Statistics for Scientists – CSC261

Counting and Probability Distributions

Dr DEBAJYOTI PAL
SCHOOL OF INFORMATION TECHNOLOGY,
KMUTT

Where We've Been

- Using probability to make inferences about populations
- Measuring the reliability of the inferences

Two Types of Random Variables

- Random Variable
  - variable that assumes numerical values associated with random outcomes of an experiment
  - One and only one numerical value is assigned to each sample point
- Two types of Random Variable
  - Discrete
  - Continuous

Two Types of Random Variables

- Discrete Random Variable
  - Random variable that has a finite, or countable number of distinct possible values
  - Example
    - Number of people born in July
    - Number of steps to the top of the Eiffel Tower
- Continuous Random Variable
  - Random variable that has an infinite number of distinct possible values
    - Average age of people born in July
    - The time a tourist stays at the top once s/he gets there

Two Types of Random Variables

- **Discrete random variables**
  - Number of sales
  - Number of calls
  - Shares of stock
  - People in line
  - Mistakes per page

- **Continuous random variables**
  - Length
  - Depth
  - Volume
  - Time
  - Weight

Probability Distributions for Discrete Random Variables

- The **probability distribution** of a discrete random variable is a graph, table or formula that specifies the probability associated with each possible outcome the random variable can assume.
- It is known as ***probability mass function***
- 2 Requirements that must be satisfied

1. $$p(x) \geq 0$$ for all values of *x.*

2. $$\sum p(x) = 1$$ where the summation of *p(x)* is over all possible values of *x.*

Probability Distributions for Discrete Random Variables

- Say a random variable x follows this pattern: $p(x) = (0.3)(0.7)^{x-1}$ for $x > 0$.
  - This table gives the probabilities (rounded to two digits) for x between 1 and 10.

x	$P(x)$
1	.30
2	.21
3	.15
4	.11
5	.07
6	.05
7	.04
8	.02
9	.02
10	.01

Probability Distributions for Discrete Random Variables

- Experiment - tossing 2 coins simultaneously
- Random variable X – number of heads observed
- X can assume values of 0, 1, and 2

- Calculate the probability associated with each value of X

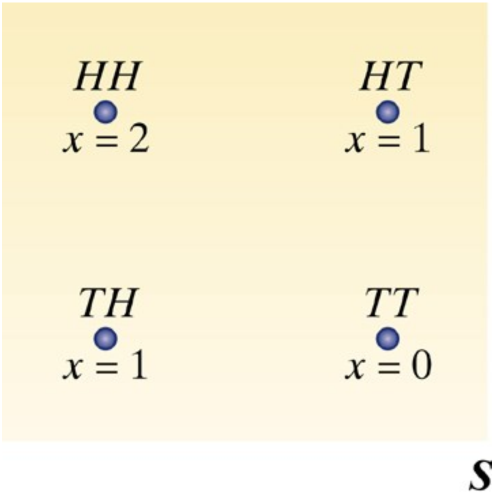


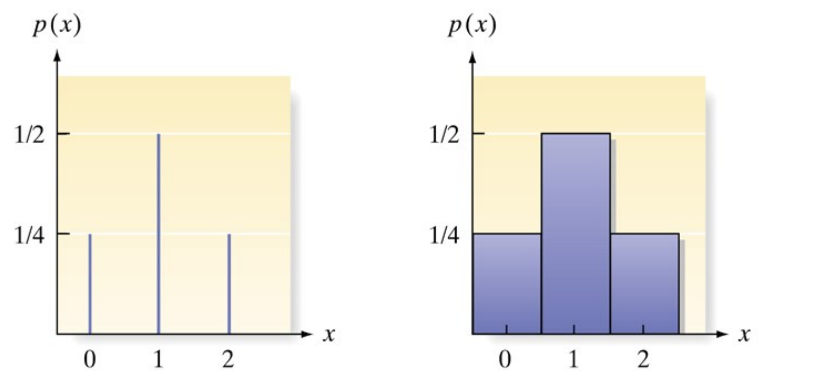
TABLE 4.1 Probability Distribution for Coin-Toss Experiment: Tabular Form

x	$p(x)$
0	$1/4$
1	$1/2$
2	$1/4$

Probability Distributions for Discrete Random Variables

- Probability Distribution of Discrete Random Variable X

$$P(x=0)=P(TT)=1/4$$
$$P(x=1)=P(TH)+P(HT)=1/4+1/4=1/2$$
$$P(x=2)=P(HH)=1/4$$



a. Point representation of $p(x)$ b. Histogram representation of $p(x)$

Expected Values of Discrete Random Variables

- The mean, or expected value of a discrete random variable is:

$$\mu = \mathrm{E}(x) = \sum xp(x)$$

Expected Value of x (number of heads observed)		
x	*p(x)*	*xp(x)*
0	¼	0
1	½	½
2	¼	½
Expected Value		1

Expected Values of Discrete Random Variables

- The variance of a discrete random variable is:

$$\sigma^2 = E[(x-\mu)^2] = \sum (x-\mu)^2 p(x) = \sum x^2 p(x) - \mu^2$$

$$= (0-1)^2 \times \frac{1}{4} + (1-1)^2 \times \frac{1}{2} + (2-1)^2 \times \frac{1}{4} = \frac{1}{2}$$

and standard deviation is

$$\sigma = \sqrt{\sigma^2} = \sqrt{1/2}$$

- The variance may also be expressed as *V(x)*.

Probability Distributions for Discrete Random Variables

	Chebyshev's Rule	Empirical Rule
	Apply to any distribution	Apply to mound-shaped and symmetric distributions
$P(\mu - \sigma < x < \mu + \sigma)$	≥ 0	$\cong 0.68$
$P(\mu - 2\sigma < x < \mu + 2\sigma)$	≥ 0.75	$\cong 0.95$
$P(\mu - 3\sigma < x < \mu + 3\sigma)$	≥ 0.89	$\cong 1.00$

Probability Distributions for Discrete Random Variables

- In a roulette wheel in a U.S. casino, a $1 bet on "even" wins $1 if the ball falls on an even number (same for "odd," or "red," or "black").
- The odds of winning this bet are 47.37%

$$P(win\,\$1) = 0.4737$$

$$P(lose\,\$1) = 0.5263$$

$$\mu = +\$1 \times 0.4737 - \$1 \times 0.5263 = -0.0526$$

$$\sigma = 0.9986$$

On average, bettors lose about a nickel (5 cents) for each dollar they put down on a bet like this.
(These are the *best* bets for patrons.)

Mean and Variance of a Discrete Random Variable

Expected Value of a Function of a Discrete Random Variable

If X is a discrete random variable with probability mass function $f(x)$, then

$$E[h(x)] = \sum_{x} h(x) f(x)$$

Discrete Uniform Distribution

Definition

A random variable X has a discrete uniform distribution if each of the n values in its range, say, $x_1, \ldots, x_n$, has equal probability. Then,

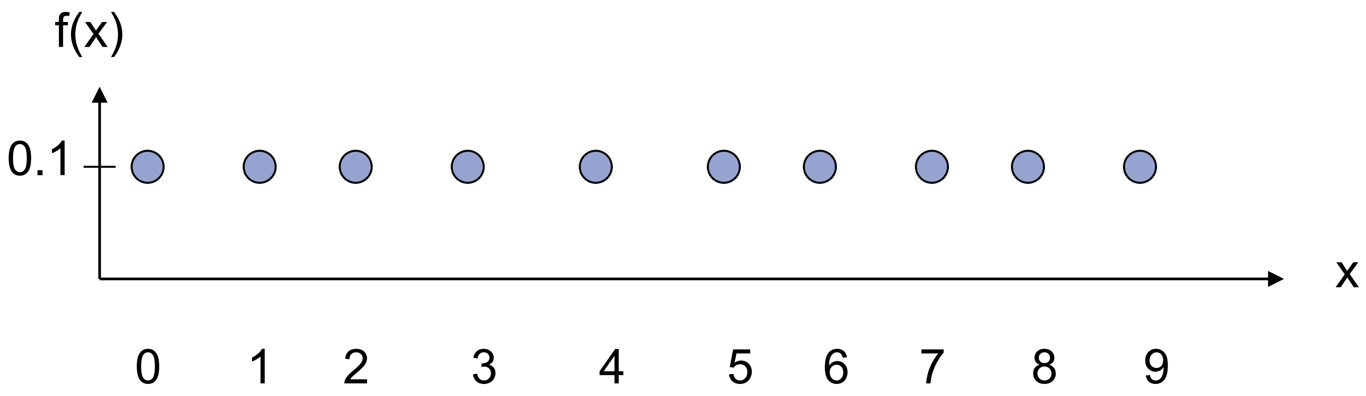
$$f(x_i) = 1/n$$

Discrete Uniform Distribution

Example

The first digit of a part's serial number is equally likely to be any one of the digits 0 through 9. If one part is selected from a large batch and X is the first digit of the serial number, X has a discrete uniform distribution with probability 0.1 for each value in $R = \{0, 1, 2, \ldots, 9\}$. That is $f(x) = 0.1$.

Discrete Uniform Distribution



Probability mass function for a discrete uniform random variable.

Discrete Uniform Distribution

Mean and Variance

Suppose X is a discrete uniform random variable on the consecutive integers *a, a+1, …, b*, for *a* ≤ *b*, then

$$\mu = E(X) = \frac{b+a}{2}$$

$$\sigma^2 = \frac{(b-a+1)^2 - 1}{12}$$

$$= \sum (x_i - \bar{x})^2 p(x_i) = \frac{1}{n} \sum (x_i - \bar{x})^2$$

Note: The variance here is not the same as the *sample variance* we have seen in Chapter 2 because there it has not considered a probability.

The Binomial Random Variable

- Binomial Random variable
  - An experiment of *n* identical trials
  - 2 possible outcomes on each trial, denoted as **S** (success) and **F** (failure), also called *Bernoulli* trial
  - Probability of success (*p*) is constant from trial to trial. Probability of failure (*q*) is *1-p*
  - Trials are independent
  - Binomial random variable, number of S's in *n* trials

The Binomial Random Variable

- A Binomial Random Variable
  - n identical trials → Flip a coin 3 times
  - Two outcomes: **S**uccess or **F**ailure → Outcomes are Heads or Tails
  - P(**S**) = p; P(**F**) = q = 1 – p → P(H) = 0.5; P(F) = 1-0.5 = 0.5
  - Trials are independent → A head on flip i doesn't change P(H) of flip i + 1
  - x is the number of **S**'s in n trials

The Binomial Random Variable

Results of 3 flips	Probability	Combined	Summary
HHH	$(p)(p)(p)$	p^3	$(1)p^3q^0$
HHT	$(p)(p)(q)$	p^2q	
HTH	$(p)(q)(p)$	p^2q	$(3)p^2q^1$
THH	$(q)(p)(p)$	p^2q	
HTT	$(p)(q)(q)$	pq^2	
THT	$(q)(p)(q)$	pq^2	$(3)p^1q^2$
TTH	$(q)(q)(p)$	pq^2	
TTT	$(q)(q)(q)$	q^3	$(1)p^0q^3$

The Binomial Random Variable

- Heart association claims that only 10% of US adults over 30 can pass the President's Physical Fitness commission's minimum requirements.
- Select 4 adults at random, administer the test. What is the probability that none of the adults passes the test?

TABLE 4.2 Sample Points for Fitness Test of Example 4.9

SSSS	*FSSS*	*FFSS*	*SFFF*	*FFFF*
	SFSS	*FSFS*	*FSFF*	
	SSFS	*FSSF*	*FFSF*	
	SSSF	*SFFS*	*FFFS*	
		SFSF		
		SSFF		

The Binomial Random Variable

- Use multiplicative rule to calculate probabilities of the possible outcomes

$$P(SSSS) = 0.1*0.1*0.1*0.1 = 0.1^4 = 0.0001$$

$$P(FSSS) = 0.9*0.1*0.1*0.1 = 0.9*0.1^3 = 0.0009$$

…..

$$P(FFFF) = 0.9*0.9*0.9*0.9 = 0.9^4 = 0.6561$$

The Binomial Random Variable

- What is the probability that 3 of the 4 adults pass the test?

P(3 of the 4 adults pass the test) = $4(0.1)^3(0.9) = 4(0.0009) = 0.0036$

- What is the probability that 3 of the 4 adults fail the test?

P(3 of the 4 adults fail the test) = $4(0.9)^3(0.1) = 4(0.0729) = 0.2916$

- Do you see a pattern?

The Binomial Random Variable

- Formula for the probability distribution p(x)

$$p(x) = \binom{n}{x} p^x q^{n-x}$$

- Where p = probability of success on single trial

  q = 1-p

  n = Number of trials

  x = number of successes in n trials

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

The Binomial Distribution

- Say 40% of the class is female.
- What is the probability that 6 of the first 10 students walking in will be female?

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

$$= \binom{10}{6} (0.4^6)(0.6^{10-6})$$

$$= 210(0.004096)(0.1296)$$

$$= 0.1115$$

The Binomial Random Variable

- Mean: $\mu = np$
- Variance: $\sigma^2 = npq$
- Standard deviation $\sigma = \sqrt{npq}$

The Binomial Distribution

- For 1,000 coin flips,

$$\mu = np = 1000 \cdot 0.5 = 500$$

$$\sigma^2 = npq = 1000 \cdot 0.5 \cdot 0.5 = 250$$

$$\sigma = \sqrt{npq} = \sqrt{250} \cong 16$$

The actual probability of getting exactly 500 heads out of 1000 flips is just over 2.5%, but the probability of getting between 484 and 516 heads (that is, within one standard deviation of the mean) is about 68%.

$$\binom{1000}{500} 0.5^{500} 0.5^{500} = 0.0252$$

The Binomial Random Variable

- Using Binomial Tables
- Binomial tables are cumulative tables, entries represent cumulative binomial probabilities
- Make use of additive and complementary properties to calculate probabilities of individual x's, or x being greater than a particular value.

The Binomial Random Variable

- If x ≤ 2, and p =0.1, n =10, then P(x≤2) = 0.930
- If x = 2, and p =0.1, n =10, then P(x=2) = P(x≤2) - P(x≤1)=0.930-0.736 = 0.194
- If x > 2, and p =0.1, n =10, then P(x>2) = 1- P(x≤2) =1-0.930 = 0.070

TABLE 4.4 Reproduction of Part of Table II of Appendix A: Binomial Probabilities for $n = 10$

k \ p	.01	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95	.99
0	.904	.599	.349	.107	.028	.006	.001	.000	.000	.000	.000	.000	.000
1	.996	.914	.736	.376	.149	.046	.011	.002	.000	.000	.000	.000	.000
2	1.000	.988	.930	.678	.383	.167	.055	.012	.002	.000	.000	.000	.000
3	1.000	.999	.987	.879	.650	.382	.172	.055	.011	.001	.000	.000	.000
4	1.000	1.000	.998	.967	.850	.633	.377	.166	.047	.006	.000	.000	.000
5	1.000	1.000	1.000	.994	.953	.834	.623	.367	.150	.033	.002	.000	.000
6	1.000	1.000	1.000	.999	.989	.945	.828	.618	.350	.121	.013	.001	.000
7	1.000	1.000	1.000	1.000	.998	.988	.945	.833	.617	.322	.070	.012	.000
8	1.000	1.000	1.000	1.000	1.000	.988	.989	.954	.851	.624	.264	.086	.004
9	1.000	1.000	1.000	1.000	1.000	1.000	.999	.994	.972	.893	.651	.401	.096

Geometric Distributions

Definition

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. The *probability mass function* is

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

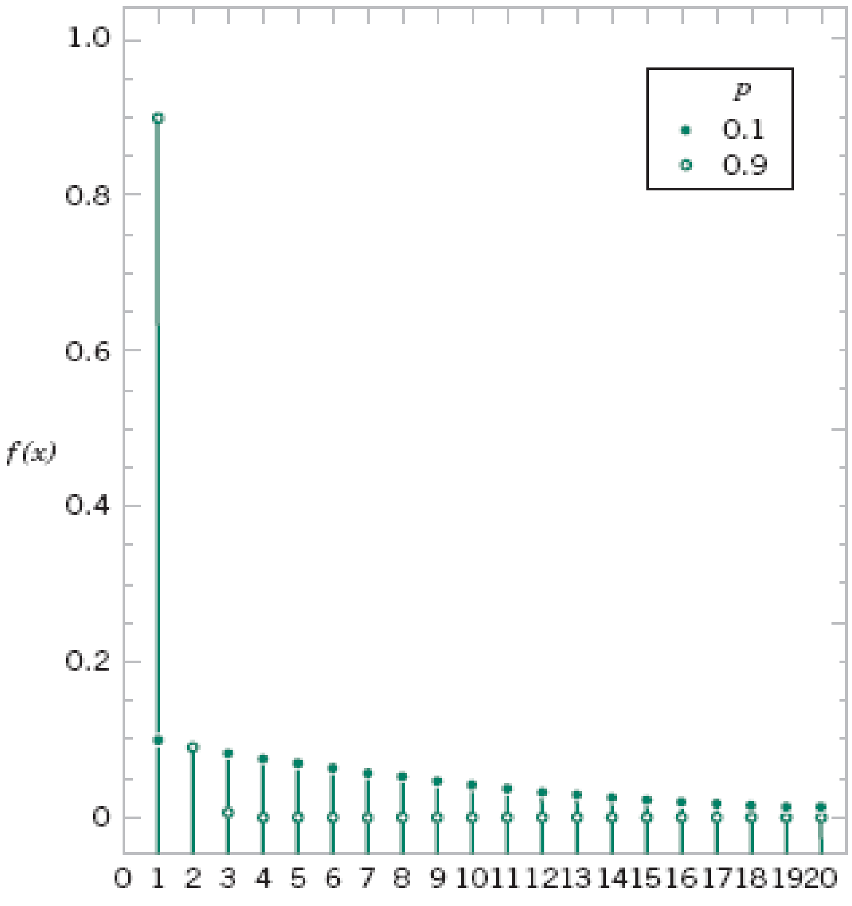
for x = 1, 2, 3, 4, ….

It has an expected value and a variance of:

$$\mu = \frac{1}{p} \quad \text{and} \quad \sigma^2 = \frac{1-p}{p^2}$$

Source: Anthony Hayter

Geometric Distributions



Geometric distributions for selected values of the parameter *p*.

Geometric Distributions

Example

The probability that a bit transmitted through a digital transmission channel is received in error is 0.1. Assume the transmissions are independent events, and let the random variable X denote the number of bits transmitted *until* the first error.

Then, $P(X = 5)$ is the probability that the first four bits are transmitted correctly and the fifth bit is in error. This event can be denoted as $\{OOOOE\}$, where O denotes an okay bit. Because the trials are independent and the probability of a correct transmission is 0.9,

$$P(X = 5) = P(OOOOE) = 0.9^4 0.1 = 0.066$$

Note that there is some probability that X will equal any integer value. Also, if the first trial is a success, $X = 1$. Therefore, the range of X is $\{1, 2, 3, \ldots\}$, that is, all positive integers.

Geometric Distributions

Example

The probability that a wafer contains a large particle of contamination is 0.01. If it is assumed that the wafers are independent, what is the probability that exactly 125 wafers need to be analyzed before a large particle is detected?

Let X denote the number of samples analyzed until a large particle is detected. Then X is a geometric random variable with $p = 0.01$. The requested probability is

$$P(X = 125) = (0.99)^{124}0.01 = 0.0029$$

Hypergeometric Distribution

Definition

It represents the distribution of the number of items of a certain kind in a random sample of size ***n*** drawn without replacement from a population of size ***N*** that contains ***r*** items of this kind. It has a *probability mass function* of:

$$P(X = x) = \frac{\binom{r}{x} \times \binom{N-r}{n-x}}{\binom{N}{n}}$$

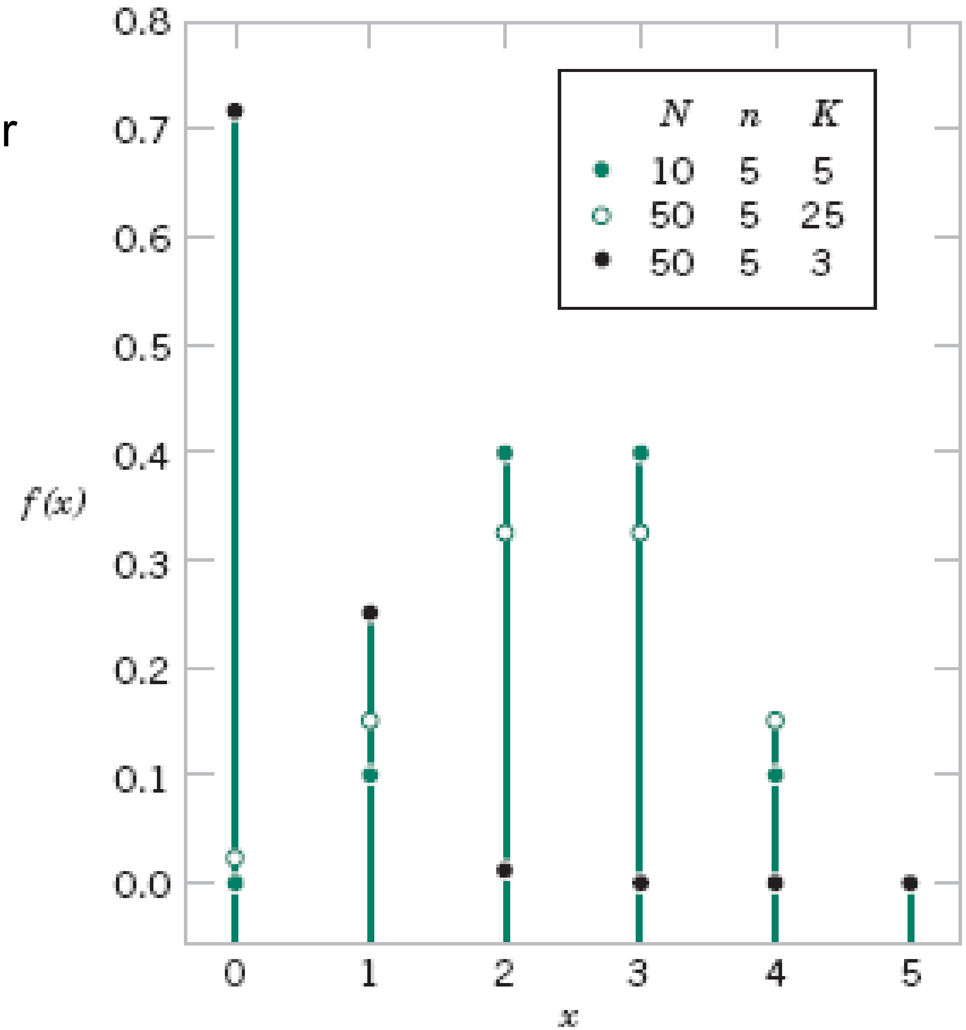
for max{ 0, $n+r-N$ } $\leq x \leq$ min{ n, r }, with

$$\mu = \frac{nr}{N} \text{ and } \sigma^2 = \left(\frac{N-n}{N-1}\right) \times n \times \frac{r}{N} \times \left(1 - \frac{r}{N}\right)$$

Source: Anthony Hayter

Hypergeometric Distribution

Hypergeometric distributions for selected values of parameters N, K, and n.



Hypergeometric Distribution

Example

A batch of parts contains 100 parts from a local supplier of tubing and 200 parts from a supplier of tubing in the next state. If four parts are selected randomly and without replacement, what is the probability they are all from the local supplier?

Let X equal the number of parts in the sample from the local supplier. Then, X has a hypergeometric distribution and the requested probability is $P(X = 4)$. Consequently,

$$P(X = 4) = \frac{\binom{100}{4}\binom{200}{0}}{\binom{300}{4}} = 0.0119$$

Hypergeometric Distribution

Example

What is the probability that two or more parts in the sample are from the local supplier?

$$P(X \geq 2) = \frac{\binom{100}{2}\binom{200}{2}}{\binom{300}{4}} + \frac{\binom{100}{3}\binom{200}{1}}{\binom{300}{4}} + \frac{\binom{100}{4}\binom{200}{0}}{\binom{300}{4}}$$

$$= 0.298 + 0.098 + 0.0119 = 0.408$$

What is the probability that at least one part in the sample is from the local supplier?

$$P(X \geq 1) = 1 - P(X = 0) = 1 - \frac{\binom{100}{0}\binom{200}{4}}{\binom{300}{4}} = 0.804$$

Poisson Distribution

Definition

A random variable X distributed as a Poisson random variable with parameter λ has a *probability mass function*

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

for x = 0, 1, 2, 3, …. The Poisson distribution is often useful to model the number of times that a certain event occurs per unit of time, distance, or volume, and it has a mean and variance both equal to the parameter value λ.

$$\mu = \sigma^2 = \lambda$$

Source: Anthony Hayter

Poisson Distribution

Example

Contamination is a problem in the manufacture of optical storage disks. The number of particles of contamination that occur on an optical disk has a Poisson distribution, and the average number of particles per centimeter squared of media surface is 0.1. The area of a disk under study is 100 squared centimeters. Find the probability that 12 particles occur in the area of a disk under study.

Let X denote the number of particles in the area of a disk under study. Because the mean number of particles is 0.1 particles per cm^2

$$E(X) = 100 \text{ cm}^2 \times 0.1 \text{ particles/cm}^2 = 10 \text{ particles}$$

Therefore,

$$P(X = 12) = \frac{e^{-10} 10^{12}}{12!} = 0.095$$

Poisson Distribution

Example

The probability that zero particles occur in the area of the disk under study is

$$P(X = 0) = e^{-10} = 4.54 \times 10^{-5}$$

Determine the probability that 12 or fewer particles occur in the area of the disk under study. The probability is

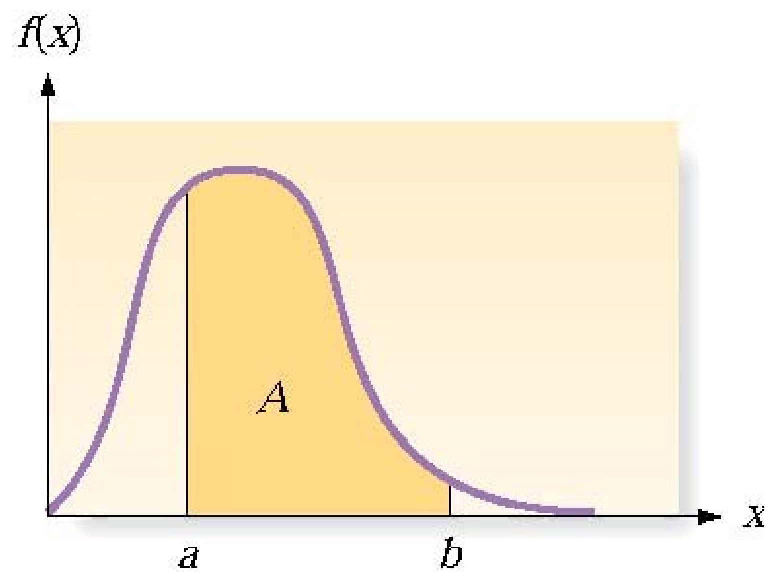
$$P(X \leq 12) = P(X = 0) + P(X = 1) + \cdots + P(X = 12) = \sum_{i=0}^{12} \frac{e^{-10} 10^i}{i!}$$

Continuous Probability Distributions

- A **continuous random variable** can assume any numerical value within some interval or intervals.
- The graph of the probability distribution is a smooth curve called a
  - ***probability density function***
  - *frequency function*
  - *probability distribution.*

Continuous Probability Distributions

- There are an infinite number of possible outcomes
  - p(x) = 0
  - Instead, find p(a<x<b) using
    - Table
    - Software
    - Integral calculus



The Normal Distribution

- Closely approximates many situations
  - Perfectly symmetrical around its mean

- The probability density function *f(x):*

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

μ = the mean of x

σ = the standard deviation of x

π = 3.1416…

e = 2.71828 …

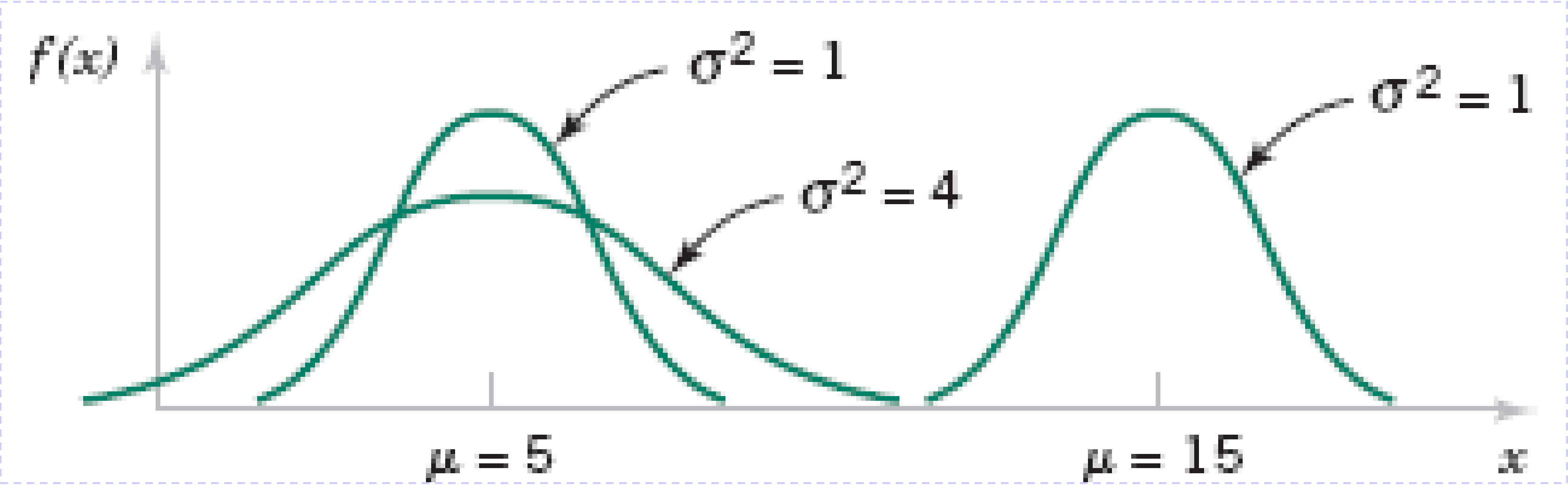
The Normal Distribution

- Each combination of μ and ? produces a unique normal curve
- The **standard normal curve** is used in practice, based on the standard normal random variable *z* (μ = 0, σ = 1), with the probability distribution

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

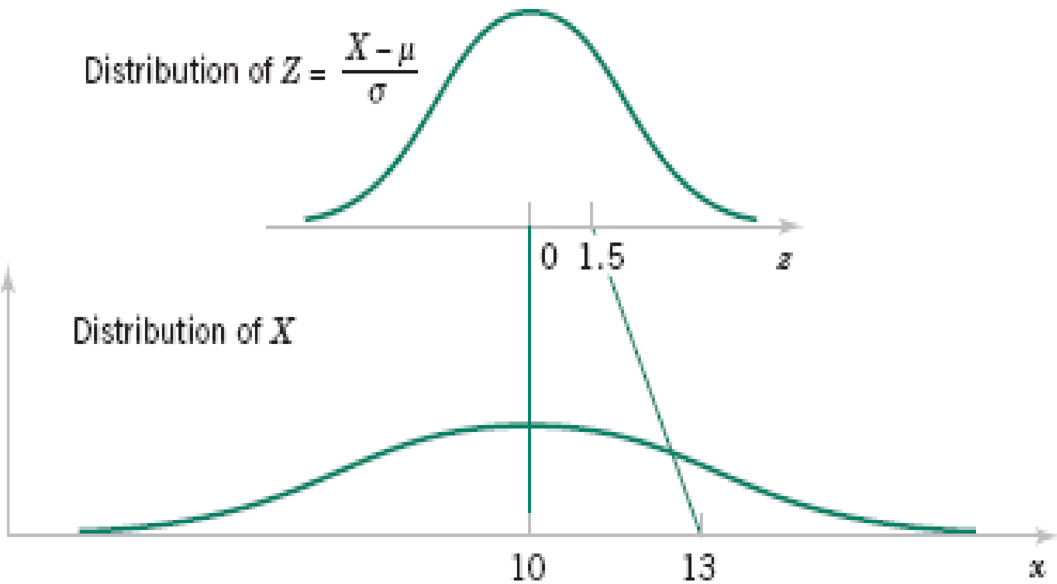
The probabilities for *z* are given in Table III

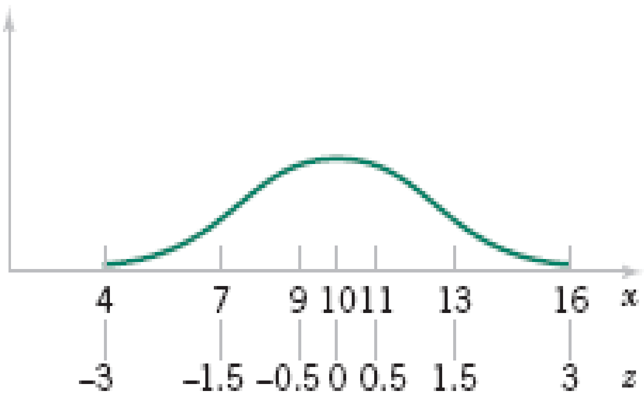
The Normal Distribution



Normal probability density functions for selected values of the parameters μ and σ^2.

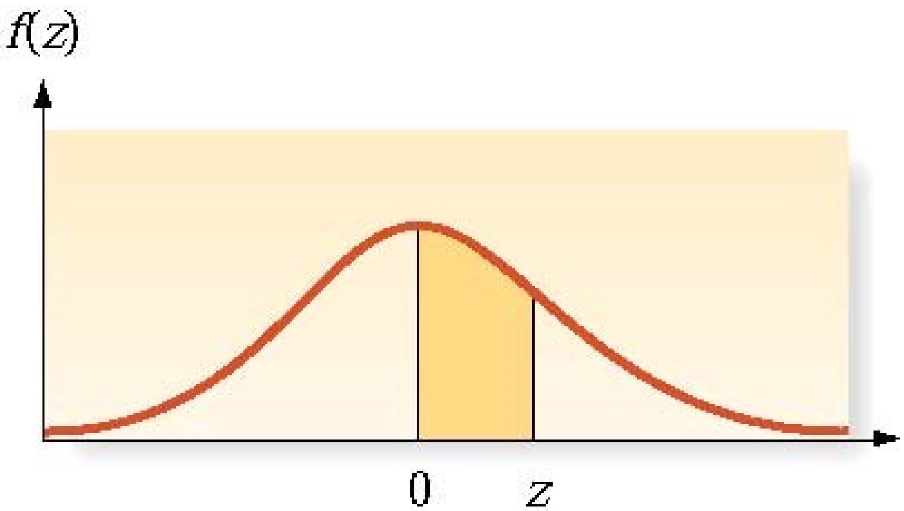
The Normal Distribution





Standardizing a normal random variable.

The Normal Distribution



$$P(0 < z < 1.00) \quad = .3413$$

$$P(-1.00 < z < 0) = .3413$$

$$P(-1 < z < 1) \quad = .3413 + .3413$$
$$= .6826$$

$$P(1 < z < 1.25) \quad =$$
$$P(0 < z < 1.25) - P(0 < z < 1.00)$$
$$= .3944 - .3413 \; = .0531$$

The Normal Distribution

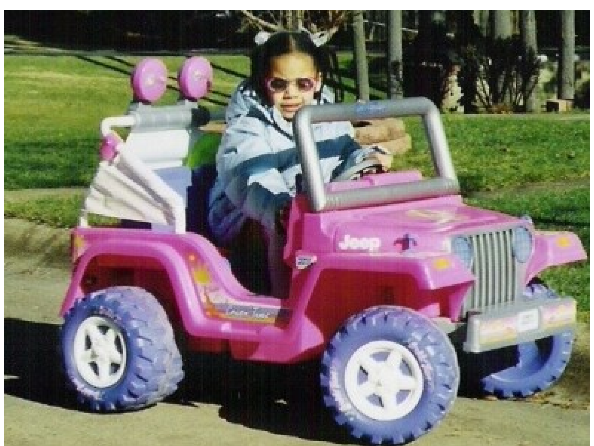
For a normally distributed random variable *x*, if we know μ and ?,

$$z_i = \frac{x_i - \mu}{\sigma}$$

So *any* normally distributed variable can be analyzed with this single distribution

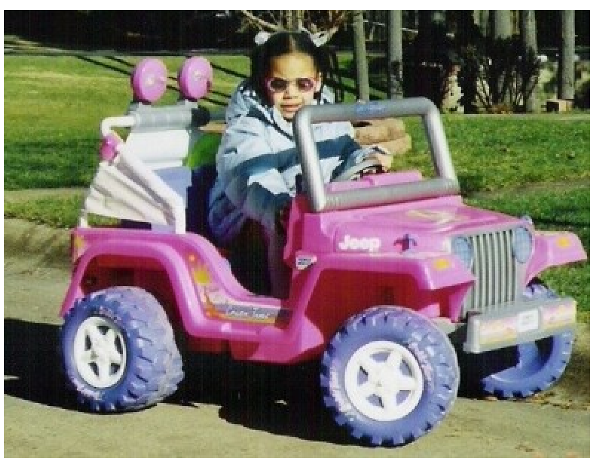
The Normal Distribution

- Say a toy car goes an average of 3,000 yards between recharges, with a standard deviation of 50 yards (i.e., $\mu = 3{,}000$ and $\sigma = 50$)
- What is the probability that the car will go more than 3,100 yards without recharging?



The Normal Distribution

- Say a toy car goes an average of 3,000 yards between recharges, with a standard deviation of 50 yards (i.e., μ = 3,000 and σ = 50)
- What is the probability that the car will go more than 3,100 yards without recharging?



$$P(x > 3100) = P\left(z > \frac{3100 - 3000}{50}\right) =$$

$$P(z > 2.00) \; = 1 - P(z < 2.00) =$$

$$1 - .5 - P(0 < z < 2.00) =$$

$$1 - .5 - .4772 = .0228$$

The Normal Distribution

- To find the probability for a normal random variable …
  - Sketch the normal distribution
  - Indicate *x*'s mean
  - Convert the *x* variables into *z* values
  - Put both sets of values on the sketch, *z* below *x*
  - Use Table III in the Appendix A to find the desired probabilities

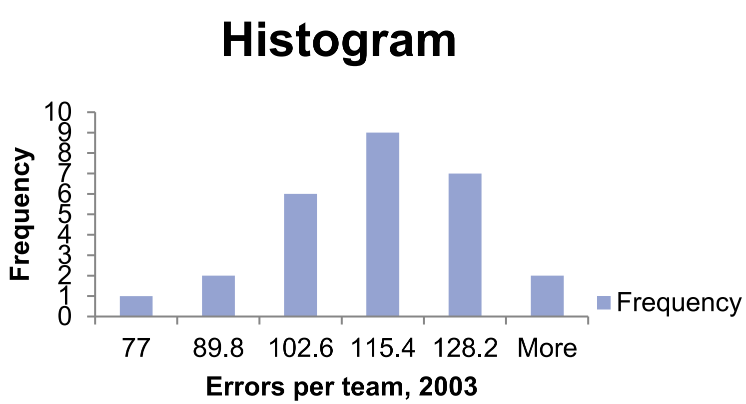
Descriptive Methods for Assessing Normality

- If the data are normal
  - A histogram or stem-and-leaf display will look like the normal curve
  - The mean ± s, 2s and 3s will approximate the empirical rule percentages
  - The ratio of the Interquartile range to the standard deviation will be about 1.3
  - A *normal probability plot* , a scatterplot with the ranked data on one axis and the expected z-scores from a standard normal distribution on the other axis, will produce close to a straight line

Descriptive Methods for Assessing Normality

Errors per MLB team in 2003

- Mean: 106
- Standard Deviation: 17
- IQR: 22



$$\frac{IQR}{s} = \frac{22}{17} = 1.29$$

$$\bar{x} \pm s = 106 \pm 17$$

$$89 \Leftrightarrow 123$$ 22 out of 30: 73%

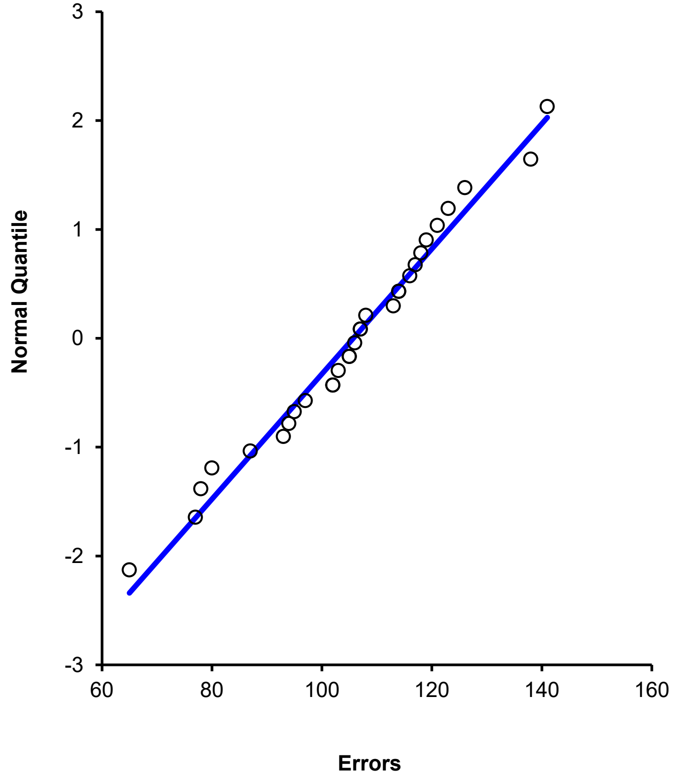
$$\bar{x} \pm 2s = 106 \pm 34$$

$$72 \Leftrightarrow 140$$ 28 out of 30: 93%

$$\bar{x} \pm 3s = 106 \pm 51$$

$$55 \Leftrightarrow 157$$ 30 out of 30: 100%

Descriptive Methods for Assessing Normality



A **normal probability plot** is a scatterplot with the ranked data on one axis and the expected z-scores from a standard normal distribution on the other axis

Approximating a Binomial Distribution with the Normal Distribution

- Discrete calculations may become very cumbersome
- The normal distribution may be used to approximate discrete distributions
  - The larger n is, and the closer p is to 0.5, the better the approximation
- Since we need a range, not a value, the **correction for continuity** must be used
  - A number r becomes $r+0.5$

Approximating a Binomial Distribution with the Normal Distribution

Calculate the mean plus/minus 3 standard deviations

$$\mu \pm 3\sigma = np \pm 3\sqrt{npq}$$

If this interval is in the range 0 to *n*, the approximation will be reasonably close

Express the binomial probability as a range of values

$$P(x \leq a)$$

$$P(x \leq b) - P(x \leq a)$$

Find the z-values for each binomial value

$$z = \frac{(a + 0.5) - \mu}{\sigma}$$

Use the standard normal distribution to find the probability for the range of values you calculated

Approximating a Binomial Distribution with the Normal Distribution

Flip a coin 100 times and compare the binomial and normal results

Binomial: $$P(x=50)=\binom{100}{50}0.5^{50}\cdot 0.5^{50}=0.0796$$

Normal: $$\mu=100\cdot 0.5=50$$

$$\sigma=\sqrt{100\cdot 0.5\cdot 0.5}=5$$

$$P(49.5\le x\le 50.5)=P\left(\frac{49.5-50}{5}\le z\le\frac{50.5-50}{5}\right)=$$

$$P(-0.10\le z\le= 0.10)=0.0796$$

Approximating a Binomial Distribution with the Normal Distribution

Flip a weighted coin *[P(H)=0.4]* 10 times and compare the results

Binomial: $P(x=5)=\binom{10}{5}0.4^5\cdot 0.6^5=0.1204$

Normal: $\mu=10\cdot 0.4=4$

$$\sigma=\sqrt{10\cdot 0.4\cdot 0.6}=1.55$$

$$P(4.5\le x\le 5.5)=P\left(\frac{4.5-4}{1.55}\le z\le\frac{5.5-4}{1.55}\right)=$$

$$P(-0.32\le z\le= 0.32)=0.1255$$

Approximating a Binomial Distribution with the Normal Distribution

Flip a weighted coin *[P(H)=0.4]* 10 times and compare the results

Binomial: $$P(x=5)=\binom{10}{5}0.4^5\cdot 0.6^5=0.1204$$

Normal: $$\mu=10\cdot 0.4=4$$

$$\sigma=\sqrt{10\cdot 0.4\cdot 0.6}=1.55$$

$$P(4.5\le x\le 5.5)=P\left(\frac{4.5-4}{1.55}\le z\le \frac{5.5-4}{1.55}\right)=$$

$$P(-0.32\le z\le 0.32)=0.1255$$

The more *p* differs from 0.5, and the smaller n is, the less precise the approximation will be

Sampling Distributions

- In practice, sample statistics are used to estimate population parameters.
  - A **parameter** is a numerical descriptive measure of a population. Its value is almost always unknown.
  - A **sample statistic** is a numerical descriptive measure of a sample. It can be calculated from the observations.

Sampling Distributions

	Parameter	Statistic
Mean	μ	$\bar{x}$
Variance	σ^2	s^2
Standard Deviation	σ	s
Binomial proportion	p	$\hat{p}$

Sampling Distributions

- Since we could draw many different samples from a population, the sample statistic used to estimate the population parameter is itself a random variable.
- The **sampling distribution** of a sample statistic calculated from a sample of *n* measurements is the probability distribution of the statistic.

Sampling Distributions

Imagine a very small population consisting of the elements 1, 2 and 3. Below are the possible samples that could be drawn, along with the means of the samples and the *mean of the means*.

n = 1	$\bar{x}$
1	1
2	2
3	3

n = 2	$\bar{x}$
1, 2	1.5
1, 3	2
2, 3	2.5

n = 3 (= N)	$\bar{x}$
1, 2, 3	2

$$\frac{\sum \bar{x}}{3} = 2$$

$$\frac{\sum \bar{x}}{3} = 2$$

$$\frac{\sum \bar{x}}{1} = 2$$

The Central Limit Theorem

- A **point estimator** is a single number based on sample data that can be used as an *estimator* of the population parameter

$\bar{x} \longrightarrow \mu$

$\hat{p} \longrightarrow p$

$s^2 \longrightarrow \sigma^2$

The Central Limit Theorem

Properties of the Sampling Distribution of $\bar{x}$

The mean of the sampling distribution equals the mean of the population

$$\mu_{\bar{x}} = E(\bar{x}) = \mu$$

The standard deviation of the sampling distribution [the **standard error (of the mean)**] equals the population standard deviation divided by the square root of *n*

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

The Central Limit Theorem

Here's our small population again, this time with the standard deviations of the sample means. Notice the mean of the sample means in each case equals the population mean and the standard error falls as *n* increases.

n = 1	$\bar{x}$
1	1
2	2
3	3

n = 2	$\bar{x}$
1, 2	1.5
1, 3	2
2, 3	2.5

n = 3 (= N)	$\bar{x}$
1, 2, 3	2

$$\frac{\sum \bar{x}}{3} = 2$$

$$\sigma_{\bar{x}} = .82$$

$$\frac{\sum \bar{x}}{3} = 2$$

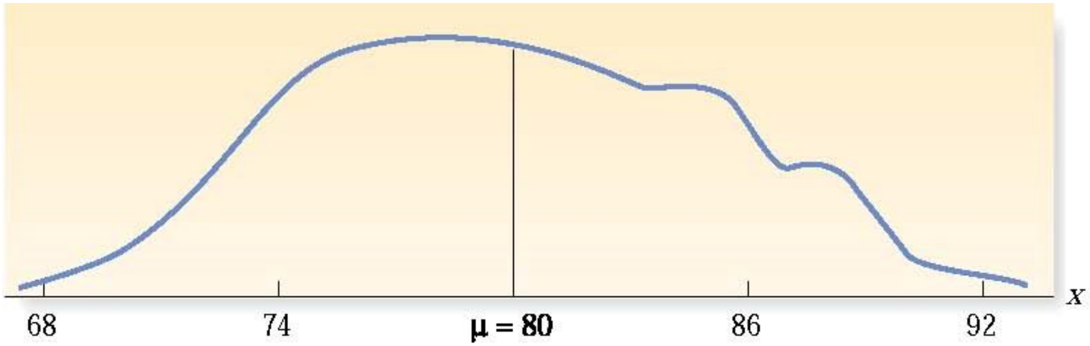
$$\sigma_{\bar{x}} = .41$$

$$\frac{\sum \bar{x}}{1} = 2$$

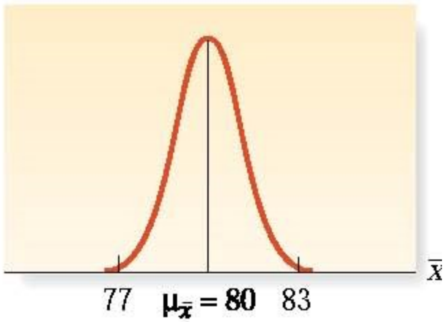
$$\sigma_{\bar{x}} = 0$$

The Central Limit Theorem

- If a random sample of *n* observations is drawn from a normally distributed population, the sampling distribution of $\bar{x}$ will be normally distributed



a. Population relative frequency distribution



b. Sampling distribution of $\bar{x}$

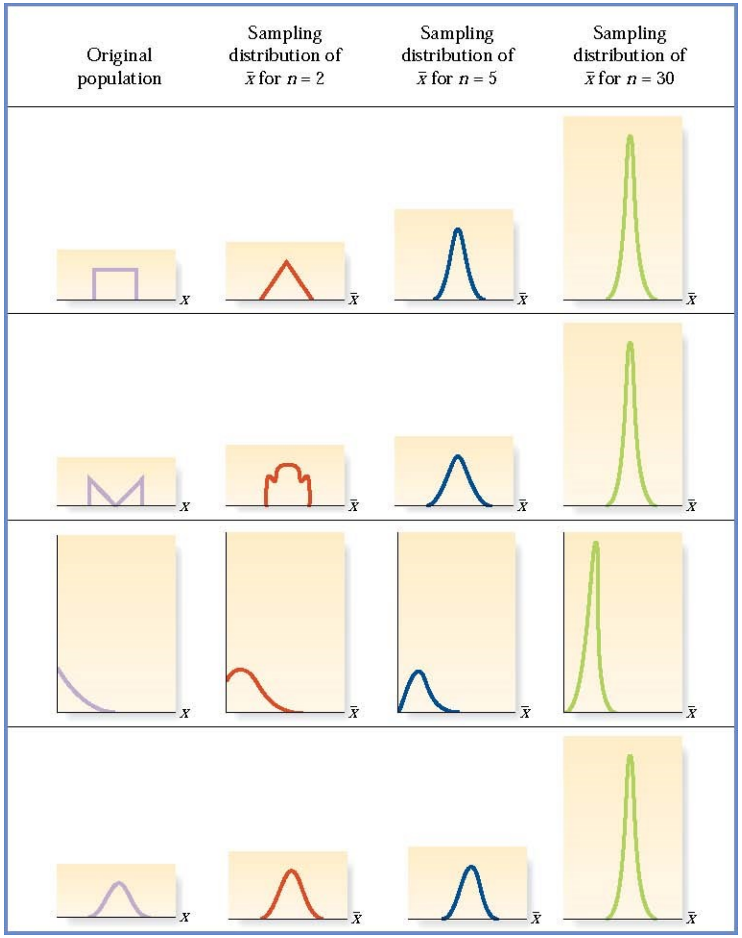
The Central Limit Theorem

- The Central Limit Theorem

  The sampling distribution of $\bar{x}$, based on a random sample of *n* observations, will be approximately normal with

$$\mu_{\bar{x}} = \mu \text{ and } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

  The larger the sample size, the better the sampling distribution will approximate the normal distribution.



The Central Limit Theorem

Suppose existing houses for sale average 2200 square feet in size, with a standard deviation of 250 ft^2.

What is the probability that a randomly selected house will have at least 2300 ft^2 ?

The Central Limit Theorem

Suppose existing houses for sale average 2200 square feet in size, with a standard deviation of 250 ft^2.

What is the probability that a randomly selected house will have at least 2300 ft^2 ?

$$P(x \geq 2300) =$$

$$P\left(z \geq \frac{2300 - 2200}{250}\right) =$$

$$P(z \geq 0.40) = .3446$$

The Central Limit Theorem

Suppose existing houses for sale average 2200 square feet in size, with a standard deviation of 250 ft^2.

What is the probability that a randomly selected sample of 16 houses will average at least 2300 ft^2 ?

The Central Limit Theorem

Suppose existing houses for sale average 2200 square feet in size, with a standard deviation of 250 ft^2.

What is the probability that a randomly selected sample of 16 houses will average at least 2300 ft^2 ?

$$P(\bar{x} \geq 2300) =$$

$$P\left(z \geq \frac{2300 - 2200}{250 / \sqrt{16}} \right) =$$

$$P(z \geq 1.60) = .0548$$