

4. Classical Probability Distributions

4.1 Discrete Models

FACT: Random variables can be used to define events that involve measurement!

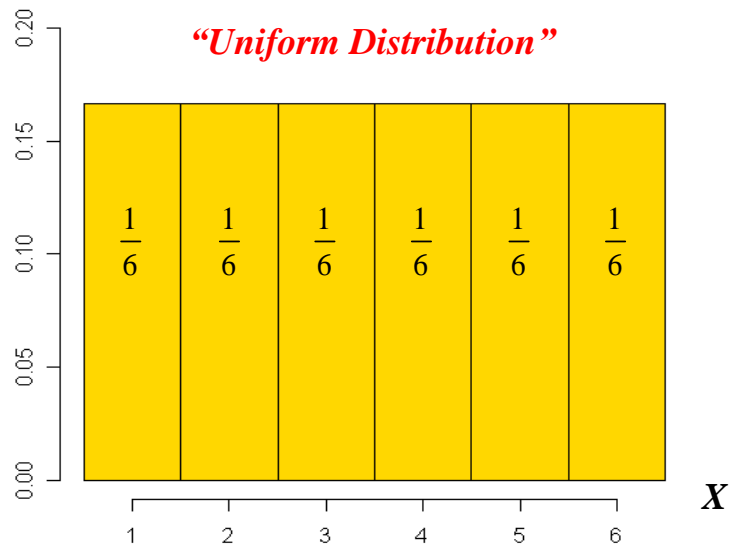


Experiment 3a: Roll one fair die... *Discrete random variable* $X =$ “value obtained”

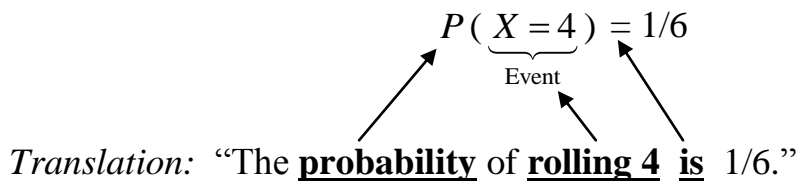
Sample Space: $S = \{1, 2, 3, 4, 5, 6\}$ $\#(S) = 6$

Because the die is fair, each of the six faces has an *equally likely* probability of occurring, i.e., $1/6$. The **probability distribution** for X can be defined by a so-called **probability mass function** (pmf) $p(x)$, organized in a **probability table**, and displayed via a corresponding **probability histogram**, as shown.

Event	Probability
x	$p(x) = P(X = x)$
1	$1/6$
2	$1/6$
3	$1/6$
4	$1/6$
5	$1/6$
6	$1/6$



Comment on notation:



Likewise for the other probabilities $P(X = 1), P(X = 2), \dots, P(X = 6)$ in this example. A mathematically succinct way to write such probabilities is by the notation $P(X = x)$, where $x = 1, 2, 3, 4, 5, 6$. In general therefore, since this *depends* on the value of x , we can also express it as a mathematical *function* of x (specifically, the pmf; see above), written $p(x)$. Thus the two notations are synonymous and interchangeable. The previous example could just as well have been written $f(4) = 1/6$.

Experiment 3b: Roll two distinct, fair dice. \Rightarrow **Outcome** = (Die 1, Die 2)



Sample Space: $S = \{(1, 1), \dots, (6, 6)\}$ $\#(S) = 6^2 = 36$

(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

Discrete random variable $X =$ “Sum of the two dice (**2, 3, 4, ..., 12**).”

Events:	“ $X = 2$ ” = $\{(1, 1)\}$	$\#(X = 2) = 1$
	“ $X = 3$ ” = $\{(1, 2), (2, 1)\}$	$\#(X = 3) = 2$
	“ $X = 4$ ” = $\{(1, 3), (2, 2), (3, 1)\}$	$\#(X = 4) = 3$
	“ $X = 5$ ” = $\{(1, 4), (2, 3), (3, 2), (4, 1)\}$	$\#(X = 5) = 4$
	“ $X = 6$ ” = $\{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}$	$\#(X = 6) = 5$
	“ $X = 7$ ” = $\{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$	$\#(X = 7) = 6$
	“ $X = 8$ ” = $\{(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\}$	$\#(X = 8) = 5$
	“ $X = 9$ ” = $\{(3, 6), (4, 5), (5, 4), (6, 3)\}$	$\#(X = 9) = 4$
	“ $X = 10$ ” = $\{(4, 6), (5, 5), (6, 4)\}$	$\#(X = 10) = 3$
	“ $X = 11$ ” = $\{(5, 6), (6, 5)\}$	$\#(X = 11) = 2$
	“ $X = 12$ ” = $\{(6, 6)\}$	$\#(X = 12) = 1$

Recall that, *by definition*, each event “ $X = x$ ” (where $x = 2, 3, 4, \dots, 12$) corresponds to a specific subset of outcomes from the sample space (of ordered pairs, in this case). Because we are still assuming *equal likelihood* of each die face appearing, the probabilities of these events can be easily calculated by the “shortcut” formula

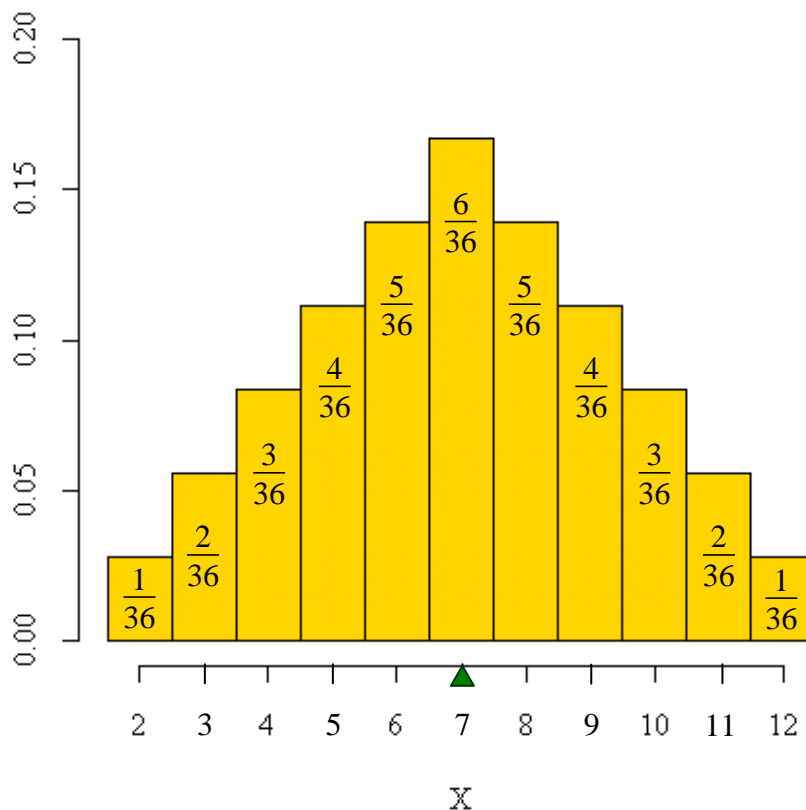
$$P(A) = \frac{\#(A)}{\#(S)}. \quad \text{Question for later: What if the dice are “loaded” (i.e., biased)?}$$

Again, the **probability distribution** for X can be organized in a **probability table**, and displayed via a **probability histogram**, both of which enable calculations to be done easily:

x	$p(x) = P(X = x)$
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

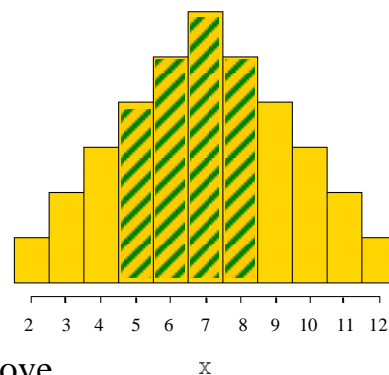
1

Probability Histogram



$$\begin{aligned}
 &\blacktriangleright P(X = 7 \text{ or } X = 11) && \text{Note that “}X = 7\text{” and “}X = 11\text{” are } \underline{\text{disjoint!}} \\
 &= P(X = 7) + P(X = 11) && \text{via Formula (3) above} \\
 &= 6/36 + 2/36 = 8/36
 \end{aligned}$$

$$\begin{aligned}
 &\blacktriangleright P(5 \leq X \leq 8) \\
 &= P(X = 5 \text{ or } X = 6 \text{ or } X = 7 \text{ or } X = 8) \\
 &= P(X = 5) + P(X = 6) + P(X = 7) + P(X = 8) \\
 &= 4/36 + 5/36 + 6/36 + 5/36 \\
 &= 20/36
 \end{aligned}$$



$$\begin{aligned}
 &\blacktriangleright P(X < 10) = 1 - P(X \geq 10) \text{ via Formula (1) above} \\
 &= 1 - [P(X = 10) + P(X = 11) + P(X = 12)] \\
 &= 1 - [3/36 + 2/36 + 1/36] = 1 - 6/36 = 30/36
 \end{aligned}$$

Exercise: How could event $E = \text{“Roll doubles”}$ be characterized in terms of a **random variable**? (Hint: Let $Y = \text{“Difference between the two dice.”}$)

The previous example motivates the important topic of...

Discrete Probability Distributions

In general, suppose that all of the distinct population values of a *discrete* random variable X are sorted in increasing order: $x_1 < x_2 < x_3 < \dots$, with corresponding probabilities of occurrence $p(x_1), p(x_2), p(x_3), \dots$. Formally then, we have the following.

Definition: $p(x)$ is a **probability mass function** for the *discrete* random variable X if, for all x ,

$$p(x) \geq 0 \quad \text{AND} \quad \sum_{\text{all } x} p(x) = 1.$$

In this case, $p(x) = P(X = x)$, the *probability* that the value x occurs in the population.

The **cumulative distribution function** (cdf) is defined as, for all x ,

$$F(x) = P(X \leq x) = \sum_{\text{all } x_i \leq x} p(x_i) = p(x_1) + p(x_2) + \dots + p(x).$$

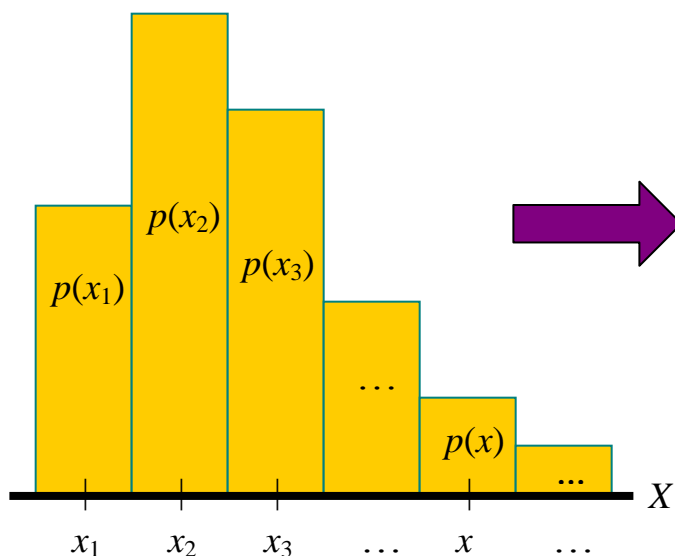
Therefore, F is *piecewise constant*, increasing from 0 to 1.

Furthermore, for any two population values $a < b$, it follows that

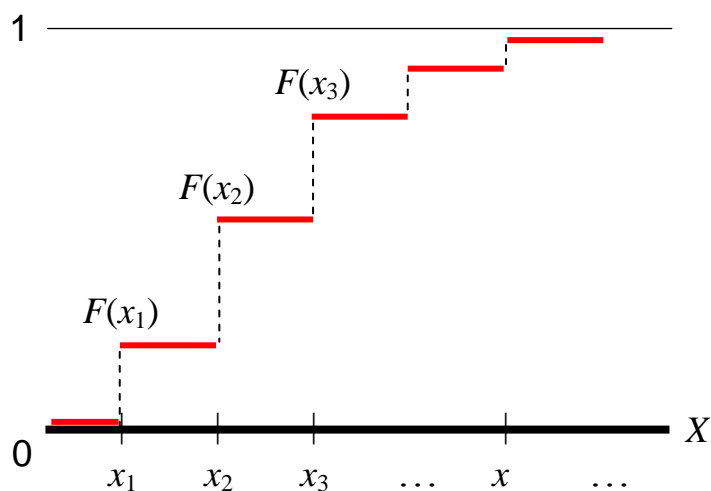
$$P(a \leq X \leq b) = \sum_a^b p(x) = F(b) - F(a^-)$$

where a^- is the value just preceding a in the sorted population.

Total Area = 1



Exercise: Sketch the cdf $F(x)$ for Experiments 3a and 3b above.



Population Parameters μ and σ^2 (vs. Sample Statistics \bar{x} and s^2)

- **population mean** = the “expected value” of the random variable X
= the “arithmetic average” of *all* the population values

If X is a *discrete* numerical random variable, then...

$$\mu = E[X] = \sum x p(x), \text{ where pmf } p(x) = P(X = x), \text{ the probability of } x.$$

Compare this with the *relative frequency* definition of **sample mean** given in §2.3.

Properties of Mathematical Expectation

1. For any constant c , it follows that $E[cX] = c E[X]$.
2. For any two random variables X and Y , it follows that
 - $E[X + Y] = E[X] + E[Y]$ and, via Property 1,
 - $E[X - Y] = E[X] - E[Y]$.

Any “operator” on variables satisfying 1 and 2 is said to be **linear**.

- **population variance** = the “expected value” of the squared deviation of the random variable X from its mean (μ)

If X is a *discrete* numerical random variable, then...

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

Equivalently,*

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

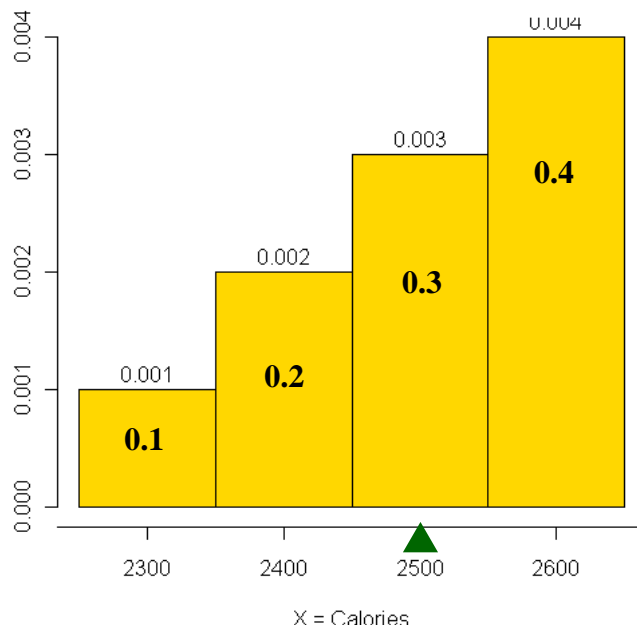
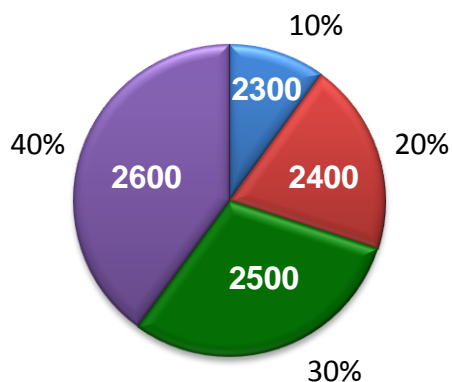
where pmf $p(x) = P(X = x)$, the *probability* of x .

Compare the first with the definition of **sample variance** given in §2.3. (The second is the analogue of the *alternate computational formula*.) Of course, the **population standard deviation** σ is defined as the square root of the variance.

***Exercise:** Algebraically expand the expression $(X - \mu)^2$, and use the properties of expectation given above.

Experiment 4: Two populations, where the daily number of calories consumed is designated by X_1 and X_2 , respectively.

Population 1



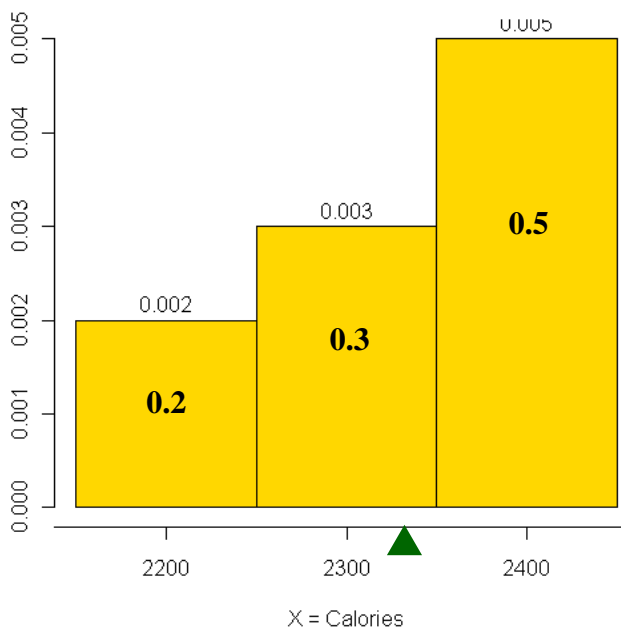
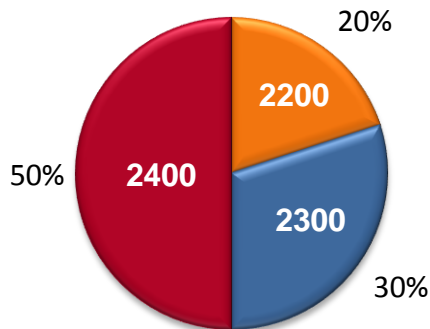
Probability Table

x	$p_1(x)$
2300	0.1
2400	0.2
2500	0.3
2600	0.4

➤ $\text{Mean}(X_1) = \mu_1 = (2300)(0.1) + (2400)(0.2) + (2500)(0.3) + (2600)(0.4) = 2500 \text{ cal}$

➤ $\text{Var}(X_1) = \sigma_1^2 = (-200)^2(0.1) + (-100)^2(0.2) + (0)^2(0.3) + (+100)^2(0.4) = 10000 \text{ cal}^2$

Population 2



Probability Table

x	$p_2(x)$
2200	0.2
2300	0.3
2400	0.5

➤ $\text{Mean}(X_2) = \mu_2 = (2200)(0.2) + (2300)(0.3) + (2400)(0.5) = 2330 \text{ cal}$

➤ $\text{Var}(X_2) = \sigma_2^2 = (-130)^2(0.2) + (-30)^2(0.3) + (70)^2(0.5) = 6100 \text{ cal}^2$

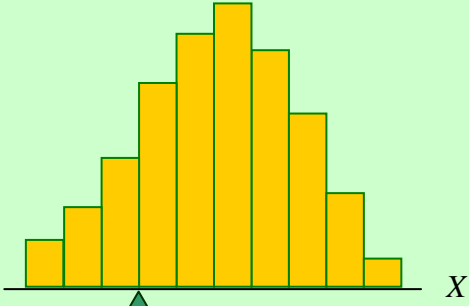
Summary (Also refer back to [2.4 - Summary](#))

POPULATION

Discrete random variable X

Probability Table → Probability Histogram

x	$p(x) = P(X = x)$
x_1	$f(x_1)$
x_2	$f(x_2)$
\cdot	\cdot
\cdot	\cdot
\cdot	\cdot
1	



$\mu = E[X] = \sum x p(x)$

$\sigma^2 = \begin{cases} E[(X - \mu)^2] = \sum (x - \mu)^2 p(x) \\ \text{or} \\ E[X^2] - \mu^2 = \sum x^2 p(x) - \mu^2 \end{cases}$

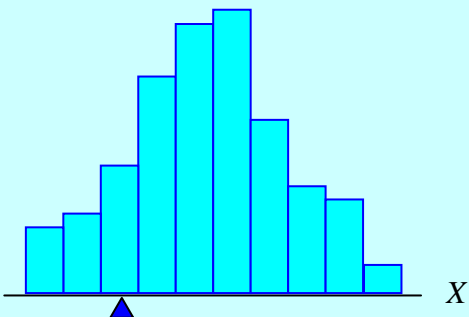
Parameters



SAMPLE, size n

Relative Frequency Table → Density Histogram

x	$p(x) = \frac{\text{freq}(x)}{n}$
x_1	$p(x_1)$
x_2	$p(x_2)$
\cdot	\cdot
\cdot	\cdot
\cdot	\cdot
x_k	$p(x_k)$
1	



$\bar{x} = \sum x p(x)$

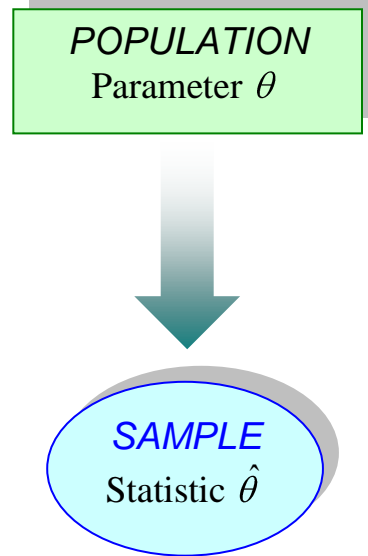
$s^2 = \begin{cases} \frac{n}{n-1} \sum (x - \bar{x})^2 p(x) \\ \text{or} \\ \frac{n}{n-1} [\sum x^2 p(x) - \bar{x}^2] \end{cases}$

Statistics

\bar{X} and S^2 can be shown to be **unbiased** estimators of μ and σ^2 , respectively. That is, $E[\bar{X}] = \mu$, and $E[S^2] = \sigma^2$. (In fact, they are **MVUE**.)

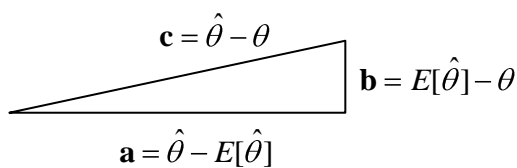
~ Some Advanced Notes on General Parameter Estimation ~

Suppose that θ is a *fixed* population parameter (e.g., μ), and $\hat{\theta}$ is a sample-based estimator (e.g., \bar{X}). Consider all the random samples of a given size n , and the resulting “**sampling distribution**” of $\hat{\theta}$ values. Formally define the following:



- **Mean** (of $\hat{\theta}$) = $E[\hat{\theta}]$, the expected value of $\hat{\theta}$.
- **Bias** = $E[\hat{\theta}] - \theta$, the difference between the expected value of $\hat{\theta}$, and the “target” parameter θ .
- **Variance** (of $\hat{\theta}$) = $E\left[\left(\hat{\theta} - E[\hat{\theta}]\right)^2\right]$, the expected value of the squared deviation of $\hat{\theta}$ from its mean $E[\hat{\theta}]$, or equivalently, $\text{Var}(\hat{\theta}) = E[\hat{\theta}^2] - E[\hat{\theta}]^2$.
- **Mean Squared Error (MSE)** = $E\left[(\hat{\theta} - \theta)^2\right]$, the expected value of the squared difference between estimator $\hat{\theta}$ and the “target” parameter θ .

Exercise: Prove* that $\text{MSE} = \text{Variance} + \text{Bias}^2$.



Vector interpretation
 $\mathbf{c} = \mathbf{a} + \mathbf{b}$
 $E[\mathbf{c}^2] = E[\mathbf{a}^2] + E[\mathbf{b}^2]$

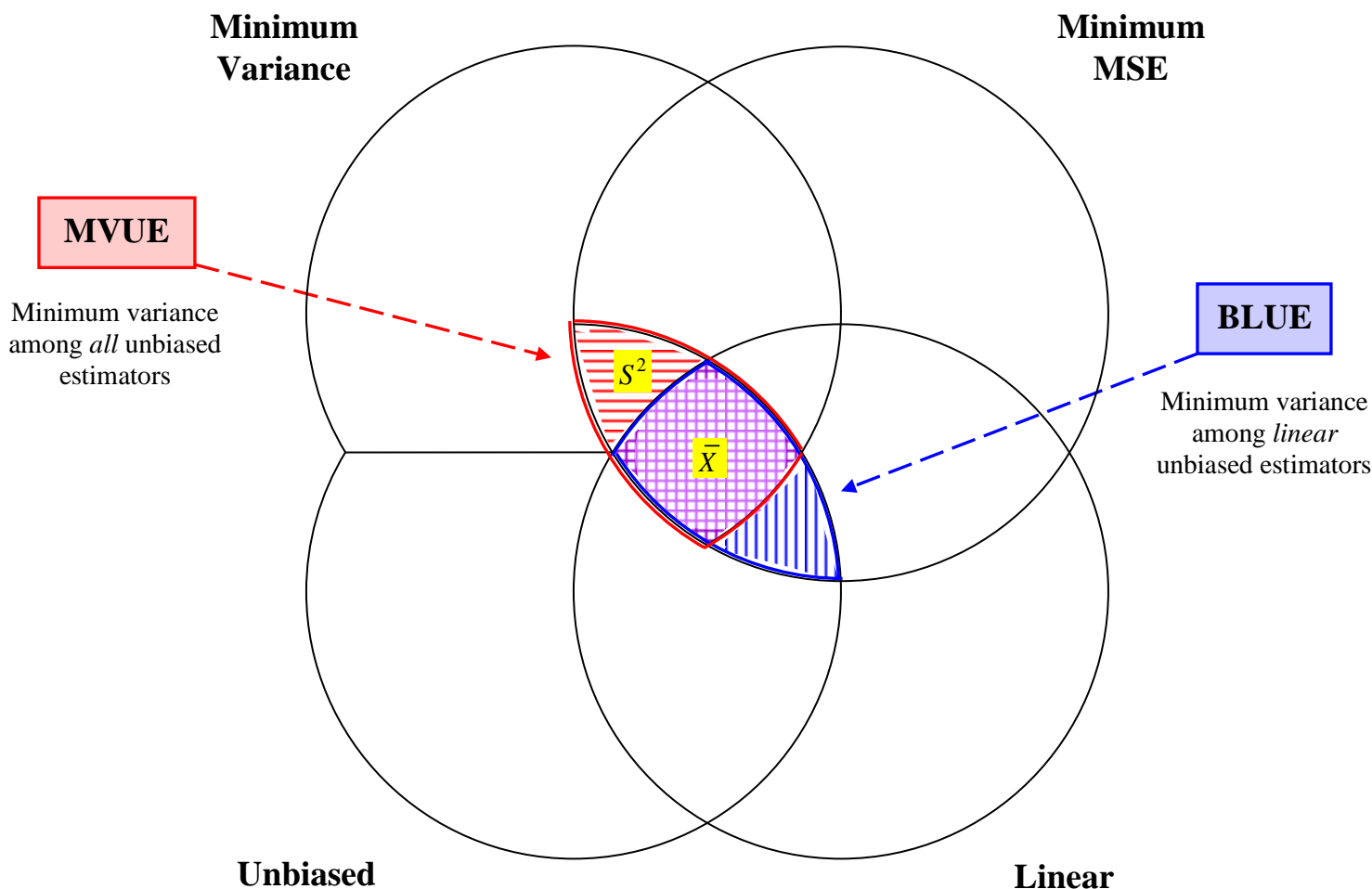
Comment: A parameter estimator $\hat{\theta}$ is defined to be **unbiased** if $E[\hat{\theta}] = \theta$, i.e., Bias = 0. In this case, MSE = Variance, so that if $\hat{\theta}$ minimizes MSE, it then follows that it has the smallest variance of *any* estimator. Such a highly desirable estimator is called MVUE (Minimum Variance Unbiased Estimator). It can be shown that the estimators \bar{X} and S^2 (of μ and σ^2 , respectively) are MVUE, but finding such an estimator $\hat{\theta}$ for a general parameter θ can be quite difficult in practice. Often, one must settle for either not having minimum variance or having a small amount of bias.

* using the basic properties of **mathematical expectation** given earlier

Related (but not identical) to this is the idea that of all **linear** combinations $c_1x_1 + c_2x_2 + \dots + c_nx_n$ of the data $\{x_1, x_2, \dots, x_n\}$ (such as \bar{X} , with $c_1 = c_2 = \dots = c_n = 1/n$) which are also **unbiased**, the one that minimizes MSE is called **BLUE** (Best Linear Unbiased Estimator). It can be shown that, in addition to being MVUE (as stated above), \bar{X} is also **BLUE**. To summarize,

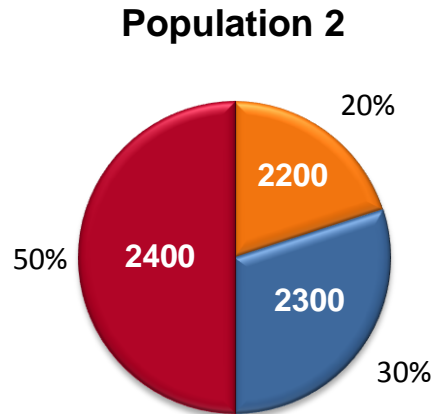
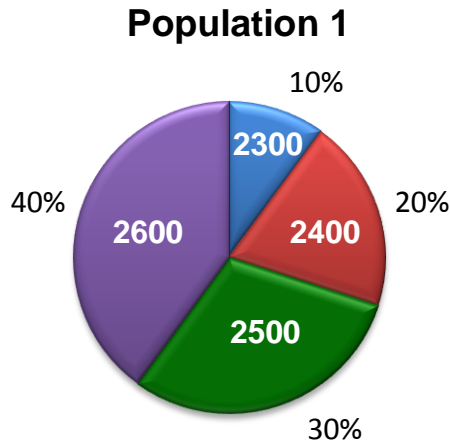
MVUE gives: Min Variance among *all* unbiased estimators
 \leq Min Variance among linear unbiased estimators
 $=$ Min MSE among linear unbiased estimators (since $\text{MSE} = \text{Var} + \text{Bias}^2$),
 given by **BLUE** (by def).

The Venn diagram below depicts these various relationships.



Comment: If $\text{MSE} \rightarrow 0$ as $n \rightarrow \infty$, then $\hat{\theta}$ is said to have **mean square convergence** to θ . This in turn implies “**convergence in probability**” (via “Markov's Inequality,” also used in proving Chebyshev's Inequality), i.e., $\hat{\theta}$ is a **consistent** estimator of θ .

Experiment 4 - revisited: Recall the previous example, where X_1 and X_2 represent the daily number of calories consumed in two populations, respectively.



x	$p_1(x)$
2300	0.1
2400	0.2
2500	0.3
2600	0.4

Mean(X_1) = μ_1 = 2500 cal;

Var(X_1) = σ_1^2 = 10000 cal²

x	$p_2(x)$
2200	0.2
2300	0.3
2400	0.5

Mean(X_2) = μ_2 = 2330 cal;

Var(X_2) = σ_2^2 = 6100 cal²

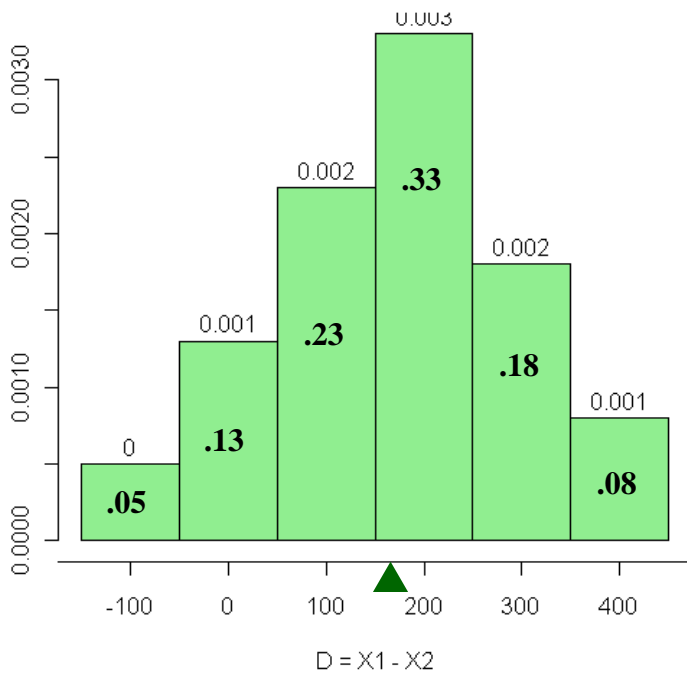
Case 1: First suppose that X_1 and X_2 are **statistically independent**, as shown in the **joint probability distribution** given in the table below. That is, each cell probability is equal to the product of the corresponding row and column marginal probabilities. For example, $P(X_1 = 2300 \cap X_2 = 2200) = .02$, but this is equal to the product of the column marginal $P(X_1 = 2300) = .1$ with the row marginal $P(X_2 = 2200) = .2$. Note that the marginal distributions for X_1 and X_2 remain the same as above, as can be seen from the single-underlined values for X_1 , and respectively, the double-underlined values for X_2 .

		$X_1 = \# \text{ calories for Pop 1}$				
		<u>2300</u>	<u>2400</u>	<u>2500</u>	<u>2600</u>	
$X_2 = \# \text{ calories for Pop 2}$	<u>2200</u>	.02	.04	.06	.08	<u>.20</u>
	<u>2300</u>	.03	.06	.09	.12	<u>.30</u>
	<u>2400</u>	.05	.10	.15	.20	<u>.50</u>
		<u>.10</u>	<u>.20</u>	<u>.30</u>	<u>.40</u>	1.00

Now imagine that we wish to compare the two populations, by considering the probability distribution of the calorie *difference* $D = X_1 - X_2$ between them. (The sum $S = X_1 + X_2$ is similar, and left as an exercise.)

<u>Events</u> $D = d$	<u>Sample Space</u> Outcomes in the form of ordered pairs (X_1, X_2)	<u>Probabilities</u> from joint distribution
$D = -100$:	(2300, 2400)	.05
$D = 0$:	(2300, 2300), (2400, 2400)	.13 = .03 + .10
$D = +100$:	(2300, 2200), (2400, 2300), (2500, 2400)	.23 = .02 + .06 + .15
$D = +200$:	(2400, 2200), (2500, 2300), (2600, 2400)	.33 = .04 + .09 + .20
$D = +300$:	(2500, 2200), (2600, 2300)	.18 = .06 + .12
$D = +400$:	(2600, 2200)	.08

As an example, there are two possible ways that $D = 300$ can occur, i.e., two possible outcomes corresponding to the event $D = 300$: *Either* $A = "X_1 = 2500 \text{ and } X_2 = 2200"$ *or* $B = "X_1 = 2600 \text{ and } X_2 = 2300,"$ that is, $A \cup B$. For its probability, recall that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. However, events A and B are disjoint, for they cannot both occur simultaneously, so that the last term is $P(A \cap B) = 0$. Thus, $P(A \cup B) = P(A) + P(B)$ with $P(A) = .06$ and $P(B) = .12$ from the joint distribution.



$$\begin{aligned} \text{Mean}(D) &= \mu_D = \\ &(-100)(.05) + (0)(.13) + (100)(.23) + \\ &(200)(.33) + (300)(.18) + (400)(.08) \\ &= \mathbf{170 \text{ cal}} \end{aligned}$$

i.e., $\mu_D = \mu_1 - \mu_2$ (Check this!)

$$\begin{aligned} \text{Var}(D) &= \sigma_D^2 = \\ &(-270)^2(.05) + (-170)^2(.13) + (-70)^2(.23) \\ &+ (30)^2(.33) + (130)^2(.18) + (230)^2(.08) \\ &= \mathbf{16100 \text{ cal}^2} \end{aligned}$$

i.e., $\sigma_D^2 = \sigma_1^2 + \sigma_2^2$ (Check this!)

Case 2: Now assume that X_1 and X_2 are **not statistically independent**, as given in the joint probability distribution table below.

		$X_1 = \# \text{ calories for Pop 1}$				
		<u>2300</u>	<u>2400</u>	<u>2500</u>	<u>2600</u>	
$X_2 = \# \text{ calories for Pop 2}$	<u>2200</u>	.01	.03	.07	.09	<u>.20</u>
	<u>2300</u>	.02	.05	.10	.13	<u>.30</u>
	<u>2400</u>	.07	.12	.13	.18	<u>.50</u>
		<u>.10</u>	<u>.20</u>	<u>.30</u>	<u>.40</u>	1.00

The events “ $D = d$ ” and the corresponding sample space of outcomes remain unchanged, but the last column of probabilities has to be recalculated, as shown. This results in a slightly different probability histogram (Exercise) and parameter values.

<u>Events</u> $D = d$	<u>Sample Space</u> <i>Outcomes in the form of ordered pairs (X_1, X_2)</i>	<u>Probabilities</u> from joint distribution
$D = -100$:	(2300, 2400)	.07
$D = 0$:	(2300, 2300), (2400, 2400)	.14 = .02 + .12
$D = +100$:	(2300, 2200), (2400, 2300), (2500, 2400)	.19 = .01 + .05 + .13
$D = +200$:	(2400, 2200), (2500, 2300), (2600, 2400)	.31 = .03 + .10 + .18
$D = +300$:	(2500, 2200), (2600, 2300)	.20 = .07 + .13
$D = +400$:	(2600, 2200)	.09

$$\begin{aligned} \text{Mean}(D) = \mu_D &= (-100)(.07) + (0)(.14) + (100)(.19) + (200)(.33) + (300)(.18) + (400)(.08) \\ &= \mathbf{170 \text{ cal}}, \text{ i.e., } \mu_D = \mu_1 - \mu_2. \end{aligned}$$

$$\begin{aligned} \text{Var}(D) = \sigma_D^2 &= (-270)^2(.07) + (-170)^2(.14) + (-70)^2(.19) + (30)^2(.31) + (130)^2(.20) + (230)^2(.09) \\ &= \mathbf{18517 \text{ cal}^2} \end{aligned}$$

It seems that “the mean of the difference is equal to the difference in the means” still holds, even when the two populations are dependent. But the variance of the difference is no longer necessarily equal the sum of the variances, as with independent populations.

These examples illustrate a general principle that can be rigorously proved with mathematics.

GENERAL FACT ~

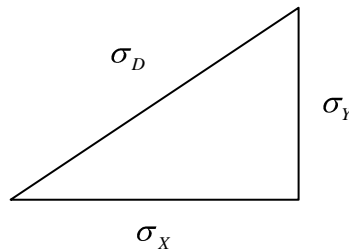
$$\text{Mean}(X + Y) = \text{Mean}(X) + \text{Mean}(Y) \quad \text{and} \quad \text{Mean}(X - Y) = \text{Mean}(X) - \text{Mean}(Y)$$

In addition, if X and Y are *independent* random variables,

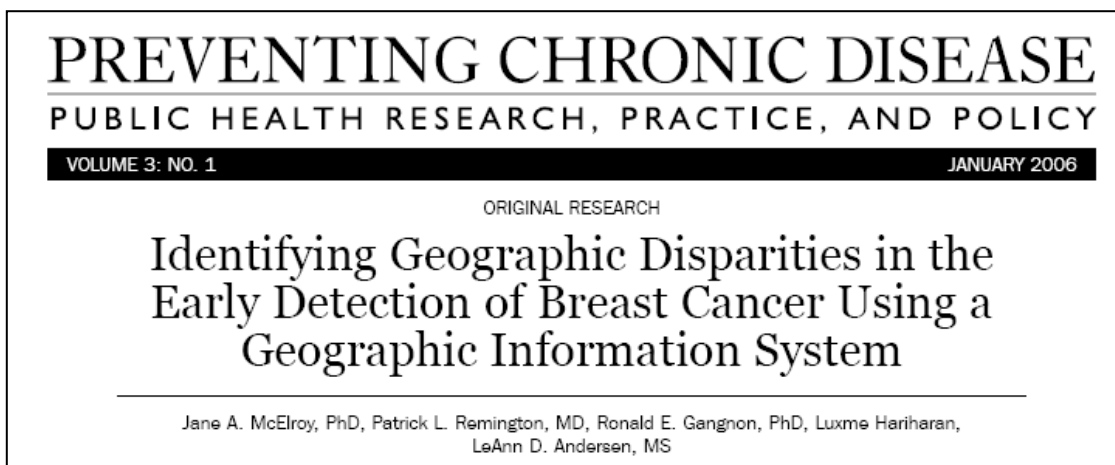
$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) \quad \text{and} \quad \text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y).$$

Comments:

- These formulas actually apply to both *discrete* and *continuous* variables (next section).
- The **difference** relations will play a crucial role in [6.2 - Two Samples](#) inference.
- If X and Y are *dependent*, then the two bottom relations regarding the variance also involve an additional term, $\text{Cov}(X, Y)$, the population **covariance** between X and Y . See problems 4.3/29 and 4.3/30 for details.
- The variance relation can be interpreted visually via the Pythagorean Theorem, which illustrates an important *geometric* connection, expanded in the Appendix.]



Certain discrete distributions (or discrete models) occur so frequently in practice, that their properties have been well-studied and applied in many different scenarios. For instance, suppose it is known that a certain population consists of **45%** males (and thus 55% females). If a random sample of **250** individuals is to be selected, then what is the probability of obtaining *exactly* **100** males? *At most* 100 males? *At least* 100 males? What is the “expected” number of males? This is the subject of the next topic:



POPULATION = Women diagnosed with breast cancer in Dane County, 1996-2000

Among other things, this study estimated that the rate of “breast cancer *in situ* (BCIS),” which is diagnosed almost exclusively via mammogram, is approximately 12-13%. That is, for any individual randomly selected from this population, we have a binary variable

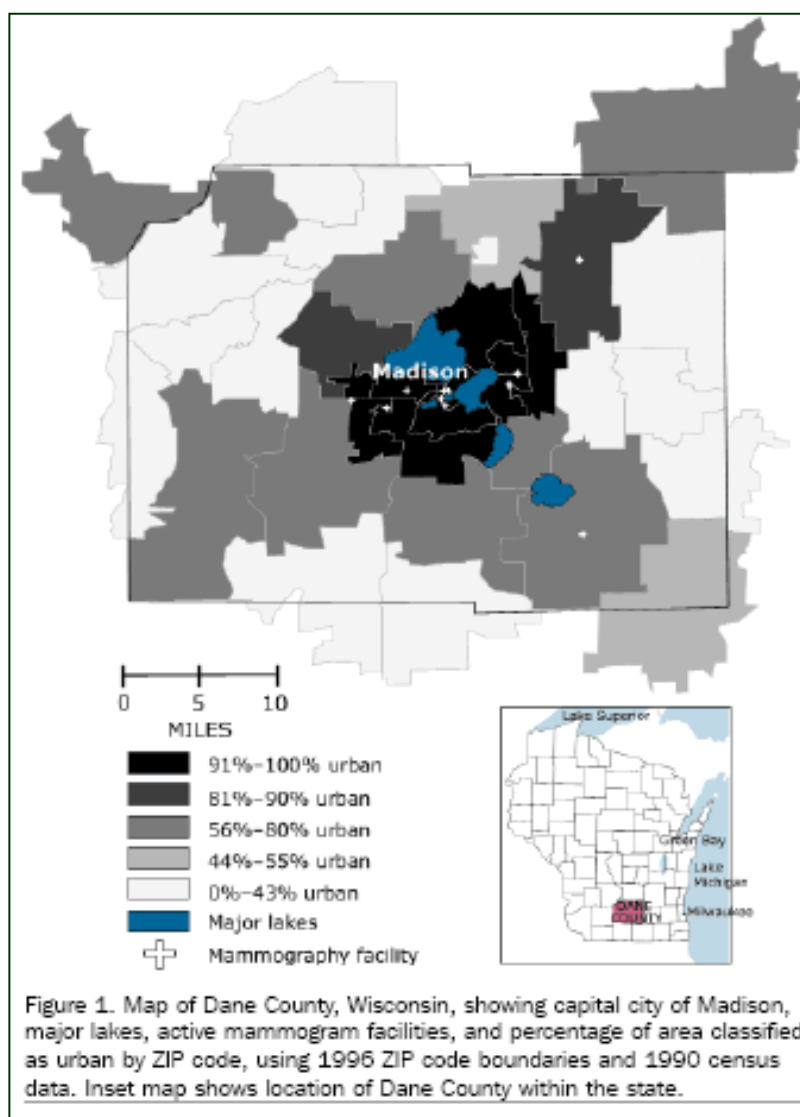
$$BCIS = \begin{cases} 1, & \text{with probability } 0.12 \\ 0, & \text{with probability } 0.88. \end{cases}$$

In a random sample of $n=100$ breast cancer diagnoses, let

$X = \#$ BCIS cases (0,1,2,...,100).

Questions:

- How can we model the **probability distribution** of X , and under what assumptions?
- **Probabilities of events**, such as $P(X=0)$, $P(X=20)$, $P(X \leq 20)$, etc.?
- **Mean** # BCIS cases = ?
- **Standard deviation** of # BCIS cases = ?



Full article available online at this [link](#).

Binomial Distribution (Paradigm model = coin tosses)

<p>Binary random variable:</p> $Y = \begin{cases} 1, & \text{Success (Heads)} \\ 0, & \text{Failure (Tails)} \end{cases}$	<p>Probability:</p> <p style="text-align: center;">with $P(\text{Success}) = \pi$</p> <p style="text-align: center;">with $P(\text{Failure}) = 1 - \pi$</p>
--	--

↓
Experiment: $n = 5$ independent coin tosses
 ↓

Sample Space S = {(H H H H H), ..., (T T T T T)} **#(S) = 2⁵ = 32**

(H H H H H)	(H H T H H)	(H T H H H)	(H T T H H)	(T H H H H)	(T H T H H)	(T T H H H)	(T T T H H)
(H H H H T)	(H H T H T)	(H T H H T)	(H T T H T)	(T H H H T)	(T H T H T)	(T T H H T)	(T T T H T)
(H H H T H)	(H H T T H)	(H T H T H)	(H T T T H)	(T H H T H)	(T H T T H)	(T T H T H)	(T T T T H)
(H H H T T)	(H H T T T)	(H T H T T)	(H T T T T)	(T H H T T)	(T H T T T)	(T T H T T)	(T T T T T)

Random Variable: $X =$ “# Heads in $n = 5$ independent tosses (0, 1, 2, 3, 4, 5)”

Events:	“ $X = 0$ ” = Exercise	$\#(X = 0) = \binom{5}{0} = 1$
	“ $X = 1$ ” = Exercise	$\#(X = 1) = \binom{5}{1} = 5$
	“ $X = 2$ ” = Exercise	$\#(X = 2) = \binom{5}{2} = 10$
	“ $X = 3$ ” = see above	$\#(X = 3) = \binom{5}{3} = 10$
	“ $X = 4$ ” = Exercise	$\#(X = 4) = \binom{5}{4} = 5$
	“ $X = 5$ ” = Exercise	$\#(X = 5) = \binom{5}{5} = 1$

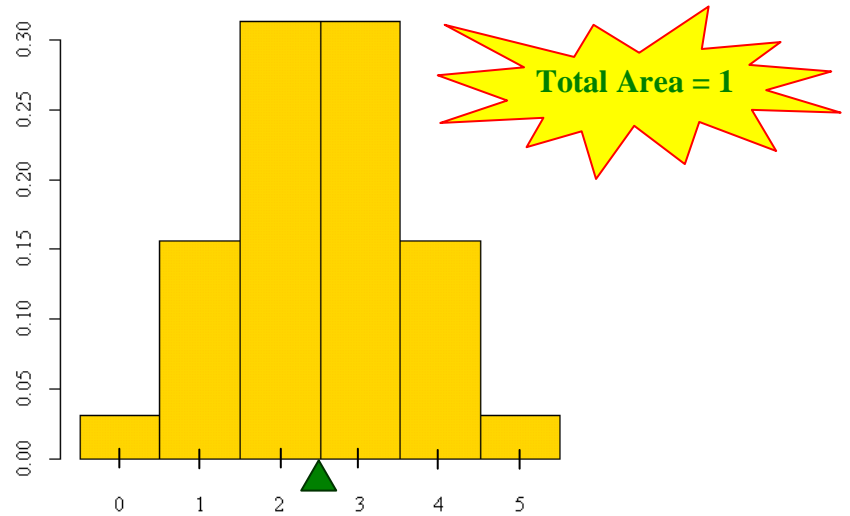
Recall: For $x = 0, 1, 2, \dots, n$, the **combinatorial symbol** $\binom{n}{x}$ – read “ n -choose- x ” – is defined as the value $\frac{n!}{x!(n-x)!}$, and counts the number of ways of rearranging x objects among n objects. See [Appendix > Basic Reviews > Perms & Combos](#) for details.

Note: $\binom{n}{r}$ is computed via the mathematical function “nCr” on most calculators.

Probabilities:

First assume the coin is **fair** ($\pi = 0.5 \Rightarrow 1 - \pi = 0.5$), i.e., **equally likely** elementary outcomes H and T on a single trial. In this case, the probability of any event A above can thus be easily calculated via $P(A) = \#(A) / \#(S)$.

x	$P(X = x) = \frac{1}{2^5} \binom{5}{x}$
0	$1/32 = 0.03125$
1	$5/32 = 0.15625$
2	$10/32 = 0.31250$
3	$10/32 = 0.31250$
4	$5/32 = 0.15625$
5	$1/32 = 0.03125$



Now consider the case where the coin is **biased** (e.g., $\pi = 0.7 \Rightarrow 1 - \pi = 0.3$). Calculating $P(X = x)$ for $x = 0, 1, 2, 3, 4, 5$ means summing $P(\text{all its outcomes})$.

Example: $P(X = 3) =$

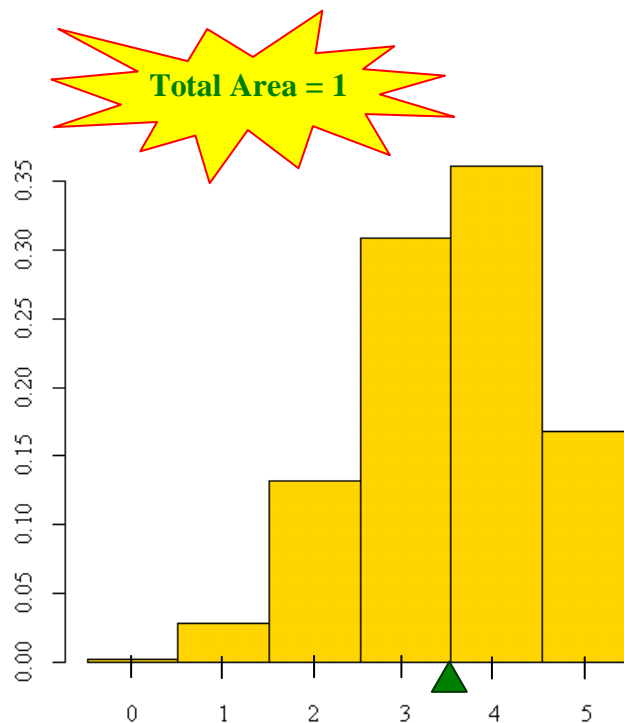
$$\begin{aligned}
 & \text{outcome} \quad \text{via independence of H, T} \\
 & P(\underline{\mathbf{H}}\underline{\mathbf{H}}\underline{\mathbf{H}}\underline{\mathbf{T}}\underline{\mathbf{T}}) = \overbrace{(0.7)(0.7)(0.7)(0.3)(0.3)}^{\text{via independence of H, T}} = (0.7)^3 (0.3)^2 \\
 & + P(\underline{\mathbf{H}}\underline{\mathbf{H}}\underline{\mathbf{T}}\underline{\mathbf{H}}\underline{\mathbf{T}}) = (0.7)(0.7)(0.3)(0.7)(0.3) = (0.7)^3 (0.3)^2 \\
 & + P(\underline{\mathbf{H}}\underline{\mathbf{H}}\underline{\mathbf{T}}\underline{\mathbf{T}}\underline{\mathbf{H}}) = (0.7)(0.7)(0.3)(0.3)(0.7) = (0.7)^3 (0.3)^2 \\
 & + P(\underline{\mathbf{H}}\underline{\mathbf{T}}\underline{\mathbf{H}}\underline{\mathbf{H}}\underline{\mathbf{T}}) = (0.7)(0.3)(0.7)(0.7)(0.3) = (0.7)^3 (0.3)^2 \\
 & + P(\underline{\mathbf{H}}\underline{\mathbf{T}}\underline{\mathbf{H}}\underline{\mathbf{T}}\underline{\mathbf{H}}) = (0.7)(0.3)(0.7)(0.3)(0.7) = (0.7)^3 (0.3)^2 \\
 & + P(\underline{\mathbf{H}}\underline{\mathbf{T}}\underline{\mathbf{T}}\underline{\mathbf{H}}\underline{\mathbf{H}}) = (0.7)(0.3)(0.3)(0.7)(0.7) = (0.7)^3 (0.3)^2 \\
 & + P(\underline{\mathbf{T}}\underline{\mathbf{H}}\underline{\mathbf{H}}\underline{\mathbf{H}}\underline{\mathbf{T}}) = (0.3)(0.7)(0.7)(0.7)(0.3) = (0.7)^3 (0.3)^2 \\
 & + P(\underline{\mathbf{T}}\underline{\mathbf{H}}\underline{\mathbf{H}}\underline{\mathbf{T}}\underline{\mathbf{H}}) = (0.3)(0.7)(0.7)(0.3)(0.7) = (0.7)^3 (0.3)^2 \\
 & + P(\underline{\mathbf{T}}\underline{\mathbf{H}}\underline{\mathbf{T}}\underline{\mathbf{H}}\underline{\mathbf{H}}) = (0.3)(0.7)(0.3)(0.7)(0.7) = (0.7)^3 (0.3)^2 \\
 & + P(\underline{\mathbf{T}}\underline{\mathbf{T}}\underline{\mathbf{H}}\underline{\mathbf{H}}\underline{\mathbf{H}}) = (0.3)(0.3)(0.7)(0.7)(0.7) = (0.7)^3 (0.3)^2
 \end{aligned}$$

via *disjoint* outcomes,

$$\left. \begin{aligned} & \dots \end{aligned} \right\} = \binom{5}{3} (0.7)^3 (0.3)^2$$

Hence, we similarly have...

x	$P(X = x) = \binom{5}{x} (0.7)^x (0.3)^{5-x}$
0	$\binom{5}{0} (0.7)^0 (0.3)^5 = 0.00243$
1	$\binom{5}{1} (0.7)^1 (0.3)^4 = 0.02835$
2	$\binom{5}{2} (0.7)^2 (0.3)^3 = 0.13230$
3	$\binom{5}{3} (0.7)^3 (0.3)^2 = 0.30870$
4	$\binom{5}{4} (0.7)^4 (0.3)^1 = 0.36015$
5	$\binom{5}{5} (0.7)^5 (0.3)^0 = 0.16807$



Example: Suppose that a certain medical procedure is known to have a 70% successful recovery rate (assuming independence). In a random sample of $n = 5$ patients, the probability that *three or fewer* patients will recover is:

$$\begin{aligned} \text{Method 1: } P(X \leq 3) &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) \\ &= 0.00243 + 0.02835 + 0.13230 + 0.30870 = 0.47178 \end{aligned}$$

$$\begin{aligned} \text{Method 2: } P(X \leq 3) &= 1 - [P(X = 4) + P(X = 5)] \\ &= 1 - [0.36015 + 0.16807] = 1 - 0.52822 = 0.47178 \end{aligned}$$

Example: The **mean** number of patients expected to recover is:

$$\begin{aligned} \mu = E[X] &= 0(0.00243) + 1(0.02835) + 2(0.13230) + 3(0.30870) + 4(0.36015) + 5(0.16807) \\ &= 3.5 \text{ patients} \end{aligned}$$

This makes perfect sense for $n = 5$ patients with a $\pi = 0.7$ recovery probability, i.e., their product. In the probability histogram above, the “balance point” fulcrum indicates the mean value of 3.5.

General formulation:

The Binomial Distribution

Let the *discrete* random variable $X = \text{"# Successes in } n \text{ independent Bernoulli trials } (0, 1, 2, \dots, n)\text{"}$, each having constant probability $P(\text{Success}) = \pi$, and hence $P(\text{Failure}) = 1 - \pi$. Then the probability of obtaining any specified number of successes $x = 0, 1, 2, \dots, n$, is given by the pmf $p(x)$:

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

We say that X has a **Binomial Distribution**, denoted $X \sim \text{Bin}(n, \pi)$.

Furthermore, the mean $\mu = n \pi$, and the standard deviation $\sigma = \sqrt{n \pi (1 - \pi)}$.

Example: Suppose that a certain spontaneous medical condition affects 1% (i.e., $\pi = 0.01$) of the population. Let $X = \text{"number of affected individuals in a random sample of } n = 300\text{"}$. Then $X \sim \text{Bin}(300, 0.01)$, i.e., the probability of obtaining any specified number $x = 0, 1, 2, \dots, 300$ of affected individuals is:

$$P(X = x) = \binom{300}{x} (0.01)^x (0.99)^{300-x}.$$

The **mean** number of affected individuals is $\mu = n\pi = (300)(0.01) = 3$ expected cases, with a **standard deviation** of $\sigma = \sqrt{(300)(0.01)(0.99)} = 1.723$ cases.

Probability Table for Binomial Dist.

x	$p(x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}$
0	$\binom{n}{0} \pi^0 (1 - \pi)^{n-0}$
1	$\binom{n}{1} \pi^1 (1 - \pi)^{n-1}$
2	$\binom{n}{2} \pi^2 (1 - \pi)^{n-2}$
etc.	etc.
n	$\binom{n}{n} \pi^n (1 - \pi)^{n-n}$

Exercise: In order to be a valid distribution, the sum of these probabilities must = 1. Prove it.

Hint: First recall the **Binomial Theorem**: How do you expand the algebraic expression $(a+b)^n$ for any $n = 0, 1, 2, 3, \dots$? Then replace a with π , and b with $1 - \pi$. Voilà!

Comments:

- The assumption of **independence** of the trials is *absolutely critical!* If not satisfied – i.e., if the “success” probability of one trial influences that of another – then the Binomial Distribution model can fail miserably. (Example: X = “number of children in a particular school infected with the flu”) The investigator must decide whether or not independence is appropriate, which is often problematic. If violated, then the **correlation** structure between the trials may have to be considered in the model.
- As in the preceding example, if the sample size n is very large, then the computation of $\binom{n}{x}$ for $x = 0, 1, 2, \dots, n$, can be intensive and impractical. An approximation to the Binomial Distribution exists, when n is large and π is small, via the **Poisson Distribution** (coming up...).
- Note that the standard deviation $\sigma = \sqrt{n \pi (1 - \pi)}$ depends on the value of π . (Later...)

How can we estimate the **parameter** π , using a sample-based **statistic** $\hat{\pi}$?

POPULATION

Binary random variable

$$Y = \begin{cases} 1, & \text{Success with probability } \pi \\ 0, & \text{Failure with probability } 1 - \pi \end{cases}$$

Experiment: n independent trials

SAMPLE

$$0/1 \ 0/1 \ 0/1 \ 0/1 \ 0/1 \ 0/1 \ \dots \ 0/1 \\ (y_1, y_2, y_3, y_4, y_5, y_6, \dots, y_n)$$

$$y_1 + y_2 + y_3 + y_4 + y_5 + \dots + y_n$$

Let $X = \# \text{ Successes in } n \text{ trials} \sim \text{Bin}(n, \pi)$
 ($n - X = \# \text{ Failures in } n \text{ trials}$).

Therefore, dividing by n ...

$$\frac{X}{n} = \text{proportion of Successes in } n \text{ trials}$$

$$\hat{\pi} = p \quad (= \bar{y}, \text{ as well})$$

and hence...

$$q = 1 - p = \text{proportion of Failures in } n \text{ trials.}$$

Example: If, in a sample of $n = 50$ randomly selected individuals, $X = 36$ are female, then the statistic $\hat{\pi} = \frac{X}{n} = \frac{36}{50} = \mathbf{0.72}$ is an estimate of the true probability π that a randomly selected individual *from the population* is female. The probability of selecting a male is therefore estimated by $1 - \hat{\pi} = \mathbf{0.28}$.

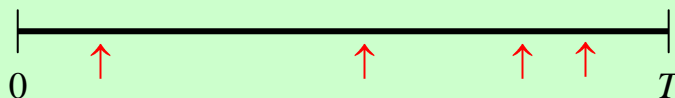


Poisson Distribution

(Models *rare* events)

Discrete Random Variable:

X = # occurrences of a (rare) event E , in a given interval of time or space, of size T . (0, 1, 2, 3, ...)



Assume:

1. All the occurrences of E are independent in the interval.
2. The **mean** number μ of expected occurrences of E in the interval is proportional to T , i.e., $\mu = \alpha T$. This constant of proportionality α is called the **rate** of the resulting **Poisson process**.

Then...

The Poisson Distribution

The probability of obtaining any specified number $x = 0, 1, 2, \dots$ of occurrences of event E is given by the pmf $p(x)$:

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

where $e = 2.71828\dots$ (“Euler’s constant”).

We say that X has a **Poisson Distribution**, denoted $X \sim \text{Poisson}(\mu)$. Furthermore, the mean is $\mu = \alpha T$, and the variance is $\sigma^2 = \alpha T$ also.

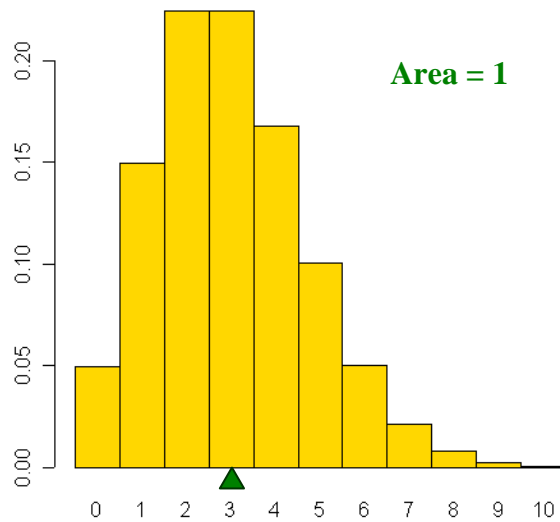
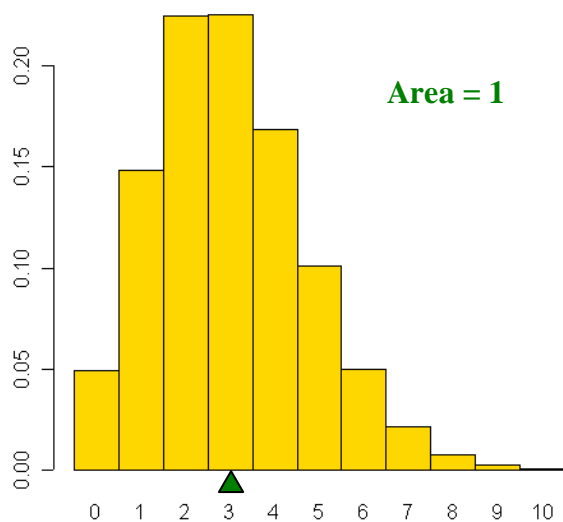
Examples: # bee-sting fatalities per year, # spontaneous cancer remissions per year, # accidental needle-stick HIV cases per year, hemocytometer cell counts

Example (see above): Again suppose that a certain spontaneous medical condition E affects 1% (i.e., $\alpha = 0.01$) of the population. Let $X =$ “number of affected individuals in a random sample of $T = 300$.” As before, the **mean** number of expected occurrences of E in the sample is $\mu = \alpha T = (0.01)(300) = 3$ cases. Hence $X \sim \text{Poisson}(3)$, and the probability that any number $x = 0, 1, 2, \dots$ of individuals are affected is given by:

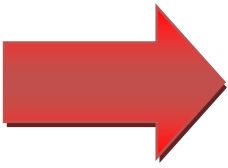
$$P(X = x) = \frac{e^{-3} 3^x}{x!}$$

which is a much easier formula to work with than the previous one. This fact is sometimes referred to as the **Poisson approximation to the Binomial Distribution**, when T (respectively, n) is large, and α (respectively, π) is small. Note that in this example, the variance is also $\sigma^2 = 3$, so that the standard deviation is $\sigma = \sqrt{3} = 1.732$, very close to the exact Binomial value.

x	Binomial $P(X = x) = \binom{300}{x} (0.01)^x (0.99)^{300-x}$	Poisson $P(X = x) = \frac{e^{-3} 3^x}{x!}$
0	0.04904	0.04979
1	0.14861	0.14936
2	0.22441	0.22404
3	0.22517	0.22404
4	0.16888	0.16803
5	0.10099	0.10082
6	0.05015	0.05041
7	0.02128	0.02160
8	0.00787	0.00810
9	0.00258	0.00270
10	0.00076	0.00081
etc.	$\rightarrow 0$	$\rightarrow 0$



Why is the Poisson Distribution a good approximation to the Binomial Distribution, for large n and small π ?

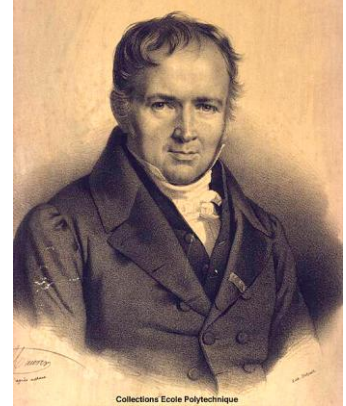


Rule of Thumb: $n \geq 20$ and $\pi \leq 0.05$; excellent if $n \geq 100$ and $\pi \leq 0.1$.

Let $p_{\text{Bin}}(x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}$ and $p_{\text{Poisson}}(x) = \frac{e^{-\lambda} \lambda^x}{x!}$, where $\lambda = n\pi$.

We wish to show formally that, for fixed λ , and $x = 0, 1, 2, \dots$, we have:

$$\lim_{\substack{n \rightarrow \infty \\ \pi \rightarrow 0}} p_{\text{Bin}}(x) = p_{\text{Poisson}}(x).$$



Siméon Poisson
(1781 - 1840)

Proof: By elementary algebra, it follows that...

$$\begin{aligned} p_{\text{Bin}}(x) &= \binom{n}{x} \pi^x (1 - \pi)^{n-x} \\ &= \frac{n!}{x! (n-x)!} \pi^x (1 - \pi)^n (1 - \pi)^{-x} \\ &= \frac{1}{x!} n(n-1)(n-2) \dots (n-x+1) \pi^x \left(1 - \frac{\lambda}{n}\right)^n (1 - \pi)^{-x} \\ &= \frac{1}{x!} \frac{n(n-1)(n-2) \dots (n-x+1)}{n^x} n^x \pi^x \left(1 - \frac{\lambda}{n}\right)^n (1 - \pi)^{-x} \\ &= \frac{1}{x!} \frac{n}{n} \left(\frac{n-1}{n}\right) \left(\frac{n-2}{n}\right) \dots \left(\frac{n-x+1}{n}\right) (n\pi)^x \left(1 - \frac{\lambda}{n}\right)^n (1 - \pi)^{-x} \\ &= \frac{1}{x!} \underbrace{1 \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \dots \left(1 - \frac{x-1}{n}\right)}_{\downarrow} \underbrace{\lambda^x}_{\downarrow} \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{\downarrow} \underbrace{(1 - \pi)^{-x}}_{\downarrow} \end{aligned}$$

As $n \rightarrow \infty$,
 $\pi \rightarrow 0$,

$$\frac{1}{x!} \quad 1(1)(1) \dots (1) = 1 \quad \lambda^x \quad e^{-\lambda} \quad 1^{-x} = 1$$

$$= \frac{e^{-\lambda} \lambda^x}{x!} = p_{\text{Poisson}}(x). \quad \text{QED}$$

Classical Discrete Probability Distributions

Binomial (probability of finding x “successes” and $n - x$ “failures” in n independent trials)

$X = \#$ successes (each with probability π) in n independent Bernoulli trials, $n = 1, 2, 3, \dots$

$$p(x) = P(X = x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}, \quad x = 0, 1, 2, \dots, n$$

Negative Binomial (probability of needing x independent trials to find k successes)

$X = \#$ independent Bernoulli trials for k successes (each with probability π), $k = 1, 2, 3, \dots$

$$p(x) = P(X = x) = \binom{x-1}{k-1} \pi^k (1 - \pi)^{x-k}, \quad x = k, k+1, k+2, \dots$$

Geometric: $X = \#$ independent Bernoulli trials for $k = 1$ success

$$p(x) = P(X = x) = \pi (1 - \pi)^{x-1}, \quad x = 1, 2, 3, \dots$$

Hypergeometric (modification of Binomial to sampling *without* replacement from “small” finite populations, relative to n .)

$X = \#$ successes in n random trials taken from a population of size N containing d successes, $n > \frac{N}{10}$

$$p(x) = P(X = x) = \frac{\binom{d}{x} \binom{N-d}{n-x}}{\binom{N}{n}}, \quad x = 0, 1, 2, \dots, d$$

Multinomial (generalization of Binomial to k categories, rather than just two)

For $i = 1, 2, 3, \dots, k$,

$X_i = \#$ outcomes in category i (each with probability π_i), in n independent Bernoulli trials, $n = 1, 2, 3, \dots$

$$\pi_1 + \pi_2 + \pi_3 + \dots + \pi_k = 1$$

$$p(x_1, x_2, \dots, x_k) = P(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k) = \frac{n!}{x_1! x_2! \dots x_k!} \pi_1^{x_1} \pi_2^{x_2} \dots \pi_k^{x_k},$$

$$x_i = 0, 1, 2, \dots, n \quad \text{with} \quad x_1 + x_2 + \dots + x_k = n$$

Poisson (“limiting case” of Binomial, with $n \rightarrow \infty$ and $\pi \rightarrow 0$, such that $n\pi = \lambda$, fixed)

$X = \#$ occurrences of a *rare* event (i.e., $\pi \approx 0$) among many (i.e., n large), with fixed mean $\lambda = n\pi$

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