4 - Discrete Probability Distributions

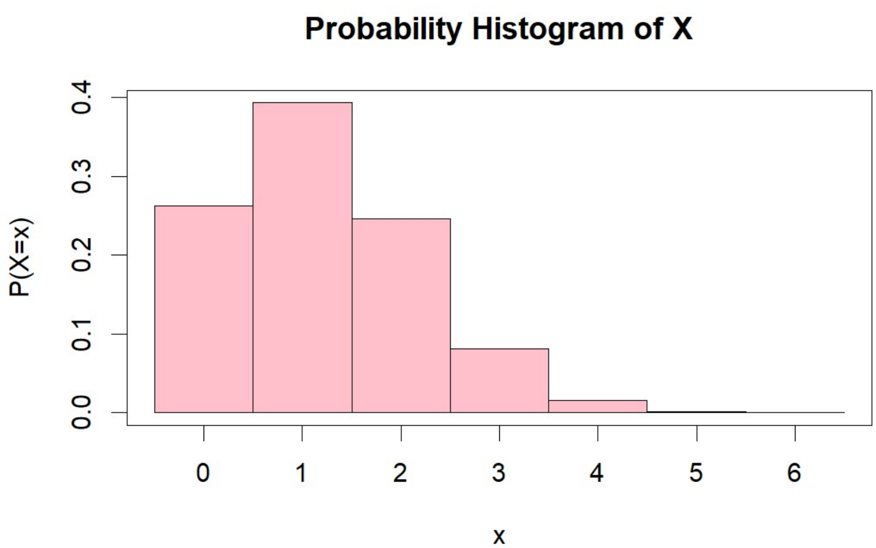
Example Suppose you roll five six-sided dice. Let $X =$ the *sum* of the five dice. Events specified in terms of X include:

$X = 30$

$X \geq 28$

Example Let $X =$ the number of CST graduates in a random sample of 6 who know how to construct a linked list in C++. Suppose the probability distribution of X is given by the following table and/or probability histogram.

x	$P(x)$
0	0.2620
1	0.3930
2	0.2459
3	0.0816
4	0.0160
5	0.0015
6	0+



What is $P(X \geq 3)$?

Definition A *discrete random variable* X is a random variable that has either a *finite* number of values or a *countable* number of values.

e.g., $X =$ the number of children a random person has in their lifetime

Definition The *discrete probability distribution* of a random variable X tells us $P(X = x)$ for any possible value x. It can be given by:

- a table/histogram
- a formula

Requirements for a Discrete Probability Distribution

For any discrete random variable X, the following must be true about the probabilities $P(x)$:

$\Sigma\, P(x) = 1$ sum of all probabilities equals 1

$0 \leq P(x) \leq 1$ for each possible value x

4.1 - Mean and Variance of a Random Variable

Example What is the mean value of $X =$ number rolled on a fair 6-sided die? Imagine many, many rolls:

```
> X.vals <- sample( 1:6, 100, replace=TRUE)
> X.vals
  [1] 5 3 3 4 4 5 4 4 5 3 3 6 2 1 5 5 1 4 2 2 5 5 3 3 1 6 4 5 3 3 4
 [32] 3 3 6 4 1 5 2 1 2 1 5 1 2 5 1 1 5 1 1 3 4 2 2 3 5 1 2 3 3 5 3
 [63] 6 1 6 1 6 2 2 5 5 5 6 2 5 3 5 1 5 5 5 1 5 2 3 3 1 6 2 5 6 4 1
 [94] 2 1 3 4 4 4 6
```

For these 100 simulated values, the mean $\bar{X}$ is:

```
> sum(X.vals) / 100
[1] 3.38
```

If we could compute the mean for *all possible* rolls of the 6-sided die, we would get:

Definition If X is a discrete random variable with probability distribution given by $P(x)$ then we define the mean and standard deviation as follows:

$$\mu = \sum_x [x \cdot P(x)]$$

The mean μ is also called the "expected value" of X and can be written as $E[X]$.

$$\sigma^2 = \sum_x [(x - \mu)^2 \cdot P(x)]$$

$$\sigma^2 = \sum_x [x^2 \cdot P(x)] - \mu^2$$

Example For the CST graduates example above, we have:

x	$P(x)$	$x \cdot P(x)$	x^2	$x^2 \cdot P(x)$
0	**0.2620**	**0.0000**	**0**	**0.0000**
1	**0.3390**	**0.3930**	**1**	**0.3930**
2	**0.2459**	**0.4918**	**4**	**0.9836**
3	**0.0816**	**0.2448**	**9**	**0.7344**
4	**0.0160**	**0.0640**	**16**	**0.2560**
5	**0.0015**	**0.0075**	**25**	**0.0375**
6	**0.0000**	**0.0000**	**36**	**0.0000**
Total	**1.000**	**1.2011**		**2.4045**

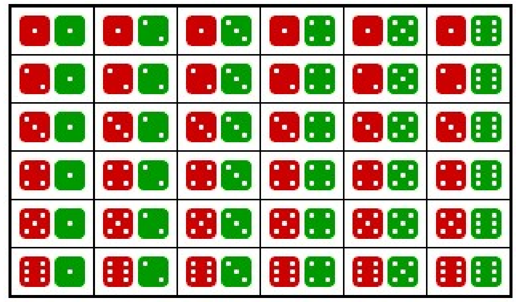
$\mu = 1.20$ CST graduates

$\sigma^2 = 2.4045 - 1.2011^2 = 0.9619$

$\sigma = 0.98$ CST graduates

Example Suppose I ask: “Pick a random number between 1 and 100.” What probability distribution am I likely thinking of? What is the mean, and what is the standard deviation?

Example (Minimum of Two Dice) The random experiment is rolling two fair six-sided die. Let X = the *minimum* of the two die values. What is the expected value of X? What is the standard deviation of X?



Example (Minimum of Three Dice) If X = the minimum of *three* fair six-sided dice, what is the mean value of X? Find the answer by simulation in R.

In these examples, we were forced to calculate μ and σ from the definitions (or simulation). There are some random experiments that are so commonly used that statisticians have developed exact formulas for μ and σ that are easy to calculate. We turn to these next.

- Binomial
- Geometric
- Hypergeometric
- Poisson

4.2 - Binomial Distribution

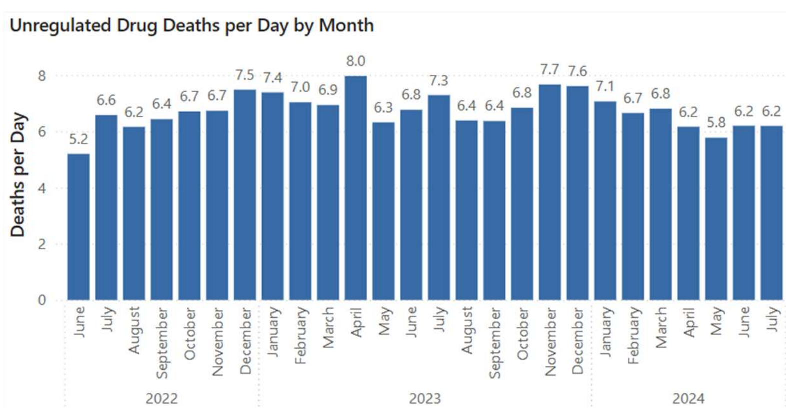
The binomial distribution arises when we are concerned with the variable:

$$X = \text{the number of “successes” in a series of } n \text{ independent trials}$$

Here a *trial* (also called a *trial*) is any random experiment that has just two possible outcomes. By convention, these two outcomes are called “success” and “failure”. For instance, they could be:

- Win/Lose
- Live/Die
- True/False
- Pass/Fail
- Within Specification/Not Within Specification

Example (Drug Deaths) Suppose $n = 10\ 000$ people consume opioid drugs today in BC. To simplify, suppose that each opioid user has the same risk $p = 0.06\%$ of dying today.

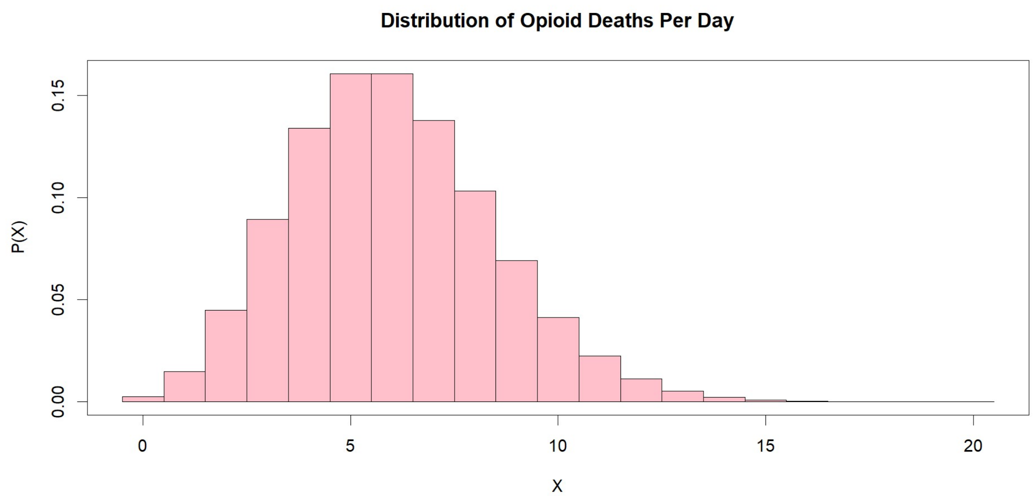


We make these assumptions:

1. **Fixed n** – the total number of users today is known and fixed.
2. **Success/Failure –** each opioid user either dies (“success”) or lives (“failure”).
3. **Equal p** - each opioid user has the same probability of dying today, $p = 0.0006$
4. **Independence -** each user’s outcome is unaffected by other users’ outcomes

Finally, let

$X =$ the number of deaths today due to opioid overdose (**example continued)** Under the above assumptions, the variable X follows a *binomial* distribution, shown here:



As an example, let's calculate just one value:

$$P(X = 5)$$

CAUTION: we need to verify that all of the conditions of the binomial distribution are satisfied before we apply our formulas in any given problem.

Example Rolling a fair die 5 times and counting the number of 3s obtained is an example of a binomial experiment. Here $X =$ the number of 3s obtained.

Check the conditions:

Calculate the probability distribution of X.

General Formula If a variable X satisfies the conditions of a binomial variable, with n trials and probability p of success, then

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

where

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

Example Suppose 5 cards are selected *with* replacement from a deck. Find the probability that 2 are jacks.

Example Suppose 5 cards are selected *with* replacement from a deck. Find the probability that at least 2 are jacks.

Mean and Variance for Binomial Distributions

The mean and variance of any probability distribution are found using:

$$\mu = \sum_x [x \cdot P(x)]$$

$$\sigma^2 = \sum_x [(x-\mu)^2 \cdot P(x)]$$

It is possible to start with these definitions and to then use the binomial distribution assumptions to derive the results:

$$\mu = np$$

$$\sigma^2 = npq$$

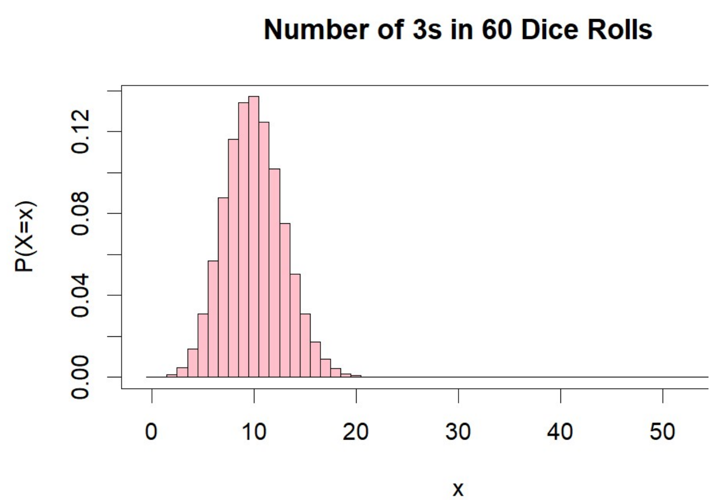
Example In the earlier example about opioid deaths, we had $n = 10\,000$ and $p = 0.0006$. Then the mean value of X (deaths in a day) is:

$$\mu = np = 10000 \times 0.0006 = 6$$

Example Suppose we roll a six-sided die 60 times. Let $X =$ the number of times we roll a 3. Then the *parameters* of the distribution of X are:

$\mu =$

$\sigma =$



Example (Empirical Rule) For a certain model of laser printer, the probability of a unit needing repairs in the first year is 10%. In an attempt to reduce this rate, modifications were made to the printer design and a year later a study of 200 randomly selected modified laser printers was conducted.

a. Assuming the modifications had no effect on the printer reliability, find the mean and standard deviation of the number of laser printers that need repair among 200.

b. Suppose that in the sample of 200 laser printers, 14 needed repair. Assuming the modifications had no effect, is this rate unusually low?

c. Does this sample provide good evidence that the modifications improved the reliability of this model of laser printer?

4.3 - Hypergeometric Distribution

The hypergeometric distribution is similar to the binomial distribution. In both cases, we are concerned with the variable

$$X = \text{number of successes in a series of } n \text{ trials}$$

The difference is shown in the table below:

Binomial	Hypergeometric
• trials are *independent* • sampling *with replacement*, or sampling from an *infinite* population • same probability p for each trial	• trials are *dependent* • sampling *without replacement* from a finite population • probability of success changes for subsequent trials

If the number of trials, n, is small compared to the population size, then the Binomial model and the Hypergeometric model give *very similar* probabilities.

Hypergeometric Distribution Formula

Suppose a population of N objects contains K "success" objects and $N - K$ "failure" objects. If you select a random sample of size n from this population, let

$$X = \text{the number of “success” objects in the sample.}$$

Then the probability of getting x "success" objects is

$$P(x) = \frac{C(K,x) \cdot C(N-K,\ n-x)}{C(N,n)}$$

for any $x = 0, 1, 2, \ldots, n$.

The mean and variance of X are:

$$\mu = n \cdot \frac{K}{N}$$

$$\sigma^2 = n \cdot \frac{K}{N} \cdot \frac{N-K}{N} \cdot \frac{N-n}{N-1}$$

Example Suppose you sample 5 cards from a deck of 52 cards. Let $X =$ the number of Jacks you get in the sample? (There are 4 Jacks in the deck). What is $P(X = 2)$?

Hypergeometric (without replacement)

$$P(X = 2) = \frac{C(4,2) \cdot C(48,3)}{C(52,5)} =$$

```
> dhyper(2, 4, 48, 5)
[1] 0.03992982
```

Binomial (with replacement)

$$P(X = 2) = C(5,2) \cdot \left(\frac{4}{52}\right)^2 \cdot \left(\frac{48}{52}\right)^3 =$$

```
> dbinom(2, 5, 4/52)
[1] 0.04654006
```

Now, suppose you combine 10 decks of cards into one “superdeck”. If you randomly select 5 cards, what is the probability of getting 2 Jacks?

Hypergeometric:

Binomial:

5% Rule

If the sample size n is 5% or less of the total population N, then probabilities calculated using Hypergeometric and Binomial distributions are practically the same.

If $\boldsymbol{n > 0.05 \times N}$ then be sure to use Hypergeometric.

Example Suppose a class of 90 students has 40 Apple Mac users. To conduct a market research survey, you randomly select a sample of 10 students.

Let $X =$ the number of Apple Mac users in the sample.

Then X is a hypergeometric variable with parameters:

$N =$

$K =$

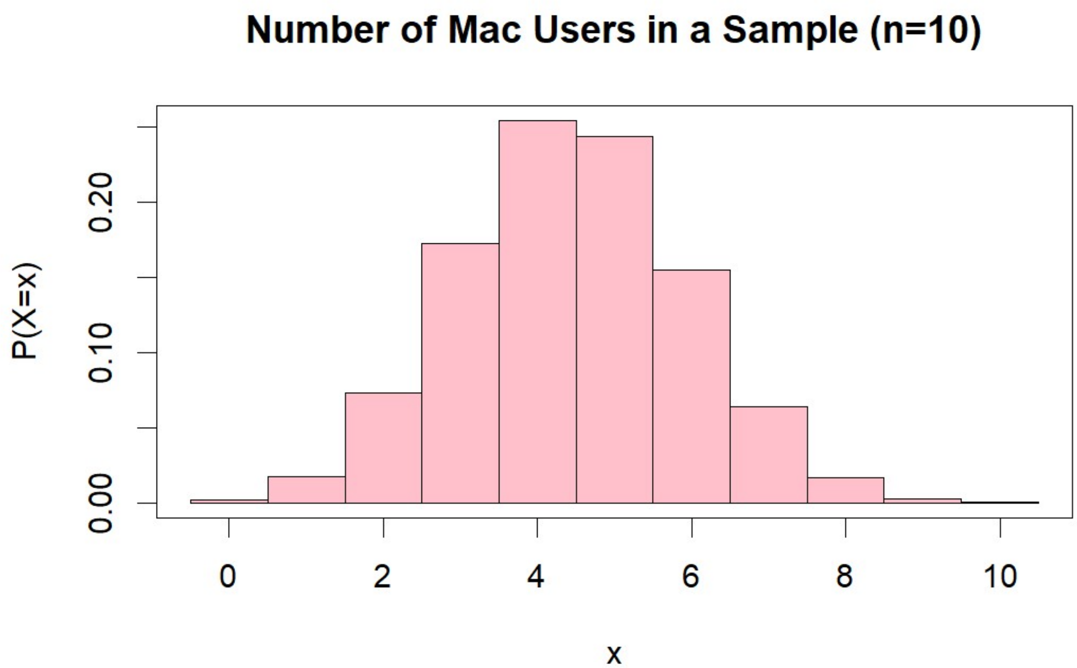
$n =$

a. What is the probability that 4 students in the sample are Apple Mac users?

b. What is the probability that 4 *or fewer* of them are Apple Mac users?

c. What are the mean and standard deviation of X?

d. Generate a probability histogram of X.



e. If you found $X = 8$ Mac users, would this be considered *unusual*?

4.4 - Geometric Distribution

The Geometric Distribution is based on the same assumptions as the binomial distribution: n trials each with probability of success p. In this case, we are working with the variable

$X =$ the number of trials it takes to obtain the first "success"

Example Suppose a tech-support telephone help line is occupied 75% of the time. Find the probability that you will have to call 3 times to gain access.

Solution: If you first gain access on the 3rd call, then you had to have:

- failure on the first call: probability $q = 0.75$
- failure on the second call: probability $q = 0.75$
- success on the third call: probability $p = 0.25$

Therefore,

$P(3) =$

Geometric Distribution Formulas

Suppose every trial of a random experiment has probability p of success and probability $q = 1 - p$ of failure, and all trials are independent.

Let $X =$ the number of trials it takes to get the first success (possible values: 1, 2, 3, ...). The probability distribution is given by:

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

The mean value and variance of X are consequently:

$$\mu = \frac{1}{p}$$

$$\sigma^2 = \frac{1-p}{p^2}$$

Derivation of μ

You are not expected to be able to perform a derivation like the following, but it may help you to be familiar with it.

$\mu = \sum_{x=1}^{\infty} x \cdot P(x)$	definition of mean value
$= \sum_{x=1}^{\infty} x \cdot q^{x-1} \cdot p$	sub formula for $P(x)$
$= p \cdot \left[\sum_{x=1}^{\infty} x \cdot q^{x-1}\right]$	factor out the p
$= p \cdot \frac{d}{dq}\left[\sum_{x=1}^{\infty} q^{x}\right]$	since $x \cdot q^{x-1} = \frac{d}{dq}[q^x]$
$= p \cdot \frac{d}{dq}\left[\frac{q}{1-q}\right]$	sum of a geometric series
$= p \cdot \frac{(1-q)+q}{(1-q)^2}$	Quotient Rule
$= p \cdot \frac{1}{(1-q)^2}$ $= \frac{p}{p^2} = \frac{1}{p}$	algebraic simplification (using $p = 1 - q$)

Example Suppose a tech-support telephone help line is occupied 75% of the time. Let $X =$ the number of times you must call until you get access.

Find $E[X] =$

Find $\sigma_X =$

(**example continued**)

Find $P(X \geq 4)$

Generate a probability histogram for the variable X using R.

4.5 - Poisson Distribution

The *Poisson Distribution* is named after the French mathematician Siméon Poisson (1781-1840). This distribution is one of the most important probability distributions in engineering and computer science. It arises whenever we are modelling events that occur randomly throughout a given time interval.

Examples The Poisson distribution could be applied for variables X like the following:

$X =$ number of jobs arriving for service at a CPU per second

$X =$ number of bits arriving in error at a network node per minute

Assumptions for Poisson Distribution

- $X =$ the number of occurrences of an event over some interval.
- Occurrences happen with uniform probability over the time interval.
- Occurrences are independent of each other.
- The *mean* number of occurrences is known:

$\lambda =$ mean number of occurrences during the time interval

Poisson Distribution Formulas

Given the assumption above, the probability of having x occurrences is:

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

Here $e = 2.71828 \ldots$ is Euler's constant from calculus. The mean and variance of X are:

$$\mu = \lambda$$

$$\sigma^2 = \lambda$$

Example Suppose a web server gets on average one request every 2 seconds. Assuming requests arrive randomly and independently over time, what is the probability that 20 requests will arrive during a one-minute interval?

What is the probability that 20 *or fewer* requests will arrive in a given minute?

Is getting 20 requests in a given minute *unusual* in the statistical sense?

(example continue) Generate the probability histogram for $X =$ the number of web requests in a given minute.