Probability Distributions

Lecture 4

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.2     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

1 Outline

- Probability Distributions of Discrete Random Variables
  - Binomial Distribution
  - Poisson Distribution
  - Other Distributions
- Probability Distributions of Continuous Random Variables
  - Uniform Distribution
  - Gaussian (Normal) Distribution
  - Other Distributions
- Applications of the Gaussian Distribution

2 Probability Distributions of Discrete Random Variables

2.1 Discrete Random Variables

Discrete Random Variable

A random variable is discrete if it can only take on a finite or countably infinite number of values.

Tip

Integers are countably infinite, hence counts are considered discrete.
Qualitative variables often take on a finite number of values.

2.2 Discrete Probability Distribution

The probability distribution of a discrete random variable is a **table, graph, formula, or other device** used to specify all possible values of a discrete random variable along with their respective probabilities.

Tip

Formulas that describe discrete probability distributions are also known as the probability mass function (pmf). The function is often denoted as $P(X = x)$, $p_X(x)$, or simply $p(x)$.

Important

Notation is important here. X is the random variable, x is the outcome/realization. $X = x$ is the event that the random variable is equal to the outcome x.

2.3 Example 1

Consider a random variable X that represents the outcome of a fair six-sided die. Prior to rolling the die, X can take on any one of the six values: {1,2,3,4,5,6}. The discrete probability distribution can be expressed in the following forms:

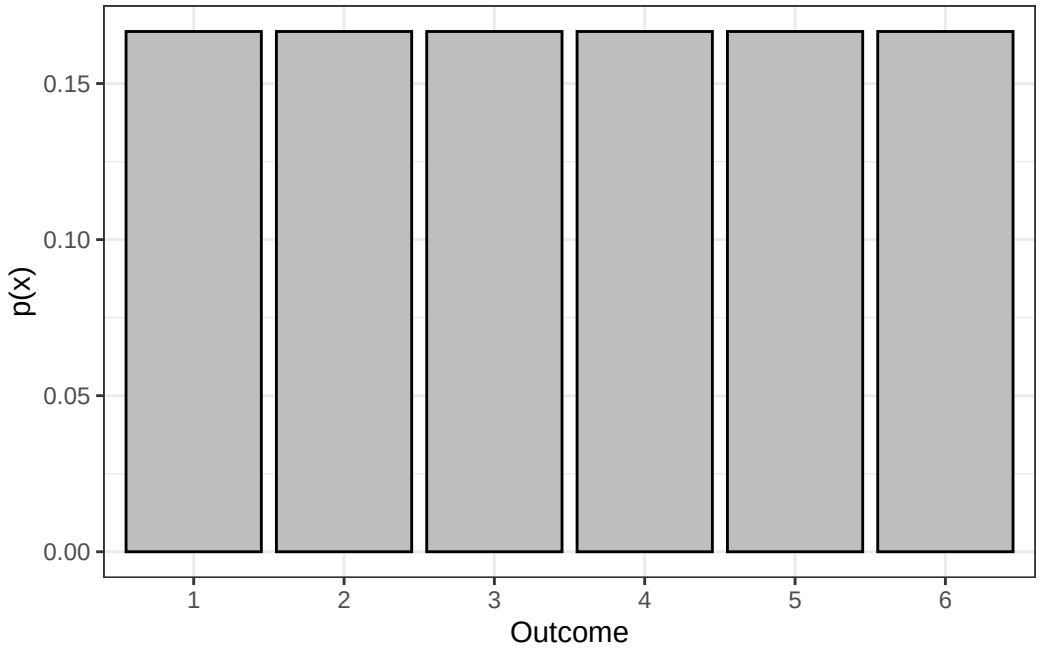
2.3.1 Table

Outcome of X	p(x)
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

2.3.2 Formula

$$p(x) = \begin{cases} 1/6, \ if \ x = 1, 2, ..., 6 \\ 0, \ otherwise \end{cases}$$

2.3.3 Graph



2.3.4 Notes

> **!** Important
>
> Here, X is the result of rolling the die, x is the outcome measured after rolling the die which can be any of the following numbers: {1,2,3,4,5,6}.

This PMF is also referred to as a **uniform discrete distribution** because the probability is uniformly distributed across the sample space.

2.4 Example 2.1

The frequency table below shows the results of a survey in which participants residing in the Appalachian region of southern Ohio were asked how many food assistance programs they had used in the last 12 months.

2.4.1 Frequency Table

Number of Programs	Frequency
1	62
2	47
3	39
4	39
5	58
6	37
7	4
8	11
Total	297

2.4.2 Probability Distribution

We can calculate the probability distribution by calculating the relative frequency of each outcome. This can be done by dividing the frequencies by the total (297).

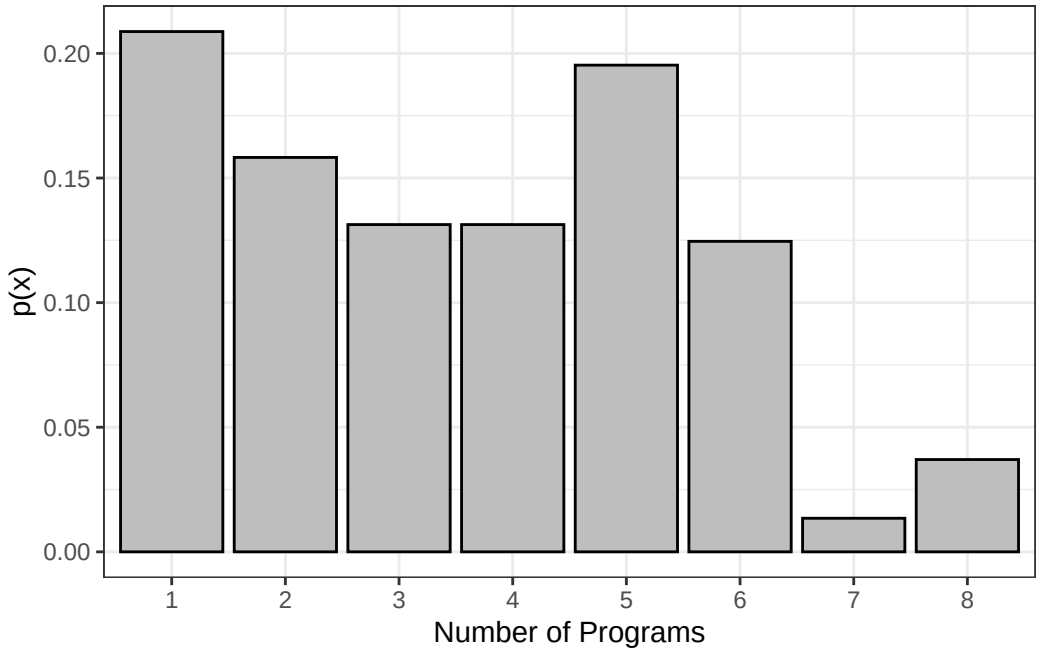
```
df <- data.frame(`Number of Programs`=factor(c(1:8), levels=c(1:8)),
Frequency = c(62,47,39,39,58,37,4,11),
Relative.Frequency = c(62,47,39,39,58,37,4,11)/297)

df
```

```
  Number.of.Programs Frequency Relative.Frequency
1                  1        62          0.20875421
2                  2        47          0.15824916
3                  3        39          0.13131313
4                  4        39          0.13131313
5                  5        58          0.19528620
6                  6        37          0.12457912
7                  7         4          0.01346801
8                  8        11          0.03703704
```

2.4.3 Graph

```
library(ggplot2)
ggplot(df, aes(x=Number.of.Programs,y=Relative.Frequency)) +
  geom_col(fill="gray",color="black") +                                    ①
  theme_bw() +                                                             ②
  labs(x="Number of Programs",y="p(x)")                                    ③
```

① Formatting the bar plot with a gray fill and black outline
② `theme_bw` using a black and white theme, which looks good aesthetically
③ Axes labels were changed.



2.5 Example 2.2

Consider the probability distribution in Example 2.1.

- What is the probability of randomly selecting a family who used four assistance programs?

$p(4) = 0.1313$

- What is the probability of randomly selecting a family who used one or three assistance programs?

> **Important**
>
> These events are disjoint, hence P(one AP AND three AP 0) = 0. Using the addition rule,
> $P(one \cup three) = P(one) + P(three) = 0.34$

2.6 PMF: Properties

The discrete probability distribution should satisfy the following properties:

1. $0 \leq p(x) \leq 1$ for all x.
2. $\sum_{all\ x} p(x) = 1$ (The sum of the probabilities of all disjoint outcomes in the sample space is 1.)

2.7 Cumulative Distributions

In addition to discrete probability distributions, discrete variables also have cumulative distribution functions (CDF)

> **Tip**
>
> The CDF for a discrete variable can be calculated by successively adding the probabilities of the outcome of interest and others before it.

> **Note**
>
> The CDF is often referred to as $F_X(x)$ or $F(x)$ to signify $P(X \leq x)$.

2.8 Example 1

Consider the PMF of the fair six-sided die. The cumulative distribution can be calculated using the cumulative relative frequency.

Outcome of X	p(x)	F(x) = P(X≤x)
1	1/6	1/6
2	1/6	2/6
3	1/6	3/6
4	1/6	4/6
5	1/6	5/6
6	1/6	6/6

2.9 Example 2

Consider the PMF for the food assistance example. We can calculate and visualize the cumulative relative frequency using `R`.

2.9.1 Table

To create a separate table as a function of existing tables, we use the `mutate(data_frame,expr)` function in the `dplyr` package available in `tidyverse`. The cumulative sum can also be calculated automatically using the `cumsum(column_name)` function

```
df <- data.frame(`Number of Programs`=1:8,
Frequency = c(62,47,39,39,58,37,4,11),
Relative.Frequency = c(62,47,39,39,58,37,4,11)/297)

mutate(df,Cumulative.Relative.Frequency = cumsum(Relative.Frequency))
```

```
  Number.of.Programs Frequency Relative.Frequency Cumulative.Relative.Frequency
1                  1        62         0.20875421                       0.2087542
2                  2        47         0.15824916                       0.3670034
3                  3        39         0.13131313                       0.4983165
4                  4        39         0.13131313                       0.6296296
5                  5        58         0.19528620                       0.8249158
6                  6        37         0.12457912                       0.9494949
7                  7         4         0.01346801                       0.9629630
8                  8        11         0.03703704                       1.0000000
```

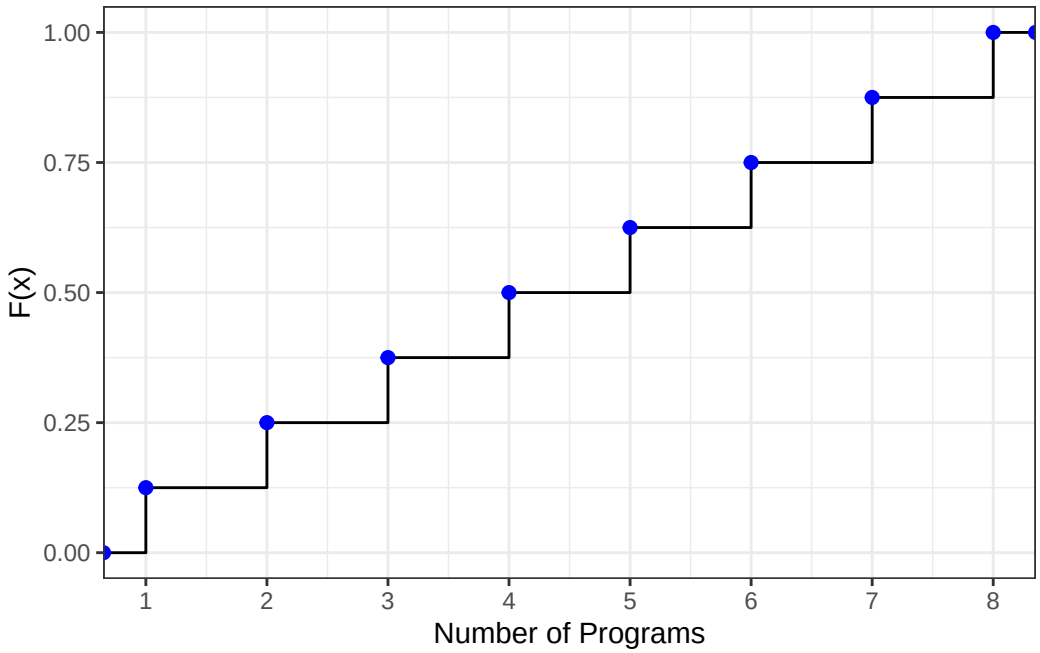
2.9.2 Graph

The plot of the cumulative distribution function is called an *ogive*. The option `stat_ecdf` provides us with a way to create ogives based on frequency values.

```
ggplot(df, aes(x=Number.of.Programs, y=Frequency)) +
  stat_ecdf(geom="step") + stat_ecdf(geom="point",color="blue",size=2) +    ①
  theme_bw() +
  scale_x_continuous(breaks=c(1:8))+                                        ②
  labs(x="Number of Programs",y="F(x)")
```

① "step" creates the lines in the ogive, while "point" shows the points corresponding to the values of F(x). The vertical lines are uninterpretable; they only provide a visual cue for the jump from one value to another.

② The `scale_x_continuous(breaks=c(1:8))` statement tells R to show all the points between 1 and 8 in the graph. This way, it is easier to interpret for other readers.



2.10 Example 2.1

Consider the food assistance example. What is the probability of randomly selecting a family that have used less than three assistance programs.

```r
df <- data.frame(`Number of Programs`=1:8,
Frequency = c(62,47,39,39,58,37,4,11),
Relative.Frequency = c(62,47,39,39,58,37,4,11)/297)

mutate(df,Cumulative.Relative.Frequency = cumsum(Relative.Frequency))
```

```
  Number.of.Programs Frequency Relative.Frequency Cumulative.Relative.Frequency
1                  1        62         0.20875421                       0.2087542
2                  2        47         0.15824916                       0.3670034
3                  3        39         0.13131313                       0.4983165
4                  4        39         0.13131313                       0.6296296
5                  5        58         0.19528620                       0.8249158
6                  6        37         0.12457912                       0.9494949
7                  7         4         0.01346801                       0.9629630
8                  8        11         0.03703704                       1.0000000
```

> **Tip**
>
> Less than three assistance programs can also mean "two or less programs" or "at most two programs". Hence, we are interested in $F(2)$
> $F(2) = P(X \leq 2) = P(X = 1) + P(X = 2) = 0.367$

2.11 Exercise

Consider the food assistance example.

2.11.1 Question

What is the probability that a randomly selected family utilized at least 4 programs?

```r
df <- data.frame(`Number of Programs`=1:8,
Frequency = c(62,47,39,39,58,37,4,11),
Relative.Frequency = c(62,47,39,39,58,37,4,11)/297)

mutate(df,Cumulative.Relative.Frequency = cumsum(Relative.Frequency))
```

```
  Number.of.Programs Frequency Relative.Frequency Cumulative.Relative.Frequency
1                  1        62         0.20875421                       0.2087542
2                  2        47         0.15824916                       0.3670034
```

3	3	39	0.13131313	0.4983165
4	4	39	0.13131313	0.6296296
5	5	58	0.19528620	0.8249158
6	6	37	0.12457912	0.9494949
7	7	4	0.01346801	0.9629630
8	8	11	0.03703704	1.0000000

2.11.2 Answer

There are two ways to approach this question.

> **Tip**
>
> Approach 1: Sum the relative frequencies from $X = 4$ to the maximum possible value ($X = 8$).
> $p(4) + p(5) + p(6) + p(7) + p(8) = 0.1313+0.1953+0.1246+0.0135+0.0370 = 0.50167$

> **Tip**
>
> Approach 2: Use the complement principle. To isolate the families who utilized at least 4 programs, we must remove those who utilized three programs or less. Then,
> $p(4) + p(5) + p(6) + p(7) + p(8) = 1\text{-F}(3) = 1\text{-}0.4983 = 0.5017$

2.12 Expectation Values of Discrete Probability Distributions

The expected value of a random variable can be thought of as the average of a very large number of observations of the random variable.

> **Tip**
>
> The expected value of X, denoted by $E(X)$, can be calculated by taking the sum of the products of each outcome $\{x_1, x_2, ...,\}$ and their corresponding probabilities $\{p(x_1), p(x_2), ...\}$. In mathematical terms,
>
> $$E(X) = \sum_i x_i p(x_i)$$

> **Important**
>
> $E(X)$ is also the **mean** of the discrete probability distribution, μ.

2.13 Example

Consider a random variable X that represents the outcome of a fair six-sided die. Prior to rolling the die, X can take on any one of the six values: {1,2,3,4,5,6}. The probability distribution is shown below:

2.13.1 Probability Distribution

Outcome of X	p(x)
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

2.13.2 Calculation

$E(X)$ can be calculated using the formula as shown below:

$$E(X) = 1(1/6) + 2(1/6) + 3(1/6) + 4(1/6) + 5(1/6) + 6(1/6)$$

The result leads to $E(X) = 3.5$

2.14 Exercise

Consider the food assistance example. What is the mean number of assistance programs used based on the PMF provided?

2.14.1 PMF

	Number.of.Programs	Frequency	Relative.Frequency
1	1	62	0.20875421
2	2	47	0.15824916
3	3	39	0.13131313
4	4	39	0.13131313
5	5	58	0.19528620
6	6	37	0.12457912

```
7                7          4          0.01346801
8                8         11          0.03703704
```

2.14.2 Answer

$E(X) = \sum_i x_i p(x_i)$ translates to:

E(X) = 1(0.2087)+2(0.1582) + 3(0.1313) + 4(0.1313) + 5(0.1953) + 6(0.1246) + 7(0.0135) + 8(0.037).

This results to $E(X) = 3.559$

2.15 Expectation Value of Functions

Expectation values can also be calculated for functions.

> **Tip**
>
> The expected value of a function $g(X)$, denoted by $E[g(X)]$, can be calculated by taking the sum of the products of each function evaluated at the outcome $\{g(x_1), g(x_2), ...,\}$ and their corresponding probabilities $\{p(x_1), p(x_2), ...\}$. In mathematical terms,
>
> $$E[g(X)] = \sum_i g(x_i)p(x_i)$$

2.16 Variance and Standard Deviation

The variance of the discrete probability distribution can be calculated using expectation values as well. The variance σ^2 can be expressed as $E[(X - \mu)^2]$.

$$\sigma^2 = E[(X - \mu)^2] = \sum_i (x_i - \mu)^2 p(x_i)$$

The standard deviation can be calculated by taking taking the square root of the variance.

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

2.17 Example

What is the variance and standard deviation of the probability distribution for rolling a fair six-sided die?

2.17.1 Notes

In calculating the variance, it would be easier to use R. First, we would need to create a data frame for the probability distribution:

```
df <- data.frame(Outcome = 1:6,
                 Prob = rep(1/6,6))

df
```

```
  Outcome      Prob
1       1 0.1666667
2       2 0.1666667
3       3 0.1666667
4       4 0.1666667
5       5 0.1666667
6       6 0.1666667
```

2.17.2 Answer

```
mu <- sum(df$Outcome*df$Prob)
mu
```

```
[1] 3.5
```

```
sigma_2 <- sum((df$Outcome-mu)^2*df$Prob)
sigma_2
```

```
[1] 2.916667
```

The variance is 2.9166667

3 Binomial Distribution

3.1 Bernoulli Trials

A random process, also referred to as a trial, can result in only one of two mutually exclusive outcomes is called a **Bernoulli trial**.

> **Note**
>
> Examples of Bernoulli trials include dichotomous random variables such as mortality (dead/alive), coin tosses (heads/tails), and infections (infected/not infected).

3.2 Bernoulli Process

> **Important**
>
> A sequence of Bernoulli trials forms a *Bernoulli process*. Bernoulli processes have the following characteristics:
>
> - The experiment consists of exactly n identical trials
> - Each trial can have only one of two outcomes (success/failure)
> - The probability of success, p is constant for every trial. The probability of failure, q, is equal to $1 - p$.
> - The trials are independent
> - The random variable of interest is the number of successes X out of the n trials.

> **Example**
>
> Imagine flipping a coin n times. The random variable X can be the number of heads observed after n flips.
>
> - Success: Flipping a head, Failure: Flipping a tail
> - Probability of success, p is 0.50 for a fair coin.

3.3 Binomial distribution

The random variable X described in the Bernoulli process follows the Binomial distribution.

Binomial Distribution

The probability mass function (PMF) of a binomial random variable X with parameters n and p is:

$$p(x) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}$$

Tip

If a random variable X follows the binomial distribution with n trials and success probability p, we can use the following notation: $X \sim BIN(n, p)$

Note

$n!$ is the **factorial** of a non-negative integer n. The factorial is the product of all the positive integers less than or equal to n.

$$n! = n(n-1)(n-2)...(2)(1)$$

With the special case $0! = 1$

3.4 Dissecting the Binomial Distribution: Independent Events

Suppose we toss a coin three times, each toss independent of each other. Assuming the probability of getting heads is p, the probability of flipping two heads out of three tosses can be calculated using the property of independent events.

$$P(2heads) = pp(1-p) = p^2(1-p)^1 = p^2(1-p)^{(3-2)}$$

3.5 Dissecting the Binomial Distribution: Counting Sample Points

However, the probability measured using independent events does not account for the multiple configurations of flipping two heads out of three tosses.

3.5.1 Results

> **Warning**
>
> Flipping two heads out of three tosses could lead to the following results: HTH, HHT, THH.

3.5.2 Notes

Note that the order in flipping the two heads does not matter; what matters is counting the number of ways that the two heads can occur after three tosses. These results are disjoint with the same probability $p^2(1-p)$. The probability that any of these results occur can be calculated using the addition rule:

$$P(2heads) = P(HTH) + P(HHT) + P(THH) = 3p^2(1-p)$$

3.6 Combination

How do we account for the number of sample points when we consider 100 tosses instead of 3?

> **Combination**
>
> The unordered selection of x successes out of n trias is called a combination, also denoted by $C(n, x)$.
>
> $$C(n, x) = \frac{n!}{x!(n-x)!}$$

> **Tip**
>
> Using combinations for the three toss example, $n = 3$, $x = 2$.
>
> $$C(3, 2) = \frac{3!}{2!(3-2)!} = 3$$

3.7 Binomial Distribution in `R`

`R` is handy for handling binomial random variables.

3.7.1 Combination

`choose(n,x)` can calculate the number of combinations of x successes in n trials.

3.7.2 PMF

`dbinom(x,n,p)` can calculate the PMF of the binomial distribution for an outcome of x successes out of n trials with success probability p.

3.7.3 CDF

`pbinom(x,n,p)` can calculate the CDF of the binomial distribution for an outcome of x successes out of n trials with success probability p.

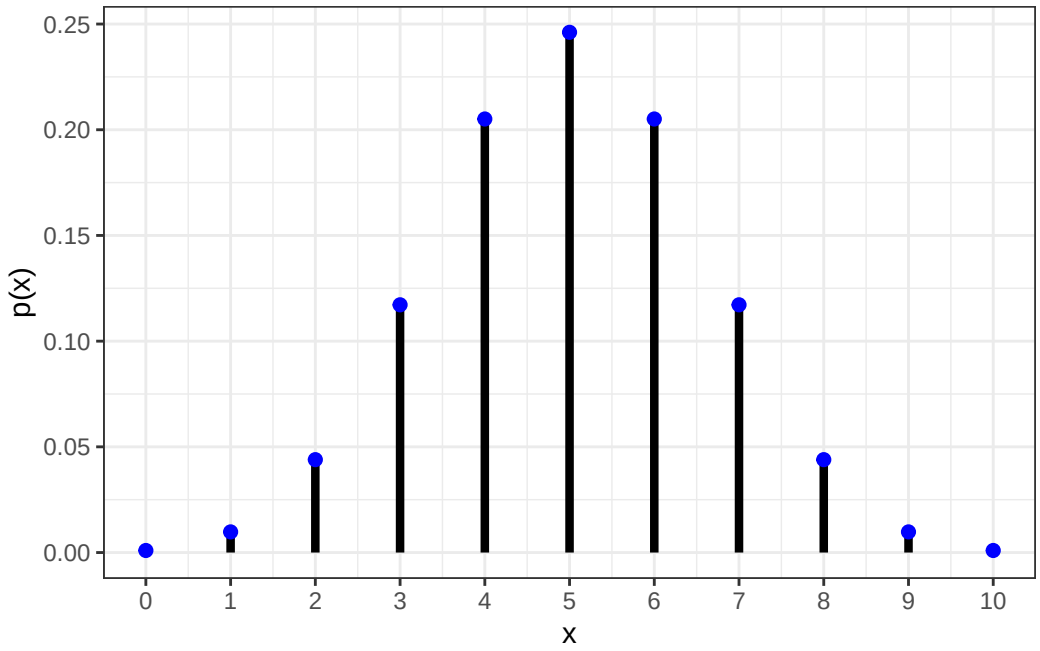
3.7.4 Random Binomial Numbers

`rbinom(k,n,p)` generates `k` numbers for a binomial distribution with number of trials n and success probability p.

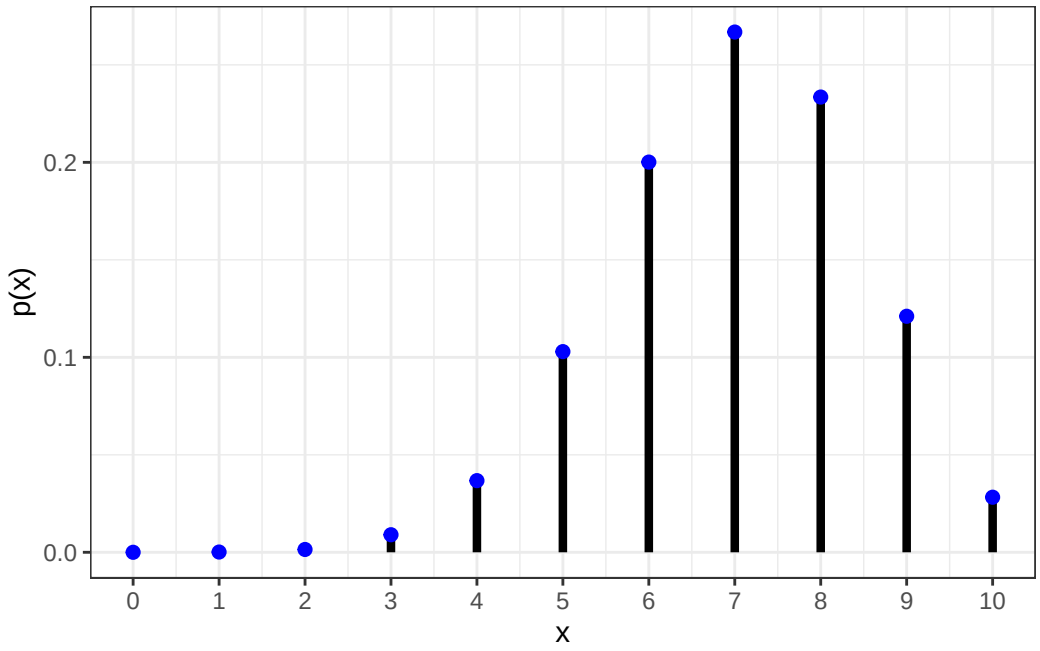
3.8 Binomial Distribution: Visualization

3.8.1 BIN(10,0.5)

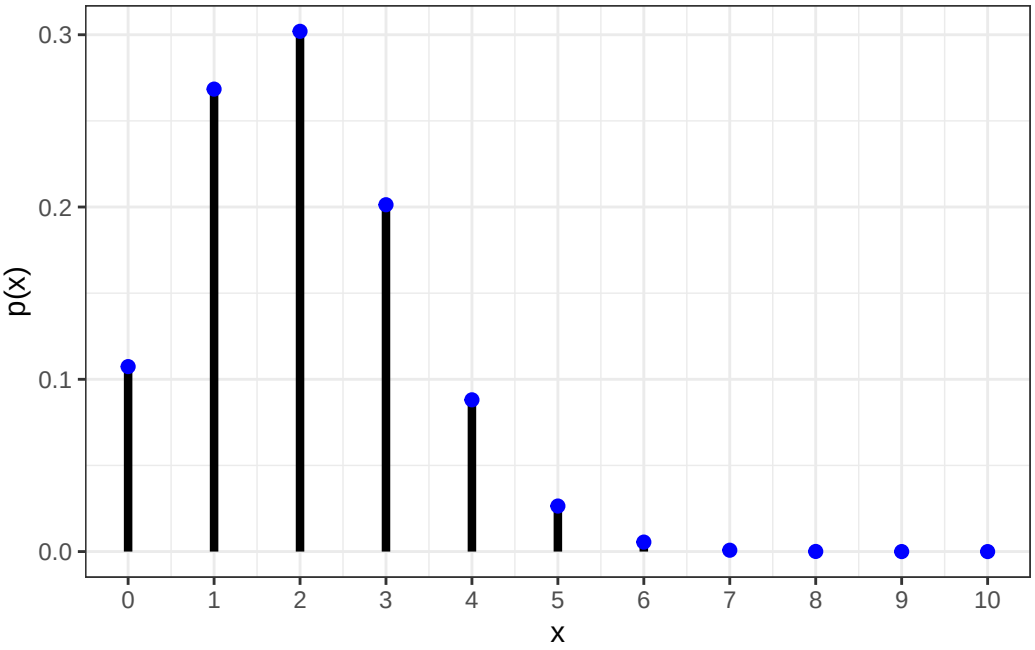
```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```


0.25
0.20
0.15
0.10
0.05
0.00
p(x)
0
1
2
3
4
5
6
7
8
9
10
x

3.8.2 BIN(10,0.7)


0.2
0.1
0.0
p(x)
0
1
2
3
4
5
6
7
8
9
10
x

3.8.3 BIN(10,0.2)



3.9 Example

Suppose we toss a fair coin (p=0.5) three times, each toss independent of each other.

3.9.1 Questions

- What is the probability of flipping heads twice?
  - Calculate manually using the equation for $p(x)$.
  - Calculate using R.
- What is the probability of flipping at most two heads? Calculate using `R`.
- Simulate this binomial process 10 times using `rbinom()`

3.9.2 PMF (Math)

$p(x) = C(3,2)(0.5)^2(1-0.5) = 3(0.5)^2(1-0.5) = 0.375$

3.9.3 PMF (R)

```
dbinom(2,3,0.5)
```

```
[1] 0.375
```

3.9.4 $P(X \leq 2)$

$P(X \leq 2)$ corresponds to the value of the CDF at $x = 2$ ($F(2)$).

```
pbinom(2,3,0.5)
```

```
[1] 0.875
```

```
dbinom(0,3,0.5) + dbinom(1,3,0.5) + dbinom(2,3,0.5)
```

```
[1] 0.875
```

You can also use the complementary principle. The complement of $X \leq 2$ is $X = 3$.

```
1- dbinom(3,3,0.5)
```

```
[1] 0.875
```

3.9.5 Simulation

```
Nsim<- 10 # Simulating 10 times
set.seed(12)                                                    ①
sims <- rbinom(Nsim,3,0.5)
sims
```

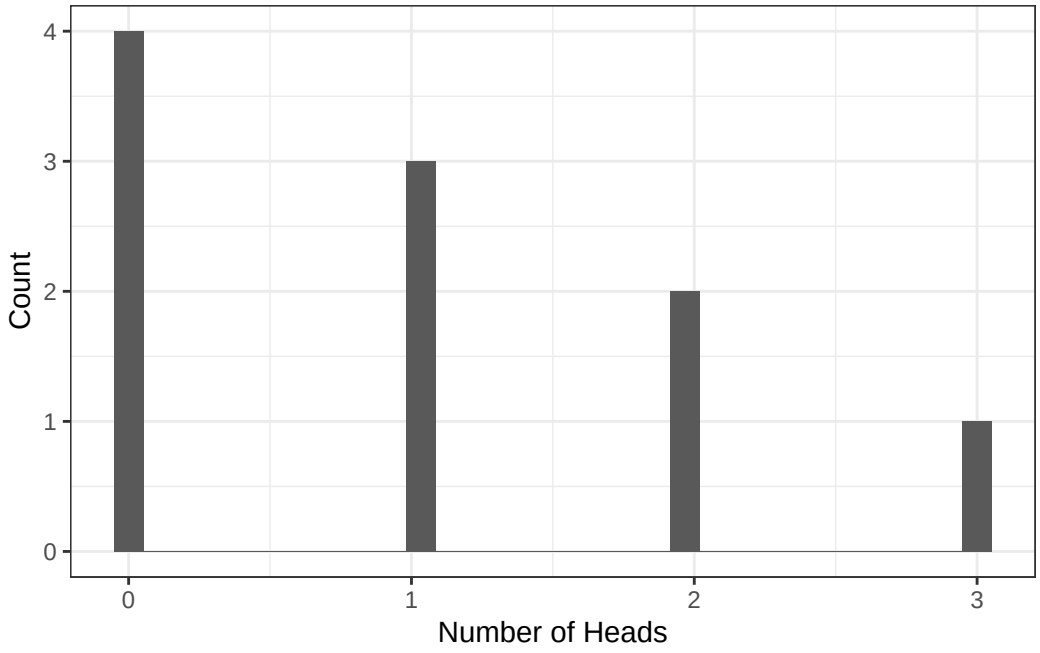
① Setting a seed enables us to draw the same numbers from the binomial distribution.

```
 [1] 0 2 3 1 1 0 1 2 0 0
```

We can create a bar plot for the simulations.

```r
df <- data.frame(simulations=sims)
ggplot(df, aes(x=simulations)) +
  geom_histogram() +
  theme_bw() +
  labs(x="Number of Heads", y="Count")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



3.10 Exercise

In 2021, 8.9% of US adults were diagnosed with diabetes (Types I and II) CDC.

3.10.1 Questions

What is the probability that out of a sample of 250 adults in 2021,

- Exactly 45 have diabetes
- Between 30 and 60, inclusive, have diabetes
- Simulate the sampling process 100 times and create a histogram of the count of participants diagnosed with diabetes out of 250 adults.

3.10.2 Answer

- Exactly 45 have diabetes

```
dbinom(45,250,0.089)
```

```
[1] 2.638007e-06
```

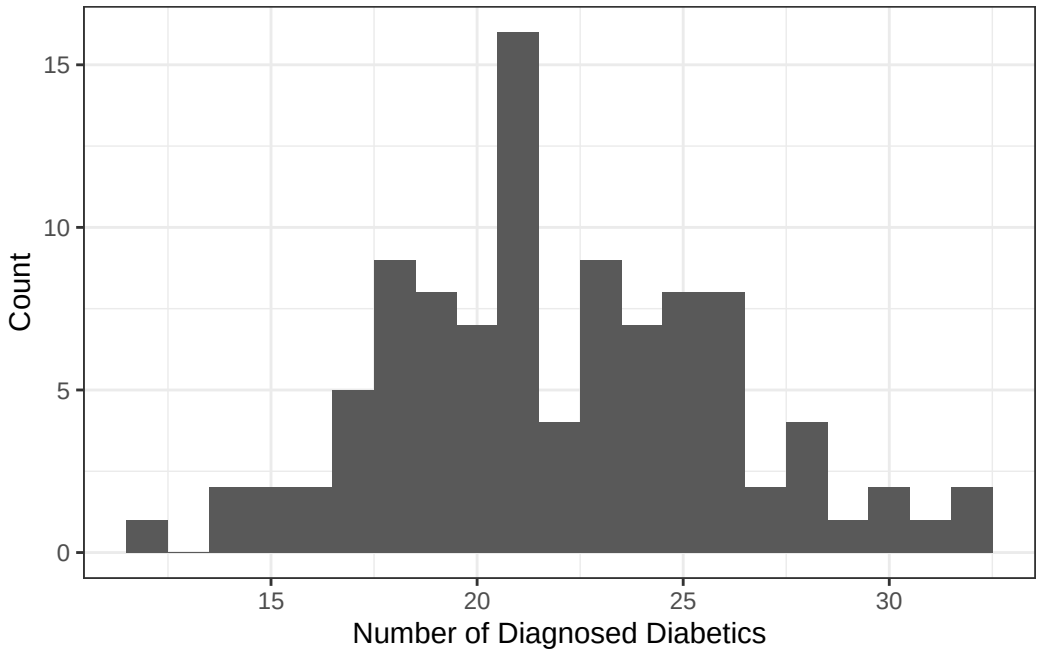
- Between 30 and 60, inclusive, have diabetes: $30 \leq x \leq 60$

```
pbinom(60,250,0.089) - pbinom(29,250,0.089)
```

```
[1] 0.05825713
```

3.10.3 Simulation

```
Nsim<- 100 # Simulating 10 times
set.seed(12)                                                    ①
sims <- rbinom(Nsim,250,0.089)
df <- data.frame(simulations=sims)
ggplot(df, aes(x=simulations)) +
  geom_histogram(binwidth=1) +
  theme_bw() +
  labs(x="Number of Diagnosed Diabetics", y="Count")
```



3.11 Expected Value and Variance

Expected Value

The expected value/mean of the binomial distribution with n trials and success probability p is $E(X) = np$.

Variance

The variance of the binomial distribution with n trials and success probability p is $V(X) = np(1-p)$.

3.12 Example

In the Philippines, the estimated percentage of individuals who received booster shots for COVID-19 is 19.7%.

3.12.1 Question

For a sample of 2,500 Philippine residents, what is the expected value of Philippine residents who received a booster? What is the variance of the associated binomial distribution?

3.12.2 Answer

Expected Value: $E(X) = 2500(0.197) = 492.5$

Variance: $V(X) = 2500(0.197)(1 - 0.197) = 395.4775$

3.13 Exercise

In 2021, 8.9% of US adults were diagnosed with diabetes (Types I and II) CDC.

3.13.1 Questions

For a sample of 250 adults in 2021, what is the expected number of diagnosed diabetics? What is the variance of the distribution?

3.13.2 Answers

Expected Value: $E(X) = 250(0.089) = 22.25$

Variance: $V(X) = 250(0.089)(1 - 0.089) = 20.26975$

3.14 Poisson Distribution

The Poisson distribution is often used to model the count of events occurring in an interval of time or space. These events are assumed to have a very low probability of occurrence in a small interval.

> **Note**
>
> Examples include number of cars stopped at an intersection for an hour, radioactive particles that decay in a given period of time, and number of hospitalizations in a day.

3.15 Poisson Process

The Poisson distribution results from a set of assumptions about an underlying process called the **Poisson process**.

> **!** Important
>
> The Poisson process has the following characteristics
>
> - The occurrence of an event in an interval of space or time has no effect on the probability of a second occurrence of the event in the same or other interval.
> - An infinite number of occurrences of the event must be possible in the interval
> - The probability of a single occurrence of the event in a given interval is proportional to the length of the interval.
> - In any infinitesimally small portion of the interval, the probability of more than one occurrence of the event is negligible.

3.16 Poisson Distribution

The PMF for the Poisson distribution is defined by the parameter λ.

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

> **i** Note
>
> e is Euler's number, λ is the Poisson rate parameter describing the average number of events per unit interval.λ must be positive.
> The notation for a Poisson distribution with a Poisson rate parameter λ is $X \sim POIS(\lambda)$.

> **!** Important
>
> The Poisson distribution can approximate a binomial distribution with large n and small p.

3.17 Poisson Distribution: `R`

`R` is also capable of calculating the PMF, CDF, and simulating data for a Poisson process.

3.17.1 Exponential

The `exp(lambda)` can calculate the $e^{-\lambda}$ term in the Poisson PMF.

3.17.2 PMF

`dpois(x,lambda)` calculates the value of the PMF for $X = x$, i.e. $P(X = x) = p(x)$.

3.17.3 CDF

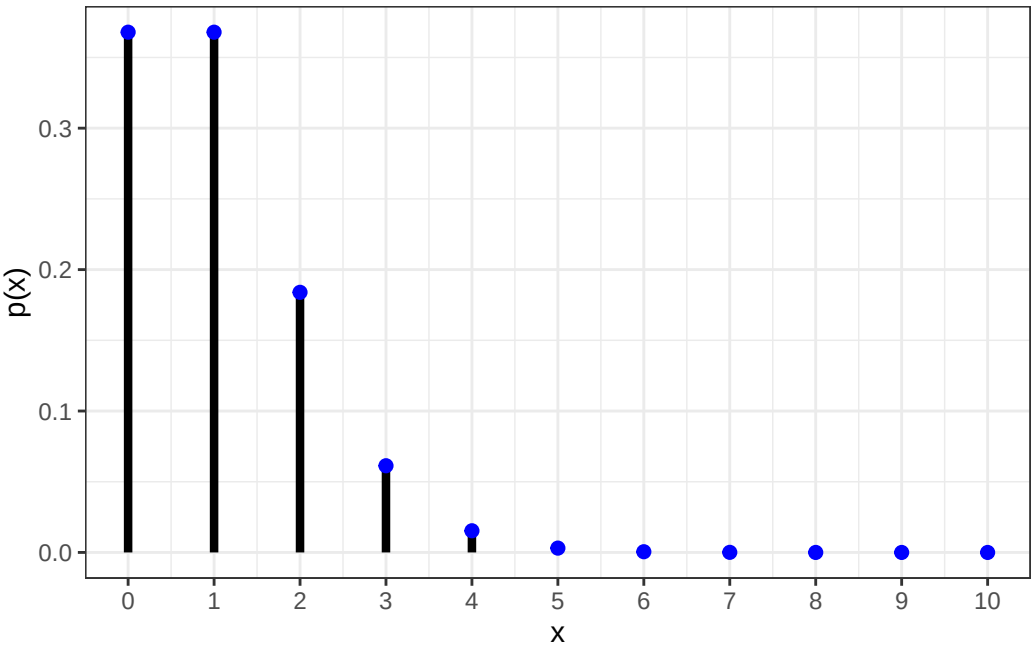
`ppois(x,lambda)` calculates the value of the CDF for $X = x$, i.e. $P(X \leq x) = F(x)$.

3.17.4 Random

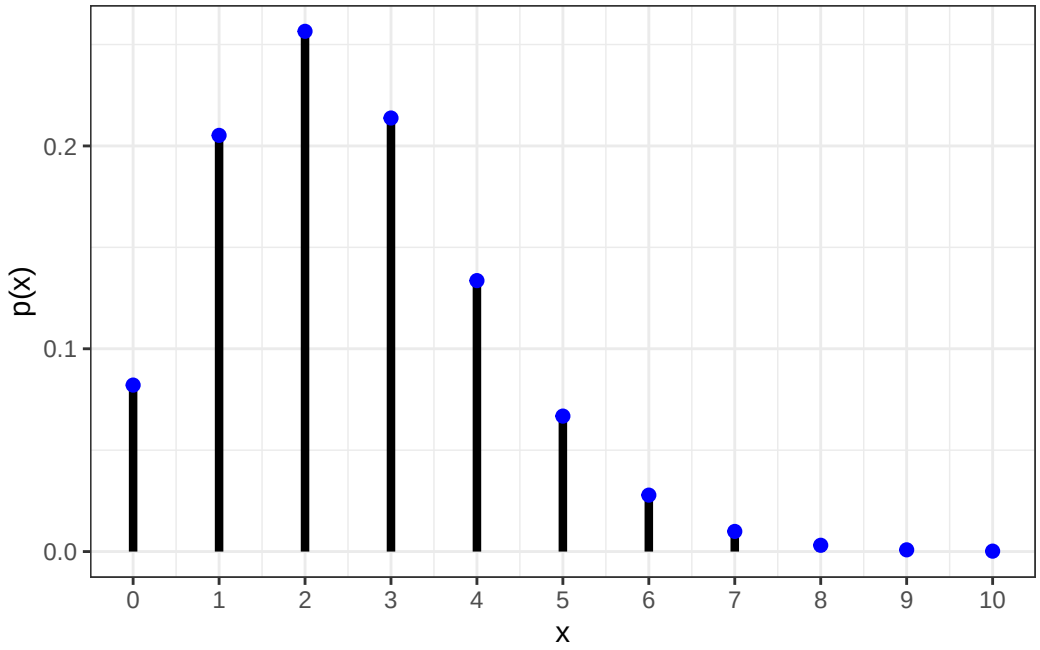
`rpois(k,lambda)` draws `k` values from a random variable $X \sim POIS(\lambda)$.

3.18 Poisson Distribution: Visualization

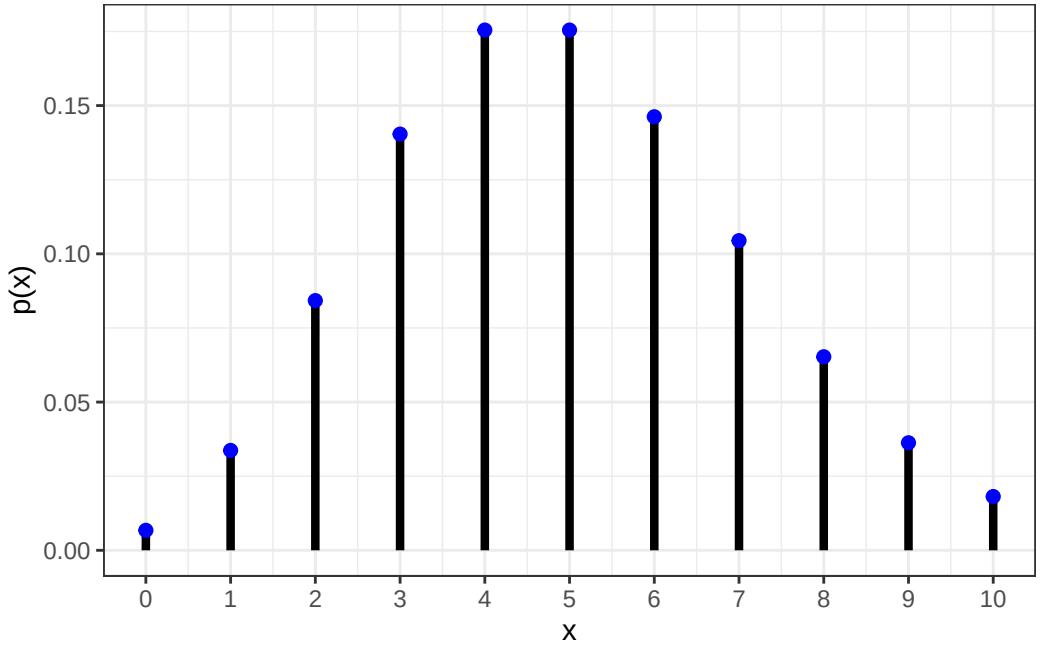
3.18.1 POIS(1)



3.18.2 POIS(2.5)



3.18.3 POIS(5)



3.19 Example

On a weekday, the expected number of people in the waiting room of a small clinic is 14.

3.19.1 Questions

- What is the probability that there will be exactly 20 people in the waiting room on a weekday?
- What is the probability that there will be at most 10 people in the waiting room on a weekday?
- Simulate 150 weekdays using `rpois()`. Create a histogram for the number of people in the waiting room from the 150 simulations.

3.19.2 Answers

- Exactly 20 people

```
dpois(20,14)
```

```
[1] 0.02859653
```

- At most 10 people

```
ppois(10,14)
```

```
[1] 0.1756812
```

3.19.3 Simulations

```
Nsim<- 150 # Simulating 10 times
set.seed(12345)                                                    ①
sims <- rpois(Nsim,14)
df <- data.frame(simulations=sims)
ggplot(df, aes(x=simulations)) +
  geom_histogram(binwidth=1) +
  theme_bw() +
  labs(x="Number of People in the Waiting Room", y="Count")
```

3.20 Exercise

The hourly number of pedestrians waiting by a bus stop is estimated to be 7.

3.20.1 Questions

- What is the probability of observing exactly 7 pedestrians at the bus stop?
- what is the probability of observing at least 10 pedestrians at the bus stop?
- Draw 1000 values from the associated Poisson distribution using `rpois`. Create a histogram for the randomly drawn values.

3.20.2 Answers

- Exactly 7 pedestrians

```
dpois(7,7)
```

```
[1] 0.1490028
```

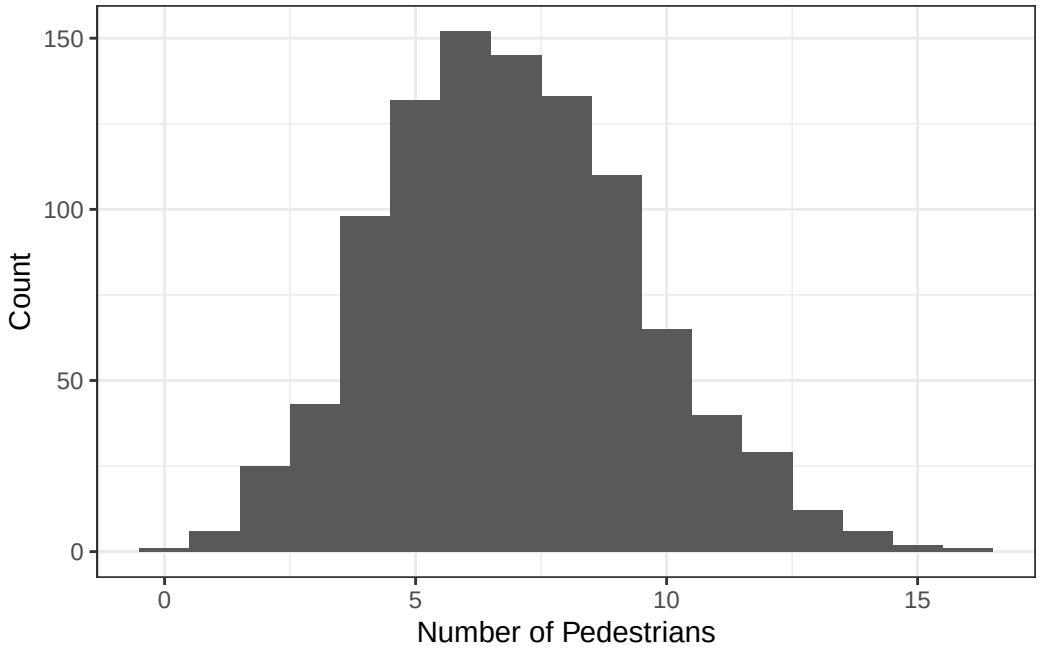
- At least 10 pedestrians: Complement of at most 9 pedestrians! Hence, the probability is equal to $1 - P(X \leq 9) = 1 - F(9)$

```r
1-ppois(9,7)
```

```
[1] 0.1695041
```

3.20.3 Simulations/Random Draws

```r
Nsim<- 1000 # Simulating 10 times
set.seed(12)                                                    ①
sims <- rpois(Nsim,7)
df <- data.frame(simulations=sims)
ggplot(df, aes(x=simulations)) +
  geom_histogram(binwidth=1) +
  theme_bw() +
  labs(x="Number of Pedestrians", y="Count")
```



3.21 Expected Value and Variance

The expected value and variance of the Poisson distribution are equal and is given by the rate parameter λ, i.e. $E(X) = \lambda, V(X) = \lambda$.

! Important

If we are interested in the expected value for t units of time, then the expected value is given by $E(X) = \lambda t$.

! Important

The equality between expected value and variance can be restrictive because data variability cannot be controlled such that this equality can hold.

3.22 Example

The hourly number of pedestrians waiting by a bus stop is estimated to be 7.

3.22.1 Questions

- What is the expected value of the number of pedestrians waiting by the bus stop after one hour?
- What is the variance of the associated Poisson distribution?
- What is the expected value of the number of pedestrians waiting by the bus stop after 30 minutes?

3.22.2 Answers

- After one hour, $E(X) = \lambda t = (7)(1) = 7$
- Variance: $V(X) = \lambda = 7$
- After 30 minutes, $E(X) = \lambda t = (7)(0.5) = 3.5$

3.23 Other Discrete Distributions

i Hypergeometric Distribution

Consider a population size N and a sample size n. Each element in the population can be categorized into a "success" or failure. If there are r successes in the population, the probability that x successes will be drawn in a sample of size n is given by the **hypergeometric distribution**.
This distribution is used in exact tests such as the Fisher's Exact Test.

Negative Binomial Distribution

The negative binomial distribution is used to model counts for overdispersed data and other biological processes. Overdispersion occurs when the data set exhibits more variability than what a statistical model expects. Recall that the Poisson distribution is restricted by the property that the mean is equal to the variance.

4 Probability Distributions of Continuous Probability Distributions

4.1 Continuous Random Variables

Continuous Random Variable

A random variable is continuous if it can only take on values in one or more intervals of the real line.

Tip

Time, distance, weight, and volume are examples of continuous variables.

4.2 Continuous Probability Distributions

Note

Recall that probabilities from discrete random variables are found by summing PMF values.

For continuous random variables, summing becomes integration.

Tip

Imagine creating a histogram with an infinite amount of narrow bins. Adding these values eventually lead to a sum of the areas under the curve.

4.3 Probability Density Functions

The continuous probability distribution can be expressed as a **probability density function (PDF)**, denoted by $f_X(x)$ or $f(x)$.

> **Warning**
>
> There are important differences between probability density functions and probability mass functions (discrete). For example, for continuous variables,
> $P(X = x) = 0 \neq f(x)$.

4.4 Cumulative Distribution Functions

Recall that the cumulative distribution function (CDF) is given by $F_X(x) = P(X \leq x)$. The CDF is mathematically defined as:

$$F_X(x) = \int_{-\infty}^{x} f(x)dx$$

> **Tip**
>
> Since $P(X = x) = 0$, it also follows that
>
> $$P(a \leq x \leq b) = P(a < x \leq b) = P(a \leq x < b) = P(a < x < b)$$
>
> and that
>
> $$F_X(x) = P(X \leq x) = P(X < x)$$

4.5 Probabilities for Continuous Random Variables

Instead of $P(X = x)$, we are more interested in the probability that a continuous random variable assumes a value between a and b, denoted by $P(a < X < b)$, which can be calculated as:

$$P(a < X < b) = \int_{a}^{b} f(x)dx = F(b) - F(a)$$

> **Note**
>
> Probability density functions have the following properties
>
> - $f(x) \geq 0$ for $-\infty < x < \infty$.
> - $\int_{-\infty}^{\infty} f(x)dx = 1$ or $F(\infty) = 1$.

4.6 Uniform Distribution: PDF

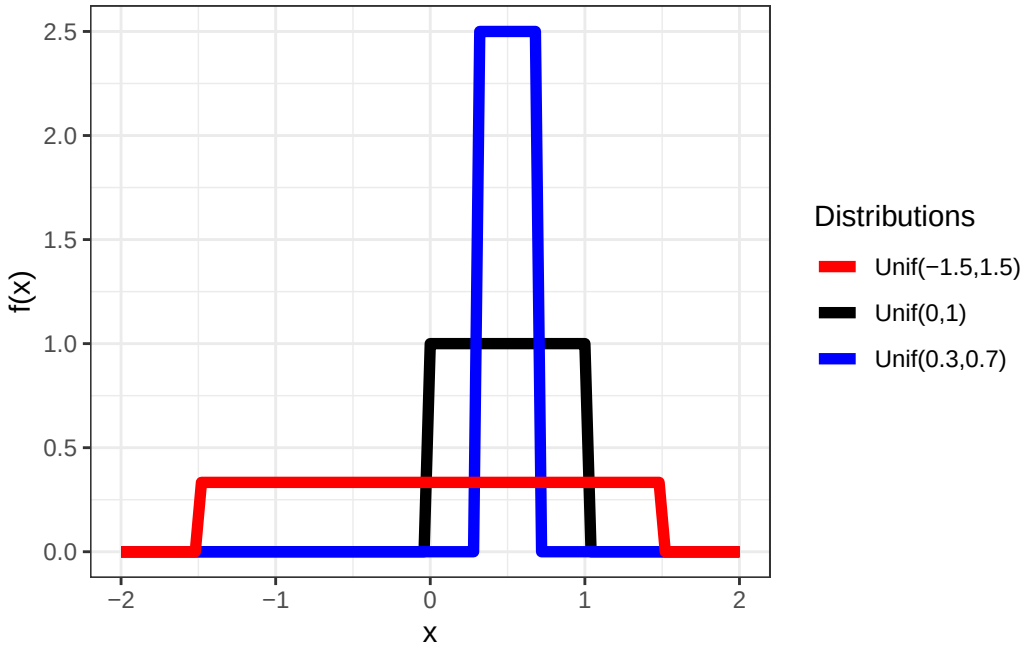
The uniform distribution has a constant PDF such that

$$f(x) = \begin{cases} \frac{1}{\theta_2 - \theta_1}, & \theta_1 \leq x \leq \theta_2; \theta_2 > \theta_1 \\ 0, & otherwise \end{cases}$$

A random variable X that has a uniform distribution on the interval $[\theta_1, \theta_2]$ is denoted as $X \sim UNIF(\theta_1, \theta_2)$.

> **Note**
>
> A uniform distribution is defined by two parameters: θ_1 and θ_2. A standard uniform distribution is a special case such that θ_1=0, θ_2=1.

4.7 Uniform Distribution: Visualization



4.8 Uniform Distribution: Probabilities

> **!** Important
>
> Lucky for you, there's no expectations of any calculus in class.

The probability $P(a < X < b)$ if $X \sim UNIF(\theta_1, \theta_2)$ can be calculated using the following equation:

$$P(a < X < b) = \frac{b-a}{\theta_2 - \theta_1}$$

where a, b are assumed to be between $\theta_1 < a \leq b < \theta_2$.

> **!** Important
>
> This result can also be demonstrated using geometry!

4.9 Uniform Distribution: Probabilities in R

In R, `punif(x,min,max)` can calculate the probability $P(X \leq x) = P(\theta_1 < X < x)$. Hence, if we want to calculate $P(a < X < b)$ when $X \sim UNIF(\theta_1, \theta_2)$,

```
P(a < X < b) = punif(b,theta_1,theta_2)-punif(a,theta_1,theta_2)
```

> **!** Important
>
> You can also draw random numbers from a uniform distribution $UNIF(\theta_1, \theta_2)$ using `runif(#sims,theta_1,theta_2)`

4.10 Example

Suppose $X \sim UNIF(0, 5)$. What is the probability that an outcome x is between 0.25 and 4?

4.10.1 Math

a = 0.25, b=4, $\theta_1 = 0, \theta_2 = 5$.

$$P(a < X < b) = \frac{b-a}{\theta_2 - \theta_1} = \frac{4 - 0.25}{5 - 0}$$

Hence, P(a < X < b) = 0.75

4.10.2 R

```
punif(4,0,5) - punif(0.25,0,5)
```

```
[1] 0.75
```

4.11 Exercise

Suppose $X \sim UNIF(0, 100)$.

4.11.1 Question

What is the probability that an outcome x is between 55 and 85?

4.11.2 Answer

$$\frac{85 - 55}{100 - 0} = 0.30$$

```
punif(85,0,100) - punif(55,0,100)
```

```
[1] 0.3
```

4.12 Uniform Distribution: Expected Value and Variance

For the uniform distribution, the mean is given by $E(X) = \frac{\theta_1 + \theta_2}{2}$. The variance is given by $E(X) = \frac{(\theta_2 + \theta_1)^2}{12}$.

4.13 Gaussian Distribution

The Gaussian distribution, also known as the normal distribution, is an important distribution in statistics.

> **Note**
>
> Many natural physical and human-related phenomena closely follow a normal distribution.

4.14 Gaussian Distribution: PDF

The probability density function of the Gaussian distribution can be expressed as:

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

> **Note**
>
> A uniform distribution is defined by two parameters: the mean μ and σ^2. A random variable X that has a Gaussian distribution with mean μ and variance σ^2 is denoted as $X \sim N(\mu, \sigma^2)$.

> **Important**
>
> A standard normal distribution is a special case such that μ=0, σ^2=1. Variables that follow a standard normal distribution are often denoted as Z.

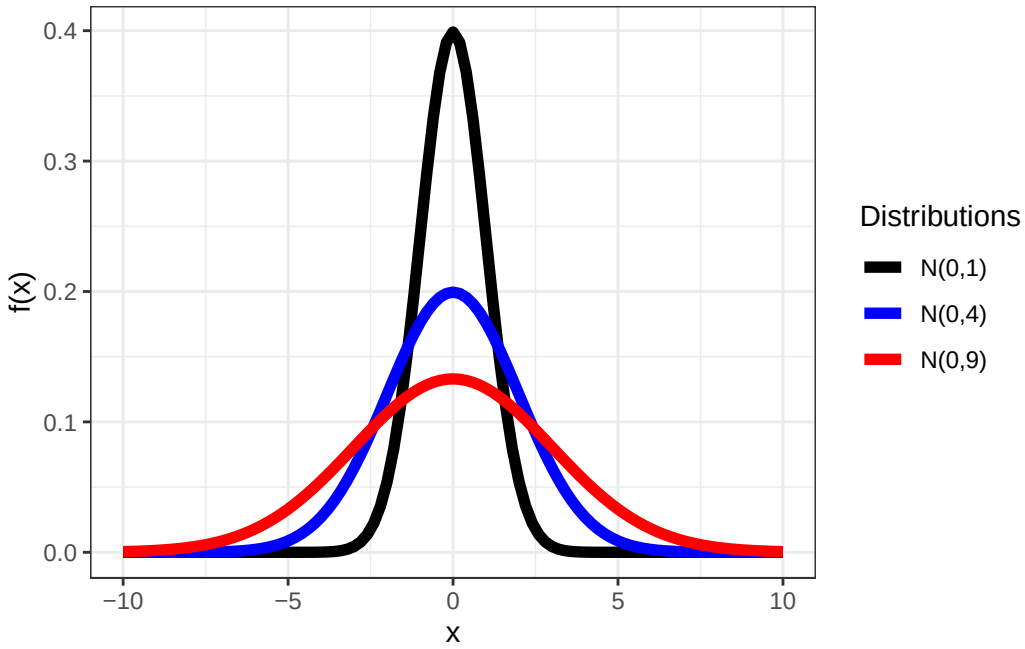
4.15 Gaussian Distribution: Characteristics

The Gaussian distribution has the following characteristics:

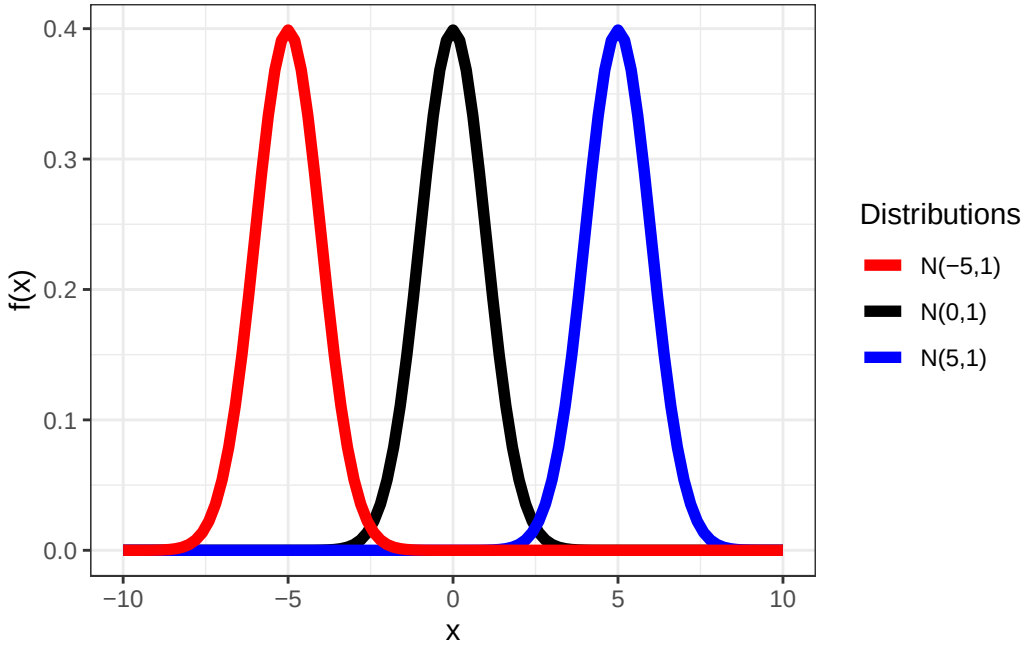
- The distribution is symmetric about the mean μ.
- The mean, median, and mode are all equal.

4.16 Gaussian Distribution: Visualization

4.16.1 Same Means, Changing Variance



4.16.2 Same Variance, Changing Means



4.17 Gaussian Distribution: `R`

> **!** Important
>
> Calculating probabilities mathematically would be impossible without calculus. However, `R` makes life easy for us.

4.17.1 PDF

You can calculate the PDF value for a normal distribution, f(x), using `dnorm(x,mean,sd)`. Note that the function uses standard deviation, not variance.

4.17.2 CDF

You can calculate the CDF value for a normal distribution, F(x), using `pnorm(x,mean,sd)`. Note that the function uses standard deviation, not variance.

4.17.3 Simulating/Generating Random Variables

You can simulate/generate random Gaussian variables using `rnorm(#sims,mean,sd)`. Note that the function uses standard deviation, not variance.

4.17.4 CDF to X/Z

If we know the value of $F(x)$, we can determine what x is using `qnorm(F(x),mean, sd)`. If the mean and standard deviation/variance is unknown, we can get the corresponding value of the standard normal distribution Z by specifying `mean=0` and `sd=1`.

4.18 Example

For a random variable Z $N(0,1)$ following the standard normal distribution

4.18.1 Question

Calculate the following probabilities using `R`

- $P(Z \leq 1)$
- $P(Z \geq 2.5)$
- $P(-2 < Z < 2)$

4.18.2 Answer

- $P(Z \leq 1) = F(1)$

```
pnorm(1,0,1)
```

```
[1] 0.8413447
```

- $P(Z \geq 2.5) = 1 - P(Z < 2.5) = 1 - F(2.5)$

```
1-pnorm(2.5,0,1)
```

```
[1] 0.006209665
```

```
pnorm(2.5,0,1,lower.tail=FALSE)
```

```
[1] 0.006209665
```

- $P(-2 < Z < 2) = F(2) - F(-2)$

```
pnorm(2,0,1) - pnorm(-2,0,1)
```

```
[1] 0.9544997
```

4.19 Exercise

Consider $X \sim N(3, 9)$.

4.19.1 Question

- $P(X \leq 6)$
- $P(X \geq 3)$
- $P(-3 < X < 9)$

4.19.2 Answer

Note that $\sigma^2 = 9$, so sd=3.

- $P(X \leq 6)$

```
pnorm(6,3,3)
```

```
[1] 0.8413447
```

- $P(X \geq 10.5)$

```
1- pnorm(10.5,3,3)
```

```
[1] 0.006209665
```

```
pnorm(10.5,3,3,lower.tail=F)
```

```
[1] 0.006209665
```

- $P(-3 < X < 9)$

```r
pnorm(9,3,3) - pnorm(-3,3,3)
```

```
[1] 0.9544997
```

> **!** Important
>
> Note that the answers are the same!

4.20 Other Continuous Distributions

> **i** Weibull Distribution
>
> The Weibull distribution is used to model random variables that are constrained to be positive, such as lifetimes, impurities, etc. This is common for epidemic modeling and survival analysis.

> **i** Statistical tests
>
> Continuous distributions that you will learn about in future chapters are the t distribution, F distribution, and the χ^2 (chi-squared) distribution.

5 Application of the Gaussian/Normal Distribution

5.1 Standardization

A common transformation employed with Gaussian distributed variables is standardization. Standardization involves rescaling the variable based on the mean, μ, and variance σ^2, of the distribution. An observation X can be standardized into a standardized variable Z using the following equation:

$$Z = \frac{X - \mu}{\sigma}$$

> **!** Important
>
> If $X \sim (\mu, \sigma^2)$, then $Z \sim N(0, 1)$. The CDF of the normal distribution is referred to by $\Phi(z) = P(Z \leq z)$.

5.2 Standard Normal Distribution: Implications

> **Note**
>
> The random variable Z calculates how many standard deviations away the observation is from the mean. The transformation also implies that
>
> $$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)$$
>
> and
>
> $$F(x) = \Phi\left(z = \frac{x-\mu}{\sigma}\right)$$

5.3 Example

The results of a biostatistics aptitude test was found to be normally distributed with a mean score of 78.5 and variance of 120.

- What is the probability that somebody scores above 92 points?
- Generate 1000 draws from the corresponding normal distribution using `rnorm()` and create a histogram from the generated values.

5.3.1 No standardization

```
pnorm(92,78.5,sqrt(120),lower.tail=F)
```

```
[1] 0.1089044
```

5.3.2 Standardization

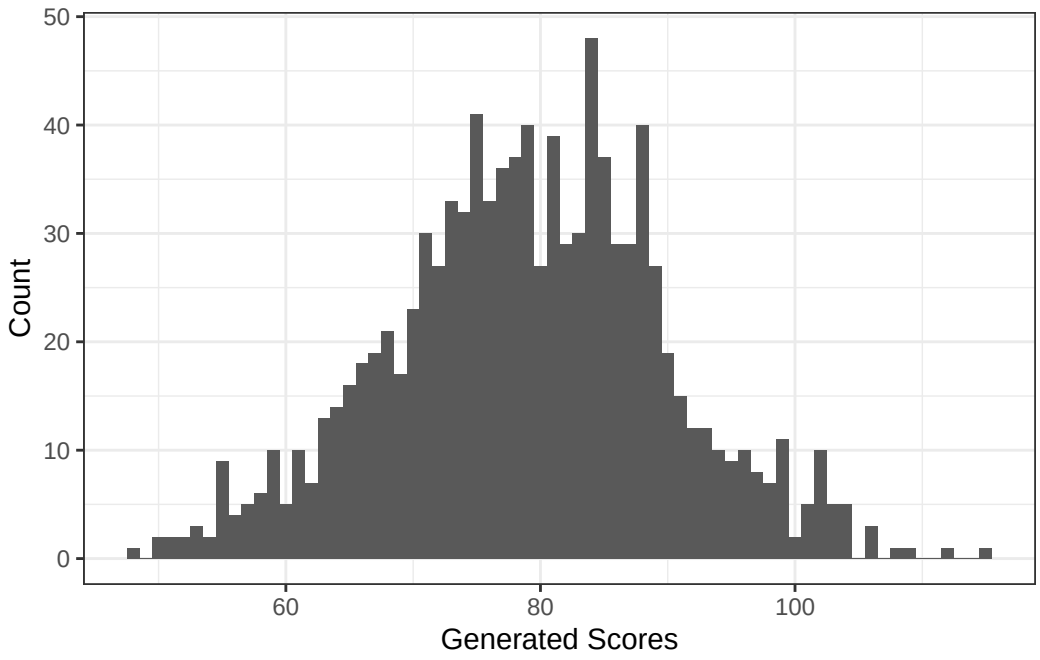
$Z = \frac{X-\mu}{\sigma} = 1.2323758$

```
pnorm((92-78.5)/sqrt(120),0,1,lower.tail=F)
```

```
[1] 0.1089044
```

5.3.3 Simulation/Generation

```
set.seed(12345)
df <- data.frame(x=rnorm(1000,mean=78.5,sd=sqrt(120)))

ggplot(df,aes(x=x)) +
  geom_histogram(binwidth=1) +
  theme_bw() +
  labs(x="Generated Scores",
       y="Count")
```



5.4 Exercise

The waiting time for physician appointments in a hospital was normally distributed with mean equal to 23.5 minutes and standard deviation of 6.7 minutes. What is the probability that a patient will wait less than five minutes for an appointment?

5.4.1 Question

- Use standardization to calculate the standardized variable z such that $P(X < 5) = P(Z < z)$.

- Calculate the probability using the standardization method.

5.4.2 Answer

$z = \frac{5-23.5}{6.7}$ = -2.761194

```
pnorm((5-23.5)/6.7,0,1)
```

```
[1] 0.002879523
```

5.5 Implications

For every random variable $X \sim N(\mu, \sigma^2)$, the probability of an outcome to be within one standard deviation of the mean can be written as:

$$P((\mu - \sigma) < X < (\mu + \sigma)) = P(-1 < Z < 1)$$

```
pnorm(1,0,1) - pnorm(-1,0,1)
```

```
[1] 0.6826895
```

> **Note**
>
> We expect 68.3% of the measurements of the random variable X is within one standard deviation of the mean ($\pm 1\sigma$)

5.6 Implications

Similarly, we can calculate the probability within $\pm 2\sigma$ and $\pm 3\sigma$. Around 95.4% of the outcomes are $\pm 2\sigma$, while 99.7% of the outcomes are $\pm 3\sigma$. This is why it is sometimes assumed that normally distributed data should be found within three standard deviations from the mean.

5.6.1 $\pm 2\sigma$

```
pnorm(2,0,1) - pnorm(-2,0,1)
```

```
[1] 0.9544997
```

5.6.2 $\pm 3\sigma$

```
pnorm(3,0,1) - pnorm(-3,0,1)
```

```
[1] 0.9973002
```

5.7 Calculating z from $\Phi(z)$

The function `qnorm(prob)` provides the value for z such that the CDF $\Phi(z) = P(Z \leq z) = prob$.

5.8 Example

Find z such that $\Phi(z) = 0.93$.

```
qnorm(0.93)
```

```
[1] 1.475791
```

5.9 Example 2

Consider a random distribution with mean 5 and standard deviation 2.7. What is the 90th percentile of the distribution?

- The 90th percentile, p_{90}, can be expressed as $P(X \leq p_{90}) = 0.90$. p_{90} can be calculated using `qnorm`.

```
qnorm(0.90,mean=5,sd=2.7)
```

```
[1] 8.460189
```

5.10 Exercise

Suppose the total cholesterol values for a certain population are approximately normally distributed with a mean of 205.7 mg/100ml with a standard deviation of 23.5 mg/100ml.

5.10.1 Questions

- What is the probability that an individual picked at random will have
  - Above 200 mg/100ml?
  - Between 150 and 200 mg/100ml?
- What is the cholesterol level such that only 5% of the population have cholesterol levels above this amount?

5.10.2 Answer

- Above 200 mg/100ml

```
pnorm(200,mean=205.7,sd=23.5,lower.tail=F)
```

```
[1] 0.5958242
```

- Between 150 and 200 mg/100ml

```
pnorm(200,mean=205.7,sd=23.5) - pnorm(150,mean=205.7,sd=23.5)
```

```
[1] 0.3952868
```

- Only 5% of the population have cholesterol levels above this amount: p_{95}

```
qnorm(0.95,mean=205.7,sd=23.5)
```

```
[1] 244.3541
```