Random Variables and Probability Models: Binomial, Geometric, Poisson, and Normal Distributions

Heads I win, tails you lose.

17th century English saying

Let the king prohibit gambling and betting in his kingdom, for these are vices that destroy the kingdom of princes.

The Code of Manu, ca. 100 A.D.

Chances rule men and not men chances

Herodotus, History Bk vii, ch. 49

IN THE REAL WORLD

(i) The advertisement below claims that consumers overwhelmingly prefer Hardee's fried chicken to KFC fried chicken. The claim is based on the results of a sample of 100 tasters who chose Hardee's 63 – 37. Is this really a landslide as claimed, or is it within the realm of reasonable outcomes if the tasters were simply tossing a fair coin to make their decision.

(ii) The newspaper article Mothers identify new babies by touch reports that 70 percent of blindfolded mothers who had spent at least an hour with their babies since birth could later choose their own children out of groups of three sleeping babies. The results showed that 22 of 32 women identified their own babies. “That's far better than the 33 percent one would expect by random guessing,” researchers said. Dr. Michael Yogman, an assistant clinical professor of pediatrics at Harvard Medical School, called the study “a pretty impressive piece of work.” Are the results really that unusual if the mothers are guessing? Just how “impressive” and convincing are these numbers? Can we quantify “far better”? 

(iii) A branch bank manager is concerned about the long lines at the ATM facility outside her bank. The excessive congestion is a hinderance to customers who need to come into the bank to conduct their business. She is convinced that the installation of another ATM would solve the problem, but she needs data to justify to her superiors the added expense of installing and maintaining another ATM. What kind of data does she need? How can she model the queue behavior at an ATM?

Mothers Identify New Babies by Touch

NEW YORK - Mothers are so attuned to their babies that most blindfolded moms in a new study could identify their newborn infants by just feeling the backs of the babies’ hands, researchers say.

Nearly 70% of mothers who had spent at least one hour with their babies since birth could later choose their own children out of groups of three sleeping babies.

That's far better than the 33 percent one would expect by random guessing, researchers said.

Dr. Michael Yogman, an assistant clinical professor of pediatrics at Harvard Medical School, called the study a pretty impressive piece of work.

Each participant's eyes and nose were covered with a heavy cotton scarf to block sight and smell. Each mother was tested with her own infant plus two others of the same sex and no more than six hours difference in age.

“We were very surprised. The women themselves are surprised.”

-- Marsha Kaitz, co-author of study

The results showed that 22 of 32 women with at least an hour of exposure to their infants identified their own babies.

Hardees Wins in a Landslide!

Hardees Press Release

TROUNCES KFC 63 - 37 IN TASTE TEST

Hardee's
Objectives
At the conclusion of this unit you will be able to:
1) list the 2 properties possessed by the probability distribution of a discrete random variable
2) apply the probability distribution of a random variable to calculate probabilities
3) calculate the expected value and standard deviation of a discrete random variable
4) list the conditions that must be satisfied for the proper application of the binomial, geometric and Poisson probability models
5) apply the binomial, geometric, and Poisson probability models to calculate probabilities
6) calculate probabilities using the normal probability model

Random Variables: Two Types

Random variables

Two types
i) discrete

Examples: number of girls in a 5-child family; number of customers that use an ATM in a 1-hour period; number of tosses of a fair coin that is required until you get 3 heads in a row

ii) continuous

Examples: time it takes to run the 100 yard dash; time between consecutive arrivals at an ATM machine; time spent waiting in the “express” checkout lane at the grocery store.

EXAMPLE: Classify the following random variables as discrete or continuous. Specify the possible values the random variable may assume.

a. $x =$ the number of customers who enter a particular bank during the noon hour on a particular day.
b. $x =$ the time (in seconds) required for a teller to serve a bank customer.
c. $x =$ the distance (in miles) between a randomly selected home in a community and the nearest pharmacy.
d. $x =$ the diameter of precision engineered “5 inch diameter” ball bearings coming off an assembly line.
e. $x =$ the number of tosses of a fair coin required to observe at least three heads in succession.

RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS

Data and data distributions (for example, past stock market behavior):
- tell us what has happened in the past

Random variables are unknown chance outcomes:
- the probability distribution of a random variable gives us an idea of what might happen in the future

EXAMPLE
An economist for Widget, Inc. has developed the following company projections for next year:
Probability distribution of a discrete random variable:
a table or formula that shows all possible values of
the random variable and the probability associated
with each value.

<table>
<thead>
<tr>
<th>Economic Scenario</th>
<th>Profit ($mill.)</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Great</td>
<td>10</td>
<td>0.20</td>
</tr>
<tr>
<td>Good</td>
<td>5</td>
<td>0.40</td>
</tr>
<tr>
<td>OK</td>
<td>1</td>
<td>0.25</td>
</tr>
<tr>
<td>Lousy</td>
<td>-4</td>
<td>0.15</td>
</tr>
</tbody>
</table>

notation: $p(x) =$

What are the chances that profits will be less than $5 million in 2011?
P(X < 5) = P(X = 1 or X = -4)
= P(X = 1) + P(X = -4)
= .25 + .15 = .40

What are the chances that profits will be less than $5 million in 2011 and less than $5 million in 2012?
P(X < 5 in 2002 and X < 5 in 2003) = P(X < 5) \times P(X < 5) = .40 \times .40 = .16

Probability Histogram:

Probability distribution requirements:

EXAMPLE: Determine whether each of the following represents a valid probability distribution. If not, explain why not.

<table>
<thead>
<tr>
<th>Example</th>
<th>$x$</th>
<th>$p(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>a.</td>
<td>0</td>
<td>.20</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>.90</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-.10</td>
</tr>
<tr>
<td>b.</td>
<td>-2</td>
<td>.3</td>
</tr>
<tr>
<td></td>
<td>-1</td>
<td>.3</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>.3</td>
</tr>
<tr>
<td>c.</td>
<td>-1</td>
<td>.25</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>.65</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>.10</td>
</tr>
</tbody>
</table>

EXAMPLE: It is known that 20% of all light bulbs produced by a certain company have a lifetime of at least 800 hours. You have just purchased two light bulbs manufactured by this company and are interested in the random variable $x =$ the number of the two bulbs which will last at least 800 hours. Construct the probability distribution for $x$, assuming the two bulbs operate independently.

Solution:

$S$: a bulb lasts at least 800 hours
$F$: a bulb fails to last 800 hours
Possible Outcomes

$$(S, S)$$

$x$ $P$(outcome)

2 $(.2)(.2) = .04$

Possible Outcomes

<table>
<thead>
<tr>
<th>$x$</th>
<th>$P$(outcome)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

**probability distribution of $x$:**

<table>
<thead>
<tr>
<th>$x$</th>
<th>$p(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

**EXAMPLE:** Every human possesses two sex chromosomes. A copy of one or the other (equally likely) is contributed to an offspring. Males have one $X$ chromosome and one $Y$ chromosome; females have two $X$ chromosomes. Suppose a couple plans to have three children and let $x$ be the number of boys. Construct the probability distribution of $x$. What is the probability that the couple will have at least one boy? at most two boys? Assume births are independent events.

**Solution:**

$x = \# \text{ of boys}$

$M = \text{offspring is male}$

$F = \text{offspring is female}$

**Possible Outcomes**

Need to find $P(M)$, $P(F)$

**Probability distribution of $x$:**

<table>
<thead>
<tr>
<th>$x$</th>
<th>$p(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

$P(x \geq 1) =$

$P(x \leq 2) =$

**Expected Value of a Discrete Random Variable**

$$E(x) = \mu = \sum_{i=1}^{k} x_i P(X = x_i)$$

**notation: $E(x)$, $\mu$, population mean**

**Comments about $E(x)$, also denoted $\mu$**

a measure of the “middle” of the probability distribution;

a “weighted average” where each value of the random variable is “weighted” by its probability;

sample mean $\overline{x}$ is a weighted average where each value $x_1$, $x_2$, ..., $x_n$ receives equal weight $\frac{1}{n}$;

$E(x)$ is not necessarily one of the possible values of the random variable.

**EXAMPLE:**

<table>
<thead>
<tr>
<th>Economic Scenario</th>
<th>Profit ($\text{mill.}$)</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Great</td>
<td>10</td>
<td>0.20</td>
</tr>
<tr>
<td>Good</td>
<td>5</td>
<td>0.40</td>
</tr>
<tr>
<td>OK</td>
<td>1</td>
<td>0.25</td>
</tr>
<tr>
<td>Lousy</td>
<td>-4</td>
<td>0.15</td>
</tr>
</tbody>
</table>

$$\mu = (10 \times 0.20) + (5 \times 0.40) + (1 \times 0.25) + ((-4) \times 0.15) = 3.65 \ ($\text{mill.}$)$$

(not one of the values of the random variable)

the **probability histogram balances** at the point on the x-axis that is equal to $\mu$

**Interpretation:**

$E(x)$ is not
E(x) is a “long run” average; if you perform the experiment many times and observe the random variable x each time, then the average (mean) of these observed x-values will get closer to E(x) as you observe more and more values of the random variable x.

**EXAMPLE:**  Green Mountain Lottery
3 digits, each between 0 and 9; repeats allowed, order counts. $10 \times 10 \times 10 = 1000$ possibilities.

<table>
<thead>
<tr>
<th>x</th>
<th>$0$</th>
<th>$500$</th>
</tr>
</thead>
<tbody>
<tr>
<td>p(x)</td>
<td>.999</td>
<td>.001</td>
</tr>
</tbody>
</table>

$E(x) =$

Interpretation: □

**EXAMPLE:** Suppose a fair coin is tossed 3 times and we let $x =$ the number of heads. Find $\mu = E(x)$.

**Solution** probability distribution of $x$:
possible values of $x$:
$p(0) = P(x = 0) = \left( \frac{1}{2} \right)^3 = \frac{1}{8}$,
$p(1) =$
$p(2) =$
$p(3) =$
$\mu = E(x) = \sum_{i=1}^{4} x_i p(x_i) =$

Note that $\mu = 1.5$ is not a possible value of the random variable.

**EXAMPLE:** US roulette wheel, expected value of a $1 bet on a single number

<table>
<thead>
<tr>
<th>x</th>
<th>-1</th>
<th>35</th>
</tr>
</thead>
<tbody>
<tr>
<td>p(x)</td>
<td>$\frac{1}{38}$</td>
<td>$\frac{1}{38}$</td>
</tr>
</tbody>
</table>

$E(x) =$

**EXAMPLE:** Refer to the above example where we toss a fair coin 3 times and let $x =$ the number of heads. What should we consider a fair charge for playing the game?

**Solution** Before we consider a fair charge for playing the game, observe that the player wins 0, 1, 4, or 9 dollars for the respective x values 0, 1, 2, 3. The probabilities $\frac{1}{8}, \frac{3}{8}, \frac{3}{8}, \frac{1}{8}$ are associated with the respective winnings 0, 1, 4, and 9 dollars. Thus the expected winnings (in dollars) are $0\left(\frac{1}{8}\right) + 1\left(\frac{3}{8}\right) + 4\left(\frac{3}{8}\right) + 9\left(\frac{1}{8}\right) = 3$.

Therefore, a fair charge for playing the game is $3$. □

**Variance and Standard Deviation of a Discrete Random Variable**

$$\sigma^2 = \sum_{i=1}^{k} (x_i - \mu)^2 \cdot P(X = x_i)$$

**EXAMPLE**

<table>
<thead>
<tr>
<th>Economic Scenario</th>
<th>Profit ($\text{mill.}$)</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Great</td>
<td>10</td>
<td>0.20</td>
</tr>
<tr>
<td>Good</td>
<td>5</td>
<td>0.40</td>
</tr>
<tr>
<td>OK</td>
<td>1</td>
<td>0.25</td>
</tr>
<tr>
<td>Lousy</td>
<td>$-4$</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Recall: $\mu = 3.65$

$$\sigma^2 = (x_1 - \mu)^2 p(x_1) + (x_2 - \mu)^2 p(x_2) + (x_3 - \mu)^2 p(x_3) + (x_4 - \mu)^2 p(x_4)$$

**Preferred Measure of Spread:**  **Standard Deviation**

$$\sigma = \sqrt{\sigma^2} \text{ or } SD(x) = \sqrt{Var(x)}$$

**EXAMPLE:** The standard deviation of the company's profit: $SD(x) = \sqrt{19.3275} = 4.40$ ($\text{million}$)
EXAMPLE: A basketball player shoots 3 free throws. \( P(\text{make}) = P(\text{miss}) = 0.5 \). Let \( X \) = number of free throws made:

<table>
<thead>
<tr>
<th>( x )</th>
<th>( 0 )</th>
<th>( 1 )</th>
<th>( 2 )</th>
<th>( 3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p(x) )</td>
<td>( \frac{1}{8} )</td>
<td>( \frac{3}{8} )</td>
<td>( \frac{3}{8} )</td>
<td>( \frac{1}{8} )</td>
</tr>
</tbody>
</table>

\( E(X) = 1.5 \)

Compute the variance \( \sigma^2 \) and the standard deviation \( \sigma \) of \( X \), the number of free throws made.

\[
\text{Var}(X) = (0 - 1.5)^2 \left( \frac{1}{8} \right) + (1 - 1.5)^2 \left( \frac{3}{8} \right) + (2 - 1.5)^2 \left( \frac{3}{8} \right) + (3 - 1.5)^2 \left( \frac{1}{8} \right)
\]

\( SD(X) = \sqrt{\text{Var}(X)} = \) 

EXAMPLE: Expected value and standard deviation of life insurance payout.

The probability model for the payout \( X \) of a particular life insurance policy is shown below. Find the expected payout \( E(X) \) and the standard deviation \( SD(X) \) of the payout.

### EXPECTED VALUE

<table>
<thead>
<tr>
<th>Policyholder outcome</th>
<th>Payout ( x )</th>
<th>Probability ( p(x) )</th>
<th>( x \cdot p(x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Death</td>
<td>100,000</td>
<td>( \frac{1}{1000} )</td>
<td>( 100,000 \cdot \frac{1}{1000} = 100 )</td>
</tr>
<tr>
<td>Disability</td>
<td>50,000</td>
<td>( \frac{2}{1000} )</td>
<td>( 50,000 \cdot \frac{2}{1000} = 100 )</td>
</tr>
<tr>
<td>Neither</td>
<td>0</td>
<td>( \frac{997}{1000} )</td>
<td>( 0 \cdot \frac{997}{1000} = 0 )</td>
</tr>
</tbody>
</table>

\( E(X) = 200 \)

### STANDARD DEVIATION

<table>
<thead>
<tr>
<th>Policyholder outcome</th>
<th>Payout ( x )</th>
<th>Probability ( p(x) )</th>
<th>Deviation ( (x - E(X)) )</th>
<th>( (x - E(X))^2 \cdot p(x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Death</td>
<td>100,000</td>
<td>( \frac{1}{1000} )</td>
<td>100,000 - 200</td>
<td>( (99,800)^2 \cdot \frac{1}{1000} = 9,960,040 )</td>
</tr>
<tr>
<td>Disability</td>
<td>50,000</td>
<td>( \frac{2}{1000} )</td>
<td>50,000 - 200</td>
<td>( (49,800)^2 \cdot \frac{2}{1000} = 4,960,080 )</td>
</tr>
<tr>
<td>Neither</td>
<td>0</td>
<td>( \frac{997}{1000} )</td>
<td>0 - 200</td>
<td>( (-200)^2 \cdot \frac{997}{1000} = 39,880 )</td>
</tr>
</tbody>
</table>

\( SD(X) = \sqrt{\text{Var}(X)} = \sqrt{14,960,000} = 3,867.82 \)

### Algebra Rules for Expected Value, Variance, and Standard Deviation

**Adding a number to or subtracting a number from a random variable**

If \( X \) is a random variable and \( c \) is a number, then

\[
E(X \pm c) = E(X) \pm c \quad Var(X \pm c) = Var(X) \quad SD(X \pm c) = SD(X)
\]

**Multiply a random variable by a number**

If \( X \) is a random variable and \( c \) is a number, then

\[
E(cX) = cE(X) \quad Var(cX) = c^2Var(X) \quad SD(cX) = |c|SD(X)
\]

### EXAMPLE

Suppose \( X \) is a random variable with expected value \( E(X) \), variance \( Var(X) \) (also denoted \( \sigma^2 \)), and standard deviation \( SD(X) \). Suppose the number \( c = -1 \) in the above identities. Then

\[
E(X - 1) = E(X) - 1 \quad Var(x - 1) = Var(x) \quad SD(x - 1) = SD(x)
\]

\[
E(-X) = -E(X) \quad Var(-x) = (-1)^2Var(x) = Var(x) \quad SD(-x) = | -1|SD(x) = SD(x)
\]
EXAMPLE  The Widget Co.’s new product, brocolli-flavored widget oil, is selling better than predicted. As a result, the company’s economist has revised projected company profits upward by $2 million for each economic scenario.

<table>
<thead>
<tr>
<th>Original projections</th>
<th>New projections</th>
</tr>
</thead>
<tbody>
<tr>
<td>Economic Scenario</td>
<td>Profit ($mill.)</td>
</tr>
<tr>
<td>Great</td>
<td>10</td>
</tr>
<tr>
<td>Good</td>
<td>5</td>
</tr>
<tr>
<td>OK</td>
<td>1</td>
</tr>
<tr>
<td>Lousy</td>
<td>-4</td>
</tr>
</tbody>
</table>

Original expected value $E(x) = 3.65$

New expected value $E(x + 2)$:

i) LONG (UNC-CH) way: $E(x + 2) = 12(0.20) + 7(0.40) + 3(0.25) + ( - 2)(0.15) = 5.65$

ii) SMART (NCSU) way: $c = 2; E(x + 2) = E(X) + 2 = 3.65 + 2 = 5.65$

Original variance $Var(x) = 19.3275$

New variance $Var(x + 2)$ and new standard deviation $SD(x + 2)$:

i) LONG (UNC-CH) way:

$Var(x + 2) = (12 - 5.65)^2(0.20) + (7 - 5.65)^2(0.40) + (3 - 5.65)^2(0.25) + ( - 2 + 5.65)^2(0.15)$

$= 19.3275$

$SD(x + 2) = \sqrt{19.3275} = 4.40$

ii) SMART (NCSU) way:

$Var(x + 2) = Var(x) = 19.3275; SD(x + 2) = SD(x) = 4.40$

EXAMPLE (Expected Value and Standard Deviation of a Linear Transformation $a + bx$)

$X$ = number of repairs a new computer needs each year; $E(X) = 0.20, SD(X) = 0.55$. Service contract: $100/yr plus $25 service charge for each repair.

What are the mean and standard deviation of the yearly cost of the service contract?

Cost = $100 + $25X

$E(cost) = E(100 + 25X) = 100 + 25E(X) = 100 + 25\times 0.20 = 100 + 5 = 105$

$SD(cost) = SD(100 + 25X) = SD(25X) = 25\times SD(X) = 25\times 0.55 = 13.75$

If $X$ and $Y$ are random variables

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

$E(X - Y) = E(X) - E(Y)$

If $X$ and $Y$ are independent random variables

$Var(X + Y) = Var(X) + Var(Y)$

$SD(X + Y) = \sqrt{Var(X + Y)} = \sqrt{Var(X) + Var(Y)}$

$Var(X - Y) = Var(X) + Var(Y)$

$SD(X - Y) = \sqrt{Var(X - Y)} = \sqrt{Var(X) + Var(Y)}$

Important!! Standard deviations do NOT add: $SD(X + Y) \neq SD(X) + SD(Y)$

Standard deviations do NOT subtract: $SD(X - Y) \neq SD(X) - SD(Y)$

Variances of Independent Random Variables follow the Pythagorean Theorem of Statistics $a^2 + b^2 = c^2$
EXAMPLE A college student on a meal plan reports that the daily amount $X$ he spends on food varies with mean $E(X) = \$13.50$ and standard deviation $SD(X) = \$7$. An offensive lineman on the football team has a jumbo meal plan; the daily amount $Y$ he spends on food has mean $E(Y) = \$24.75$ and standard deviation $SD(Y) = \$9.50$.

i) What are the mean and standard deviation of the total amount the college student spends in two consecutive days?

Let $X_1$ ($X_2$) be the amount he spends on day 1 (day 2).

\[ E(X_1 + X_2) = E(X_1) + E(X_2) = \$13.50 + \$13.50 = \$27 \]
\[ Var(X_1 + X_2) = Var(X_1) + Var(X_2) = (\$7)^2 + (\$7)^2 = \$49 + \$49 = \$98, \text{ so} \]
\[ SD(X_1 + X_2) = \sqrt{Var(X_1 + X_2)} = \sqrt{\$98} = \$9.90 \]

NOTE!! $\$9.90 = SD(X_1 + X_2) \neq SD(X_1) + SD(X_2) = \$14$ STANDARD DEVIATIONS DO NOT ADD; VARIANCES ADD WHEN and are INDEPENDENT.

ii) What are the mean and standard deviation of the amount by which the football player's daily spending exceeds the college student's spending?

Football player's spending exceeds college student's spending by an amount $Y - X$.

\[ E(Y - X) = E(Y) - E(X) = \$24.75 - \$13.50 = \$11.25 \]
\[ Var(Y - X) = Var(Y) + Var(X) = (\$9.50)^2 + (\$7)^2 = \$90.25 + \$49 = \$139.25, \text{ so} \]
\[ SD(Y - X) = \sqrt{Var(Y - X)} = \sqrt{\$139.25} = \$11.80 \]

NOTE!! $\$11.80 = SD(Y - X) \neq SD(Y) - SD(X) = \$9.50 - \$7 = \$2.50$.

EXAMPLE Does $X + X = 2X$? Maybe, but be careful!

Let $X$ be the annual payout on a life insurance policy. From mortality tables $E(X) = \$200$ and $SD(X) = \$3,867$.

i) If the payout amounts are doubled, what are the new expected value and standard deviation?

Double payout is $2X$. $E(2X) = 2E(X) = 2*\$200 = \$400$

\[ SD(2X) = 2SD(X) = 2*\$3,867 = \$7,734 \]

ii) Suppose insurance policies are sold to 2 people. The annual payouts are $X_1$ and $X_2$. Assume the 2 people behave independently. What are the expected value and standard deviation of the total payout?

\[ E(X_1 + X_2) = E(X_1) + E(X_2) = \$200 + \$200 = \$400. \]
\[ SD(X_1 + X_2) = \sqrt{Var(X_1 + X_2)} = \sqrt{Var(X_1) + Var(X_2)} = \sqrt{(\$3867)^2 + (\$3867)^2} \]
\[ = \sqrt{\$14,953,689 + \$14,953,689} = \sqrt{\$29,907,378} = \$5,468.76 \]
EXAMPLE Let the random variable $X$ denote the number of hours a student from our class slept between noon yesterday and noon today. Suppose the mean is $E(X) = 6.5$ hours and the standard deviation is $SD(X) = .34$ hours.

Let the random variable $Y$ denote the number of hours a randomly selected student from our class was awake between noon yesterday and noon today.

i) What are the mean and standard deviation of $Y$, the number of hours a student is awake?

\[
Y = 24 - X; 
E(Y) = E(24 - X) = 24 - E(X) = 24 - 6.5 = 17.5 \text{ hours}
\]

\[
SD(Y) = SD(24 - X) = SD(-X) = (-1)^2SD(X) = SD(X) = .34 \text{ hours}
\]

ii) What are the mean and standard deviation of the total hours $X + Y$ that a student is asleep and awake between noon yesterday and noon today?

\[
E(X + Y) = E(X + 24 - X) = E(24) = 24
\]

\[
SD(X + Y) = SD(X + 24 - X) = SD(24) = 0
\]

NOTE!! In ii) we don't add the variances since $X$ and $Y$ are not independent.

Covariance and Correlation of Random Variables

Recall from Chapter 7 the sample correlation coefficient $r$ between two quantitative variables $X$ and $Y$ calculated from the bivariate data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

**Correlation coefficient**

\[
r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \overline{x}}{s_x} \right) \left( \frac{y_i - \overline{y}}{s_y} \right)
\]

Can also also calculate the correlation between two random variables.

But first: COVARIANCE

Consider two random variables $X$ and $Y$ with expected values $E(X)$ and $E(Y)$, respectively.

The covariance of $X$ and $Y$ is defined as

\[
Cov(X, Y) = E\{[X - E(X)]\{Y - E(Y)\}\}
\]

The covariance measures the linear dependence between $X$ and $Y$.

Properties of covariance:

1. $Cov(X, Y) = Cov(Y, X)$
2. $Cov(X, X) = Var(X)$
3. $Cov(cX, dY) = cdCov(X, Y)$, for any constants $c$ and $d$.
4. $Cov(X, Y) = E(XY) - E(X)E(Y)$

5. if $X$ and $Y$ are independent random variables, then $Cov(X, Y) = 0$. **NOTE:** the converse is NOT true.

Variance of sum or difference of two random variables when they are not independent:

\[
Var(X \pm Y) = Var(X) + Var(Y) \pm 2Cov(X, Y).
\]

When $X$ and $Y$ are independent random variables, $Cov(X, Y) = 0$, so the above formula reduces to

\[
Var(X \pm Y) = Var(X) + Var(Y)
\]

- The covariance does not have to be between $-1$ and $+1$ like the correlation in Chapter 7.
- If $X$ and $Y$ have large values, then the covariance will be large as well.
A covariance of $-68.4$ does not give us a good sense for how negatively related the random variables are.

\[
\text{Correlation} \\
\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y}
\]

**Properties of correlation:**

1. The sign of $\text{Corr}(X, Y)$ gives the direction of the association between $X$ and $Y$.
2. $-1 \leq \text{Corr}(X, Y) \leq +1$
3. $\text{Corr}(X, Y) = \text{Corr}(Y, X)$
4. Correlation is not affected by linear transformations on either random variable.

**Example**

For a household, the monthly bill for natural gas is a random variable $X$ with 
\[E(X) = 125, \quad \sigma_X = 75\]

while the monthly bill for electricity is a random variable $Y$ with 
\[E(Y) = 174, \quad \sigma_Y = 41\]

The correlation $\text{Corr}(X, Y)$ between the two bills $-0.55$.

What are the mean and standard deviation of the total of the natural gas bill and the electric bill?

**Answer**

\[E(X + Y) = 125 + 174 = 299\]

Note that $\text{Cov}(X, Y) = \sigma_X \sigma_Y \text{Corr}(X, Y) = 75 \times 41 \times (-0.55) = -1691.25$

\[SD(X + Y) = \sqrt{\text{Var}(X + Y)} = \sqrt{\text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)}\]

\[= \sqrt{75^2 + 41^2 + 2(-1691.25)} = \sqrt{3923.5} = 62.64\]

So the total of the natural gas bill and the electric bill has mean $299$ and standard deviation $62.24$.

**Probability Models**

**The Binomial Random Variable**

classical examples
coin tossing
opinion polling
common characteristics:

**Example:**
Through 2/24/2011 NC State’s free-throw percentage is 69.6% (146th out 345 in Div. 1).

If in the 2/26/2011 game with GaTech, NCSU shoots 11 free-throws, what is the probability that:
NCSU makes exactly 8 free-throws?
NCSU makes at most 8 free throws?
NCSU makes at least 8 free-throws?
Characteristics of a binomial experiment
1. $n$ identical trials
2. only two possible outcomes on each trial
3. the probability $p$ of “success” and the probability $q = 1 - p$ of “failure” remain constant from trial to trial
4. the trials are independent
5. the binomial random variable $x$ is the number of “successes” in the $n$ trials

Note: when $n = 1$, a binomial experiment is called a Bernoulli trial; a binomial experiment can be thought of as a Bernoulli trial repeated $n$ times. A Bernoulli trial is thus an experiment with only two possible outcomes, “success” with probability $p$ and “failure” with probability $q = 1 - p$. If we let $y$ be the random variable that counts the number of successes in one Bernoulli trial, then $y = 0$ or $1$. The random variable $y$ is called a Bernoulli random variable; thus, a Bernoulli random variable is a binomial random variable with $n = 1$. Observe that if $x$ is a binomial random variable in a binomial experiment with $n$ trials, then

$$x = \sum_{i=1}^{n} y_i,$$

where $y_i$ is the Bernoulli random variable for the $i$th Bernoulli trial.

EXAMPLE: For each of the following experiments, decide whether $x$ is a binomial random variable.

a. From past records, it is known that 5% of all personal computers manufactured by a certain company will need major repairs within three months of purchase. Your company has just purchased 10 personal computers from this firm. Let $x =$ the number of these computers that will need major repairs within three months.

b. It is known that 1% of the liquid crystal display (LCD) screens made for laptop computers by a particular production process are defective. A quality control engineer wants to inspect the output from a particular production run. Let $x =$ the number of screens that must be inspected until two defective screens are found.

c. Five of the members of the dean's Council on Engineering Education are female and five are male. Three of the ten members of this Council will be randomly selected to appear before a congressional committee studying global economic competitiveness. Let $x =$ the number of females chosen.

d. Past experience indicates that 10% of the silicon chips produced by Outel Corp. are defective. Apricot Computer Corp. has just purchased 1500 of these chips for use in its computers. Let $x =$ the number of these chips that are defective.

Binomial probability distribution

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n,$$

where

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

For a binomial random variable $x$ with parameters $n$ and $p$,

$$\mu = E(x) = np,$$

$$\sigma^2 = Var(x) = np(1-p).$$

EXAMPLE: A production line produces motor housings, 5% of which have cosmetic defects. A quality control manager randomly selects 4 housings from the production line. Let $x =$ the number of housings that have a cosmetic defect. Tabulate the probability distribution for $x$.

Solution
(i). Use the methods of Lecture Unit Section 4.1 to list the possible outcomes, assign the value of the random variable to each possible outcome, and compute probabilities. (see ppt slide)

(ii). Observe that \( x \) satisfies the requirements of a binomial random variable:

\[
\begin{align*}
\text{Then } \\
p(0) &= \binom{4}{0} (0.05)^0 (0.95)^4 = \\
p(1) = \\
p(2) = \\
p(3) = \\
p(4) =
\end{align*}
\]

Thus, the probability distribution of \( x \) is:

\[
\begin{array}{c|c|c|c|c|c}
x & 0 & 1 & 2 & 3 & 4 \\ 
\hline
p(x) & & & & &
\end{array}
\]

**EXAMPLE:** Refer to the previous example.

a. What is the probability that at least two of the housings will have cosmetic defects?

b. What is the probability that at most one plate will not have a cosmetic defect?

**Solution**

**BINOMIAL TABLES**

cumulative probabilities for \( n = 20, 25, \) various values of \( p \)

\[
P(x \leq a) = \sum_{x=0}^{a} p(x) = p(0) + p(1) + \ldots + p(a).
\]

**EXAMPLE:** Suppose \( x \) is a binomial random variable with \( n = 20, p = .3 \).

a. \( P(x \leq 5) = \)

b. \( P(x > 8) = \)

c. \( P(x < 9) = \)

d. \( P(x \geq 10) = \)

e. \( P(3 \leq x \leq 7) = \)

f. \( P(2 < x \leq 9) = \)

g. \( P(x = 8) = \)
To calculate binomial probabilities using: Excel, click on the “insert function” icon \( f_x \); in the Statistical category scroll down to BINOMDIST; Statcrunch, click Stat > Calculators > Binomial; TI calculator: Calculator Appendix, p. 10-11.

**EXAMPLE:** Color Blindness

The frequency of color blindness (dyschromatopsia) in the Caucasian American male population is estimated to be about 8%. We take a random sample of size 25 from this population.

We can model this situation with a \( B(n = 25, p = 0.08) \) distribution.

1. What is the probability that five individuals or fewer in the sample are color blind?

   Use Excel’s \[ \text{=BINOMDIST(number_s,trials,probability_s,cumulative)} \] function

   \[ P(X \leq 5) = \text{BINOMDIST}(5, 25, .08, 1) = 0.9877 \]

2. What is the probability that more than five will be color blind?

   \[ P(X > 5) = 1 - P(x \leq 5) = 1 - 0.9877 = 0.0123 \]

3. What is the probability that exactly five will be color blind?

   \[ P(X = 5) = \text{BINOMDIST}(5, 25, .08, 0) = 0.0329 \]

\[
\begin{align*}
B(25, 0.08) : & \quad E(X) = np = 25 \times 0.08 = 2 \quad SD(X) = \sqrt{np(1 - p)} = \sqrt{25 \times 0.08 \times 0.92} = 1.36 \\
B(10, 0.08) : & \quad E(X) = 10 \times 0.08 = 0.8 \quad SD(X) = \sqrt{n \times p \times (1 - p)} = \sqrt{10 \times 0.08 \times 0.92} = 0.86 \\
B(75, 0.08) : & \quad E(X) = 75 \times 0.08 = 6 \quad SD(X) = \sqrt{n \times p \times (1 - p)} = \sqrt{75 \times 0.08 \times 0.92} = 2.35 
\end{align*}
\]

Geometric Random Variables and Their Probability Distributions

**Example**

Through 2/24/2011 NC State’s free-throw percentage was 69.6% (146th of 345 in Div. 1).

In the 2/26/2011 game with GaTech what was the probability that the first missed free-throw by the ‘Pack occurs on the 5th attempt?

- A geometric random variable counts the number of Bernoulli trials until the first success is observed.
- Geometric random variables are completely specified by one parameter, \( p \), the probability of success, and are denoted \( \text{Geom}(p) \).
- Unlike the binomial random variable, the number of trials is NOT fixed.
Geometric probability distribution

\[ p = \text{probability of success} \]
\[ q = 1 - p = \text{probability of failure} \]

\[ X = \text{\# of trials until the first success occurs} \]

\[ p(x) = P(X = x) = q^{x-1}p, \quad x = 1, 2, 3, \ldots \]

For a geometric random variable \( x \) with parameter \( p \),

\[ \mu = E(X) = \frac{1}{p}, \quad \sigma = \sqrt{\frac{q}{p^2}} \]

**The 10% condition:** Bernoulli trials must be independent. If that assumption is violated, it is still okay to proceed as long as the sample is smaller than 10% of the population.

**EXAMPLE**

The American Red Cross says that about 11% of the U.S. population has Type B blood. A blood drive is being held in your area.

1. How many blood donors should the American Red Cross expect to collect from until it gets a donor with Type B blood?

\[ p = .11, \quad E(X) = \frac{1}{p} = \frac{1}{.11} = 9.1 \]

2. What is the probability that the fourth blood donor is the first donor with Type B blood?

3. What is the probability that the first Type B blood donor is among the first four people in line?

\[ p(1) + p(2) + p(3) + p(4) = (.89^0 \times .11) + (.89^1 \times .11) + (.89^2 \times .11) + (.89^3 \times .11) \]

\[
\begin{align*}
P(X = 1) &= p(1) = .9^0 \times .1 = .1 \quad P(X = 3) = p(3) = .9^2 \times .1 = .081 \\
P(X = 2) &= p(2) = .9^1 \times .1 = .09 \quad P(X = 4) = p(4) = .9^3 \times .1 = .0729
\end{align*}
\]
Geometric Probability Distribution

\( p = 0.25 \)

\[
P(X = 1) = p(1) = .75^0 \times .25 = .25 \\
P(X = 2) = p(2) = .75^1 \times .25 = .1875 \\
P(X = 3) = p(3) = .75^2 \times .25 = .1406 \\
P(X = 4) = p(4) = .75^3 \times .25 = .1055
\]

Geometric Probability Distribution

\( p = 0.5 \)

\[
P(X = 1) = p(1) = .5^0 \times .5 = .5 \\
P(X = 2) = p(2) = .5^1 \times .5 = .25 \\
P(X = 3) = p(3) = .5^2 \times .5 = .125 \\
P(X = 4) = p(4) = .5^3 \times .5 = .0625
\]

Geometric Probability Distribution

\( p = 0.75 \)

\[
P(X = 1) = p(1) = .25^0 \times .75 = .75 \\
P(X = 2) = p(2) = .25^1 \times .75 = .1875 \\
P(X = 3) = p(3) = .25^2 \times .75 = .0469 \\
P(X = 4) = p(4) = .25^3 \times .75 = .0117
\]
To calculate geometric probabilities using: Excel, on our course web page http://www.stat.ncsu.edu/people/reiland/courses/st350/ click Student Resources in the left panel, under Statistical Enhancements for Excel click on "Geometric Probabilities"; ti 83/84, calculate geometric probabilities by pressing [2nd VARS] scroll down to geometpdf and geometcdf

Example 3-point attempts
Shanille O’Keal is a WNBA player who makes 25% of her 3-point attempts.

1. The expected number of attempts until she makes her first 3-point shot is what value? $E(X) = \frac{\ln(1 - p)}{p}$

2. What is the probability that the first 3-point shot she makes occurs on her 3rd attempt? $p(3) = \frac{(0.25)^3}{3!} = 0.0016$  

Poisson Random Variables and Their Probability Distributions

- The Poisson random variable counts the number of “rare” events that occur over a fixed amount of time or within a specified region
- Typical cases:
  - The number of errors a typist makes per page
  - The number of customers entering a service station per hour
  - The number of login requests received per minute by an internet server

- Properties of the Poisson random variable
  - The number of events that occur in a certain time interval is independent of the number of successes that occur in another time interval.
  - The probability of an event in a certain time interval is:
    - i) the same for all time intervals of the same size
    - ii) proportional to the length of the interval.
  - The probability that two or more events will occur in an interval approaches zero as the interval becomes smaller

Poisson Probability Distribution

$$p(x) = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, 3, ...$$

where $\mu = E(x) = \lambda; \quad \sigma^2 = Var(x) = \lambda \quad e = 2.71828...$

Example

According to Science magazine, over the period 1944 to 2000 the average number of hurricanes per year that hit the U.S mainland was 2.35. Suppose the number of hurricanes that hit the U.S mainland can be modeled by a Poisson random variable with $\lambda = 2.35$

1) What is the expected number of hurricanes next year that will hit the U.S. mainland?
2) What is the probability that no hurricanes will hit the U.S mainland next year?
3) What is the probability that 3 hurricanes hit the U.S mainland next year next year?
4) What is the probability that during the next 2 years exactly 4 hurricanes hit the U.S mainland?
Poisson Probability Distribution $\lambda=1$

\[ P(X = 0) = p(0) = \frac{e^{-1}1^0}{0!} = .3679 \quad P(X = 2) = p(2) = \frac{e^{-1}1^2}{2!} = .1839 \]
\[ P(X = 1) = p(1) = \frac{e^{-1}1^1}{1!} = .3679 \quad P(X = 3) = p(3) = \frac{e^{-1}1^3}{3!} = .0613 \]

Poisson Probability Distribution $\lambda = 2$

\[ P(X = 0) = p(0) = \frac{e^{-2}2^0}{0!} = .1353 \quad P(X = 2) = p(2) = \frac{e^{-2}2^2}{2!} = .2707 \]
\[ P(X = 1) = p(1) = \frac{e^{-2}2^1}{1!} = .2707 \quad P(X = 3) = p(3) = \frac{e^{-2}2^3}{3!} = .1804 \]

Poisson Probability Distribution $\lambda = 5$

\[ P(X = 0) = p(0) = \frac{e^{-5}5^0}{0!} = .0067 \quad P(X = 2) = p(2) = \frac{e^{-5}5^2}{2!} = .0842 \]
Poisson Probability Distribution \( \lambda = 7 \)

\[
P(X = 0) = p(0) = \frac{e^{-7} \cdot 0^0}{0!} = 0.0005 \\
P(X = 1) = p(1) = \frac{e^{-7} \cdot 1^1}{1!} = 0.0064 \\
P(X = 2) = p(2) = \frac{e^{-7} \cdot 2^2}{2!} = 0.0223 \\
P(X = 3) = p(3) = \frac{e^{-7} \cdot 3^3}{3!} = 0.0521
\]

To calculate Poisson probabilities using: Excel, click the Insert function \( f_x \) icon, in the Statistical category choose POISSON; Statcrunch, click Stat > Calculators > Poisson; ti 83/84 calculate Poisson probabilities by pressing \( \boxed{2nd} \) [VARS] scroll down to poissonpdf and poissoncdf.

**Continuous Random Variables: Normal Distributions**

Everybody believes in the normal approximation, the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact.

G. Lippmann (French physicist, 1845-1921)

**IN THE REAL WORLD** “I signed up for Schwartz in History 201,” said Ben. “Why did you sign up for her?” asked Ted incredulously. “Don't you know she has given only three A's in the last fourteen years?!” “That's just a rumor. Hunter grades on the curve,” Ben replied. “Don't give me that 'on the curve' stuff,” continued Ted, “I'll bet you don't even know what that means. Anyway, my sister-in-law had her last year and she had to drink so much Maalox\textsuperscript{TM} that . . . .”

**Probability Density Functions**

In Section 4.2 of Lecture Unit 4 we studied discrete random variables and their probability distributions. Probability histograms give us a pictorial presentation of the probability distribution of a discrete random variable. For example, below is a probability histogram showing the probability distribution of a binomial random variable with \( n = 10 \) and \( p = .5 \).
NOTE: the areas of the bars in a probability histogram for a discrete random variable give the probability corresponding to each possible value of the discrete random variable, and the sum of the areas of all the bars is 1.

Recall from earlier in this handout that a continuous random variable takes all possible values in an interval of the real line.

Since the possible values of a continuous random variable are not isolated values like those for a discrete random variable, we cannot use individual bars to pictorially represent the probability distribution of a continuous random variable.

In constructing a relative frequency histogram to describe a set of data, as we use a finer and finer grid the bars of the histogram will become narrower and narrower and trace out a curve that becomes progressively more smooth. If we carry this process to the limit we obtain a smooth curve.

Smooth curve superimposed on binomial probability distribution function: $n = 100$, $p = 0.50$
Smooth curves of this type are the means by which probability distributions for continuous random variables are displayed graphically. For a continuous random variable $X$, this curve is a function of $x$ and is typically denoted $f(x)$. $f(x)$ is called the **probability density function** of $X$.

Probability density functions come in many shapes, depending on the probability distribution of the continuous random variable that the density function represents.

### PROPERTIES OF A PROBABILITY DENSITY FUNCTION $f(x)$:

1) $f(x) \geq 0$ for all $x$
2) the total area under the graph of $f(x)$ is 1

**NOTE:**

- values of $p(x)$ for a discrete random variable $X$ are probabilities: $p(x) = P(X = x)$;
- values of $f(x)$ are **not** probabilities: it is areas under the curve that are probabilities.

**Normal Probability Distributions** (bell curves)

One type of continuous probability distribution - the probability distribution of a normal random variable $x$. Probability density function $f(x)$ for a normal random variable $x$

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

where:

- $\mu = \text{population mean (expected value) of } x$
- $\sigma = \text{standard deviation of } x$
Important Properties:

i) the standard normal density curve is symmetric around the mean $\mu$.

ii) the total area under the curve is 1.

Note: Since normal density curves are symmetric around the mean $\mu$, $\mu = \text{median}$.

Dear Abby:
You wrote in your column that a woman is pregnant for 266 days. Who said so? I carried my baby for 10 months and 5 days, and there is no doubt about it because I know the exact date my baby was conceived. My husband is in the Navy and it could not possibly have been conceived any other time because I saw him only once for an hour and I didn't see him again until the day before the baby was born. I don't drink or run around, and there is no way this baby isn't his, so please print a retraction about that 266-day carrying time because otherwise I'm in a lot of trouble.  

San Diego Reader

Dear Reader:
The average gestation period is 266 days. Some babies come early. Others come late. Yours was late.

Note: In the above expression for $f(x)$, $\mu$ can be any number between $-\infty$ and $+\infty$, $\sigma$ can be any positive number, so there are many normal random variables $x$ and normal probability distributions $f(x)$.

Notation: $X \sim N(\mu, \sigma)$ is written to denote that the random variable $X$ has a normal distribution with mean $\mu$ and standard deviation $\sigma$. 