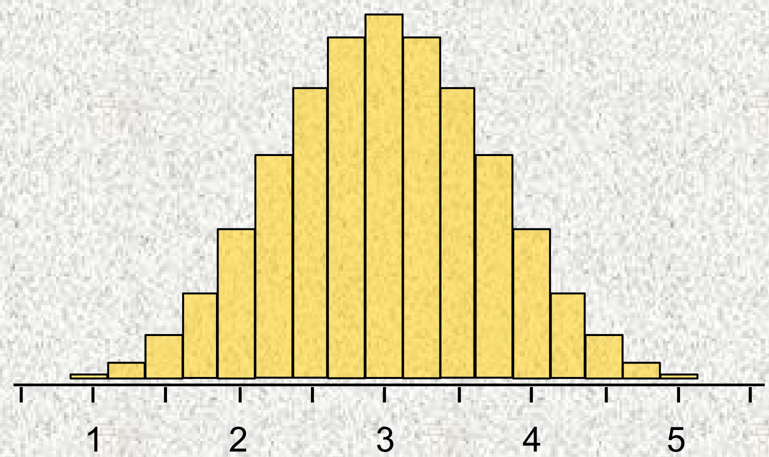
Original distribution



Sampling distribution n = 2



Sampling distribution n = 3



Sampling distribution n = 4

Sampling Distribution of $\bar{x}$

If a random sample of n measurements is selected from a population with mean μ and standard deviation σ, the sampling distribution of the sample mean $\bar{x}$ will have a mean

$$\mu_{\bar{x}} = \mu$$

and a standard deviation

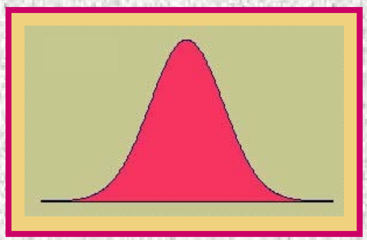
$$\sigma_{\bar{x}} = \sigma / \sqrt{n}$$

Central Limit Theorem: If random samples of n observations are drawn from a nonnormal population with finite μ and standard deviation σ , then, when n is large, the sampling distribution of the sample mean $\bar{x}$ is approximately normally distributed, with mean μ and standard deviation

$\sigma / \sqrt{n}$

. The approximation becomes more accurate as n becomes large.

Why is this Important?



- The **Central Limit Theorem** also implies that the sum of n measurements is approximately normal with mean $n\mu$ and standard deviation $\sigma\sqrt{n}$.
- Many statistics that are used for statistical inference are **sums** or **averages** of sample measurements.
- When n is large, these statistics will have approximately **normal** distributions.
- This will allow us to describe their behavior and evaluate the **reliability** of our inferences.

How Large is Large?

If the sample is **normal**, then the sampling distribution of $\bar{x}$ will also be normal, no matter what the sample size.

When the sample population is approximately **symmetric**, the distribution becomes approximately normal for relatively small values of *n*.

When the sample population is **skewed**, the sample size must be **at least 30** before the sampling distribution of $\bar{x}$ becomes approximately normal.

The Central Limit Theorem

Let $X_1, \ldots, X_n$ be a simple random sample from a population with mean μ and variance σ^2.

Let $\overline{X} = \dfrac{X_1 + \cdots + X_n}{n}$ be the sample mean.

Let $S_n = X_1 + \cdots + X_n$ be the sum of the sample observations.

Then if n is sufficiently large,

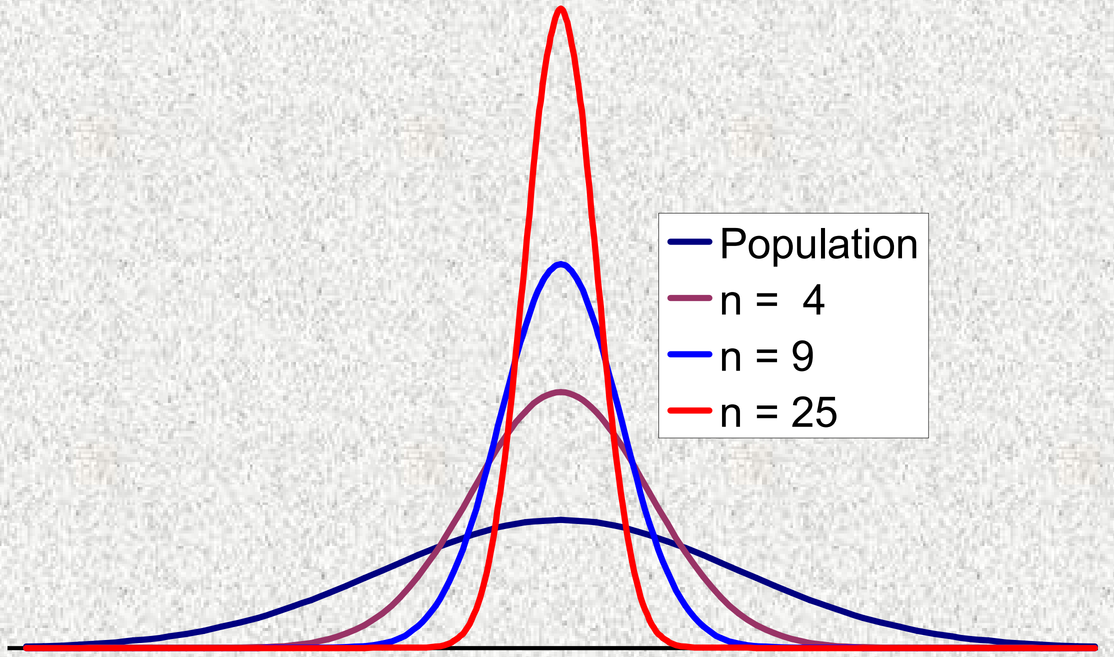
$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{approximately} \tag{4.55}$$

and

$$S_n \sim N(n\mu, n\sigma^2) \quad \text{approximately} \tag{4.56}$$

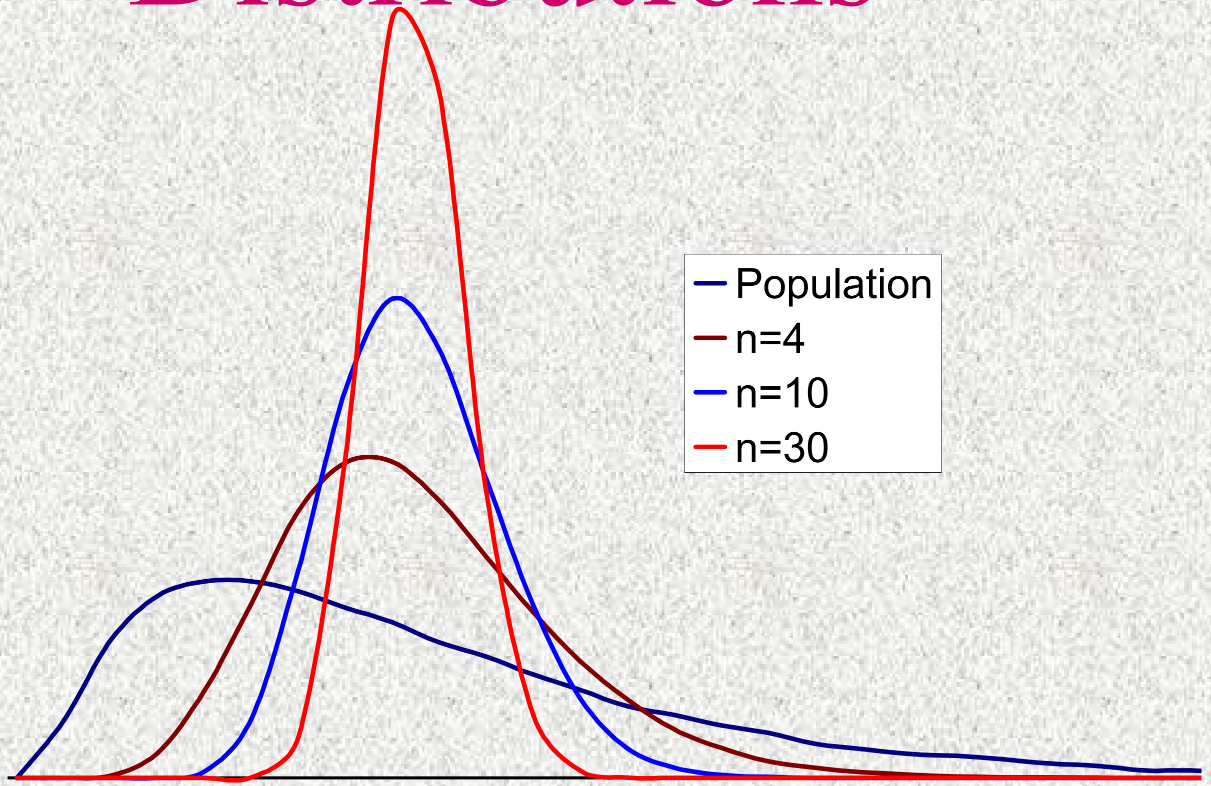
For most populations, if the sample size is greater than 30, the Central Limit Theorem approximation is good.

Illustrations of Sampling Distributions

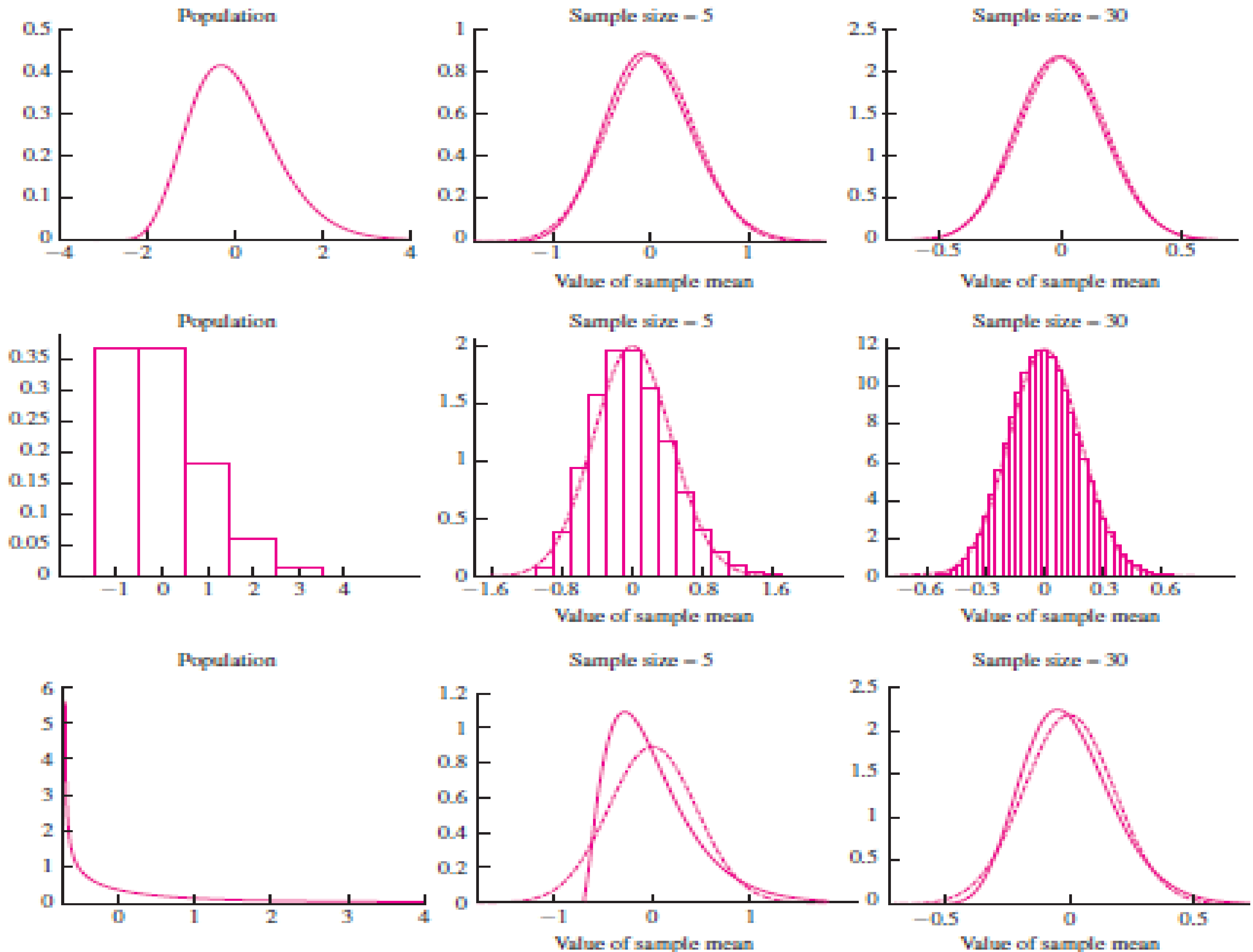


Symmetric normal like population

Illustrations of Sampling Distributions



Skewed population



Population
0.5
0.4
0.3
0.2
0.1
0
−4
−2
0
2
4
Sample size = 5
1
0.8
0.6
0.4
0.2
0
−1
0
1
Value of sample mean
Sample size = 30
2.5
2
1.5
1
0.5
0
−0.5
0
0.5
Value of sample mean
Population
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−1
0
1
2
3
4
Sample size = 5
2
1.5
1
0.5
0
−1.6
−0.8
0
0.8
1.6
Value of sample mean
Sample size = 30
12
10
8
6
4
2
0
−0.6
−0.3
0
0.3
0.6
Value of sample mean
Population
6
5
4
3
2
1
0
0
1
2
3
4
Sample size = 5
1.2
1
0.8
0.6
0.4
0.2
0
−1
0
1
Value of sample mean
Sample size = 30
2.5
2
1.5
1
0.5
0
−0.5
0
0.5
Value of sample mean

Finding Probabilities for the Sample Mean

- ✓ If the sampling distribution of $\bar{x}$ is normal or approximately normal, *standardize or rescale* the interval of interest in terms of

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

- ✓ Find the appropriate area using Table 3.

Example: A random sample of size $n = 16$ from a normal distribution with $\mu = 10$ and $\sigma = 8$.

Example

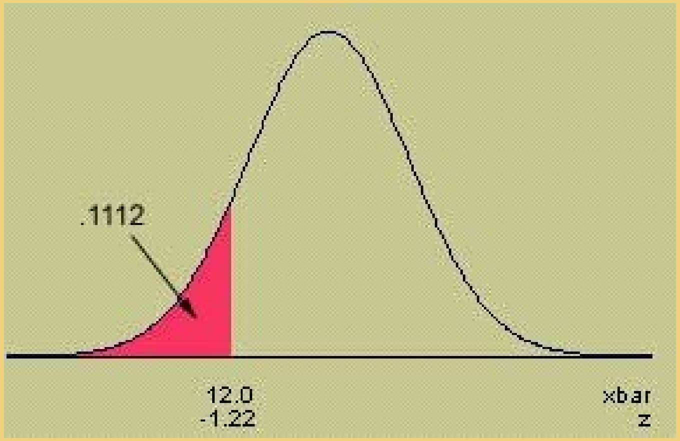
A soda filling machine is supposed to fill cans of soda with 12 fluid ounces. Suppose that the fills are actually normally distributed with a mean of 12.1 oz and a standard deviation of .2 oz.

What is the probability that the average fill for a 6-pack of soda is less than 12 oz?

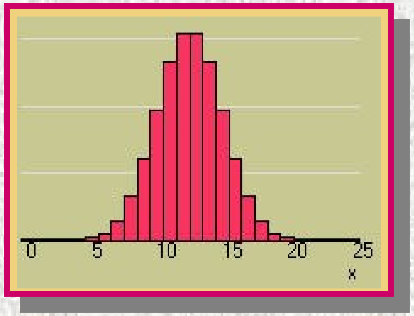
$$P(\bar{x} < 12) =$$

$$P\left(\frac{\bar{x} - \mu}{\sigma / \sqrt{n}} < \frac{12 - 12.1}{.2 / \sqrt{6}}\right) =$$

$$P(z < -1.22) = .1112$$

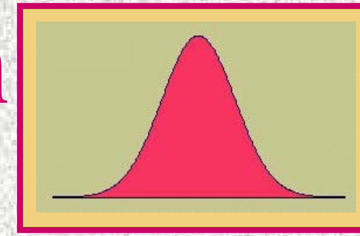


The Sampling Distribution of the Sample Proportion



- The **Central Limit Theorem** can be used to conclude that the binomial random variable x is approximately normal when n is large, with mean $\boldsymbol{np}$ and variance $\boldsymbol{npq}$.
- The sample proportion, $\hat{p} = \frac{x}{n}$ is simply a *rescaling* of the binomial random variable x, dividing it by n.
- From the Central Limit Theorem, the sampling distribution of $\hat{p}$ will also be **approximately normal,** with a *rescaled* mean and standard deviation.

The Sampling Distribution of the Sample Proportion



- ✓A random sample of size ***n*** is selected from a binomial population with parameter ***p***.
- ✓The sampling distribution of the sample proportion,

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

will have mean ***p*** and standard deviation $\sqrt{\frac{pq}{n}}$

- ✓If n is large, and p is not too close to zero or one, the sampling distribution of $\hat{p}$ will be **approximately normal.**

The standard deviation of *p*-hat is sometimes called the STANDARD ERROR (SE) of *p*-hat.

Finding Probabilities for the Sample Proportion

- ✓ If the sampling distribution of $\hat{p}$ is normal or approximately normal, *standardize or rescale* the interval of interest in terms of
$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$
- ✓ Find the appropriate area using Table 3.

If $X \sim \text{Bin}(n,\ p)$, and if $np > 10$ and $n(1-p) > 10$, then

$$X \sim N(np,\ np(1-p)) \qquad \text{approximately}$$

$$\widehat{p} \sim N\left(p,\ \frac{p(1-p)}{n}\right) \qquad \text{approximately}$$

Example

The soda bottler in the previous example claims that only 5% of the soda cans are underfilled. A quality control technician randomly samples 200 cans of soda. What is the probability that more than 10% of the cans are underfilled?

$\boldsymbol{n} = \mathbf{200}$

S: underfilled can

$\boldsymbol{p} = \mathbf{P(S) = .05}$

$\boldsymbol{q} = \mathbf{.95}$

$\boldsymbol{np} = \mathbf{10}$ $\boldsymbol{nq} = \mathbf{190}$

OK to use the normal approximation

$$P(\hat{p} > .10)$$

$$= P\left(z > \frac{.10 - .05}{\sqrt{\frac{.05(.95)}{200}}}\right) = P(z > 3.24)$$

$$< .5 - .4990 = .001$$

This would be very unusual, if indeed $p = .05$!

Example

Suppose 3% of the people contacted by phone are receptive to a certain sales pitch and buy your product. If your sales staff contacts 2000 people, what is the probability that more than 100 of the people contacted will purchase your product?

n=2000, p= 0.03, np=60, nq=1940,

OK to use the normal approximation

$$P(\hat{p} > 100/2000) = P\left(z > \frac{.05 - .03}{\sqrt{\frac{.03(.97)}{2000}}}\right) = P(z > 5.24) \approx 0$$

Let *X denote the number of flaws in a 1 in. length of copper wire. The probability* mass function of *X is presented in the following table.*

x	***P(X = x)***
0	0.48
1	0.39
2	0.12
3	0.01

One hundred wires are sampled from this population. What is the probability that the average number of flaws per wire in this sample is less than 0.5?

The population mean number of flaws is $\mu = 0.66$, *and the population variance is* $\sigma^2 = 0.5244$

We need to find $P(\bar{x} < 0.5)$.
sample size is $n = 100$, which is a large sample. It follows from the Central Limit Theorem

that $\bar{x} \sim N(0.66, 0.005244)$.
The z-score of 0.5 is therefore
$z = (0.5 - 0.66) / \sqrt{0.005244} = -2.21$

From the *z table, the area to the left of* -2.21 *is* 0.0136.
Therefore $P(\bar{x} < 0.5) = 0.0136$, *so only 1.36% of samples of size 100 will have fewer than 0.5 flaws per wire.*

At a large university, the mean age of the students is 22.3 years, and the standard deviation is 4 years. A random sample of 64 students is drawn. What is the probability that the average age of these students is greater than 23 years?

Let $X1, \ldots, X64$ *be the ages of the 64 students in the sample.*

Find $P(\bar{x} > 23)$.

Now the population from which the sample was drawn has mean $\mu = 22.3$ *and variance* $\sigma^2 = 16$.

The sample size is $n = 64$.
It follows from the Central Limit Theorem that $\bar{x} \sim N(22.3, 0.25)$.

The z-score for 23 is
$z = 23 - 22.3 / \sqrt{0.25} = 1.40$

From the *z table, the area to the right of 1.40 is 0.0808.*
Therefore $P(\bar{x} > 23) = 0.0808$.

The manufacture of a certain part requires two different machine operations. The time on machine 1 has mean 0.4 hours and standard deviation 0.1 hours. The time on machine 2 has mean 0.45 hours and standard deviation 0.15 hours. The times needed on the machines are independent. Suppose that 65 parts are manufactured. What is the distribution of the total time on machine 1? On machine 2? What is the probability that the total time used by both machines together is between 50 and 55 hours?

The Central Limit Theorem

Let $X_1, \ldots, X_n$ be a simple random sample from a population with mean μ and variance σ^2.

Let $\overline{X} = \dfrac{X_1 + \cdots + X_n}{n}$ be the sample mean.

Let $S_n = X_1 + \cdots + X_n$ be the sum of the sample observations.

Then if n is sufficiently large,

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{approximately} \tag{4.55}$$

and

$$S_n \sim N(n\mu, n\sigma^2) \quad \text{approximately} \tag{4.56}$$

For most populations, if the sample size is greater than 30, the Central Limit Theorem approximation is good.

Let *X1, . . . , X65 represent the times of the 65 parts on machine 1*.

The population from which this sample was drawn has mean $\mu_X = 0.4$ *and standard deviation* $\sigma_X = 0.1$.

Let $S_X = X1 + \cdots + X65$ *be the total time on machine 1*.

It follows from the Central Limit Theorem that

$S_X \sim N(65\mu_X , 65\sigma^2_X) = N(26, 0.65)$

Let $Y1, \ldots, Y65$ *represent the times of the 65 parts on machine 2.*

The population from which this sample was drawn has mean $\mu_Y = 0.45$ *and standard deviation* $\sigma_Y = 0.15$.

Let $S_Y = Y1 + \cdots + Y65$ *be the total time on machine 2.*

It follows from the Central Limit Theorem that

$$S_Y \sim N(65\mu_Y, 65\sigma^2_Y) = N(29.25, 1.4625)$$

let $T = S_X + S_Y$ *represent the total time on both machines.*

Since

$S_X \sim N(26, 0.65)$, $S_Y \sim N(29.25, 1.4625)$, *and* S_X *and* S_Y *are independent, it* follows that

$\mu_T = 26+29.25=55.25$, $\sigma^2_T=0.65+1.4625=2.1125$, *and*

$T \sim N(55.25, 2.1125)$

To find $P(50 < T < 55)$ *we compute the z-scores of 50 and of 55.*

$z = (50 - 55.25) / \sqrt{2.1125} = -3.61$

$z = (55 - 55.25) / \sqrt{2.1125} = -0.17$

The area to the left of $z = -3.61$ *is 0.0002*.

The area to the left of $z = -0.17$ *is 0.4325*.

The area between $z = -3.61$ *and* $z = -0.17$ *is* $0.4325 - 0.0002 = 0.4323$.

The probability that the total time used by both machines together is between 50 and 55 hours is 0.4323

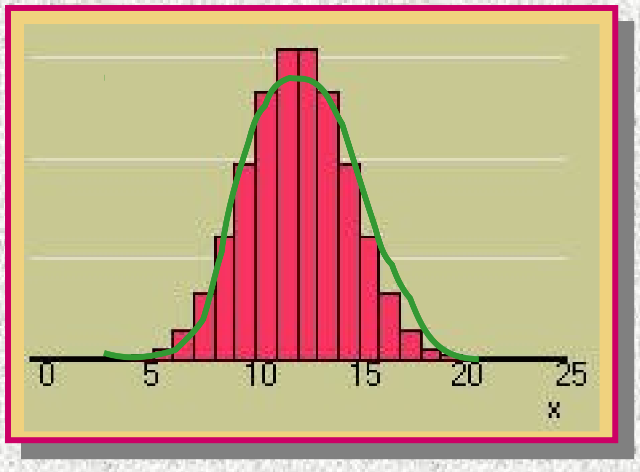
The Normal Approximation to the Binomial

We can calculate binomial probabilities using

The binomial formula

The cumulative binomial tables

When n is large, and p is not too close to zero or one, areas under the normal curve with mean np and variance npq can be used to approximate binomial probabilities.



Approximating the Binomial

- ✓Make sure to include the entire rectangle for the values of x in the interval of interest. This is called the continuity correction.
- ✓Standardize the values of x using

$$z = \frac{x - \mu}{\sigma}, \mu = np, \sigma = \sqrt{npq}$$

- ✓Make sure that np and nq are both greater than 10 to avoid inaccurate approximations! Or
- ✓n is large and $\mu \pm 2\sigma$ falls between 0 and n

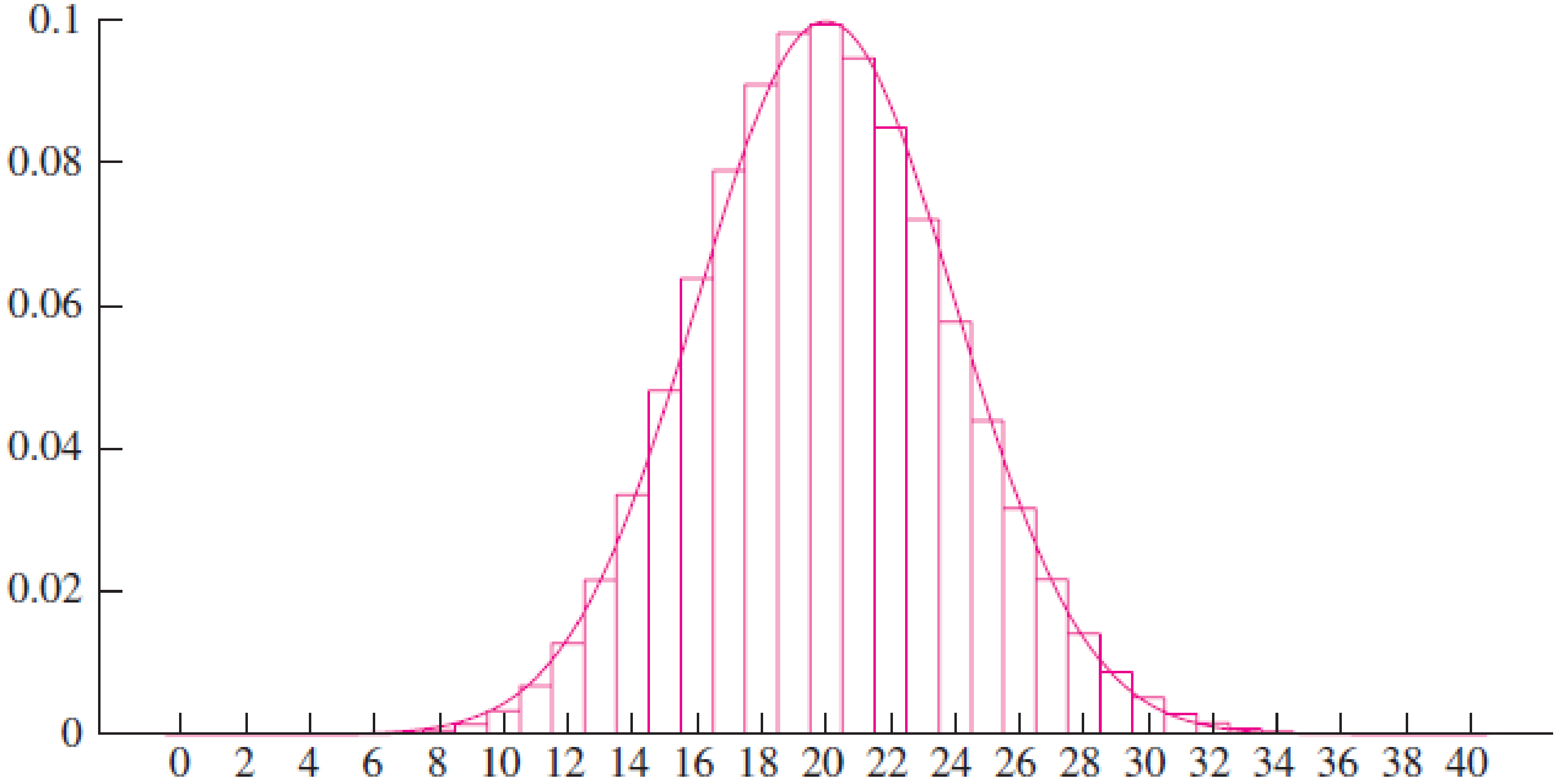
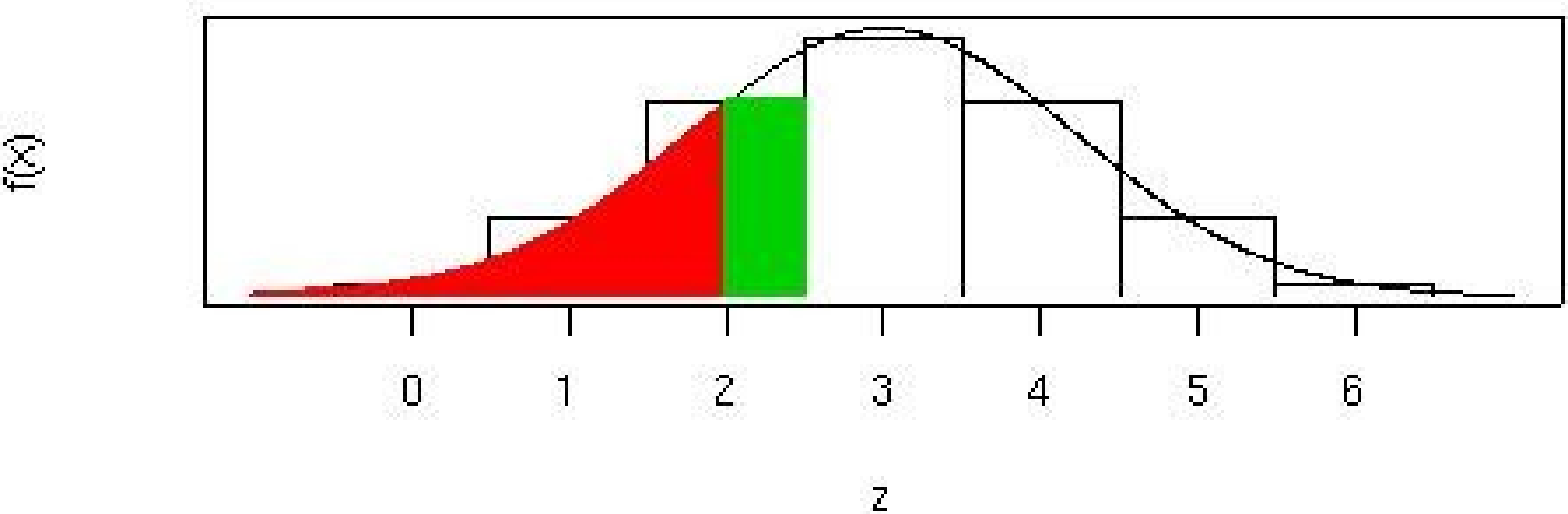


FIGURE 4.27 The Bin(100, 0.2) probability histogram, with the $N(20, 16)$ probability density function superimposed.

- What is the Continuity Correction Factor?
- A continuity correction factor is used when you use a continuous probability distribution to approximate a discrete probability distribution. For example, when you want to use the normal to approximate a binomial.

Correction for Continuity

Add or subtract .5 to include the entire rectangle. For illustration, suppose x is a Binomial random variable with n=6, p=.5. We want to compute P(x≤ 2). Using 2 directly will miss the green area. P(x≤ 2)=P(x≤ 2.5) and use 2.5.



Continuity Correction Factor Table

If $P(X=n)$ use $P(n - 0.5 < X < n + 0.5)$

If $P(X > n)$ use $P(X > n + 0.5)$

If $P(X \leq n)$ use $P(X < n + 0.5)$

If $P(X < n)$ use $P(X < n - 0.5)$

If $P(X \geq n)$ use $P(X > n - 0.5)$

Discrete	Continuous
$x = 6$	$5.5 < x < 6.5$
$x > 6$	$x > 6.5$
$x \leq 6$	$x < 6.5$
$x < 6$	$x < 5.5$
$x \geq 6$	$x > 5.5$

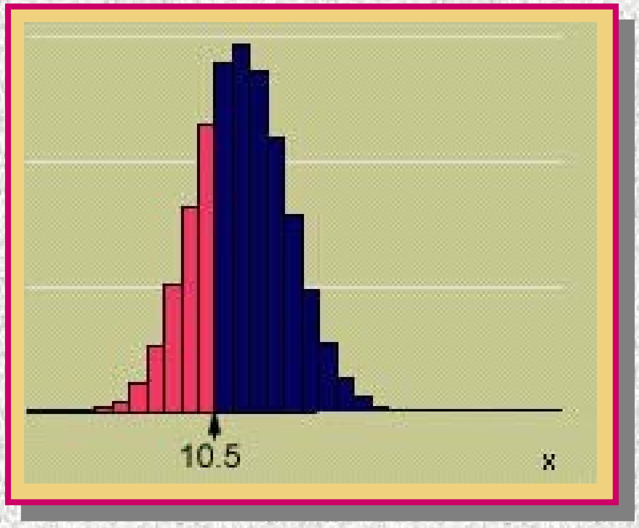
Example

Suppose x is a binomial random variable with $n = 30$ and $p = .4$. Using the normal approximation to find $P(x \leq 10)$.

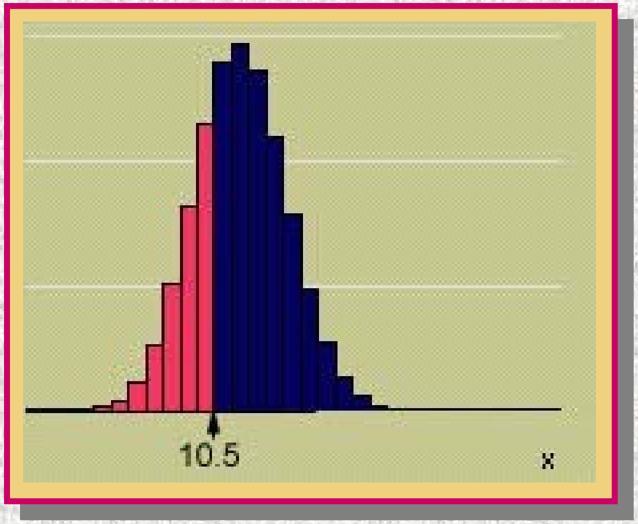
$n = 30 \quad p = .4 \qquad q = .6$

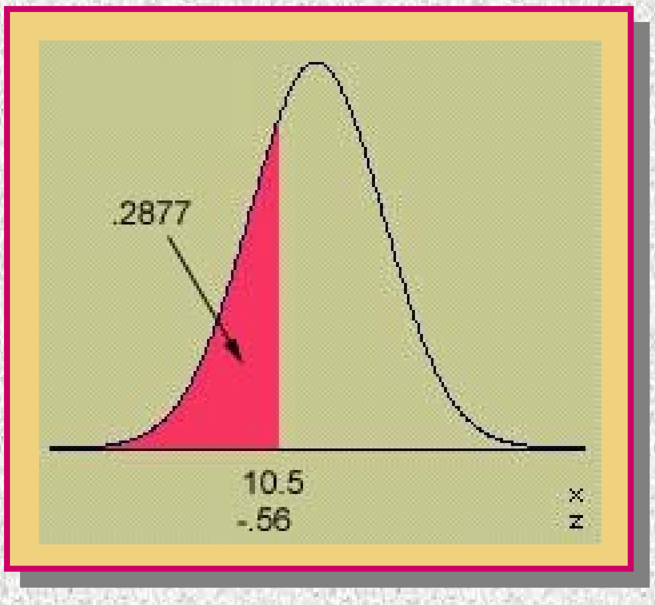
$np = 12 \qquad nq = 18$

The normal approximation is ok!



Example

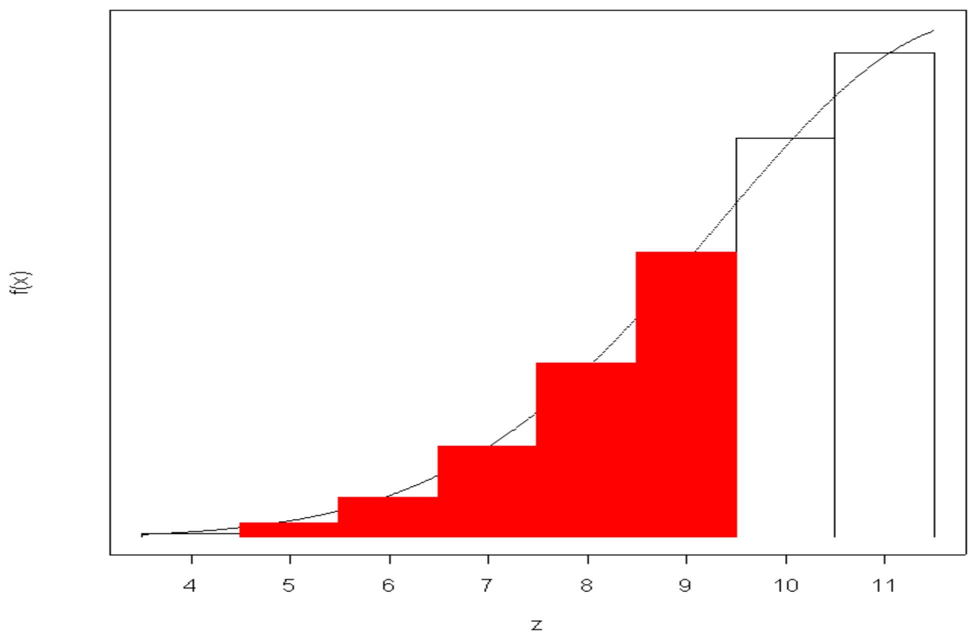




$$P(x \le 10) \approx P(z \le \frac{10.5-12}{2.683})$$

$$= P(z \le -.56) = .2877$$

Example



$$P(x < 10) = P(x < 9.5)$$

$$P(x \geq 5) = P(x \geq 4.5)$$

$$P(x > 5) = P(x > 5.5)$$

$$P(5 < x < 10) = P(5.5 < x < 9.5)$$

$$P(5 \leq x < 10) = P(4.5 < x < 9.5)$$

If a fair coin is tossed 100 times, use the normal curve to approximate the probability that the number of heads is between 45 and 55 *inclusive*.

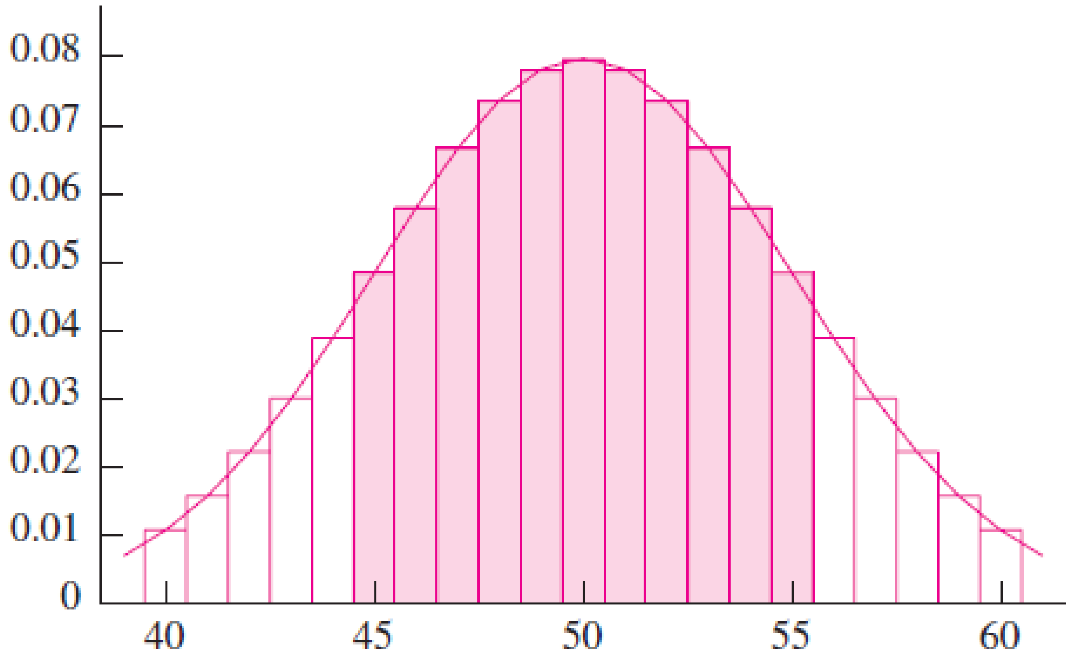
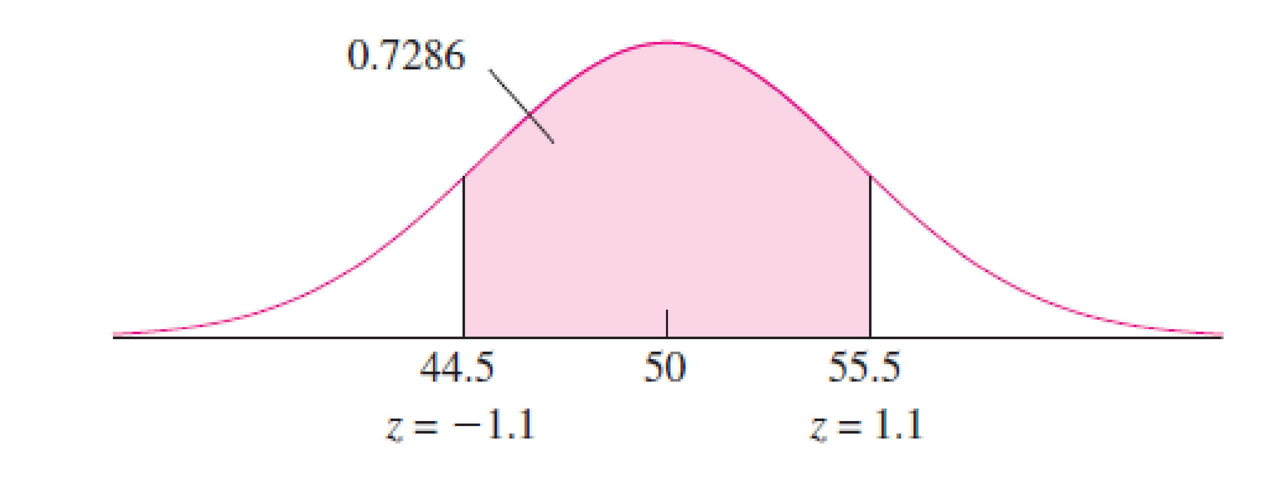


FIGURE 4.28 To compute $P(45 \leq X \leq 55)$, the areas of the rectangles corresponding to 45 and to 55 should be included. To approximate this probability with the normal curve, compute the area under the curve between 44.5 and 55.5.

Let X *be the number of heads obtained.*
Then $X \sim Bin(100, 0.5)$.
Substituting $n = 100$ *and* $p = 0.5$

we obtain the normal approximation $X \sim N(50, 25)$.

Since the endpoints 45 and 55 are to be included, we should compute the area under the normal curve between 44.5 and 55.5.
The *z-scores for 44.5 and 55.5 are*
$z = 44.5 - 50/5 = -1.1$,
$z = 55.5 - 50/5 = 1.1$
From the *z table we find that the probability is 0.7286*



If a fair coin is tossed 100 times, use the normal curve to approximate the probability that the number of heads is between 45 and 55 *exclusive.*

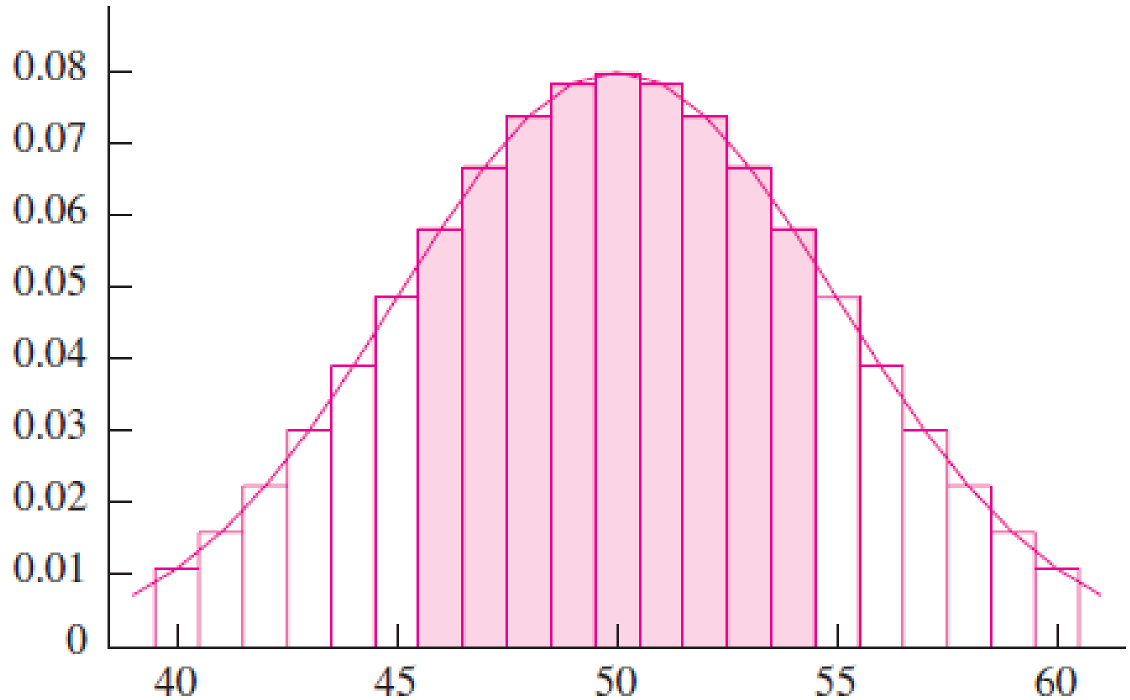


FIGURE 4.29 To compute $P(45 < X < 55)$, the areas of the rectangles corresponding to 45 and to 55 should be excluded. To approximate this probability with the normal curve, compute the area under the curve between 45.5 and 54.5.

Let *X be the number of heads obtained.*

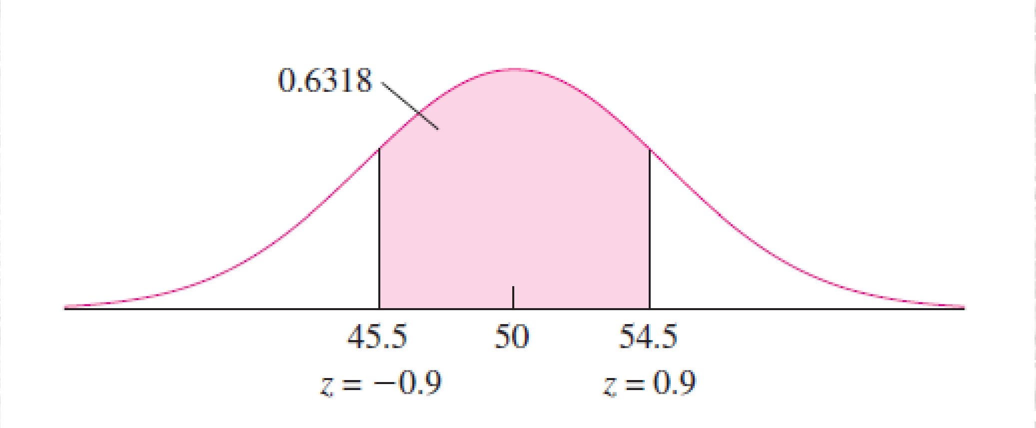
X ~ *Bin(100, 0.5), and the normal approximation is X* ~ *N(50, 25).*

Since the endpoints 45 and 55 are to be excluded, we should compute the area under the normal curve between 45.5 and 54.5.

The *z-scores for 45.5 and 54.5 are*
$z = 45.5 - 50/5 = -0.9$,

$z = 54.5 - 50/5 = 0.9$

From the *z table we find that the probability is 0.6318.*



Example

A production line produces AA batteries with a reliability rate of 95%. A sample of n = 200 batteries is selected. Find the probability that at least 195 of the batteries work.

Success = working battery $n = 200$

$p = .95 \quad np = 190 \quad nq = 10$

The normal approximation is ok!

$$P(x \geq 195) \approx P\left(z \geq \frac{194.5 - 190}{\sqrt{200(.95)(.05)}}\right)$$

$$= P(z \geq 1.46) = 1 - .9278 = .0722$$

In a certain large university, 25% of the students are over 21 years of age. In a sample of 400 students, what is the probability that more than 110 of them are over 21?

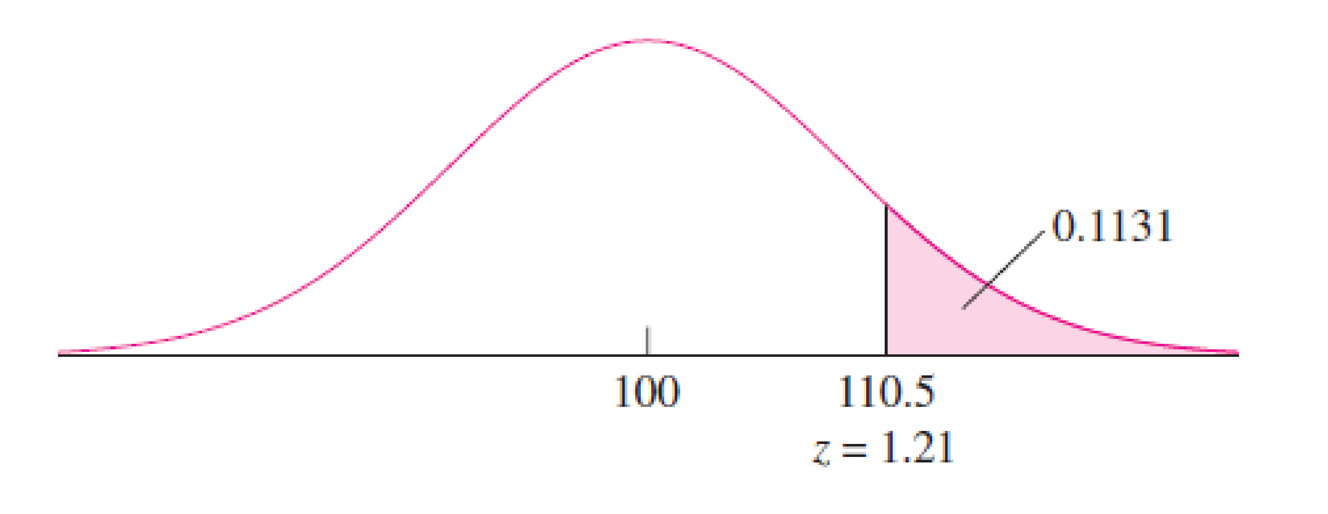
Let X *represent the number of students who are over 21*.

Then $X \sim Bin(400, 0.25)$.
Substituting $n = 400$ *and* $p = 0.25$

we obtain the normal approximation $X \sim N(100, 75)$.

Since we want to find the probability that the number of students is *more than 110, the value 110 is excluded.*

We therefore find $P(X > 110.5)$.
We compute the z-score for 110.5, which is
$z = 110.5 - 100/\sqrt{75}$
$= 1.21$
Using the z *table, we find that* $P(X > 110.5) = 0.1131$.


0.1131
100
110.5
$z = 1.21$

- **Normal Approximation to the Poisson**
- Recall
- if $X \sim$ *Poisson*(λ)*, then X is approximately binomial with n large and np* $= \lambda$.
- $\mu_X = \lambda$ *and* $\sigma^2_X = \lambda$.
- *It follows that if* λ *is sufficiently* large, i.e., $\lambda > 10$*, then X is approximately binomial, with np* > 10.
- *It follows* from the Central Limit Theorem that *X is also approximately normal, with mean and* variance both equal to λ.
- *Thus we can use the normal distribution to approximate the* Poisson.

Summary

If $X \sim \text{Poisson}(\lambda)$, where $\lambda > 10$, then

$$X \sim N(\lambda,\ \lambda) \qquad \text{approximately} \tag{4.59}$$

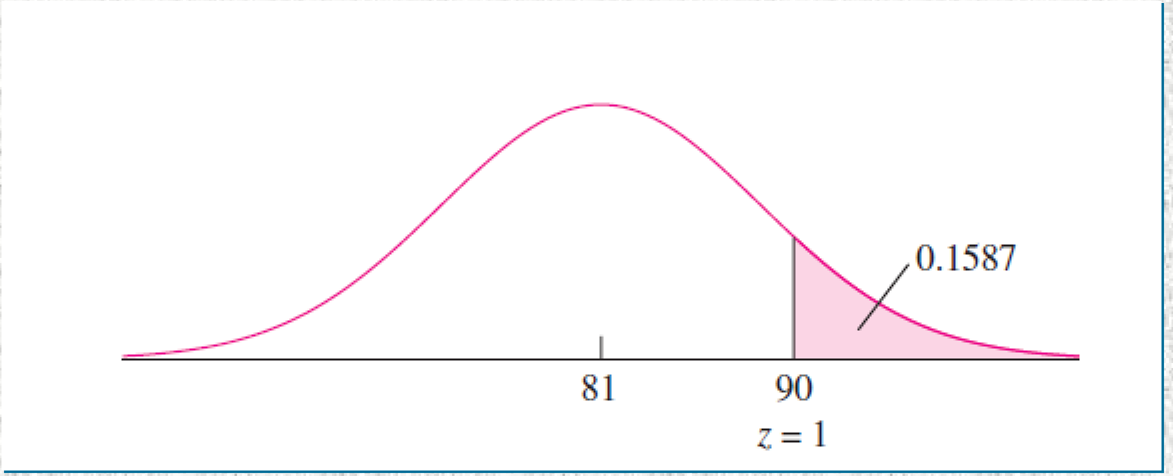
Continuity Correction for the Poisson Distribution

- Since the Poisson distribution is discrete, the continuity correction can in principle be applied when using the normal approximation.
- For areas that include the central part of the curve, the continuity correction generally improves the normal approximation,
- but for areas in the tails the continuity correction sometimes makes the approximation worse.
- Hence, the continuity correction is not used for the Poisson distribution.

- The number of hits on a website follows a Poisson distribution, with a mean of 27 hits per hour. Find the probability that there will be 90 or more hits in three hours.

Solution

- Let X *denote the number of hits on the website in three hours.*
- *The mean number* of hits in three hours is 81, so $X \sim \text{Poisson}(81)$.
- *Using the normal approximation,* $X \sim N(81, 81)$.



- *find $P(X \geq 90)$. We compute the z-score of 90,*
- which is $z = (90 - 81)/\sqrt{81} = 1.00$
- Using the *z table, we find that* $P(X \geq 90) = 0.1587$.

Key Concepts

I. Sampling Plans and Experimental Designs

Simple random sampling: Each possible sample is equally likely to occur.

II. Statistics and Sampling Distributions

1. Sampling distributions describe the possible values of a statistic and how often they occur in repeated sampling.
2. The **Central Limit Theorem** states that sums and averages of measurements from a nonnormal population with finite mean μ and standard deviation σ have approximately normal distributions for large samples of size n.

Key Concepts

III. Sampling Distribution of the Sample Mean

1. When samples of size n are drawn from a normal population with mean μ and variance σ^2, the sample mean $\bar{x}$ has a normal distribution with mean μ and variance σ^2/n.

2. When samples of size n are drawn from a nonnormal population with mean μ and variance σ^2, the Central Limit Theorem ensures that the sample mean $\bar{x}$ will have an approximately normal distribution with mean μ and variance σ^2/n when n is large ($n \geq 30$).

3. Probabilities involving the sample mean μ can be calculated by standardizing the value of $\bar{x}$ using $$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Key Concepts

IV. Sampling Distribution of the Sample Proportion

1. When samples of size n are drawn from a binomial population with parameter p, the sample proportion $\hat{p}$ will have an approximately normal distribution with mean p and variance pq/n as long as $np > 5$ and $nq > 5$.

2. Probabilities involving the sample proportion can be calculated by standardizing the value $\hat{p}$ using

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