Lecture 5: Distributions Continued; Expectation and Variance

Today's Agenda

- Announcements / Class 4 Review (34)
  - 6:00 ~ 6:50pm
- Bio Break
  - 6:50 ~ 7:00pm
- Expectation & Variance; Bernoulli & Binomial (78)
  - 7:00 ~ 7:50pm
- Bio Break
  - 7:50 ~ 8:00pm
- More Distributions! (133)
  - 8:00 ~ 8:50pm

Probability Functions Review and Cumulative Distribution Functions

Last week we learned about two types of random variables:

- **discrete random variables** are described by **probability mass functions** (PMFs)
- **continuous random variables** are described by **probability density functions** (PDFs)

Probability functions (both PMFs and PDFs) relate possible values of a random variable to the probability of those values being observed. Specifically,

- **probability mass functions** (PMFs) relate possible values to probability. You can think of this function as a table that maps specific values to probability.
- **probability density functions** (PDFs) relate possible values to density. You can think of this function as a curve that, when integrated, returns probability.

Notation: Random Variables

Random variables are denoted by capital letters:

$$X$$

whereas specific values of a random variable are denoted by the corresponding lowercase letter. i.e.: x

This distinction between the random variable an an observation of the random variable allows notation like the following:

$$Pr[X = x]$$

which denotes probability of a random variable X being equal to a specific value x.

Notation: Probability Functions

For a discrete random variable X, $Pr[X = x]$ *is* the probability mass function

$$Pr[X = x] = f_X(x)$$

In the context of statistics, probability functions are often denoted as $f(x)$ or $f_X(x)$.

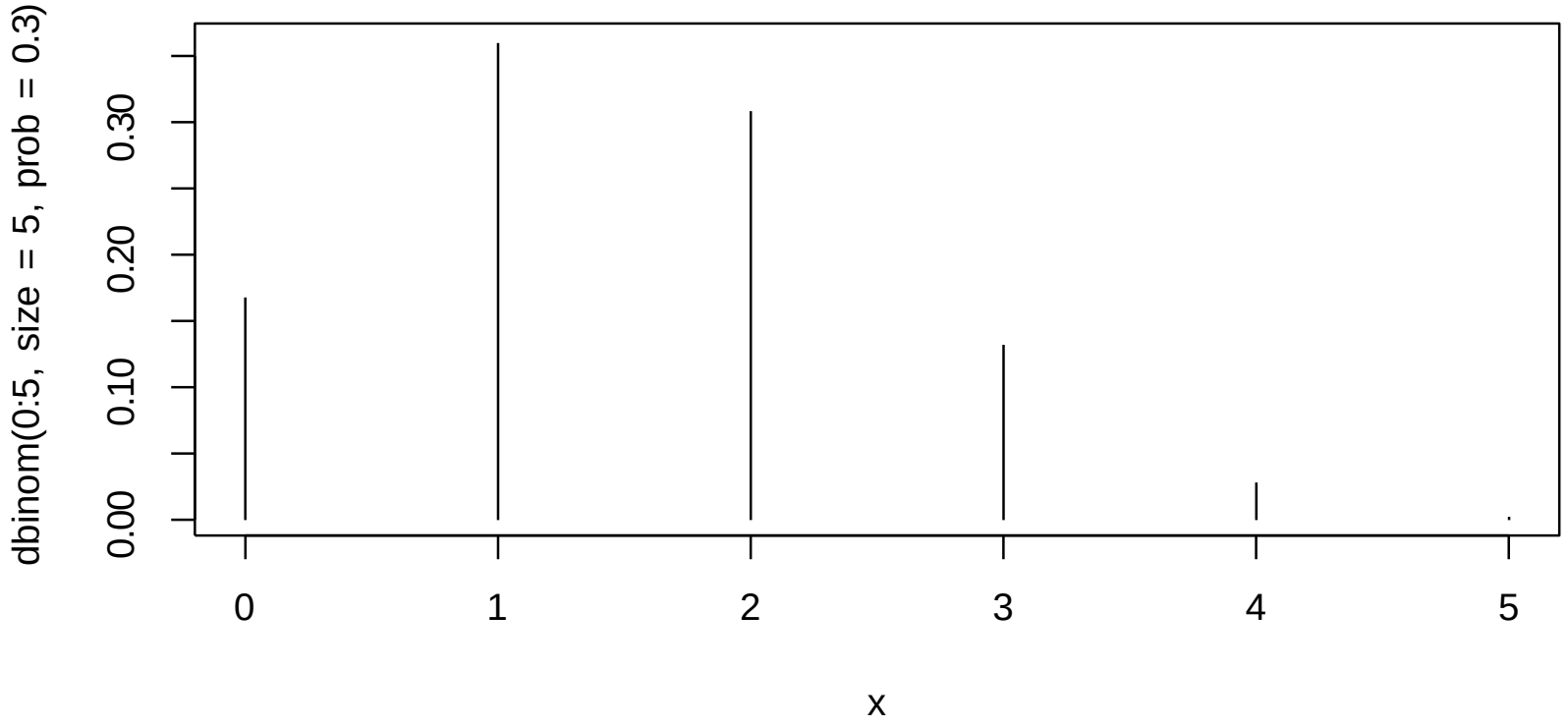
For a continuous random variable X, the probability density function is also represented as $f(x)$ or $f_X(x)$:

$$\Pr[a \leq X \leq b] = \int_a^b f_X(x)dx$$

In R, probability mass functions are named according to a common naming scheme: `d*` where * indicates the distribution basename. Examples include `dbinom`, `dgeom`...

To plot the binomial distribution, we just plot probability ($f(x)$ returned by dbinom) as a function of the sample space.

```
> x <- 0:5
> plot(x, dbinom(0:5, size = 5, prob = 0.3), type = "h")
```



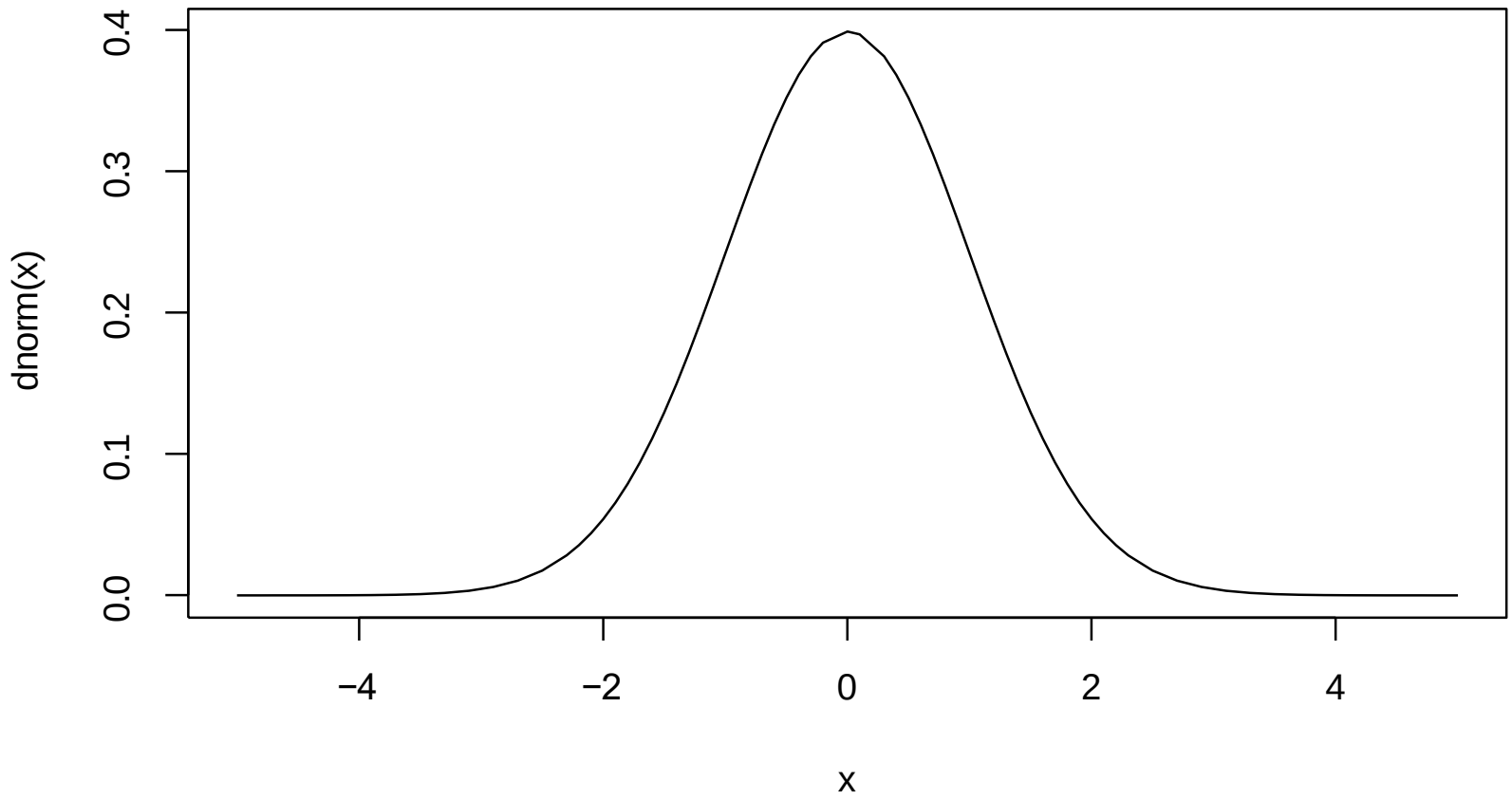
type=”h” gives us the vertical-line style plot.

In R, probability density functions are *also* named according to the same naming scheme: `d*` where * indicates the distribution basename.

Examples include `dnorm`, `dunif`...

How would we plot a probability density function in R?

```
> x <- seq(-5, 5, by = 0.1)  #pick a range of the sample space
> plot(x, dnorm(x), type = "l")  #plot x and f(x).
```



Did we plot the entire distribution?

The cumulative distribution function (continuous)

Recall that the PDF is defined as the function that satisfies

$$\Pr[a \leq X \leq b] = \int_a^b f_X(x)dx$$

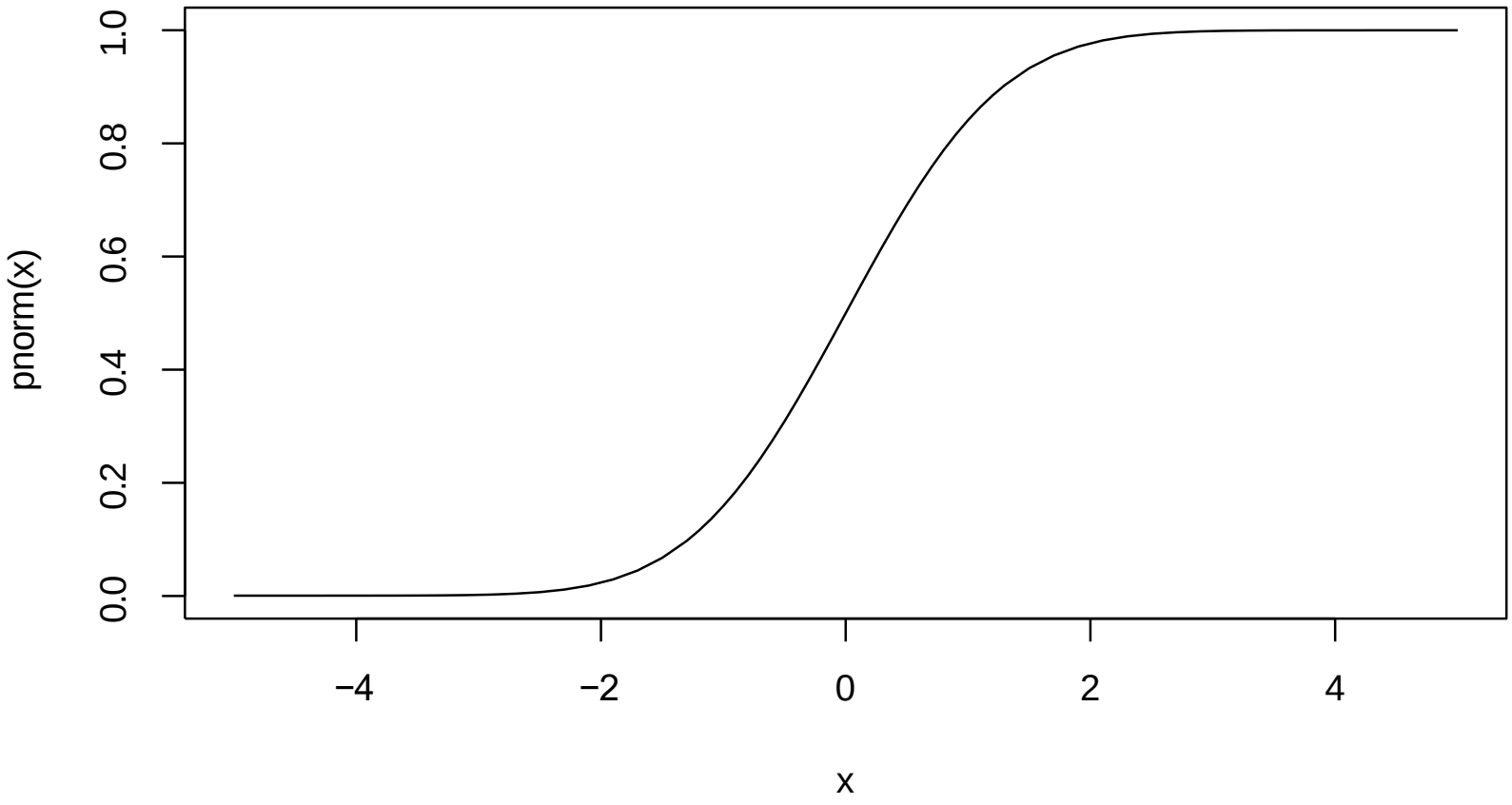
There is another way of representing probability model: The **cumulative distribution function** (CDF) returns the probability of observing a value less than or equal to x:

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du$$

In R, the cumulative distribution function for $X \sim N(0,1)$ is given by

pnorm

```
> x <- seq(-5, 5, by = 0.1)
> plot(x, pnorm(x), type = "l")
```



Recall the sea lion length data we discussed last week.

Assuming normality with a population mean of 112.5 and a population standard deviation of 5.46, how can we use the *cumulative distribution function* (CDF) to estimate the probability of observing a sea lion with length less than 110?

Exercise: Understanding distributions

A factory moves 10,000 boxes a day. Each box has a .0005 probability of getting dropped. So on a given day, some number of boxes get dropped.

What sort of distribution might describe the total number of boxes that get dropped in a day?

I'm just asking for the **name of the distribution** which describes this sort of random phenomenon.

Exercise: Understanding distributions

A factory moves 10,000 boxes a day. Each box has a .0005 probability of getting dropped. So on a given day, some number of boxes get dropped.

There are infinitely many versions of the binomial distribution. Which specific version of the binomial distribution is occuring here?

In other words, what makes our factory example a different binomial distribution from any other binomial distribution? (eg here's another different binomial distribution: the number of baskets in 10 tries with Shaq's free throw probability of 0.37)

Exercise: Understanding distributions

A factory moves 10,000 boxes a day. Each box has a .0005 probability of getting dropped. So on a given day, some number of boxes get dropped.

There are infinitely many versions of the binomial distribution. Which specific version of the binomial distribution is occuring here?

Distributions have **parameters.** Parameters distinguish different versions of the same type of distribution. So I'm asking for the parameters of this binomial distribution.

Exercise: Understanding distributions

A factory moves 10,000 boxes a day. Each box has a .0005 probability of getting dropped. So on a given day, some number of boxes get dropped.

How many boxes could be dropped on any given day? That is, list out each possible value for the number of boxes that could get dropped tomorrow.

Exercise: Understanding distributions

A factory moves 10,000 boxes a day. Each box has a .0005 probability of getting dropped. So on a given day, some number of boxes get dropped.

How many boxes could be dropped on any given day? That is, list out each possible value for the number of boxes that could get dropped tomorrow.

Distributions have **sample spaces.** The sample space is the list of all possible values that can occur in that distribution. To find the sample space you can use reasoning about what the binomial distribution describes (number of successes out of n tries).

Exercise: Understanding distributions

A factory moves 10,000 boxes a day. Each box has a .0005 probability of getting dropped. So on a given day, some number of boxes get dropped.

If I wanted to know what the probability of dropping 10 or fewer boxes in a day, how would I answer this question?

Exercise: Understanding distributions

A factory moves 10,000 boxes a day. Each box has a .0005 probability of getting dropped. So on a given day, some number of boxes get dropped.

If I wanted to know what the probability of dropping 10 or fewer boxes in a day, how would I answer this question?

Distributions have **probability functions** which relate values in the sample space (ie integers 0 to 10) to probabilities. If you're asked to estimate the probability of an event and you know the distribution, you can just use the probability function.

Exercise: Understanding distributions

A factory moves 10,000 boxes a day. Each box has a .0005 probability of getting dropped. So on a given day, some number of boxes get dropped.

If I wanted to know what the probability of dropping 10 or fewer boxes in a day, what code would I use to calculate this in R?

Exercise: Understanding distributions

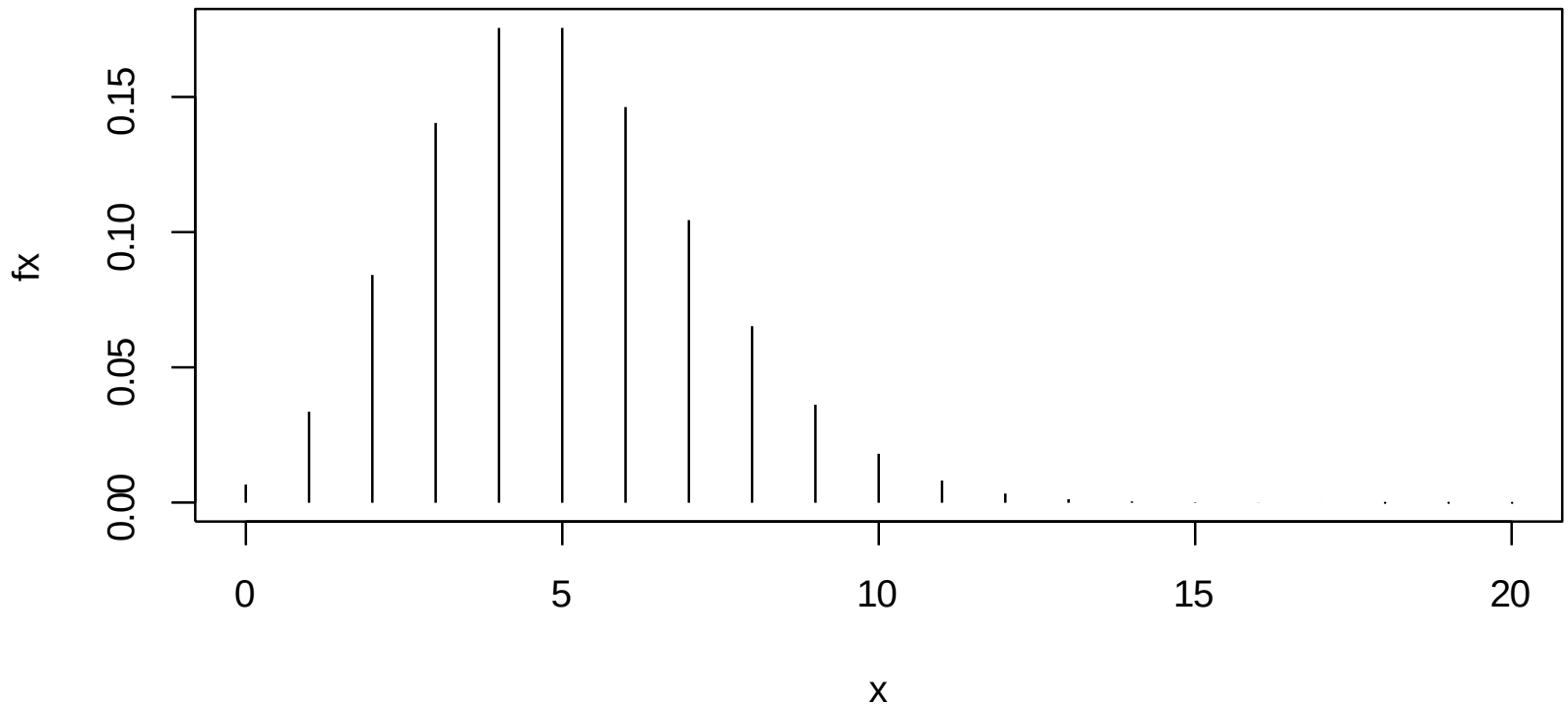
A factory moves 10,000 boxes a day. Each box has a .0005 probability of getting dropped. So on a given day, some number of boxes get dropped.

The probability mass function for the binomial distribution is dbinom. We can use this to calculate the probability of individual values in the sample space, then sum over them. `dbinom(0, size=10000, p=.0005) + dbinom(1, size=10000, p=.0005) + etc.` Or, since dbinom is vectorized: `sum(dbinom(0:10, size=10000, p=.0005))`

Plotting the PMF

Here's the probability mass function for our distribution.

```
> x <- 0:10000
> fx <- dbinom(x, size = 10000, p = 5e-04)
> plot(x, fx, type = "h", xlim = c(0, 20))
```



The cumulative distribution function (discrete)

Recall the PMF gives the probability for each value of x in the sample space.

$$f_X(x) = \Pr(X = x)$$

The **cumulative distribution function (**CDF) denoted $F(x)$ (with a capital F) returns the sum of the probabilities of all values in the sample space less than or equal to some value x:

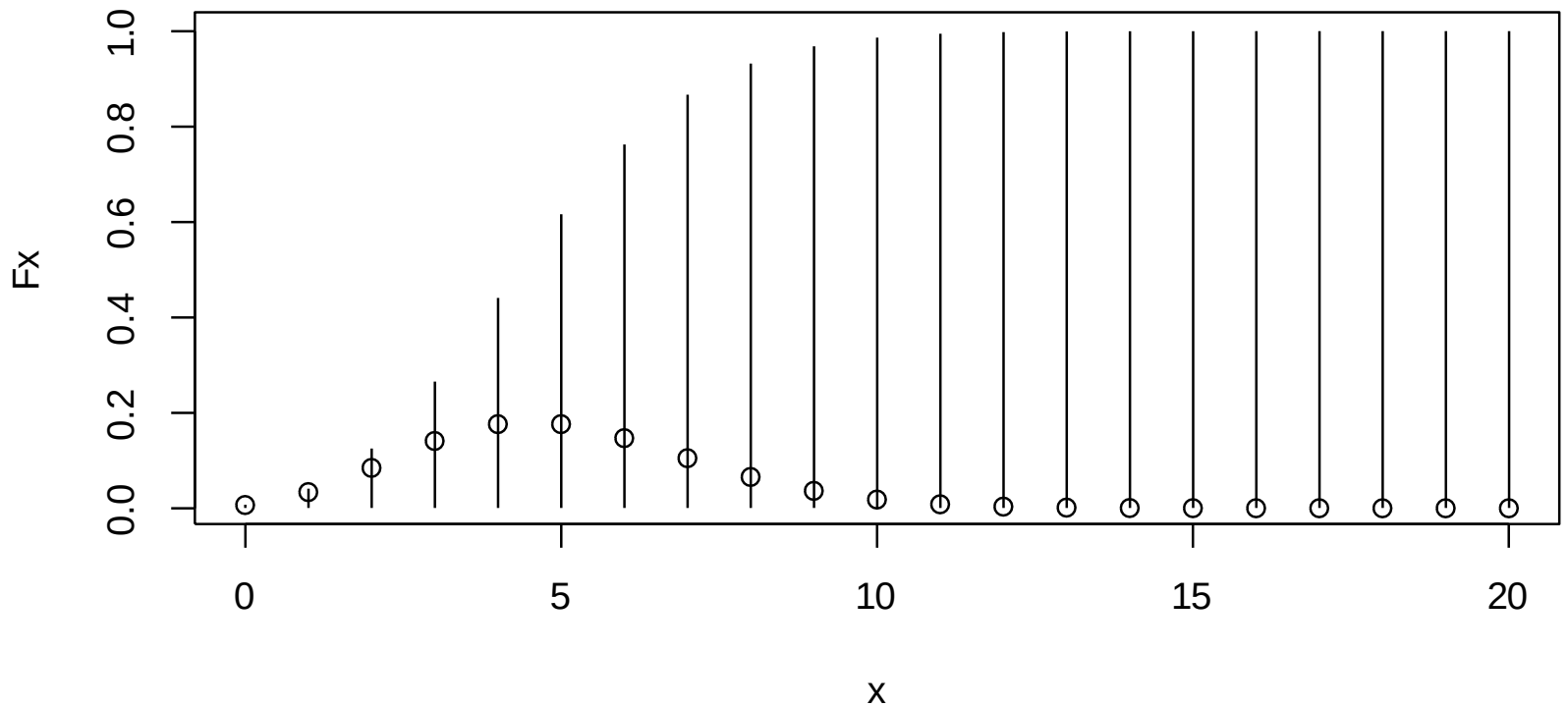
$$F_X(x) = \Pr[X \leq x] = \sum_{y \leq x} \Pr(X = y) = \sum_{y \leq x} f_X(y)$$

In R, density or mass functions start with a ”d” and cumulative distribution functions start with a ”p” (`pnorm`, `pbinom`, `punif`).

Plotting the cumulative distribution function

Here's cumulative distribution function for our distribution.

```
> x <- 0:10000  #vector corresponding to the sample space
> Fx <- pbinom(x, size = 10000, p = 5e-04)
> plot(x, Fx, type = "h", xlim = c(0, 20))
> points(x, fx)
```



Exercise: Understanding distributions

A factory moves 10,000 boxes a day. Each box has a .0005 probability of getting dropped. So on a given day, some number of boxes get dropped.

If I wanted to know what the probability of dropping 10 or fewer boxes in a day, what code would I use to calculate this in R using the cumulative distribution function (CDF)?

Exercise: Understanding distributions

A factory moves 10,000 boxes a day. Each box has a .0005 probability of getting dropped. So on a given day, some number of boxes get dropped.

```
pbinom(10, size=10000, p=.0005)
```

this takes the sum of the probabilities for values less than or equal to 10. It does exactly the same thing as:

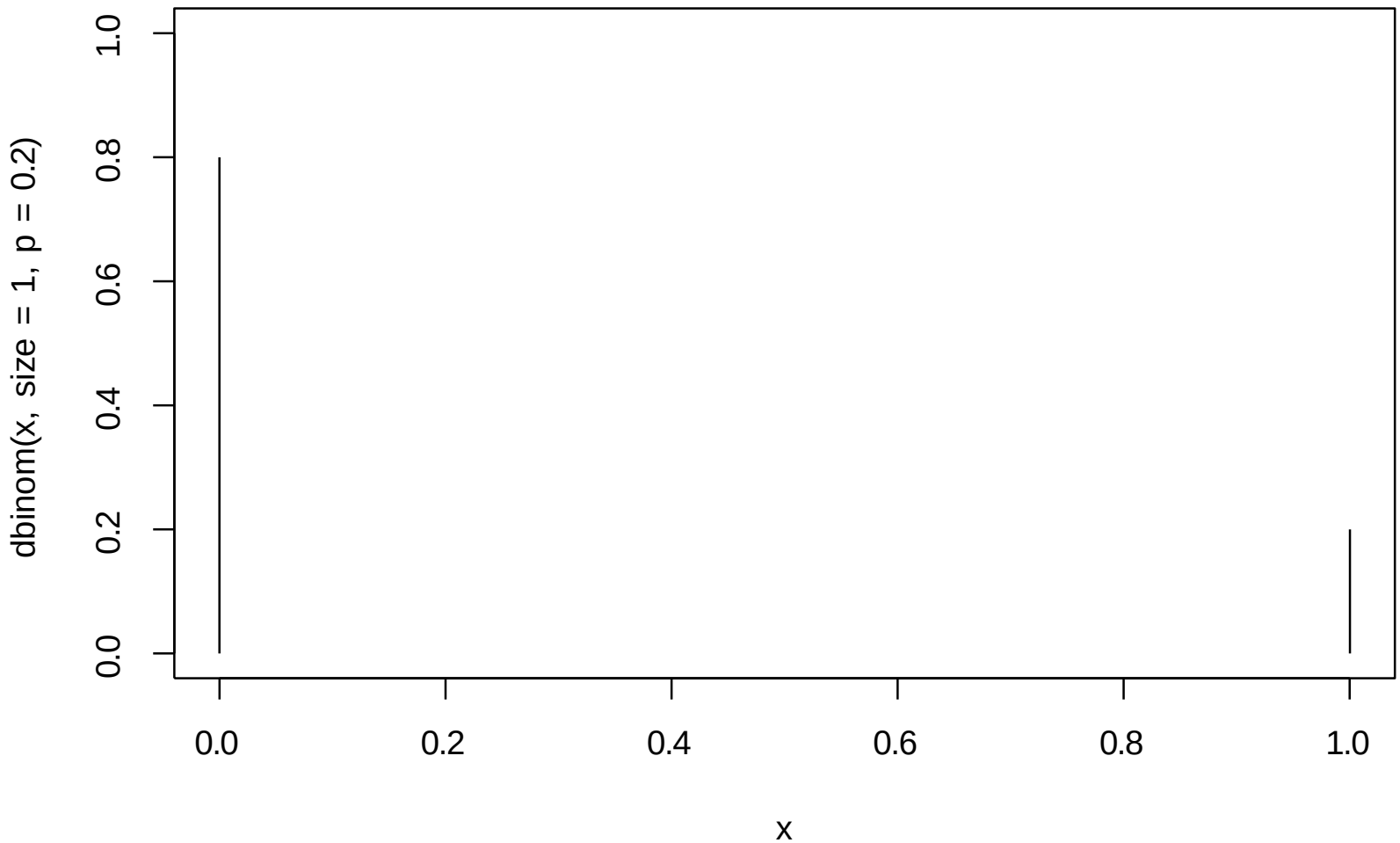
```
sum(dbinom(0:10, size=10000,p=.0005))
```

Lecture 4 Review Exercises

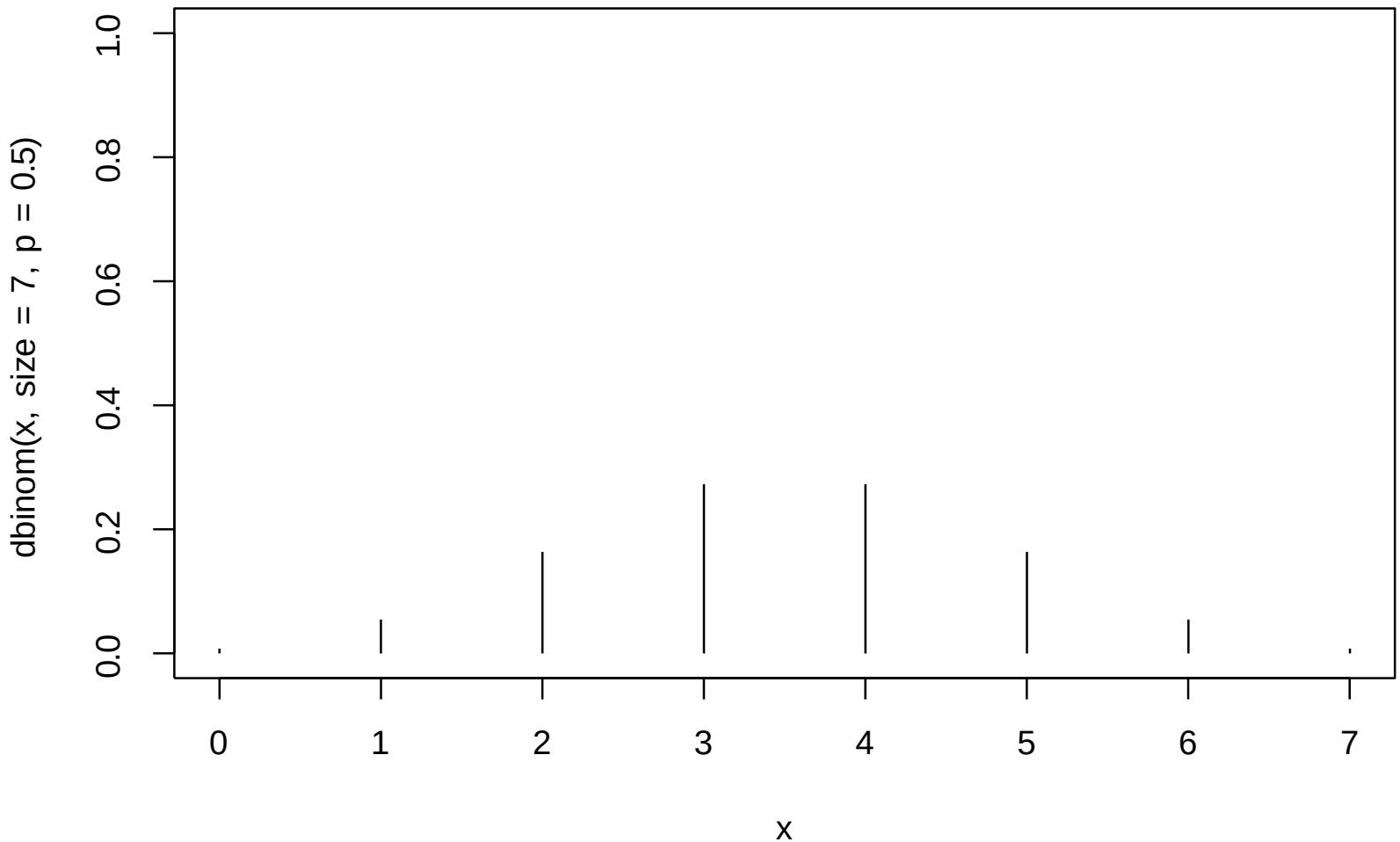
Complete the following:

1. Generate a set of 7 randomly generated values from the Bernoulli p = 0.2 distribution (hint: read the first couple paragraphs of the Wikipedia for "Binomial Distribution").
2. How would we generate the PMF of the Bernoulli p = 0.2 distribution? Plot it.
3. Generate 7 random values from the distribution corresponding to the sum of heads after 10 coin flips (fair coin)
4. Plot the PMF and CDF for this distribution

```
> rbinom(7, size = 1, p = 0.2)
[1] 0 0 1 0 0 0 0
> x <- 0:1
> plot(x, dbinom(x, size = 1, p = 0.2), type = "h", ylim = c(0,
+     1))
```



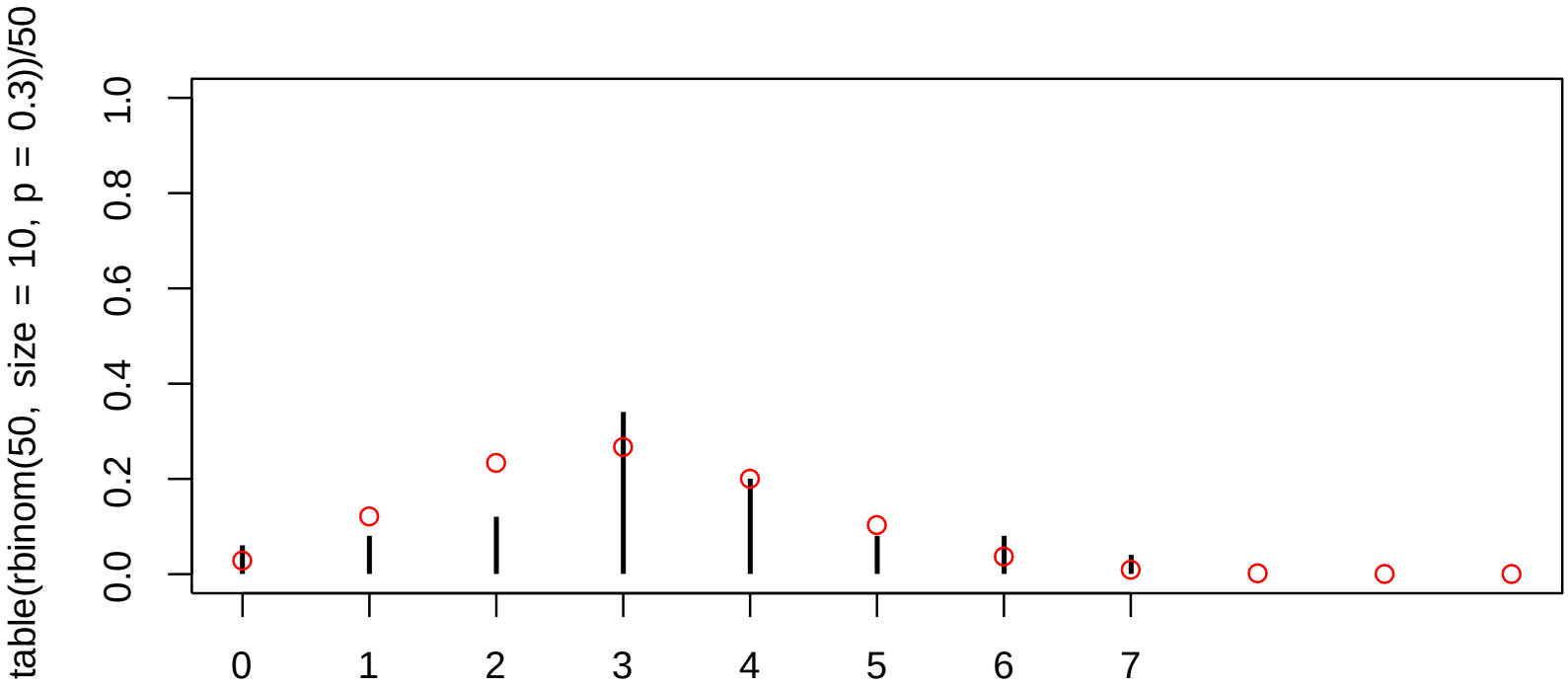
```
> rbinom(7, size = 10, p = 0.5)
[1] 4 3 7 5 3 4 7
> x <- 0:7
> plot(x, dbinom(x, size = 7, p = 0.5), type = "h", ylim = c(0,
+     1))
```



Exercise: Generate 50 values from the Binomial(n = 10, p = 0.3) distribution. Create a plot with probability on the y-axis, sample space on the x-axis, and vertical bars to represent the sample distribution obtained from this sample. Then plot the true distribution as points on top of these lines.

(Hint: plotting the result of table() makes something like a discrete histogram)

```
> plot(table(rbinom(50, size = 10, p = 0.3))/50, xlim = c(0, 10)
+      ylim = c(0, 1))
> points(0:10, dbinom(0:10, size = 10, p = 0.3), col = "red")
```



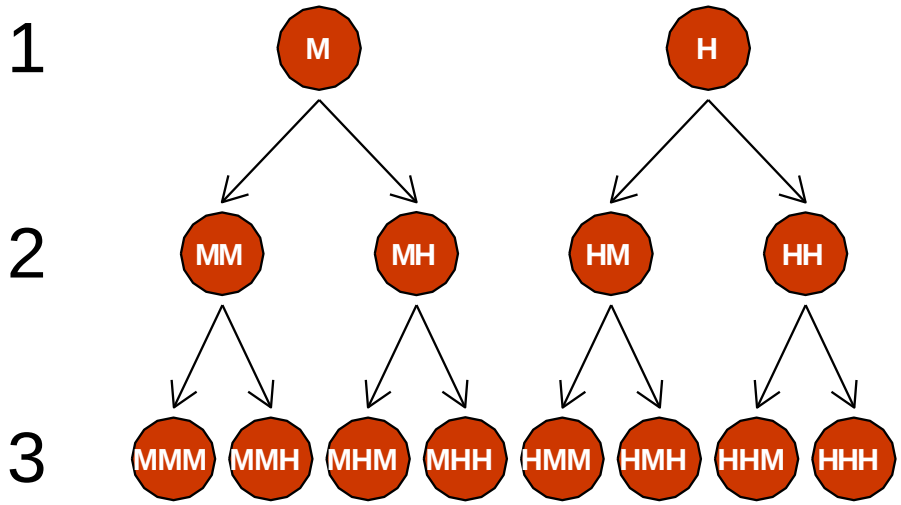
Expectation and Variance

We've learned about ways to describe the *center* and *spread* of sample.

How do we describe the center and spread of a random variable?

Some sample spaces are not "numerical"

Example: Sequence of three free throws



- S = {MMM, MMH, MHM, MHH, HMM, HMH, HHM, HHH}

There's no natural way to order this state space.

Some sample spaces are numerical

Example: Sum of points after three free throws.

- Question II: How many baskets will the basketball player make total?

try 0:	0
try 1:	0 1
try 2:	0 1 2
try 3:	0 1 2 3

- S = {0, 1, 2, 3}

- This is a naturally "numerical" sample space.
- Every outcome can be assigned a value

Definition

A Random Variable ...

... is a variable whose value is a numerical outcome of a random phenomenon.

Or (more technically) a **random variable** X is a function that takes each element of a sample space S and assigns it to a real number.

Discrete random variable

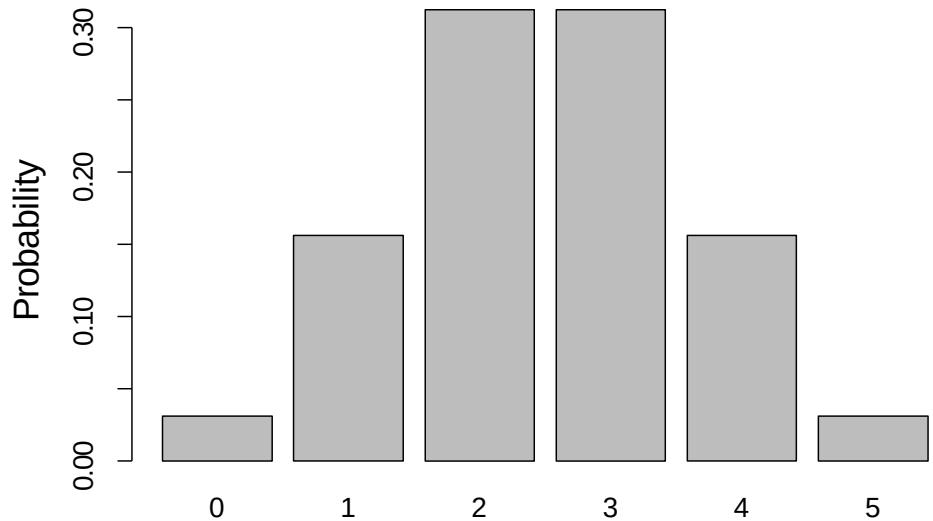
- Every value that a random variable can take is associated with a probability
- The probabilities sum to 1

Value of X	Probability
x_1	p_1
x_2	p_2
x_3	p_3
...	...
x_k	p_k

Example: Number of heads after 5 coin flips

- X is the total number of heads after 5 coin flips
- Possible values of X are: $\{0,1,2,3,4,5\}$
- Probability distribution of X is:
  $P(X = k) = \text{Binomial}(k|n = 5, p = 0.5)$

x	$P(X = x)$
0	0.03125
1	0.15625
2	0.31250
3	0.31250
4	0.15625
5	0.03125



Question: How many heads do we *expect* to get?

- One way to think of this problem is that if we repeated the experiment infinity times - what would the average score be.
- Or, if we performed it some large amount of times but "non-randomly" - i.e. we got the exact probabilities out.

Question: How many heads do we *expect* to get?

- One way to think of this problem is that if we repeated the experiment infinity times - what would the average score be.
- Or, if we performed it some large amount of times but "non-randomly" - i.e. we got the exact probabilities out.

Consider 100,000 realizations of the 5 coin toss:

x	$P(X = x)$	N_i
0	0.03125	3,125
1	0.15625	15,625
2	0.31250	31,250
3	0.31250	31,250
4	0.15625	15,625
5	0.03125	3,125
Sum	**1**	100, 000

Question: How many heads do we *expect* to get?

- One way to think of this problem is that if we repeated the experiment infinity times - what would the average score be.
- Or, if we performed it some large amount of times but "non-randomly" - i.e. we got the exact probabilities out.

Consider 100,000 realizations of the 5 coin toss:

x	$P(X = x)$	N_i
0	0.03125	3,125
1	0.15625	15,625
2	0.31250	31,250
3	0.31250	31,250
4	0.15625	15,625
5	0.03125	3,125
Sum	**1**	100, 000

The mean of these realizations is:

$$\bar{X} = \frac{(3{,}125 * 0 + 15{,}625 * 1 + \cdots + 15{,}625 * 4 + 3{,}125 * 5)}{100{,}000}$$

$$= \sum_{i=1}^{6} \frac{(100{,}000) * (\Pr[X = x_i]) * x_i}{100{,}000}$$

$$= \sum_{i=1}^{6} x_i f_X(x_i)$$

Definition

Expectation

The **expected value** or **expectation** of a discrete random variable X with probability function $f(x)$ is

$$E[X] = \sum_{i=1}^{n} x_i f_X(x_i)$$

where $\{x_1, x_2, \ldots x_n\}$ is the set of all values that X can take.
In statistics, $E(X)$ it is often denoted μ (which should give you a clue as to what it "means").

Question: How many heads do we *expect* to get?

$$E[X] = \sum_{i=1}^{n} x_i f_X(x_i)$$

x	$f(x)$	$xf(x)$
0	0.03125	
1	0.15625	
2	0.31250	
3	0.31250	
4	0.15625	
5	0.03125	
Sum		

Question: How many heads do we *expect* to get?

$$E[X] = \sum_{i=1}^{n} x_i f_X(x_i)$$

x	$f(x)$	$xf(x)$
0	0.03125	0
1	0.15625	0.15625
2	0.31250	0.6250
3	0.31250	0.9375
4	0.15625	0.6250
5	0.03125	0.15625
Sum		

Question: How many heads do we *expect* to get?

$$E[X] = \sum_{i=1}^{n} x_i f_X(x_i)$$

x	$f(x)$	$xf(x)$
0	0.03125	0
1	0.15625	0.15625
2	0.31250	0.6250
3	0.31250	0.9375
4	0.15625	0.6250
5	0.03125	0.15625
Sum	1.0	2.5

Question: How many heads do we *expect* to get?

$$E[X] = \sum_{i=1}^{n} x_i f_X(x_i)$$

x	$f(x)$	$xf(x)$
0	0.03125	0
1	0.15625	0.15625
2	0.31250	0.6250
3	0.31250	0.9375
4	0.15625	0.6250
5	0.03125	0.15625
Sum	1.0	2.5

So, the **expected value** of X is **2.5**.

Expectation of the binomial distribution

Binomial distribution:

$$f_X(x) = f_X(x|n,p) = \Pr[X = x] = \binom{n}{x} p^x (1-p)^{n-x}$$

Expectation of the binomial distribution

Binomial distribution:

$$f_X(x) = f_X(x|n,p) = \Pr[X = x] = \binom{n}{x} p^x (1-p)^{n-x}$$

Solving for the expectation of the binomial distribution:

$$E[X] = \sum_{i=0}^{n} x f_X(x|n,p) = \sum_{i=0}^{n} x_i \binom{n}{x_i} p^{x_i} (1-p)^{n-x_i}$$

$$= \sum_{i=0}^{n} x_i \frac{n!}{x_i!\,(n-x_i)!} p^{x_i} (1-p)^{n-x_i}$$

$$= \sum_{i=0}^{n} x_i \frac{n(n-1)!}{x_i(x_i-1)!\,(n-x_i)!} p p^{x_i-1} (1-p)^{n-x_i}$$

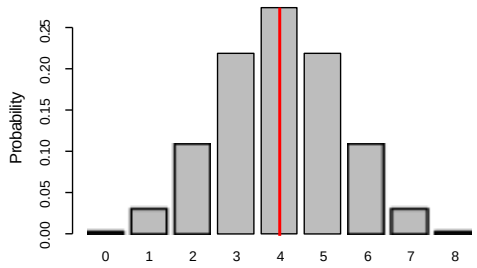
$$= np \sum_{i=0}^{n} \frac{(n-1)!}{(x_i-1)!\,[(n-1)-(x_i-1)]!} p^{x_i-1} (1-p)^{[(n-1)-(x_i-1)]}$$

$$= np \sum_{i=0}^{n} f_X(x-1|n-1,p) = np(1) = np$$

Example of the binomial expectation

- How many heads to we expect after 8 tosses?
  - $p = 0.5$, $n = 8$

    $E(X) = np = 4$

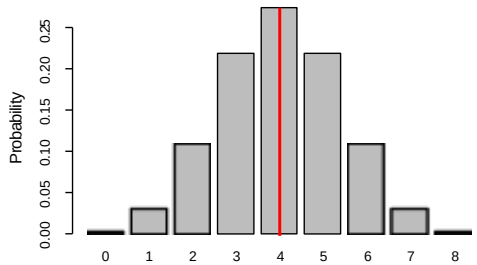


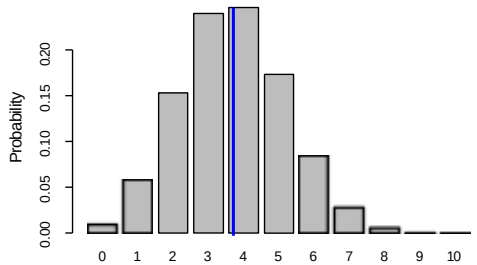
Example of the binomial expectation

- How many heads to we expect after 8 tosses?
  - $p = 0.5, n = 8$
  - $E(X) = np = 4$
- How many baskets do we expect Shaq to make in 10 attempts?

  $p = 0.37, n = 10$

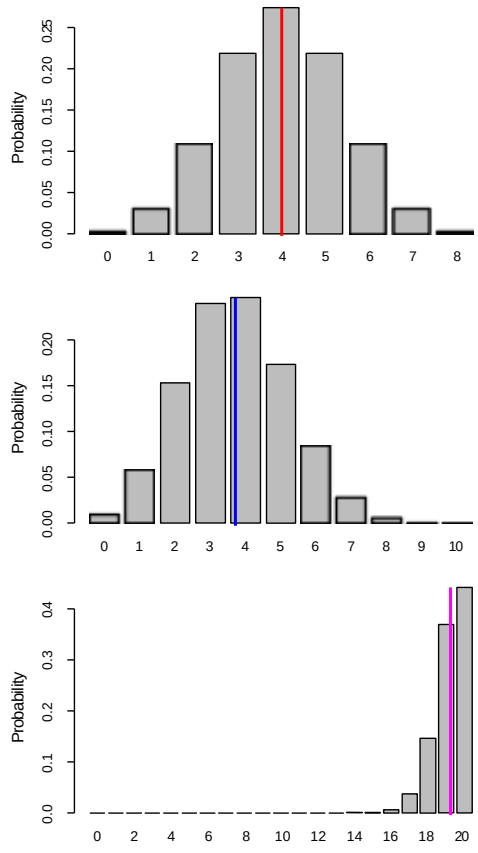
  $E(X) = 3.7$





Example of the binomial expectation

- How many heads to we expect after 8 tosses?
  - $p = 0.5$, $n = 8$
  - $E(X) = np = 4$
- How many baskets do we expect Shaq to make in 10 attempts?
  - $p = 0.37$, $n = 10$
  - $E(X) = 3.7$
- How many baskets do we expect Ray Allen to make in 20 attempts?
  - $p = 0.96$, $n = 20$
  - $E(X) = 19.2$



The **Expectation**

is often denoted μ and called the **mean** of a distribution.

- The expectation tells you the **true mean** of any known, theoretical, distribution
- It is not quite the same as the **sample mean** which we obtain empirically for data.

Some basic arithmetical property of expectations

$$E(A + B) = E(A) + E(B)$$
$$E(kA) = kE(A); \text{ where k is a constant}$$
$$E(AB) = E(A)E(B); \text{ only if } A \text{ and } B \text{ are independent}$$

Variance

- Another very important quantity is the **variance** of a distribution.
- It is the **Expected Squared Deviation from the Mean**
  - Convert that to math notation:

$$Var[X] = E[(X - E[X])^2]$$

Variance

- Another very important quantity is the **variance** of a distribution.
- It is the **Expected Squared Deviation from the Mean**
  - Convert that to math notation:

$$Var[X] = E[(X - E[X])^2]$$

Use the properties of expectation to simplify:

$$\begin{aligned} Var[X] &= E[(X - E[X])^2] \\ &= E[X^2 - 2XE(X) + E[X]^2] \\ &= E[X^2] - 2E[XE(X)] + E[X]^2] \\ &= E[X^2] - 2E(X)E[X] + E[X]^2] \\ &= E[X^2] - 2E[X]^2 + E[X]^2] \\ &= E[X^2] - E[X]^2 \end{aligned}$$

Example: Variance of 4 coin flips

- X is the total number of heads after 4 coin flips
- Possible values of X are: $x = \{0,1,2,3,4\}$
- Probability distribution of X is:
  $P(X = x) = \text{Binomial}(k|n = 4, p = 1/2)$
- Expected value of X is: $E(X) = \mu = np = 2$

Example: Variance of 4 coin flips

- X is the total number of heads after 4 coin flips
- Possible values of X are: $x = \{0,1,2,3,4\}$
- Probability distribution of X is:
  $P(X = x) = \text{Binomial}(k|n = 4, p = 1/2)$
- Expected value of X is: $E(X) = \mu = np = 2$

x	$P(X = x)$	$x - \mu$	$(x - \mu)^2$	$(x - \mu)^2 P(x)$
0	1/16	-2	4	1/4
1	1/4	-1	1	1/4
2	3/8	0	0	0
3	1/4	1	1	1/4
4	1/16	2	4	1/4
Σ	1			**1**

$$\text{Var}(X) = 1$$

The **Variance**

- The **variance** of a random variable X is defined by the following expressions:

$$Var[X] = E[(X - E[X])^2]$$

$$Var[X] = E[X^2] - E[X]^2]$$

- For **discrete random variables,** $X \in \{x_1, x_2, x_3 \ldots x_n\}$, with known probability function $Pr(X = x) = f(x)$:

$$Var[X] = \sum_{i=1}^{n} \left(x_i - \sum_{i=1}^{n} x_i f(x_i) \right)^2 f(x_i)$$

$$Var[X] = \sum_{i=1}^{n} {x_i}^2 f(x_i) - \left(\sum_{i=1}^{n} x_i f(x_i) \right)^2$$

Variance of the binomial distribution

Binomial distribution:

$$f_X(x) = f_X(x|n,p) = \Pr[X = x] = \binom{n}{x} p^x (1-p)^{n-x}$$

Variance of the binomial distribution

Binomial distribution:

$$f_X(x) = f_X(x|n,p) = \Pr[X = x] = \binom{n}{x} p^x (1-p)^{n-x}$$

Solving for the variance of the binomial distribution:

$$Var[X] = \sum_{i=0}^{n} {x_i}^2 f(x_i|n,p) - (E[X])^2$$

$$\ldots hands\ wave\ wildly\ while\ doing\ algebra \ldots$$

$$= np(1-p)$$

The **Variance**

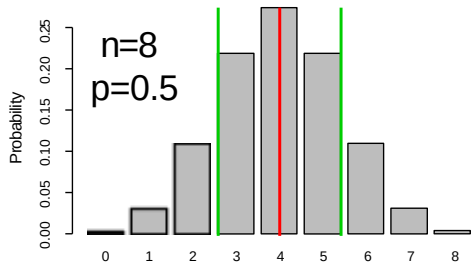
...is often denoted σ^2.

- It tells you something quantitative about the amount of spread in a distribution from the **mean**.
- The square root of the variance, σ is the standard deviation - which is the units of the random variable X.

- This quantity is the **true variance** of any known, theoretical, distribution
- It is not exactly the same as the **sample variance** which we obtain empirically for data.

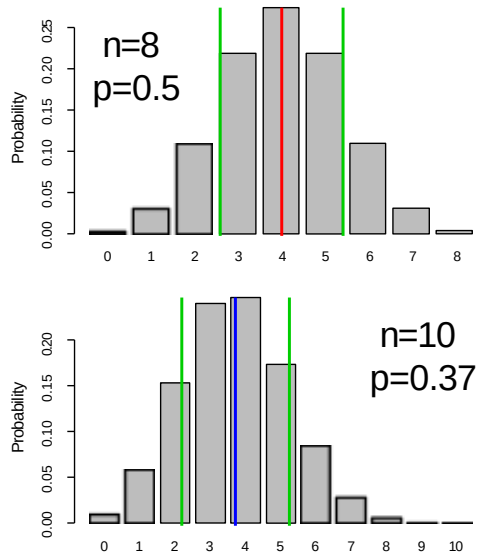
Example of the binomial variance

- What's the variance of heads after 8 coin tosses?
  - $p = 0.5$, $n = 8$, $\mathrm{E}(X) = 4$
    $\mathrm{Var}(X) = np(1-p) = 2$



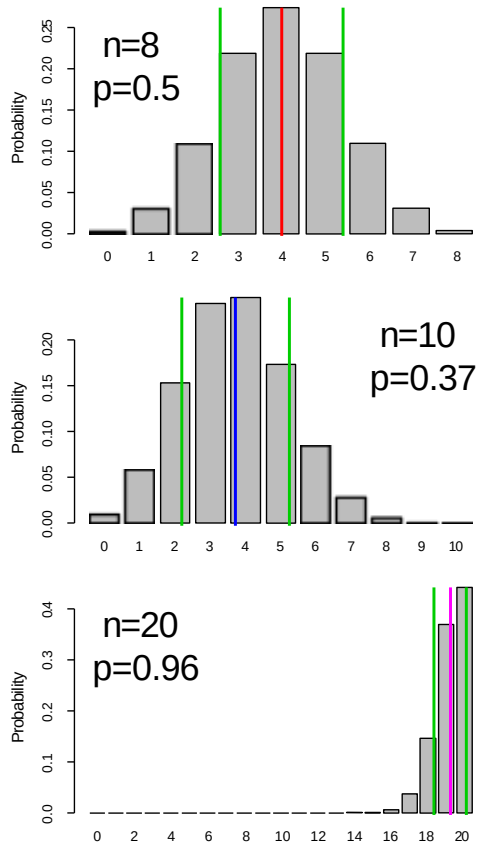
Example of the binomial variance

- What's the variance of heads after 8 coin tosses?
  - $p = 0.5$, $n = 8$, $E(X) = 4$
    $Var(X) = np(1-p) = 2$
- What's the variance of Shaq's 10 FT attempts?
  - $p = 0.37$, $n = 10$, $E(X) = 3.7$
    $Var(X) = 2.331$



Example of the binomial variance

- What's the variance of heads after 8 coin tosses?
  - $p = 0.5$, $n = 8$, $E(X) = 4$
    $Var(X) = np(1-p) = 2$
- What's the variance of Shaq's 10 FT attempts?
  - $p = 0.37$, $n = 10$, $E(X) = 3.7$
    $Var(X) = 2.331$
- What's the variance of Ray Allen's 20 FT attempts?
  - $p = 0.96$, $n = 20$, $E(X) = 19.2$
    $Var(X) = 0.768$



Basic arithmetical property of variances

- $\mathrm{Var}(kA) = k^2\mathrm{Var}(A)$
- If A and B are independent,

$$\mathrm{Var}(A + B) = \mathrm{Var}(A) + \mathrm{Var}(B)$$

Components of a probability distribution

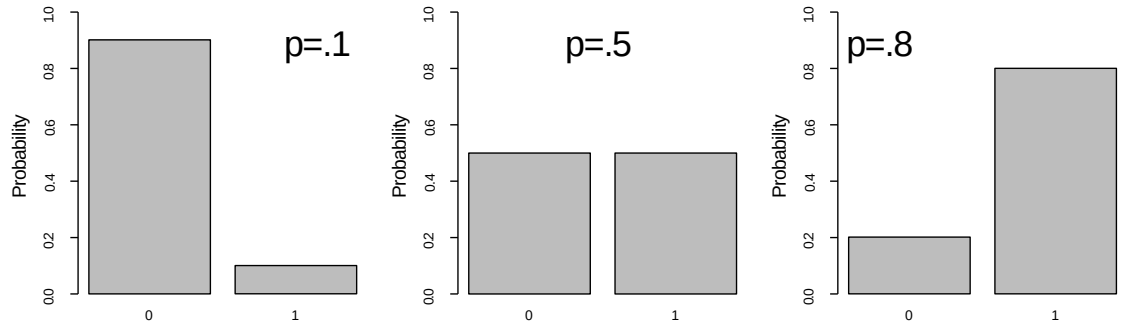
- X: The **random variable** (r.v.)
- x: The possible values (or **support**) of X
  - $X \in \{x_1, x_2, x_3 \ldots x_n\}$
- $P(X = x|\theta) = f(x, \theta)$: The **probability mass functions**
  - Often contracted to "p.m.f." of "pmf"
  - for continuous r.v.'s, called "**probability density function**" (p.d.f.)
- θ: The **parameters** of the pdf.
- $\mathrm{E}(X) = \mu$: The theoretical **mean** of the pdf.
- $\mathrm{Var}(X) = \sigma^2$: The theoretical **variance** of the pdf.

$X \sim \text{Binomial}(n, p)$

Name:	Binomial Distribution	Total number of successes with probability p after n tries.
Support:	x	0,1,2,3,...,n
Parameters:	$n \in \mathbf{N}$ (positive integers) $p \in [0, 1]$	number of trials (positive integer) probability of success
pmf	$P(X = x \mid n, p)$	$f(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$
mean	$E(X)$	np
variance	$\text{Var}(X)$	$np(1-p)$

Bernoulli distribution

- The Bernoulli distribution is the distribution of a **single** event with probability p.[f]
- $P(X = x) = \begin{cases} 1 - p & \text{for } x = 0 \\ p & \text{for } x = 1 \end{cases}$
- Examples: a single coin flip, a single FT.



Question: What is the expected value and variance of the Bernoulli distribution?

$X \sim \text{Bernoulli}(p)$

Name:	Bernoulli Distribution	Models number of successes after one trial
Support:	x	0,1
Parameters:	$p \in [0, 1]$	probability of success
pmf	$P(X = x \mid p)$	$\begin{cases} 1 - p & \text{for } x = 0 \\ p & \text{for } x = 1 \end{cases}$
mean	$E(X)$	p
variance	$\text{Var}(X)$	$p(1 - p)$

Relationship between Binomial and Bernoulli distribution

- A Binomial(n, p) is the sum of n Bernoulli(p) trials
- Formally, if X_1, X_2, X_3 … X_n are each independent variables with:

$$X_i \sim Bernoulli\ (p)$$

then

$$Y = \sum_{i=1}^{n} X_i \sim Binomial\ (n, p)$$

Example Problem: The Lottery

Rules

- Buy a ticket for $1
- Pick 6 numbers from 1, ..., 53
- Win $1,000,000 if your numbers are identical to winning numbers (in any order).

Example Problem: The Lottery

Rules

- Buy a ticket for $1
- Pick 6 numbers from 1, ..., 53
- Win $1,000,000 if your numbers are identical to winning numbers (in any order).

- Q1: What is the probability of winning the jackpot?
  - Uniform sample space: Choose 6 from 53
  - $\binom{53}{6} = \frac{53!}{6!\,(53-6)!} = \frac{53!}{6!\,47!} = \frac{53 * 53 * \cdots * 49 * 48}{6 * 5 * \cdots * 2 * 1} = 22{,}957{,}480$
  - $\Pr(Win) = \frac{1}{22{,}957{,}480} \sim 0.000000043558788$

Example Problem: The Lottery

$P(Winning) = 1/22957480 \sim 0.00000004355878$

- Q2: What are your expected winnings?

Example Problem: The Lottery

P(*Winning*) = 1/22957480

- Q2: What are your expected winnings?
- Call X the dollar amount of your winnings
  - $X \in \{0, 1{,}000{,}000\}$
  - $$\Pr[X = x] = \begin{cases} \frac{22{,}957{,}479}{22{,}9597{,}480} & for\ x = 0 \\ \frac{1}{22{,}9597{,}480} & for\ x = 1{,}000{,}000 \end{cases}$$
  - $$E[X] = \left(0 * \frac{22{,}957{,}479}{22{,}9597{,}480}\right) + \left(1{,}000{,}000 * \frac{1}{22{,}9597{,}480}\right) = 0.04$$
- So: You will lose 0.96 cents, on average, every time you play.
- **Good luck!**

Discrete Distribution Continued: The Uniform and Geometric distributions

We've introduced two important discrete distributions:

- The Bernoulli distribution: representing a binary outcome where succcess has a probability of p
- The Binomial distribution: the sum of successes in n tries, where each success has an independent probability of p. (this is the sum of n bernoulli random variables)

Two other important discrete distributions are worth discussing: the discrete uniform distribution and the geometric distribution.

The discrete uniform distribution

Consider a standard 6-sided die. All sides are equally likely meaning that the probability of observing any specific value in the set $\{1, 2, 3, 4, 5, 6\}$ is 1/6.

This is an incredibly commonly used distribution referred to as the **discrete uniform distribution.** It can be easily simulated in R:

```
> sample(1:6, 10, replace = TRUE)  #10 rolls of a 6-sided die:
 [1] 4 1 3 1 4 5 3 3 5 4
```

The discrete uniform distribution

- The **parameters** of the discrete uniform distribution are a and b and represent the lower and upper bounds.
- The **support** of the discrete uniform distribution are the integers in the interval $[a, b]$.
- The **probability mass function** of the discrete uniform distribution:

  let n be the number of integers in $[a, b]$.

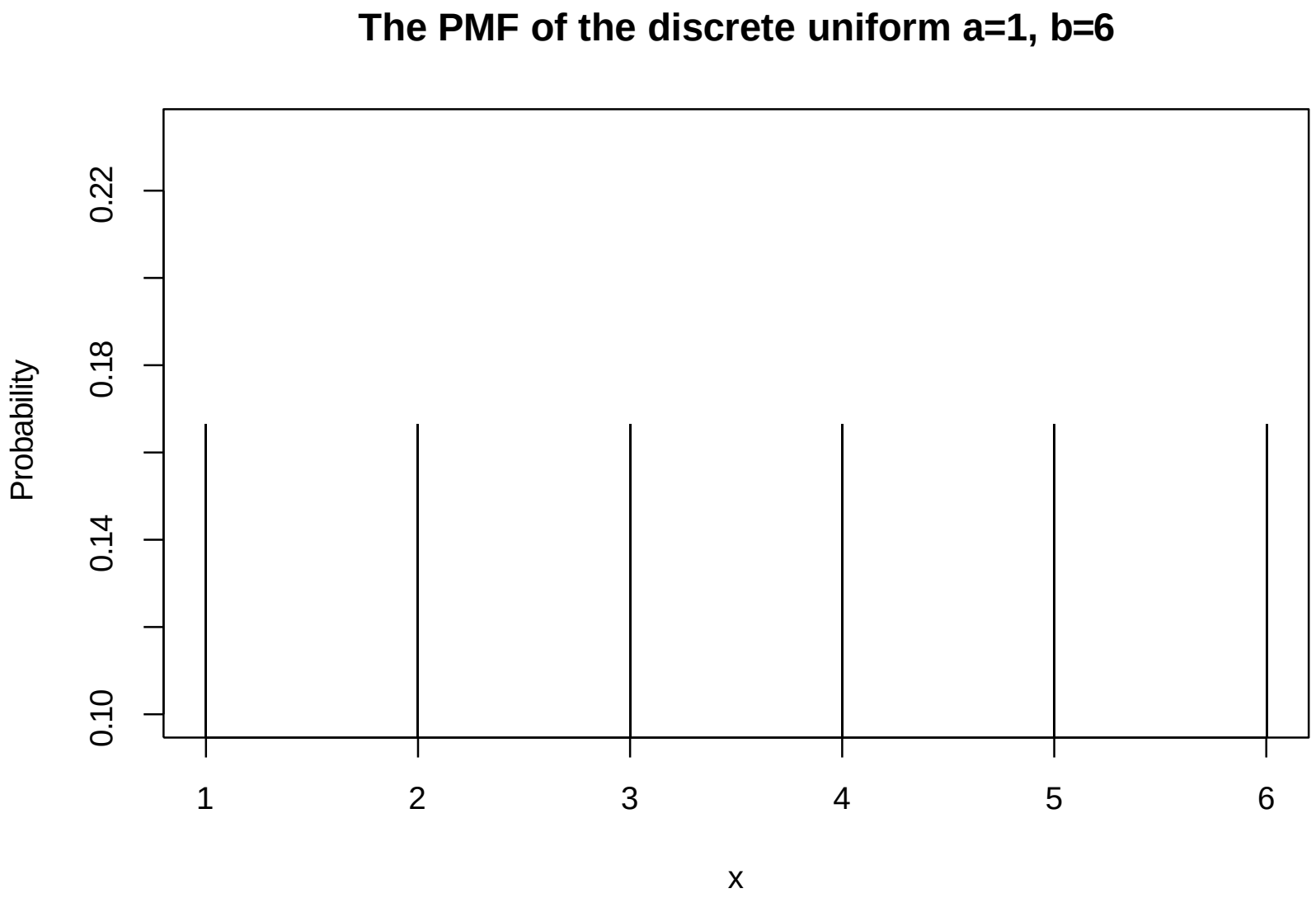
  $$f_X(x) = \Pr(X = x) = 1/n$$

  when x is an integer in $[a, b]$.

  For all other values of x:

  $$f_X(x) = \Pr(X = x) = 0$$

```
> x <- 1:6
> plot(x, rep(1/6, 6), type = "h", ylab = "Probability")
> title("The PMF of the discrete uniform a=1, b=6")
```



Note: the discrete uniform distribution is different than the continuous uniform distribution. The continuous uniform distribution has functions in R: runif, dunif, punif

R does not have functions for the discrete uniform distribution.

random numbers can be generated using `sample()` as seen above.

Complete the following:

5. Mathematically derive the expected value of a discrete uniform distribution with parameters (1,b). Note that

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$$

$$E[X] = \sum_{i=1}^{n} x_i f_X(x_i)$$

6. Write an R function for the PMF of the discrete uniform distribution with parameters (1,b) (hint: make sure it returns 0 for x values not in the support)
7. Use R to find the expected value of a discrete uniform distribution with parameters(1,10)

the expected value of a discrete uniform distribution with parameters (1,b)

$$E[X] = \sum_{x=1}^{b} xf(x) = \sum_{x=1}^{b} x\frac{1}{b} = \frac{1}{b}\sum_{x=1}^{b} x$$

$$= \frac{1}{b}\frac{b(b+1)}{2} = \frac{(b+1)}{2}$$

an R function for the PMF of the discrete uniform distribution with parameters (1,b)

What's wrong with this PMF?

```
> ddunif <- function(x, b = 10) {
+     1/b
+ }
```

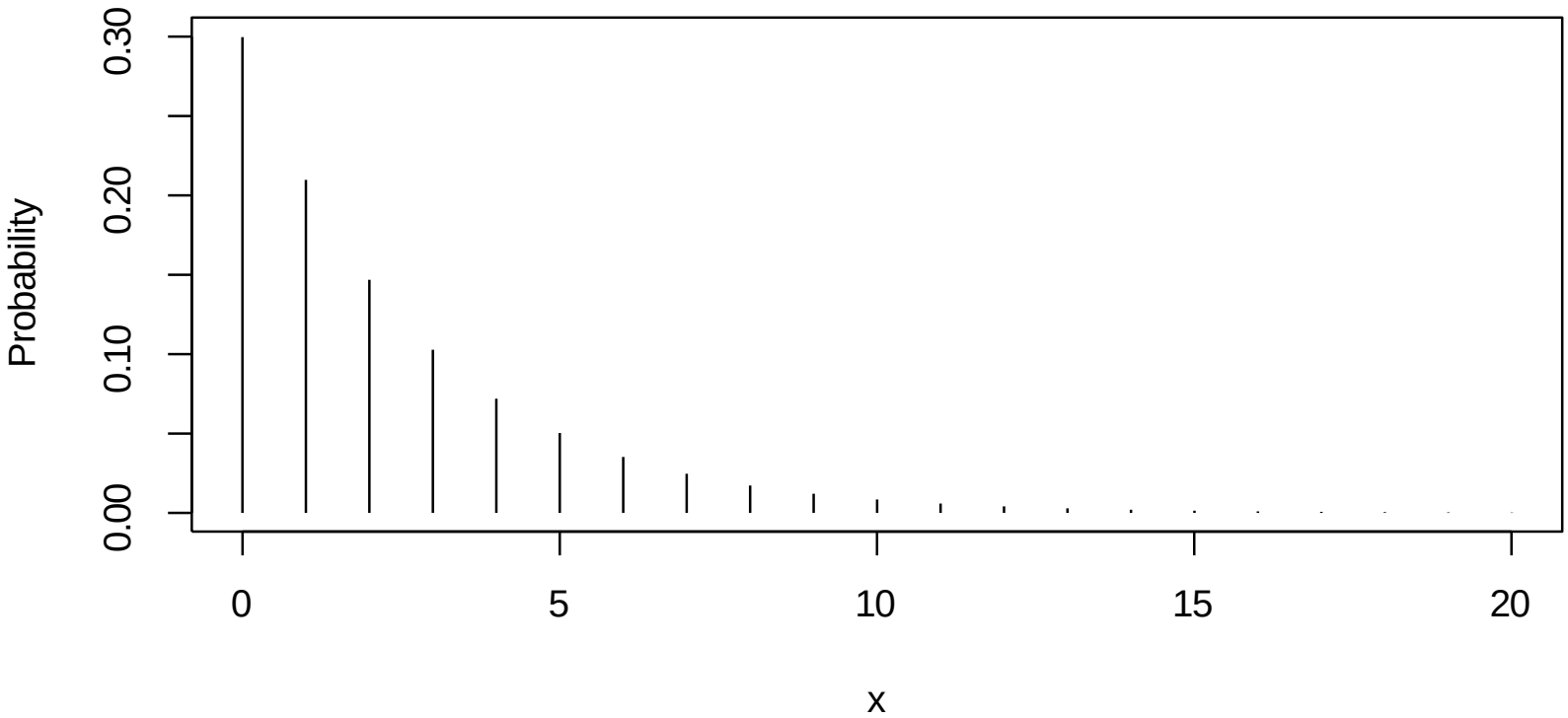
```
> ddunif <- function(x, b = 10) {
+     1/b
+ }
> ddunif(3, 10)
[1] 0.1
> ddunif(25, 10)
[1] 0.1
```

`ddunif` returns nonzero probability for values not in the support of the discrete uniform!

```
> ddunif <- function(x, b = 10) {
+     (x %in% 1:b) * 1/b
+ }
> ddunif(3, 10)
[1] 0.1
> ddunif(25, 10)
[1] 0
```

```
> x <- 1:10
> (Ef <- sum(x * ddunif(x)))
[1] 5.5
```

The Geometric Distribution



Geometric distribution

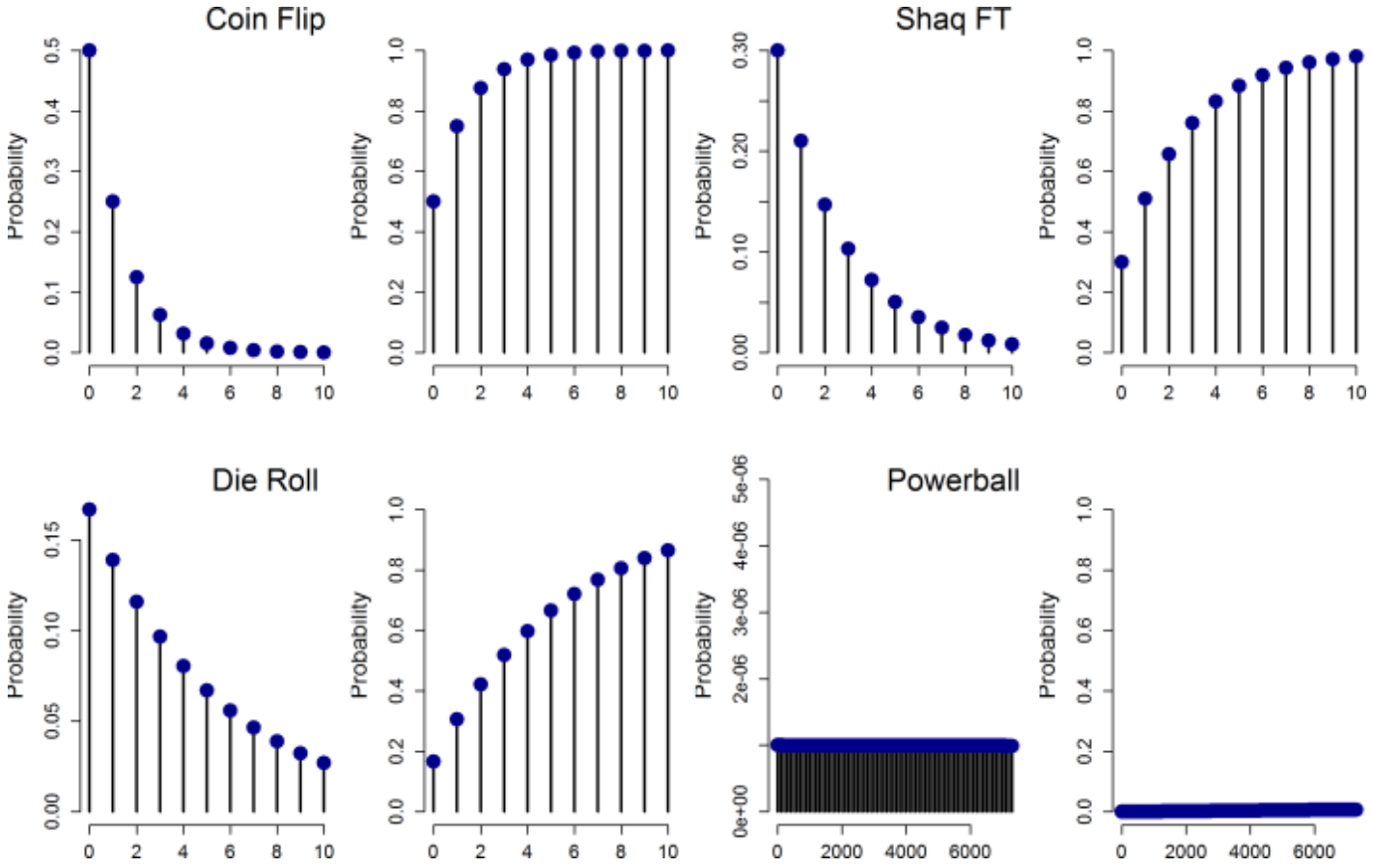
Questions:

- How long (how many rolls) will it take me to roll a 6 on average?
- How long will it take Shaq to make a free throw?
- If I bought a lottery ticket every day, how long would it take me to win?

The Geometric Distribution

- The geometric distribution described the number of Bernoulli trials with probability p before a success - the *waiting time* of a distribution.
- $P(X = k) = p(1-p)^k$ where $k = \{0, 1, 2, 3, \ldots\}$
- Note: one p is for success, k $(1-p)$'s for failure.

Geometric distribution: examples



Geometric distribution: examples

- What is the probability that Shaq (p=0.3) will miss exactly three times before making it? $p(1-p)^3$ = dgeom(3, 0.3) = 10.3%
- What is the probability that Shaq (p=0.3) will miss less than three times in a row?
  pgeom(2, 0.3) = 65%
- What is the probability that Shaq (p=0.3) will miss more than three times in a row?
  1 - pgeom(4, 0.3) = 17%
- What is the probability that you will have won SuperLotto at least once after 20 years of playing daily?
  pgeom(365*20, p = 1/22e6) = 0.03%

Geometric distribution: Memorylessness

- After 3 misses, what is the probability Shaq will miss 3 more times?
  ALSO 10.3%!
- After 20 years of trying, what is the probability you might win after another 20 years?
  ALSO 0.03%!
- It does not matter how long you have been trying to get a success, the waiting time will always have the same distribution.
- This is called "memorylessness" and is very special.

$$\Pr[X > m + n \mid X > m] = \Pr[X > n]$$

$$\Pr[X > 10 + 5 \mid X > 10] = \Pr[X > 5]$$

Geometric distribution

$$E[X] = \sum_{i=1}^{\infty} kp(1-p)^k = \frac{(1-p)}{p}$$

$$Var[X] = \sum_{i=1}^{\infty} k^2 p(1-p)^k - E[X]^2 = \frac{(1-p)}{p^2}$$

Geometric distribution

$$E[X] = \sum_{i=1}^{\infty} kp(1-p)^k = \frac{(1-p)}{p}$$

$$Var[X] = \sum_{i=1}^{\infty} k^2 p(1-p)^k - E[X]^2 = \frac{(1-p)}{p^2}$$

- How long (how many flips) will it take me before I get a head from a fair coin on average?
  Answer: μ_x = (1 – 1/2)/1/2 = 1 flip
- How long will it take me to roll a 6 on average?
  Answer: μ_x = (1 – 1/6)/1/6 = 5 rolls
- How long will it take Shaq to make a free throw?
  Answer: μ_x = (1 – 0.3)/0.3 = 2.333 attempts, σ_x = 2.789
- How long would it take me to win Powerball?
  Answer: μ_x = (1 – 1/23e6)/23e6 = 22e6 days = 63,013 years, σ_x = 22e6 days

$X \sim \text{Geometric}(p)$

Name:	Geometric Distribution	Waiting time of success for Bernoulli trials
Support:	x	0,1,2,3 ...
Parameters:	$p \in [0, 1]$	probability of success
pmf	$P(X = x \mid p)$	$p(1-p)^x$
mean	$E(X)$	$\frac{1-p}{p}$
variance	$\text{Var}(X)$	$\frac{1-p}{p^2}$
Special Feature:	**Memorylessness!**	

Continuous Distributions

The Volcano: Part I

UW vulcanologist says: "Mt. Rainier will definitely erupt in the next 100 years, but it could happen any time."

Continuous random variable

- Eruption time X is a **continuous random variable**.

Continuous random variable

- Eruption time X is a **continuous random variable**.
- If eruption is equally likely to occur at any moment between now and the 100 year interval, then:
  - $P(X < 50 \text{ years}) = 0.5$
  - $P(X < 10 \text{ years}) = 0.1$
  - $P(X > 80 \text{ years}) = 0.2$
  - $P(50 < X < 80 \text{ years}) = 0.3$

Continuous random variable

- Eruption time X is a **continuous random variable**.
- If eruption is equally likely to occur at any moment between now and the 100 year interval, then:
  - $P(X < 50 \text{ years}) = 0.5$
  - $P(X < 10 \text{ years}) = 0.1$
  - $P(X > 80 \text{ years}) = 0.2$
  - $P(50 < X < 80 \text{ years}) = 0.3$
- BUT
  - $P(X = 50 \text{ years}) = P(X = 10 \text{ years}) = P(X = 86.593832715 \text{ years}) = 0$
- How do you calculate these numbers?

Probability density function $f(x)$ (pdf)

- $f(x) \geq 0$ for all values of x.
- Total area under the curve of $f(x)$ is 1.

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

- $\Pr[a < X < b] = \int_{a}^{b} f(x)dx$

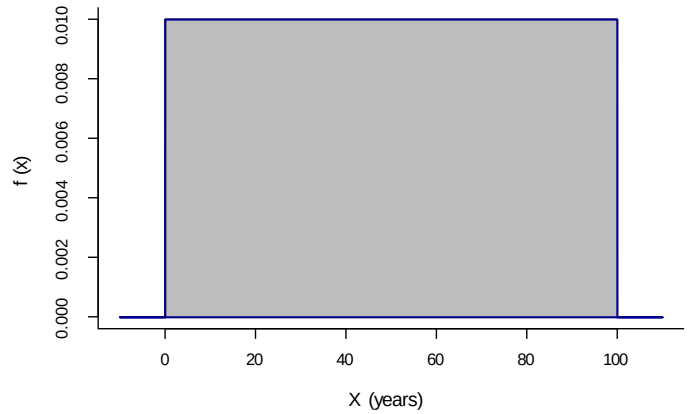
Uniform random variable

Models an event that can happen with equal probability at *any* moment between time *a* and time *b*.

- $X \sim Unif(a, b)$
- $$f(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & otherwise \end{cases}$$
- $$\Pr[c < X < d] = \int_c^d \frac{1}{b-a}\,dx = \frac{d-c}{b-a}$$
- Known as: **continuous uniform distribution**.

Uniform distribution

- $X \sim Unif(0, 100)$
- $f(x) = \begin{cases} \frac{1}{100} & 0 \leq x \leq 100 \\ 0 & otherwise \end{cases}$



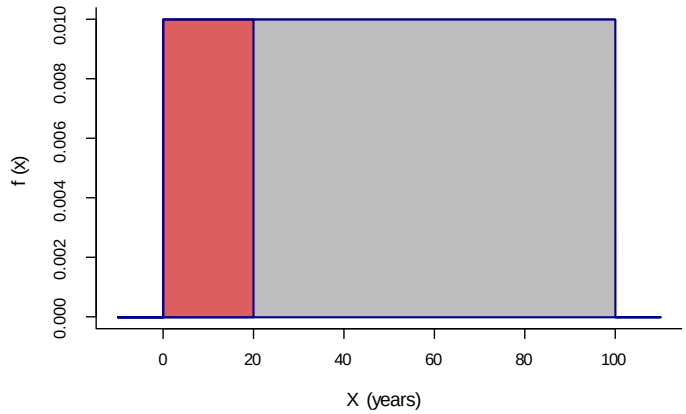
Uniform distribution

- $X \sim Unif(0, 100)$
- $f(x) = \begin{cases} \frac{1}{100} & 0 \le x \le 100 \\ 0 & otherwise \end{cases}$

$$\Pr[X < 20] = \int_{-\infty}^{20} \frac{1}{100} dx$$

$$= \frac{20}{100} - \frac{0}{100}$$

$$= 0.20$$



punif(20,0,100)

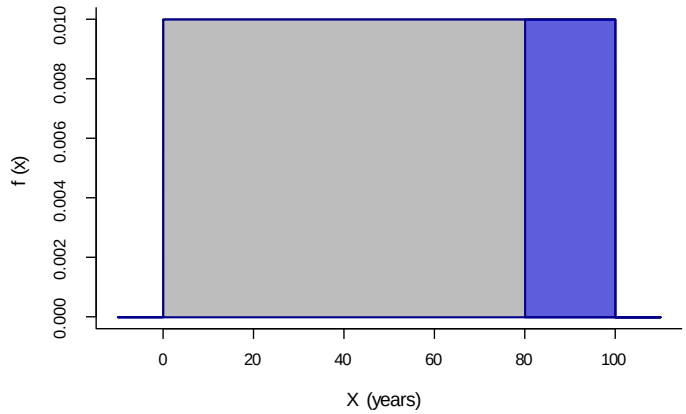
Uniform distribution

- $X \sim Unif(0, 100)$
- $f(x) = \begin{cases} \frac{1}{100} & 0 \le x \le 100 \\ 0 & otherwise \end{cases}$

$$\Pr[X > 80] = \int_{80}^{\infty} \frac{1}{100} dx$$

$$= \frac{100}{100} - \frac{80}{100}$$

$$= 0.20$$




```
1 - punif(80,0,100)
```

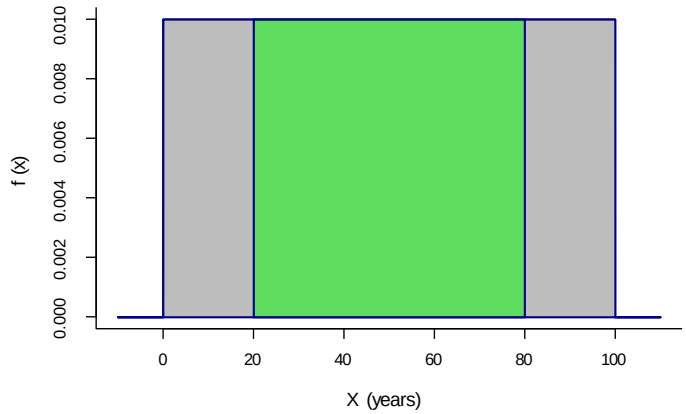
Uniform distribution

- $X \sim Unif(0, 100)$
- $f(x) = \begin{cases} \frac{1}{100} & 0 \le x \le 100 \\ 0 & otherwise \end{cases}$

$$\Pr[20 < X < 80] = \int_{20}^{80} \frac{1}{100} dx$$

$$= \frac{80}{100} - \frac{20}{100}$$

$$= 0.60$$




```
punif(80,0,100) - punif(20,0,100)
```

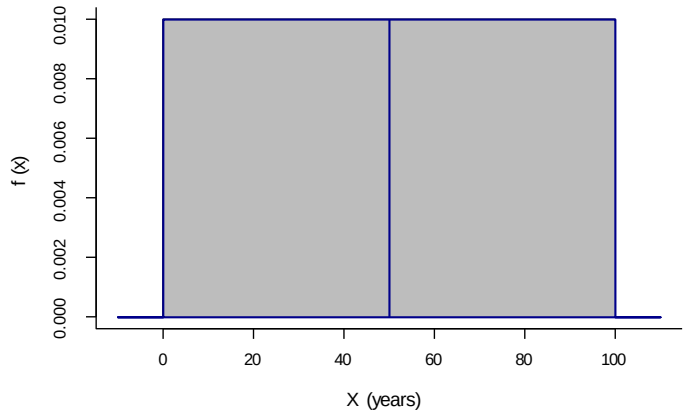
Uniform distribution

- $X \sim Unif(0, 100)$
- $f(x) = \begin{cases} \frac{1}{100} & 0 \leq x \leq 100 \\ 0 & otherwise \end{cases}$

$$\Pr[X = 50] = \int_{50}^{50} \frac{1}{100} dx$$

$$= \frac{50}{100} - \frac{50}{100}$$

$$= 0$$



Expected value and variance

- Refresher: Let X be a discrete random variable with possible values $x_1, x_2, x_3, \ldots, x_n$ and probability mass function $f(x)$
- $E[X] = \sum_{i=1}^{n} x_i f_X(x_i)$
- $Var[X] = E[(X - E[X])^2]] = E[X^2] - E[X]^2 = \sum_{i=1}^{n} (x_i - E[X])^2 f(x_i)$

Expected value and variance

- Refresher: Let X be a discrete random variable with possible values $x_1, x_2, x_3, \ldots, x_n$ and probability mass function $f(x)$
- $E[X] = \sum_{i=1}^{n} x_i f_X(x_i)$
- $Var[X] = E[(X - E[X])^2]] = E[X^2] - E[X]^2 = \sum_{i=1}^{n}(x_i - E[X])^2 f(x_i)$

Expectation and variance for continuous random variables

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

$$Var[X] = E[(X - E[X])^2]]$$

$$= \int_{-\infty}^{\infty} (x - E[X])^2 f(x) dx$$

Uniform random variable

$$X \sim Unif(a,b), f(x) = \frac{1}{b-a}$$

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx = \int_{a}^{b} x\frac{1}{b-a}dx$$

$$= \frac{1}{b-a}\frac{x^2}{2}\Big|_a^b = \frac{b^2-a^2}{2(b-a)}$$

$$= \frac{(b-a)(b+a)}{2(b-a)} = \frac{b+a}{2}$$

$$Var[X] = \int_{-\infty}^{\infty} (x - E[X])^2 f(x)dx$$

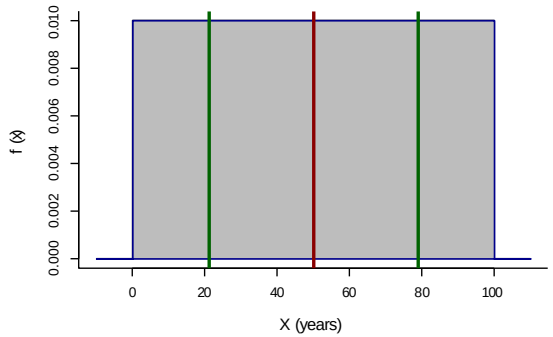
$$= \int_{-\infty}^{\infty} (x - \frac{b+a}{2})^2 \frac{1}{b-a}dx$$

... hands wave wildly while doing algebra ...

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

Uniform random variable

- Eruption time $X \sim \text{Unif}(0, 100)$
- $E(X) = 100 / 2 = 50$
- $\text{Var}(X) = 100^2 / 12 = 833.33$
- $\text{SD}(X) = 28.87$



$X \sim \text{Unif}(a, \beta)$

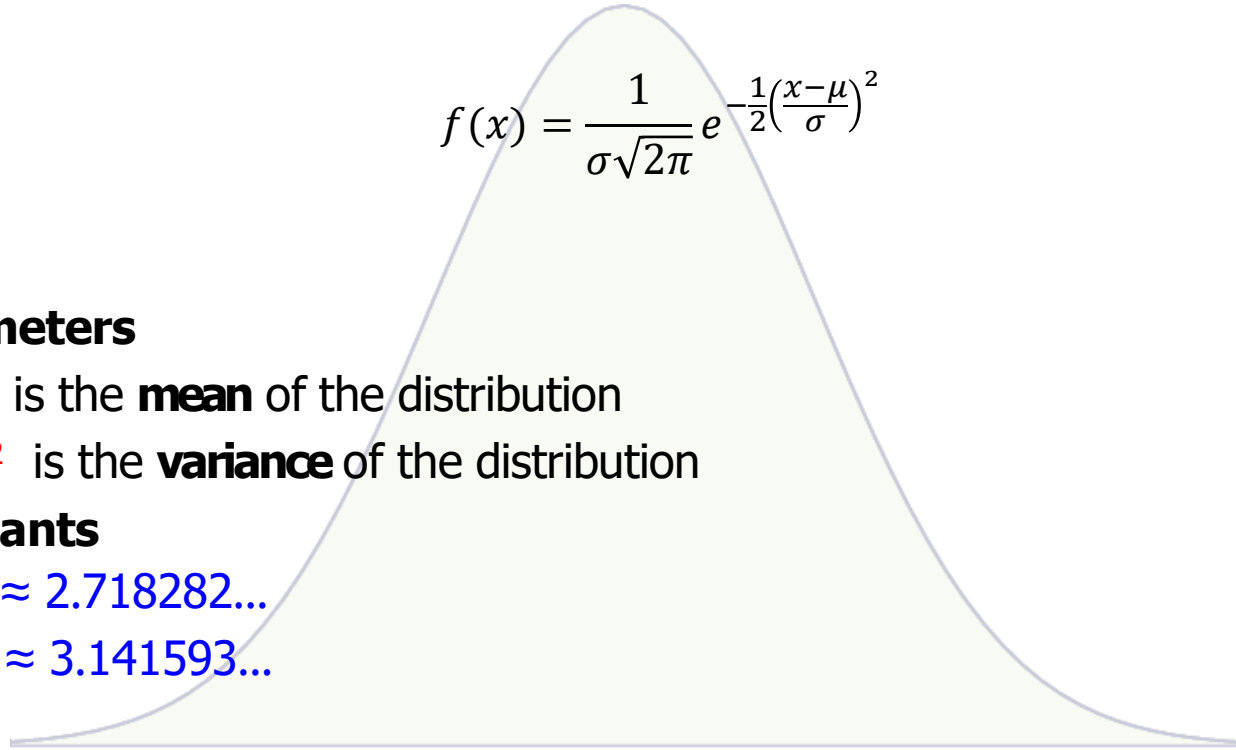
Name:	Uniform distribution	Models equal probability events within a continuous range of values
Support:	x	$(-\infty, \infty)$
Parameters:	$\alpha \in (-\infty, \beta)$ $\beta \in (\alpha, \infty)$	minimum maximum
pdf	$f(x \mid \alpha, \beta) =$	$\frac{1}{\beta - \alpha}$
mean	$E(X)$	$\frac{\beta + \alpha}{2}$
variance	$\text{Var}(X)$	$\frac{1}{12}(\beta - \alpha)^2$

The Volcano Part II

Consider a volcano that erupts with a normal distribution with mean 50 and standard deviation 20 years...

Normal distribution: The formula

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



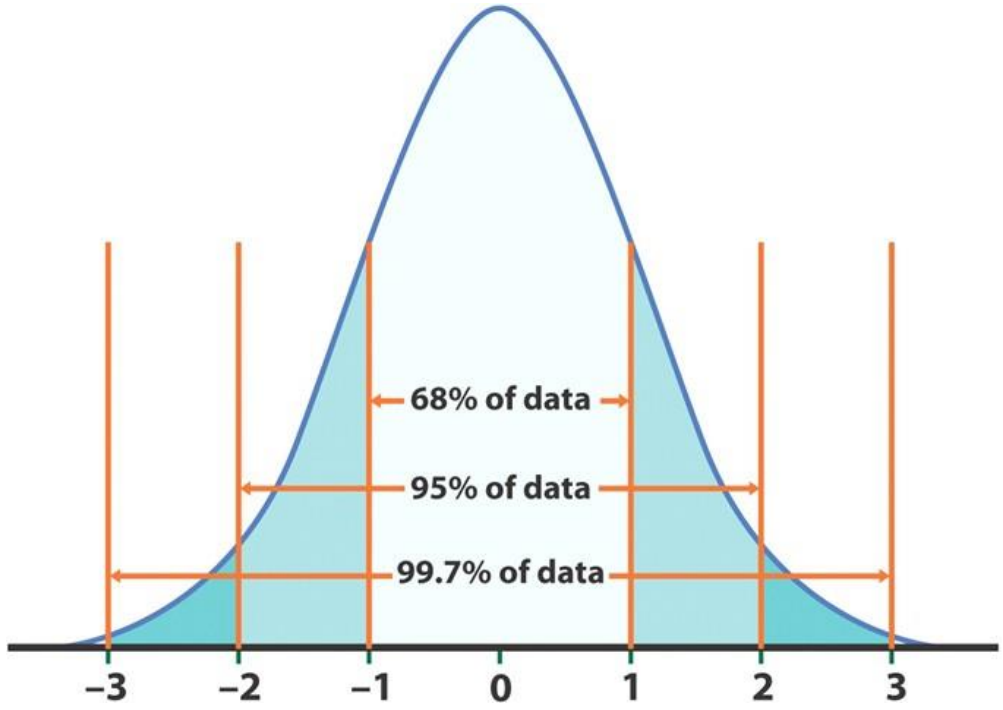
Parameters

- μ is the **mean** of the distribution
- σ^2 is the **variance** of the distribution

Constants

- $e \approx 2.718282\ldots$
- $\pi \approx 3.141593\ldots$

No matter what the parameters are...

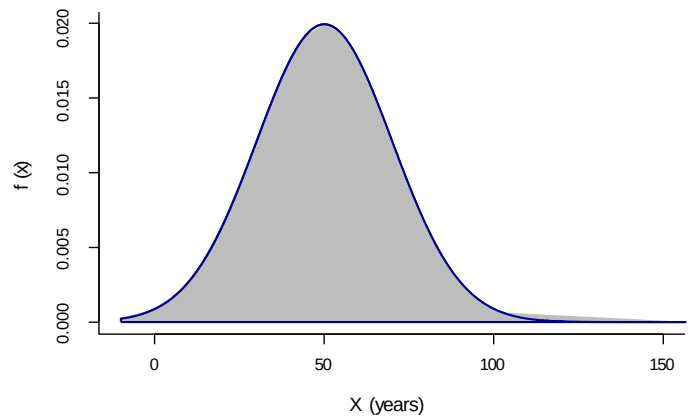


Approximately:

- 68% of the probability lies "within one standard deviations" ($\mu \pm \sigma$).
- 95% of the probability lies "within 2 standard deviations"($\mu \pm 2\sigma$).
- 99.7% of the probability lies "within 3 standard deviations"($\mu \pm 3\sigma$).

Normal distribution

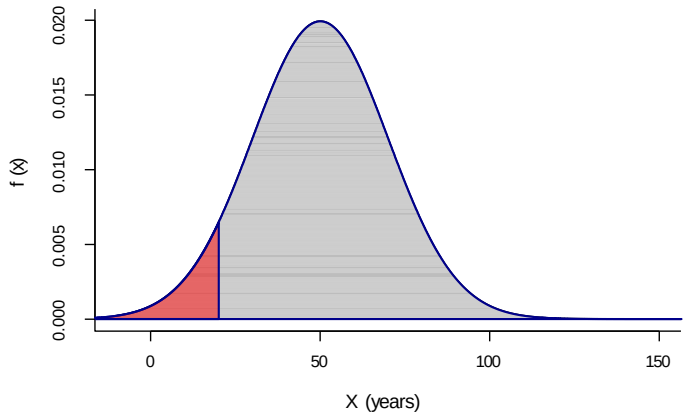
- $X \sim N(\mu = 50, \sigma = 20)$
- $f(x) = \frac{1}{20\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-50}{20}\right)^2}$



Normal distribution

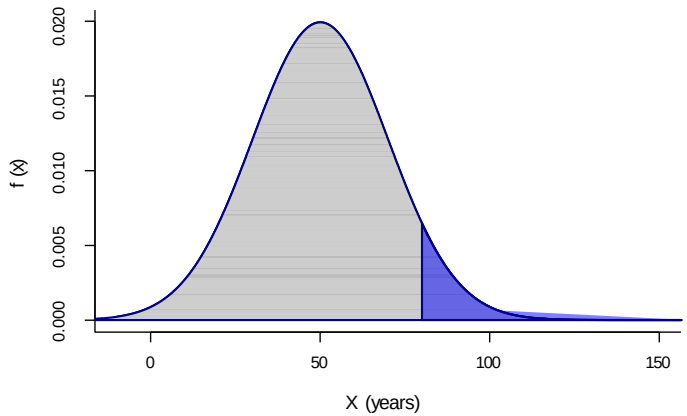
- $X \sim N(\mu = 50, \sigma = 20)$
- $f(x) = \frac{1}{20\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-50}{20}\right)^2}$

- $\Pr[X < 20] = \int_{-\infty}^{20} f(x)dx = 0.067$



Normal distribution

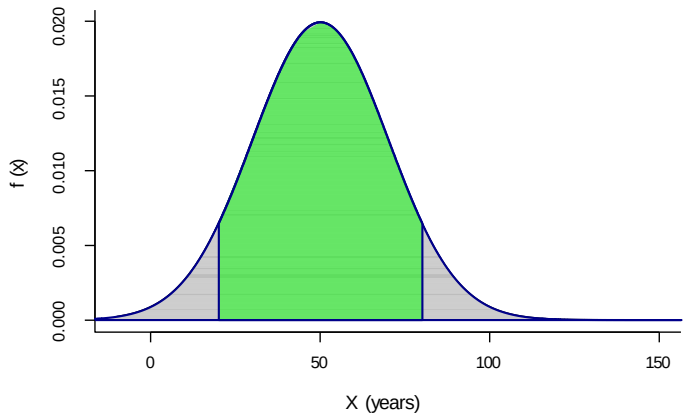
- $X \sim N(\mu = 50, \sigma = 20)$
- $f(x) = \frac{1}{20\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-50}{20}\right)^2}$
- $\Pr[X > 80] = \int_{80}^{\infty} f(x)dx = 0.067$



Normal distribution

- $X \sim N(\mu = 50, \sigma = 20)$
- $f(x) = \frac{1}{20\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-50}{20}\right)^2}$

- $\Pr[20 < X < 80] = 0.866$



$X \sim \text{Normal}(\mu, \sigma)$

Name:	Normal distribution	Models bell-shaped continuous variables.
Support:	x	$(-\infty, \infty)$
Parameters:	$\mu \in (-\infty, \infty)$ $\sigma \in (0, \infty)$	measure of center measure of spread
pdf	$f(x\|\mu, \sigma) =$	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
mean	$E(X)$	μ
variance	$\text{Var}(X)$	σ^2

Note on the normal distribution

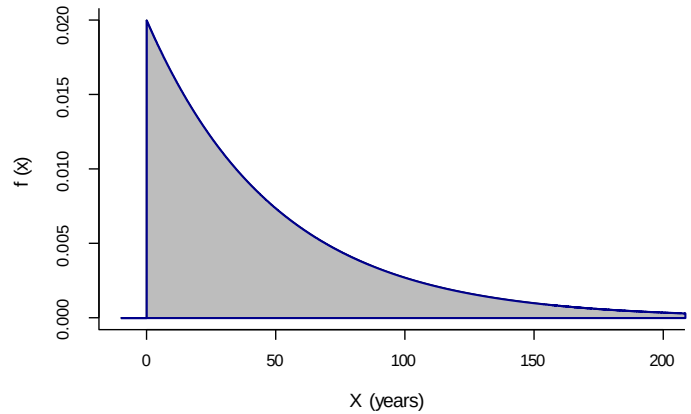
The normal distribution is a fantastic distribution for many things, but it is a lousy model for volcano eruptions!

The Volcano Part III

Russian vulcanologist says: "Mt. Avacha has been erupting at totally random times, but on average every 50 years."

Exponential distribution

- $X \sim Exp(\gamma = 50)$
- $f(x) = \frac{1}{\gamma} e^{-\frac{x}{\gamma}}, \qquad for\ x > 0$



$$\Pr[X > 80] = \int_{80}^{\infty} f(x)dx = 0.067$$

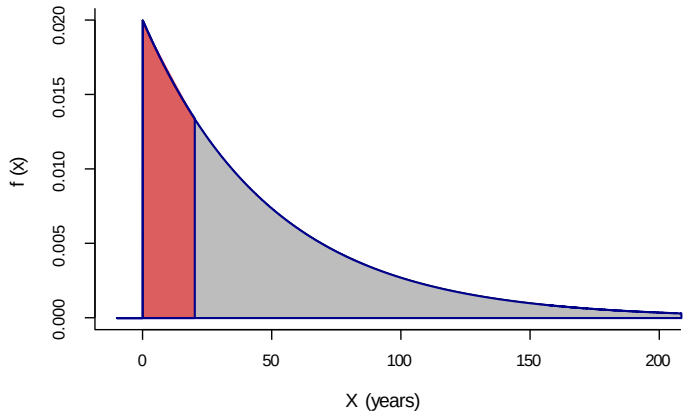
Exponential distribution

- $X \sim Exp(\gamma = 50)$
- $f(x) = \frac{1}{\gamma} e^{-\frac{x}{\gamma}}, \qquad for\ x > 0$

$$\Pr[X < 20] = \int_{-\infty}^{20} f(x)dx = 0.33$$

```
pexp(20, rate=1/50)
```

Note, the use of "rate = 1/γ" instead of scale.

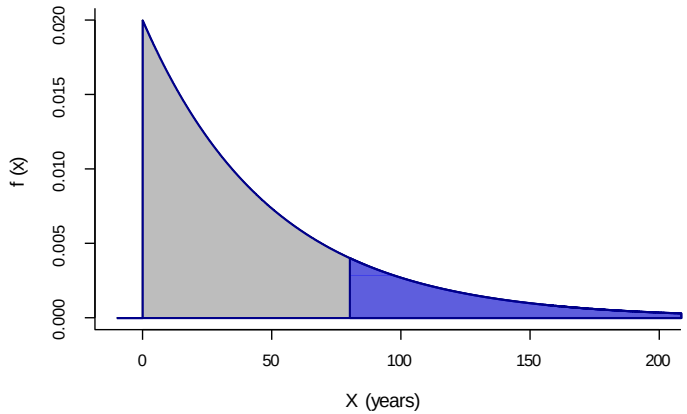


Exponential distribution

- $X \sim Exp(\gamma = 50)$
- $f(x) = \frac{1}{\gamma} e^{-\frac{x}{\gamma}}, \qquad for\ x > 0$

$$\Pr[X < 80] = \int_{80}^{\infty} f(x)dx = 0.20$$

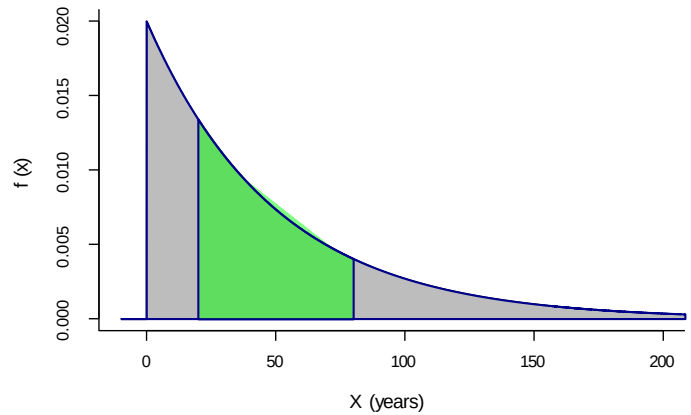
```
1-pexp(80, rate=1/50)
```



Exponential distribution

- $X \sim Exp(\gamma = 50)$
- $f(x) = \frac{1}{\gamma} e^{-\frac{x}{\gamma}}, \qquad for\ x > 0$

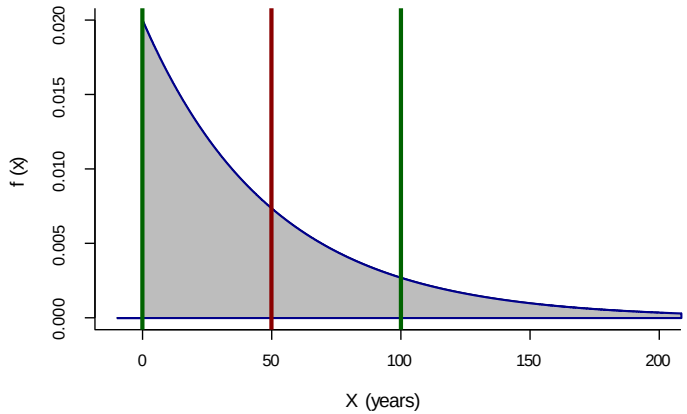
$\Pr[20 < X < 80] = 0.47$




```
pexp(80, 1/50) - pexp(20, 1/50)
```

Exponential Distribution

- Eruption time
  $X \sim \text{Exp}(\gamma = 50)$
- $E(X) = 50$
- $\text{Var}(X) = 50^2$
- $\text{SD}(X) = 50$



Exponential random variable

$$\mathrm{X} \sim \mathrm{Exp}(\gamma), f(x) = \frac{1}{\gamma} e^{-\frac{x}{\gamma}}$$

$$E[X] = \int_0^\infty x f(x) dx = \int_0^\infty \frac{x}{\gamma} e^{-\frac{x}{\gamma}}$$

$$= e^{-\frac{x}{\gamma}}(\gamma + x) \Big|_0^\infty$$

$$= \gamma$$

$$Var[X] = \int_0^\infty (x - \gamma)^2 f(x) dx$$

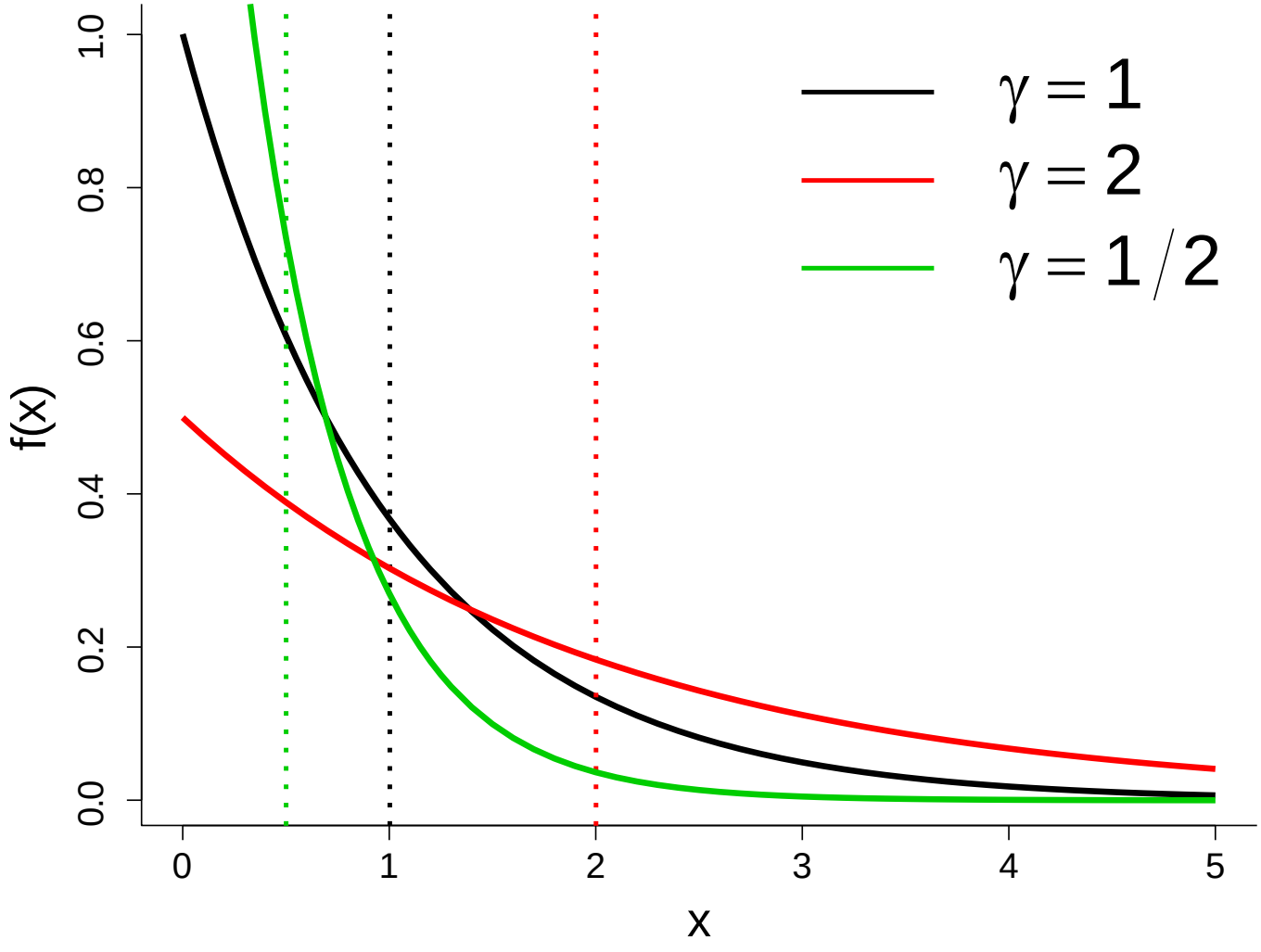
$$\ldots \textit{hands wave wildly while doing algebra} \ldots$$

$$= \gamma^2$$

$X \sim \text{Exp}(\gamma)$

Name:	Exponential distribution	Models time-independent random events
Support:	x	$(0, \infty)$
Parameters:	$\gamma \in (0, \infty)$	scale parameter
pdf	$f(x\|\mu, \sigma) =$	$\frac{1}{\gamma} e^{-\frac{x}{\gamma}}$
mean	$\text{E}(X)$	γ
variance	$\text{Var}(X)$	γ^2

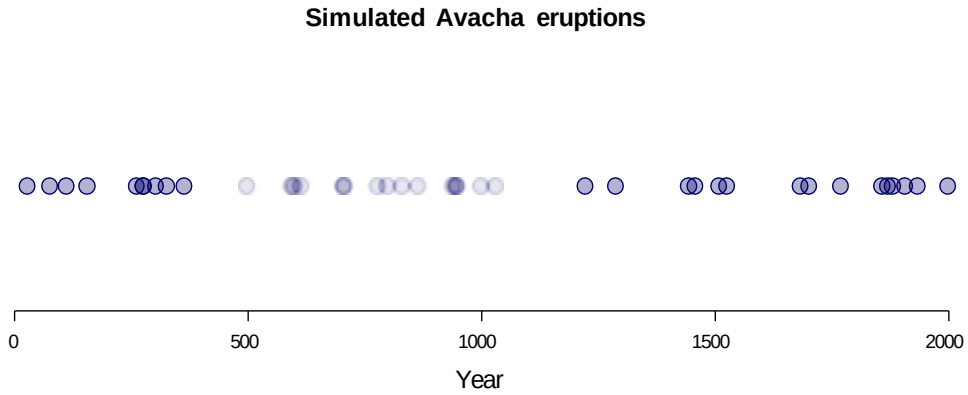
Exponential Distribution: $X \sim \text{Exp}(\gamma)$



Important comments on the exponential distribution

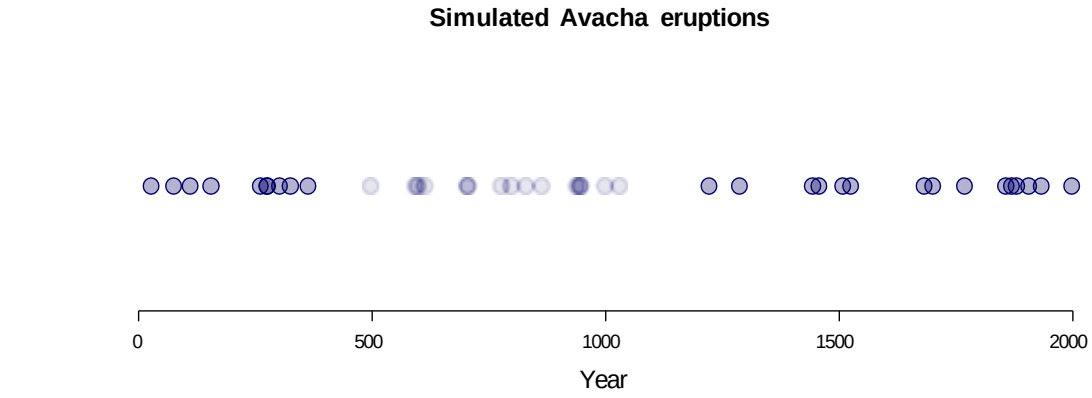
- The exponential distribution models the waiting time to any event which can occur with equal probability in time.
- It is the continuous analogue of the geometric distribution, and is also memoryless:
  - If Avacha erupts today, the expected time to next eruption is 50 years.
  - If Avacha hasn't erupted in 50 years, the expected time to next eruption is 50 years.
  - If Avacha hasn't erupted in 500 years, the expected time to next eruption is 50 years (though you might consider updating your Exp(50) model).
- It is readily identified by the standard deviation being similar to the mean.

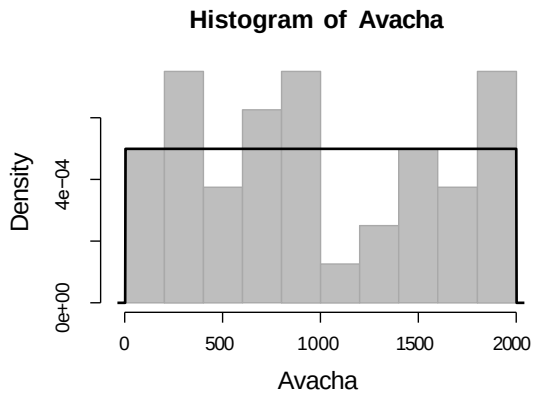
Relationship between exponential and uniform distributions




```
> Avacha <- runif(40,0,1000)
```

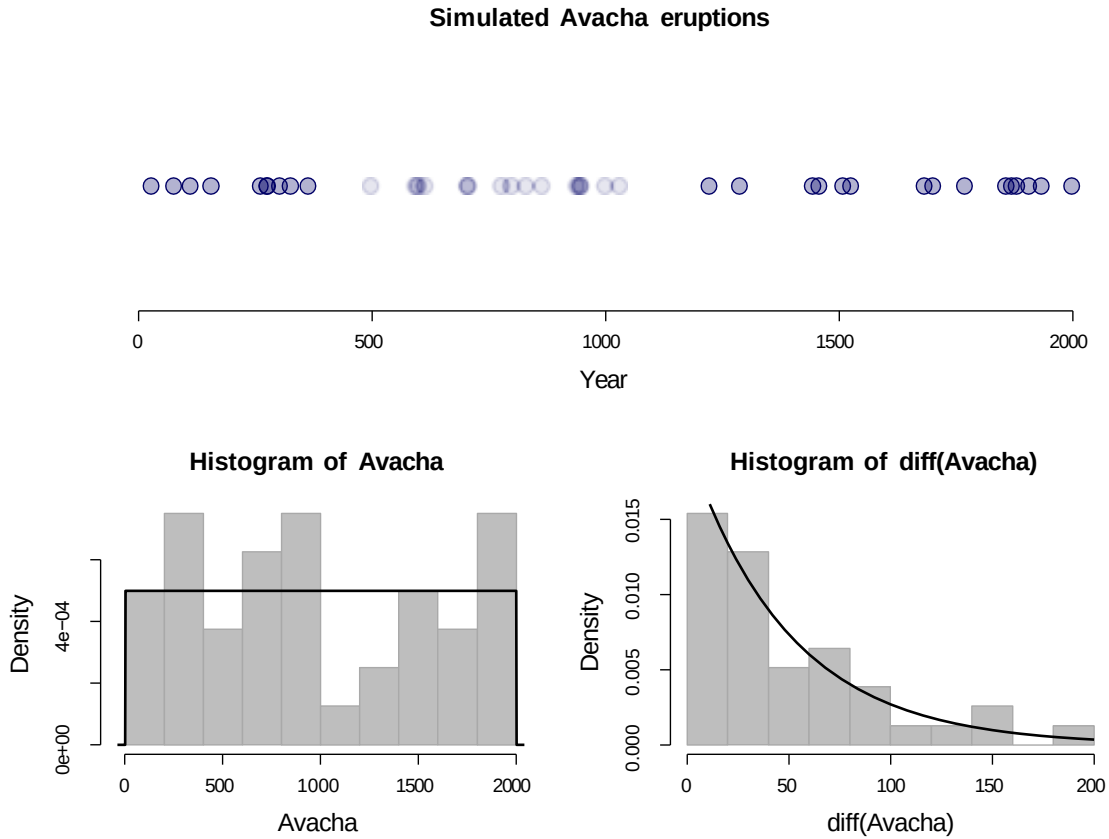
Relationship between exponential and uniform distributions





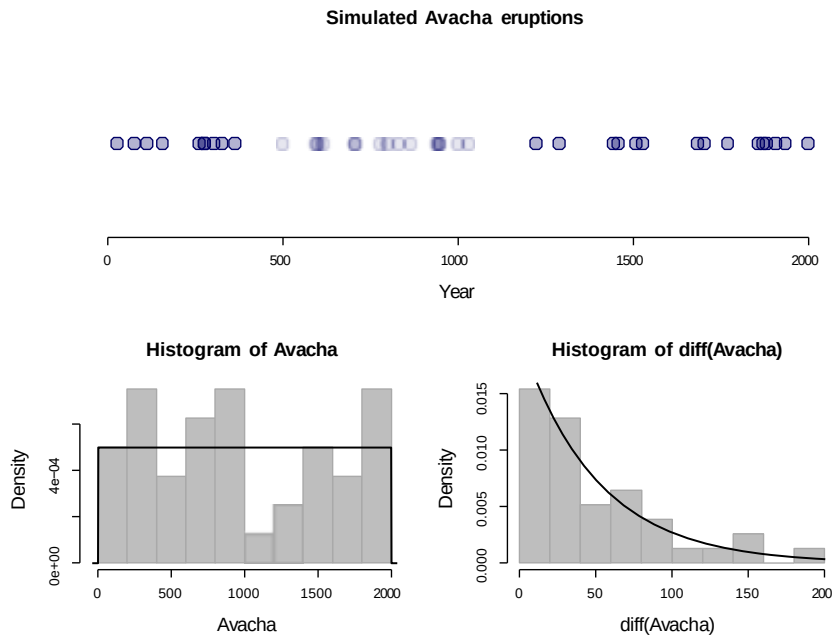

```
> Avacha <- runif(40,0,1000)
```

Relationship between exponential and uniform distributions




```
> Avacha <- runif(40,0,1000)
```

Relationship between exponential and uniform distributions



Important fact: If $X_{(1)}, X_{(2)}, X_{(3)} \ldots X_{(n)}$ are ordered uniform r.v.'s: $X \sim \text{Unif}(0, l)$, then:

$$W_i = X_{i+1} - X_i \sim Exp\left(\gamma = \frac{n}{l}\right)$$

Complete the following

- using runif–generate random numbers from the exponential distribution with a mean of 50 (that is, a process with an average of 50 years between events)
- generate random numbers from the exponential distributions with a mean of 50 using rexp
- plot the density histograms and overlay a line corresponding to the PDF (using dexp)

to make this question easier to understand, let's say we're modeling volcano eruptions. Let's pick an arbitrary year range over which to generate some eruptions, say 1 million years.

If volcanoes erupt on average every 50 years over this million year period, how many eruptions would we expect over this million year period?

If volcanoes erupt on average every 50 years over this million year period, how many eruptions would we expect over this million year period?

$$\frac{1000000\ years}{50\ years/event} = 20000\ events$$

so how would we generate 20000 events randomlly over a 1 million year period using an R function?

```
> x <- runif(1e+06/50, 0, 1e+06)
> head(x)
[1] 281444.9 483535.9 460420.5 283742.0 665149.2
[6] 418981.8
```

so if the goal is to generate waiting time, we need to order these events then take the difference of successive events. How might we do that in R?

```
> w1 <- diff(sort(x))
```

I call this variable w1 for waiting period.

How would we generate waiting periods from the exponential distribution directly (using the same average waiting time of 50)? note the r functions are rexp, dexp, pexp

Note that in R, the exponential distribution is parameterized by "rate"

```
> args(rexp)
function (n, rate = 1)
NULL
```

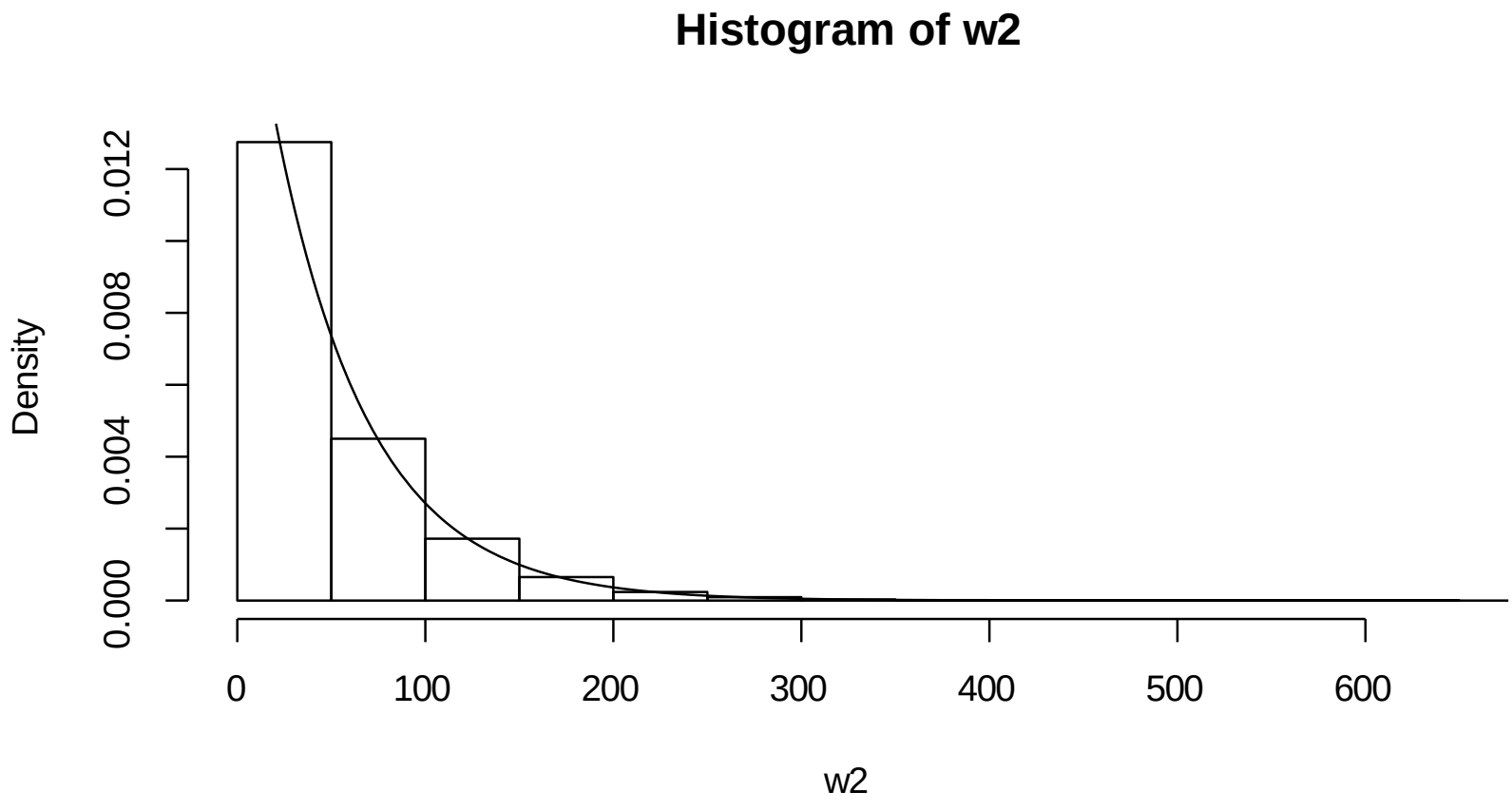
which corresponds to 1/(average wait time between events).

Lesson: some people parameterize the exponential distribution by the mean and some use 1/mean. Be careful!

```
> w2 <- rexp(20000, 1/50)
> head(w2)
[1]  30.158589   4.751406  14.046046 104.192785
[5]  19.588154  49.245431
```

How would we plot the histogram of w2 with the true density fuction overlaid on that histogram?

```
> x <- 0:1000
> hist(w2, freq = FALSE)
> lines(x, dexp(x, 1/50))
```

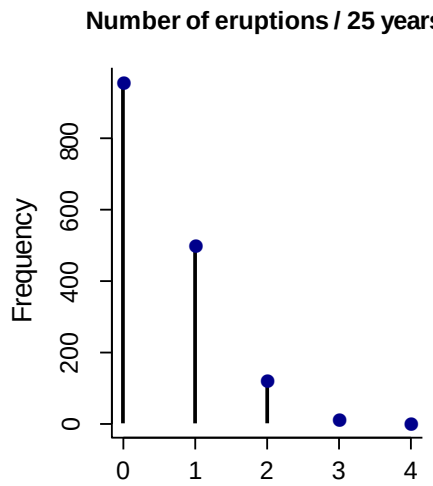


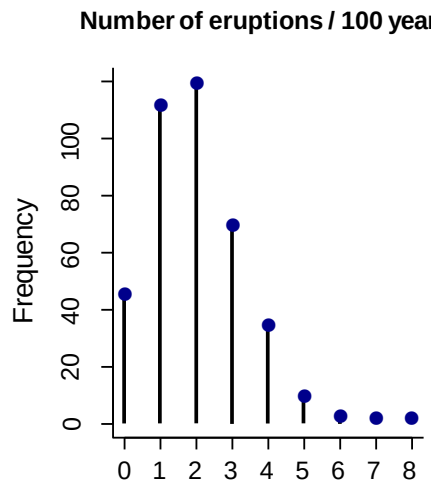
We were able to generate a useful distribution describing waiting times from a continuous uniform distribution.

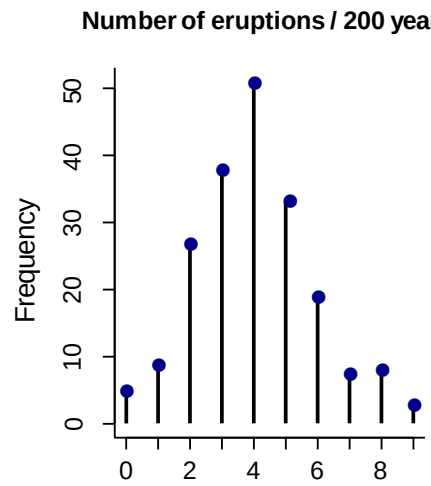
Next, we'll show another useful distributiod derived from the continuous uniform distribution.

More questions about exponential distribution

Ok, if Avacha erupts randomly with mean 50 years. How many eruptions can we *expect* in a century? 200 years? 25 years?







Complete the following

- Using randomly generated data (from runif) representing a 10000-year period with an average wait time of 50 years, create a plot showing the distribution of number of events in 100-year periods.

Given a vector of time-points, how would we divide this data into 100 year periods

```
> set.seed(12)
> times <- runif(200, 0, 10000)
> head(times)
[1]  693.6092 8177.7520 9426.2173 2693.8188
[5] 1693.4812  338.9562
```

```
> cuttime <- cut(times, breaks = seq(0, 10000, by = 100))
> head(cuttime)
[1] (600,700]         (8.1e+03,8.2e+03]
[3] (9.4e+03,9.5e+03] (2.6e+03,2.7e+03]
[5] (1.6e+03,1.7e+03] (300,400]
100 Levels: (0,100] (100,200] ... (9.9e+03,1e+04]
```

Each observation is now represented as categorical 100-interval. So 693 becomes 600-700, 8177 becomes 8100-8200, etc

How would we count up the number of observations occuring in each 100-year interval?

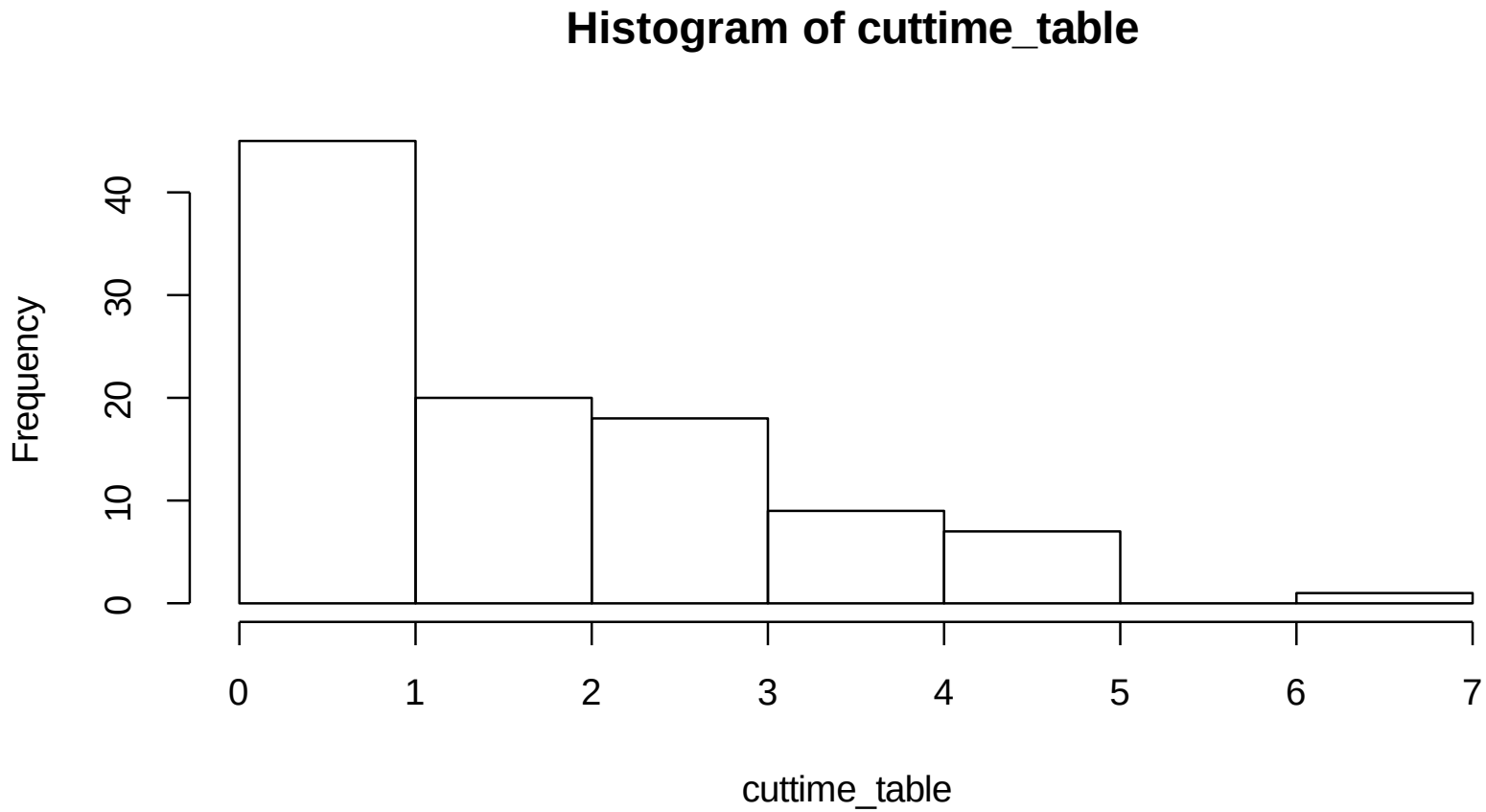
```
>head(table(cuttime))
cuttime
  (0,100] (100,200] (200,300] (300,400] (400,500]
        1         1         1         1         2
(500,600]
        1
> cuttime_table <- tabulate(cuttime)
> head(cuttime_table)
[1] 1 1 1 1 2 1
```

Our new variable is now represents the counts of occurances in every 100-year interval.

Note that the `tabulate` function in R performs the same operation as `table` but returns a simple unnamed vector.

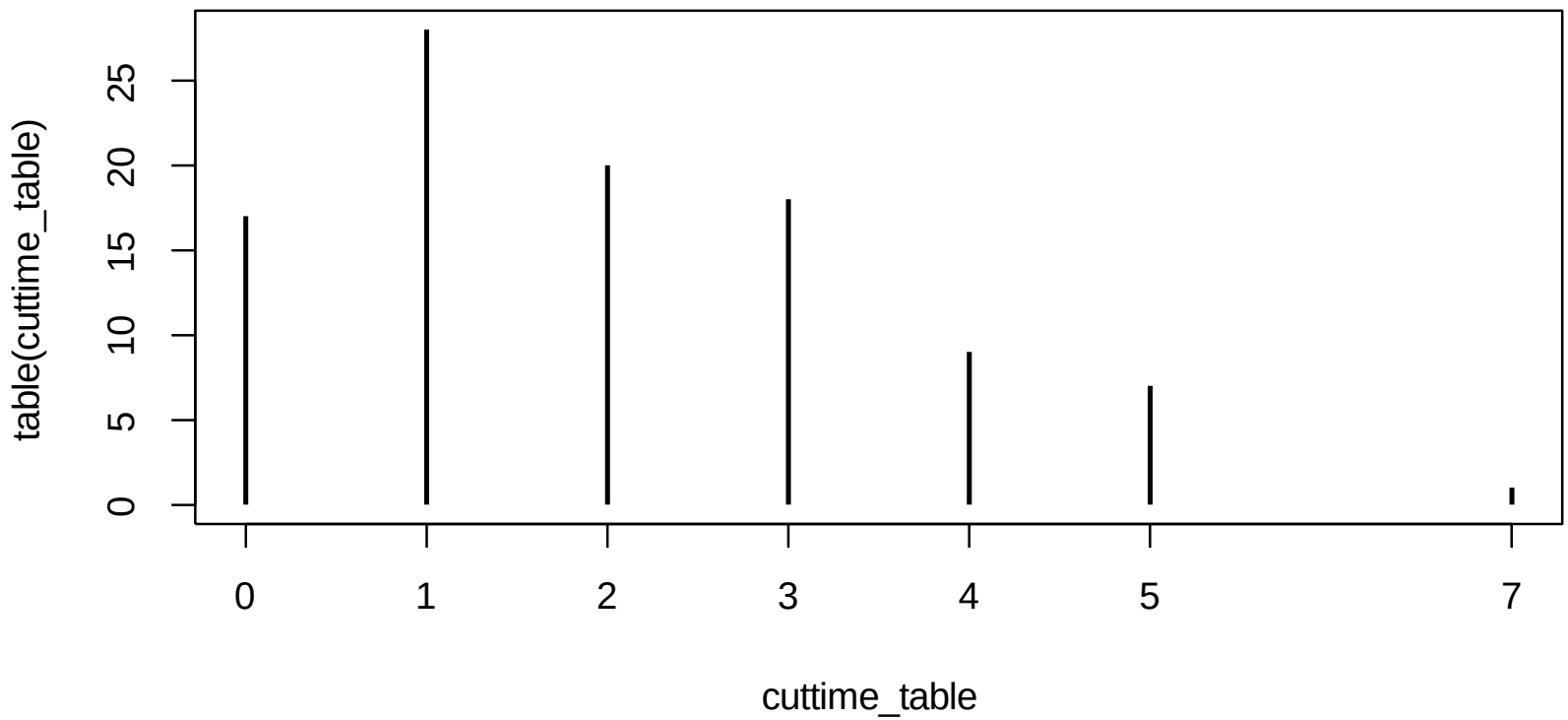
this variable cutttime_table follows a distribution, where some proportion of intervals have 0 events, some have 1 event, etc

```
> hist(cuttime_table)
```



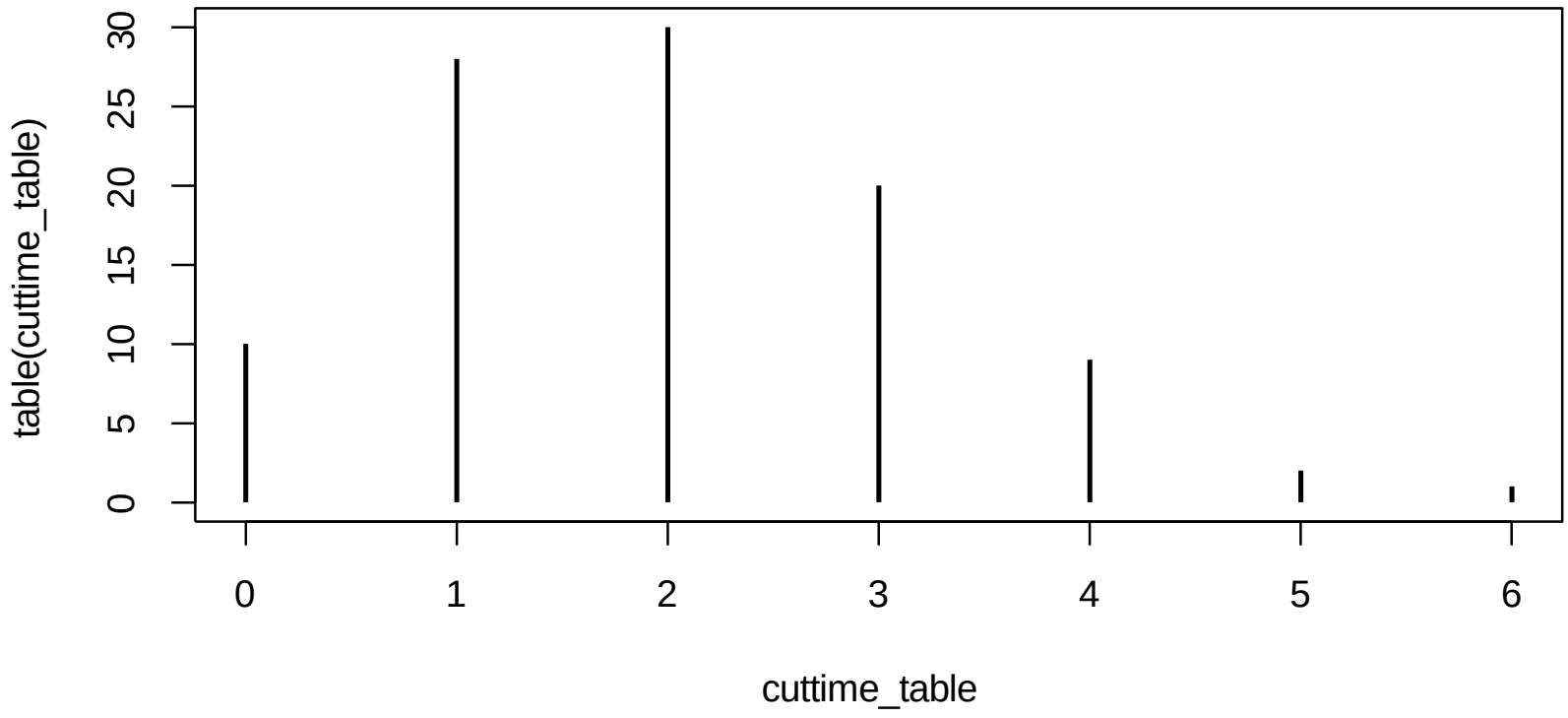
If we take table(cuttime_table) then plot it, we basically get something like a discrete histogram

```
> plot(table(cuttime_table))
```



All the code put together:

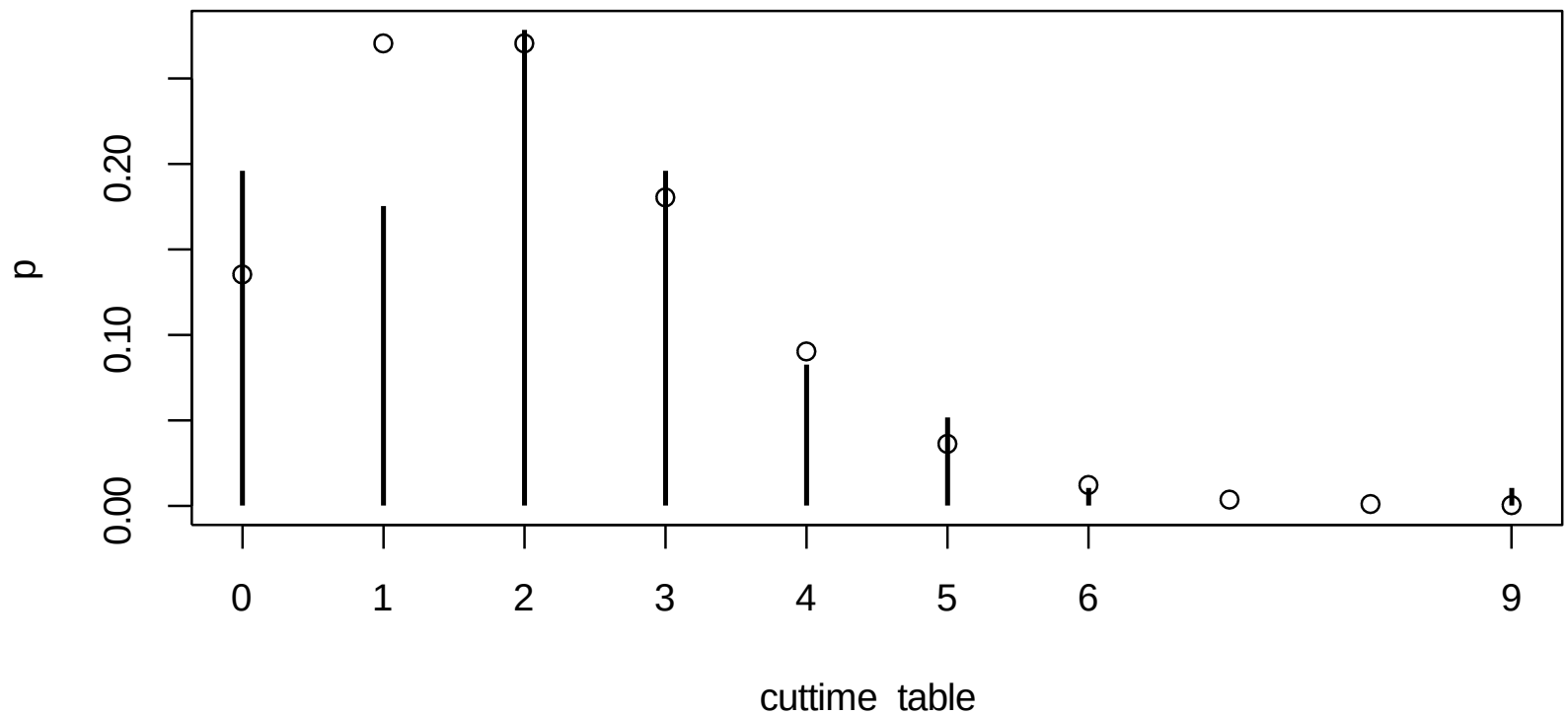
```
> times <- runif(200, 0, 10000)
> cuttime <- cut(times, breaks = seq(0, 10000, by = 100))
> cuttime_table <- tabulate(cuttime)
> plot(table(cuttime_table))
```



How might we add the true distribution as points overlayed on top

All the code put together:

```
> times <- runif(200, 0, 10000)
> cuttime <- cut(times, breaks=seq(0,10000,by=100))
> cuttime_table <- tabulate(cuttime)
> plot(table(cuttime_table)/sum(table(cuttime_table)),
+      ylab="p")
> points(0:10, dpois(0:10,lambda=100/50))
```



Note that

$$\lambda = \frac{100}{50} = 2$$

Where 100 was the interval size and 1/50 was the number of events per unit period (200 events over a 10000 year period).

2 is the mean of this distribution

```
> mean(cuttime_table)
[1] 2.061856
```

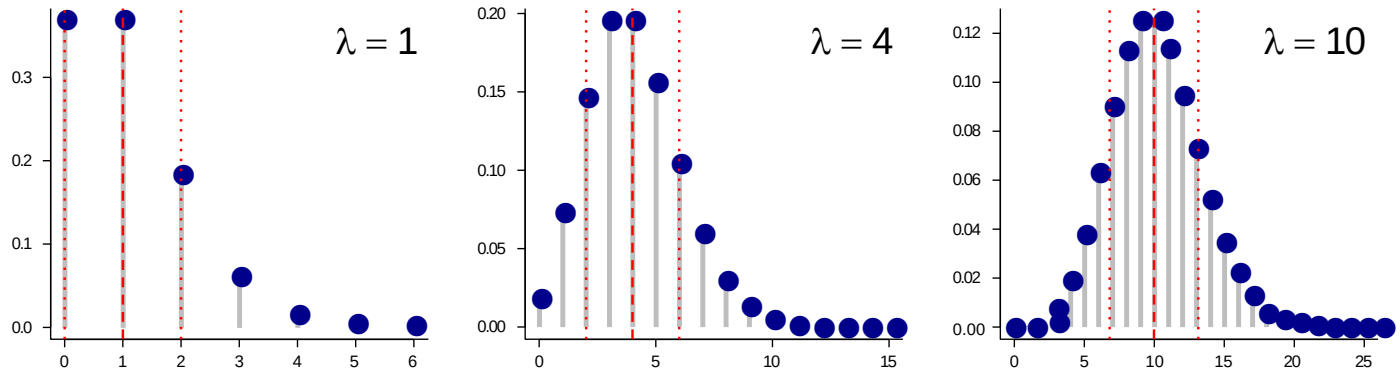
Poisson distribution

- Describes the number of times a random, independent event occurring with constant intensity in a given interval of time or space

$$\Pr[X = k] = \frac{\lambda^k}{k!} e^{-\lambda}$$

where λ = rate or intensity of occurrence

Poisson process



For $X \sim \text{Poisson}(\lambda)$

$$E[X] = \lambda$$
$$\text{Var}[X] = \lambda$$

i.e.: **Variance = Mean = Intensity!**
This is a very useful property for determining whether the Poisson distribution is an appopriate model.

$X \sim \text{Poisson}(\lambda)$

Name:	Poisson distribution	Models frequency of discrete random events
Support:	x	$\{0, 1, 2, 3, ...\}$
Parameters:	$\lambda \in (0, \infty)$	intensity parameter
pdf	$P(X = k \mid \lambda) =$	$\frac{\lambda^k}{k!} e^{-\lambda}$
mean	$E(X)$	λ
variance	$\text{Var}(X)$	λ

Historical Aside

RECHERCHES

PROBABILITÉ DES JUGEMENTS

EN MATIÈRE CRIMINELLE

ET EN MATIÈRE CIVILE,

Par S.-D. POISSON.

Le calcul des probabilités s'applique également aux choses de toute espèce, morales ou physiques, et ne dépend aucunement de leur nature, pourvu que dans chaque cas, l'observation fournisse les données numériques, nécessaires à ses applications.

Siméon Denis Poisson (1781-1840) - French physicist and mathematician, developed the Poisson distribution to model the number of convictions in the civil courts in France, noting in 1837 that:

"The science of probability can be applied to any subject - be they moral or physical - regardless of their nature, as long as the observations provide the numerical data required for its application."

While Loops

- While loops are blocks of code that get rerun until a condition is met (technically, until the while statement evaluates to false)
- While loops are useful for situations where you don’t know how many times a block of code needs to run before stopping, but there exists a clear stopping point.
- If you’re not careful, a while loop can run forever

Structure of a while loop:

- There must be a length-1 object that is used as an argument to the while() function.
- that object must start out as TRUE
- the code that goes into the ”body” of the while loop will be evaluated repeatedly until it changes the argument of the while function to FALSE

```
> z <- TRUE
> while(z){
+   #code to be repeatedly evaluated
+   #this code should eventually changes the value of z to FALSE
+ }
```

If I were to flip a coin until it landed on ”Heads”, how many tries would it take?

```
> x <- 0
> counter <- 0
> while (x < 0.5) {
+     x <- runif(1)  #getting x >= .5 would correspond to 'Heads
+     counter <- counter + 1
+ }
> counter  #a single value from a geometric(p=.5) distribution
[1] 2
```

It took 2 ”Bernoulli trials” (coin flips, for example) to get a success (Heads).

Exercise: find the first 20 prime numbers using a while loop and the is.prime function. start at 2.

```
> is.prime <- function(x) {
+     x == 2 | all(x%%2:(x - 1) != 0)
+ }
```

```
> is.prime <- function(x) {
+     x == 2 | all(x%%2:(x - 1) != 0)
+ }
> primes <- vector()
> n <- 2
> while (length(primes) < 20) {
+     if (is.prime(n))
+         primes[length(primes) + 1] <- n
+     n <- n + 1
+ }
> primes
 [1]  2  3  5  7 11 13 17 19 23 29 31 37 41 43 47
[16] 53 59 61 67 71
```