

This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike License](https://creativecommons.org/licenses/by-nc-sa/4.0/). Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.



Copyright 2006, The Johns Hopkins University and Karl Broman. All rights reserved. Use of these materials permitted only in accordance with license rights granted. Materials provided "AS IS"; no representations or warranties provided. User assumes all responsibility for use, and all liability related thereto, and must independently review all materials for accuracy and efficacy. May contain materials owned by others. User is responsible for obtaining permissions for use from third parties as needed.

Where are we going?

Deer ticks: Are they attracted by deer-gland-substance?

Suppose that 21 out of 30 deer ticks go to the deer-gland-substance-treated rod, while the other 9 go to the control rod.

Would this be a reasonable result if the deer ticks were choosing between the rods completely at random?

Mouse survival following treatment: does the treatment have an effect?

Suppose that 15/30 control mice die, while 8/30 treatment mice die.

Is the probability that a control mouse dies the same as the probability that a treatment mouse dies?

A brief review

Random experiment: A well-defined process with an uncertain outcome.

Sample space (\mathcal{S}): The set of all possible outcomes of the experiment.

Event: A subset of \mathcal{S} .

Probabilities are assigned to events.

Conditional probability: $\Pr(A \mid B) = \Pr(A \text{ and } B) / \Pr(B)$

A and B are **independent** if $\Pr(A \text{ and } B) = \Pr(A) \Pr(B)$.

Random variables

Random variable: A number assigned to each outcome of a random experiment.

Example 1: **I toss a brick at my neighbor's house.**

D = distance the brick travels

X = 1 if I break a window; 0 otherwise

Y = cost of repair

T = time until the police arrive

N = number of people injured

Example 2: **Treat 10 spider mites with DDT.**

X = number of spider mites that survive

P = proportion of mites that survive.

Further examples

Example 3: **Pick a random student in the School.**

S = 1 if female; 0 otherwise

H = his/her height

W = his/her weight

Z = 1 if Canadian citizen; 0 otherwise

T = number of teeth he/she has

Example 4: **Sample 20 students from the School**

H_i = height of student i

\bar{H} = mean of the 20 student heights

S_H = sample SD of heights

T_i = number of teeth of student i

\bar{T} = average number of teeth

Random variables are ...

Discrete: Take values in a “countable” set (e.g., the positive integers).

For example: number of teeth, number of gall stones, number of birds, number of cells responding to a particular antigen, number of heads in 20 tosses of a coin.

Continuous: Take values in an interval (e.g., $[0,1]$ or the real line).

For example: height, weight, mass, measure of gene expression, blood pressure.

Random variables may also be **partly discrete and partly continuous** (for example, mass of gall stones, concentration of infecting bacteria).

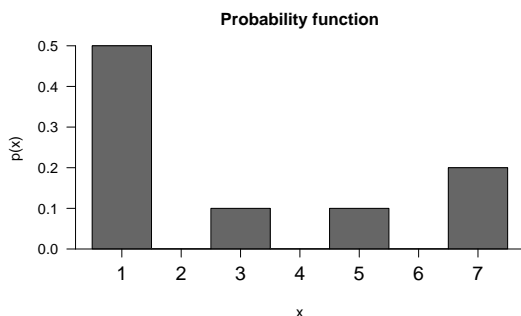
Probability function

Consider a *discrete* random variable, X .

The **probability function** (or probability distribution, or probability mass function) of X is

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

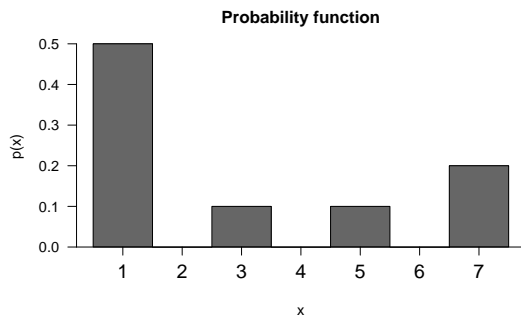
Note that $\{X = x\}$ is an *event*—a set of possible outcomes. Also $p(x) \geq 0$ for all x and $\sum p(x) = 1$.



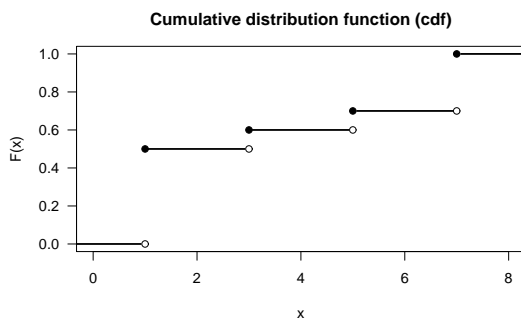
x	p(x)
1	0.5
3	0.1
5	0.1
7	0.3

Cumulative distribution function (cdf)

The cdf of X is $F(x) = \Pr(X \leq x)$



x	p(x)
1	0.5
3	0.1
5	0.1
7	0.2



x	F(x)
$(-\infty, 1)$	0
$[1, 3)$	0.5
$[3, 5)$	0.6
$[5, 7)$	0.7
$[7, \infty)$	1.0

Binomial random variable

Prototype:

The number of heads in n **independent** tosses of a coin, where $\Pr(\text{heads}) = p$ **for each toss**.

(n and p are called *parameters*.)

Alternatively, imagine an urn containing red balls and black balls, and suppose that p is the proportion of red balls. Consider the number of red balls in n random draws *with replacement* from the urn.

Example 1:

Sample n people at random from a large population, and consider the number of people with some property (e.g., that are graduate students or that have exactly 32 teeth).

Example 2:

Apply a treatment to n mice and count the number of survivors (or the number that are dead).

Example 3:

Apply a large dose of DDT to 30 groups of 10 spider mites. Count the number of groups with at least two surviving spider mites.

Binomial distribution

Consider the binomial(n, p) distribution. (The number of red balls in n draws with replacement from an urn for which the proportion of red balls is p .)

What is its probability function?

Consider the case $p = 20\%$ and $n = 9$.

Let $X \sim \text{binomial}(n = 9, p = 0.2)$.

We seek $p(x) = \Pr(X = x)$ for $x = 0, 1, 2, \dots, 9$.

$$p(0) = \Pr(X = 0) = \Pr(\text{no red balls}) = (1 - p)^n = 0.8^9 \approx 13\%.$$

$$p(9) = \Pr(X = 9) = \Pr(\text{all red balls}) = p^n = 0.2^9 \approx 5 \times 10^{-7}$$

$$p(1) = \Pr(X = 1) = \Pr(\text{exactly one red ball}) = \dots$$

Binomial distribution

$$p(1) = \Pr(X = 1) = \Pr(\text{exactly one red ball})$$

$$= \Pr(\text{RBBBBBBBBB or BRBBBBBBBB or ... or BBBBBBBBBR})$$

$$\begin{aligned} &= \Pr(\text{RBBBBBBBBB}) + \Pr(\text{BRBBBBBBBB}) + \Pr(\text{BBRBBBBBBB}) \\ &\quad + \Pr(\text{BBBRBBBBBB}) + \Pr(\text{BBBBRBBBBB}) \\ &\quad + \Pr(\text{BBBBBBRBBB}) + \Pr(\text{BBBBBBRBBB}) \\ &\quad + \Pr(\text{BBBBBBBBRB}) + \Pr(\text{BBBBBBBBBR}) \end{aligned}$$

$$= p(1 - p)^8 + p(1 - p)^8 + \dots + p(1 - p)^8 = 9p(1 - p)^8 \approx 30\%.$$

How about $p(2) = \Pr(X = 2)$?

How many outcomes have 2 red balls among the 9 balls drawn? (This is a problem of **combinatorics**; that is, counting.)

Getting at $\Pr(X = 2)$

RRBBBBBBB RBRBBBBBBB RBBRBBBBBB RBBBRBBBBB
RBBBBBRBBB RBBBBBBRBB RBBBBBBBRB RBBBBBBBRR
BRRBBBBBBB BRBRBBBBBB BRBBRBBBBB BRBBBRBBBB
BRBBBBRBBB BRBBBBBBRB BRBBBBBBBR BBRRBBBBBB
BBRBRBBBBB BBRBBRBBBB BBRBBBRBBB BBRBBBBRBB
BBRBBBBBBR BBBRRBBBBB BBBRBRBBBB BBBRBBRBBB
BBBRBBBBRB BBBRBBBBBR BBBBRRBBBB BBBBRBRBBB
BBBBRBBRBB BBBBRBBBBR BBBBBRRBBB BBBBBRBRBB
BBBBBRBBRR BBBBBBRRRB BBBBBBRBRB BBBBBBRRR

How many are there?

$$9 \times 8 / 2 = 36.$$

The binomial coefficient

“the number of possible samples of size k
from a population of size n ”

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

where $n! = n \times (n-1) \times (n-2) \times \dots \times 3 \times 2 \times 1$

with $0! = 1$

For a binomial(n, p) random variable,

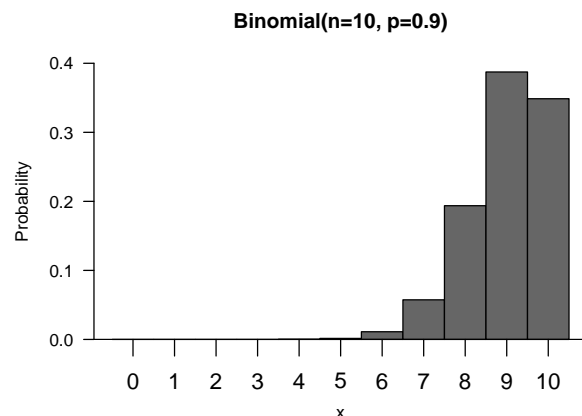
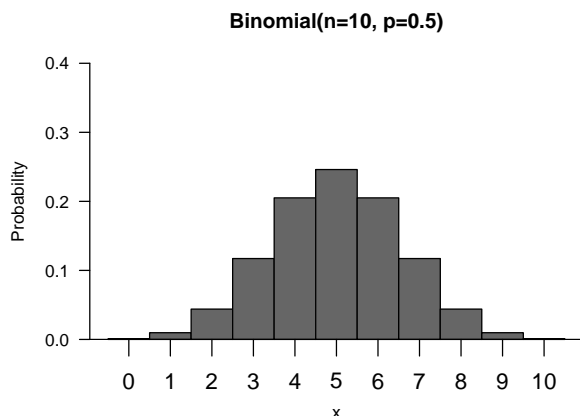
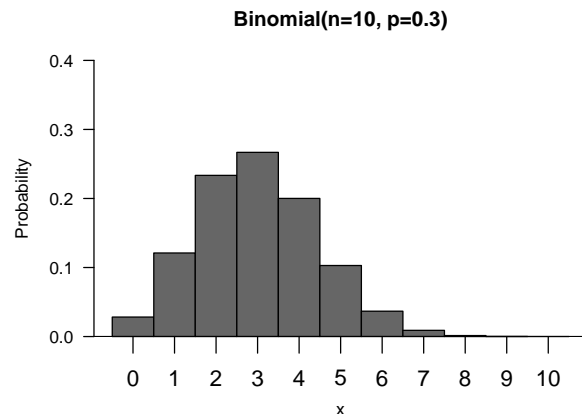
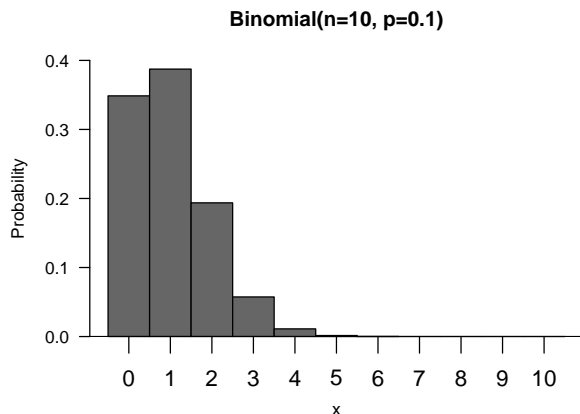
$$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{(n-k)}$$

Example

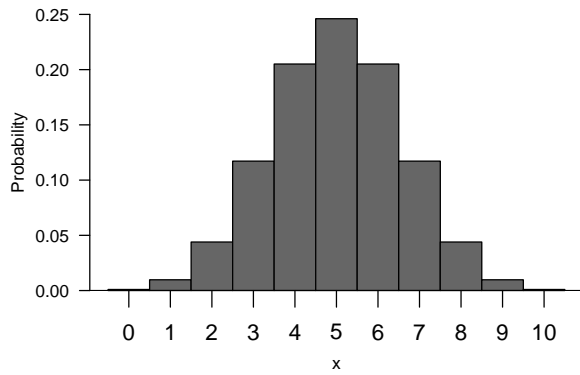
Suppose $\Pr(\text{mouse survives treatment}) = 90\%$, and we apply the treatment to 10 random mice.

$$\begin{aligned}\Pr(\text{exactly 7 mice survive}) &= \binom{10}{7}(0.9)^7(0.1)^3 \\ &= \frac{10 \times 9 \times 8}{3 \times 2} (0.9)^7(0.1)^3 \\ &= 120 (0.9)^7(0.1)^3 \\ &\approx 5\%.\end{aligned}$$

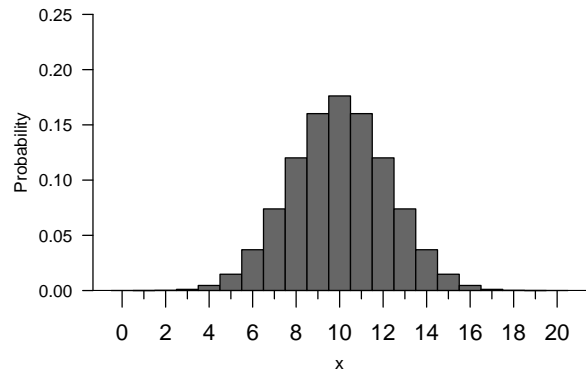
$$\begin{aligned}\Pr(\text{fewer than 9 survive}) &= 1 - p(9) - p(10) \\ &= 1 - 10 (0.9)^9(0.1) - (0.9)^{10} \\ &\approx 26\%.\end{aligned}$$



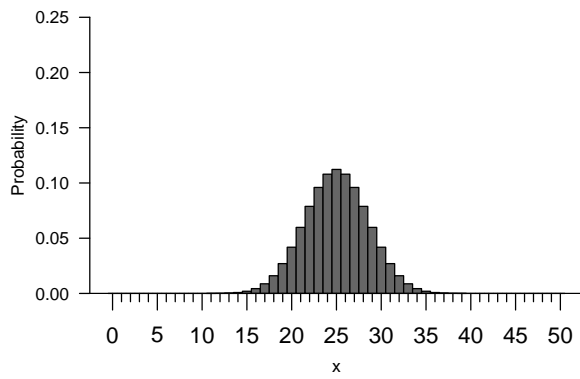
Binomial(n=10, p=0.5)



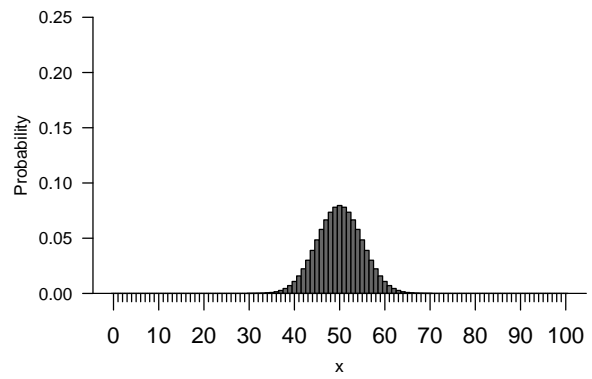
Binomial(n=20, p=0.5)



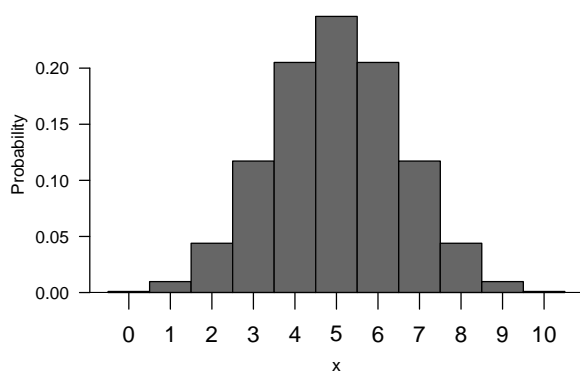
Binomial(n=50, p=0.5)



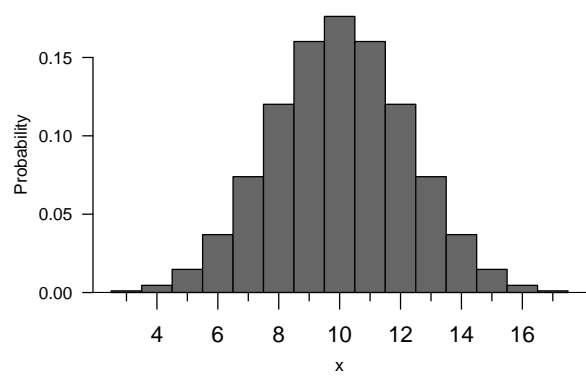
Binomial(n=100, p=0.5)



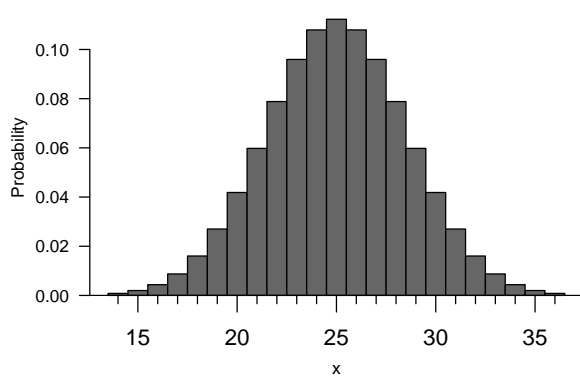
Binomial(n=10, p=0.5)



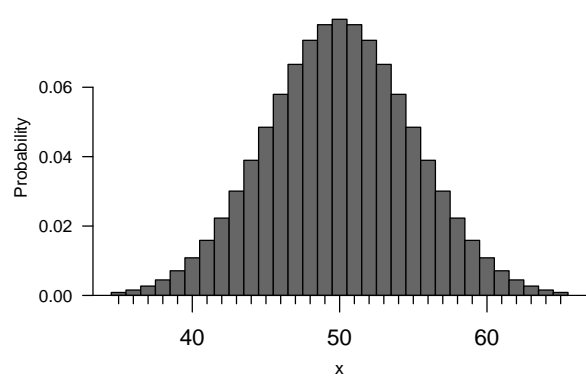
Binomial(n=20, p=0.5)

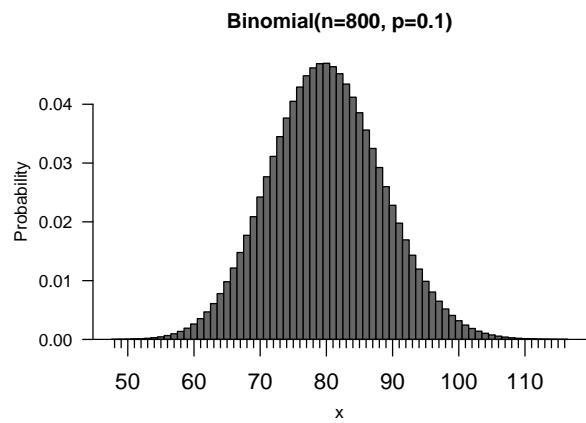
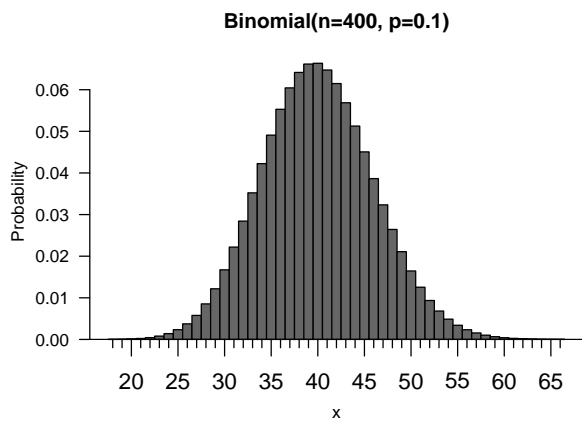
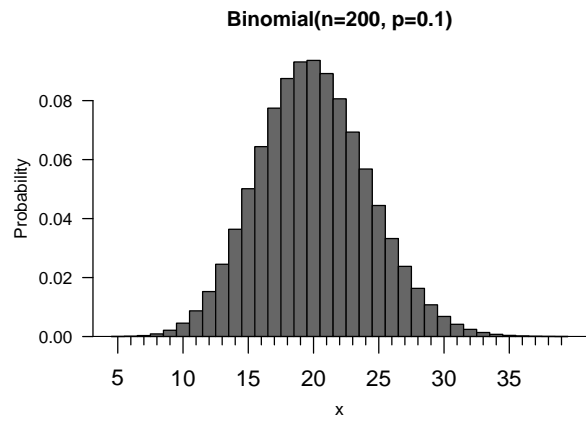
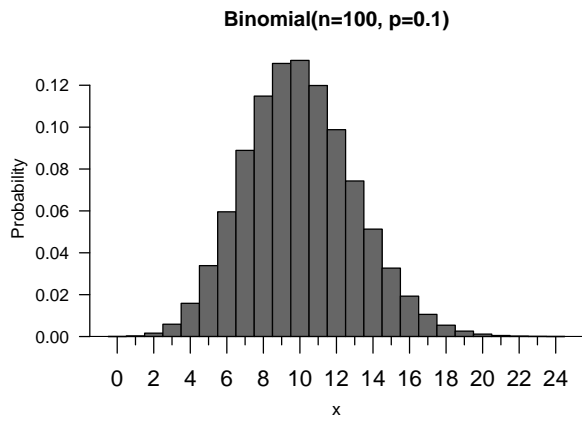
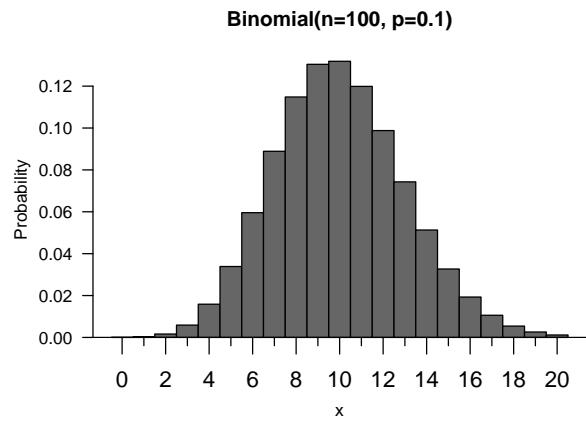
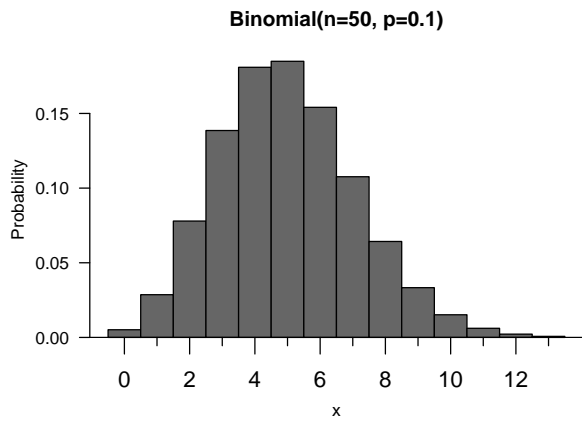
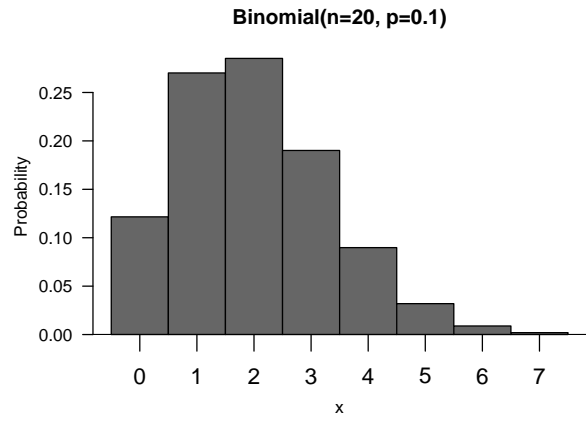
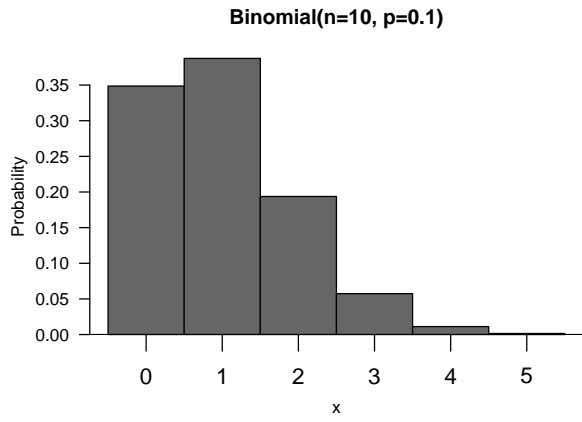


Binomial(n=50, p=0.5)



Binomial(n=100, p=0.5)





Expected value and SD

The **expected value** (or mean) of a discrete random variable, X , with prob. fctn. $p(x)$, is

$$\mu = \mathbf{E}(X) = \sum_x x p(x)$$

The **variance** of a discrete random variable, X , with prob. fctn. $p(x)$, is

$$\sigma^2 = \mathbf{var}(X) = \sum_x (x - \mu)^2 p(x)$$

The **standard deviation (SD)** of X is

$$\mathbf{SD}(X) = \sqrt{\mathbf{var}(X)}.$$

Mean and SD of binomial RV

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

$$\mathbf{E}(X) = n p$$

$$\mathbf{SD}(X) = \sqrt{n p (1 - p)}$$

Examples:

n	p	mean	SD
10	10%	1	0.9
10	30%	3	1.4
10	50%	5	1.6
10	90%	9	0.9

Calculations within R

<code>rbinom(m, size, prob)</code>	Simulate binomial random variables
<code>dbinom(x, size, prob)</code>	The binomial probability function: $\Pr(X = x)$
<code>pbinom(q, size, prob)</code>	The binomial CDF: $\Pr(X \leq q)$
<code>qbinom(p, size, prob)</code>	The inverse of the CDF the smallest q such that $\Pr(X \leq q) \geq p$

Binomial random variable

Number of **successes** in n **trials** where:

- Trials **independent**
- $p = \Pr(\text{success})$ is **constant**

The number of successes in n trials does not necessarily follow a binomial distribution.

Deviations from the binomial:

- Varying p
- Clumping or repulsion (non-independence)

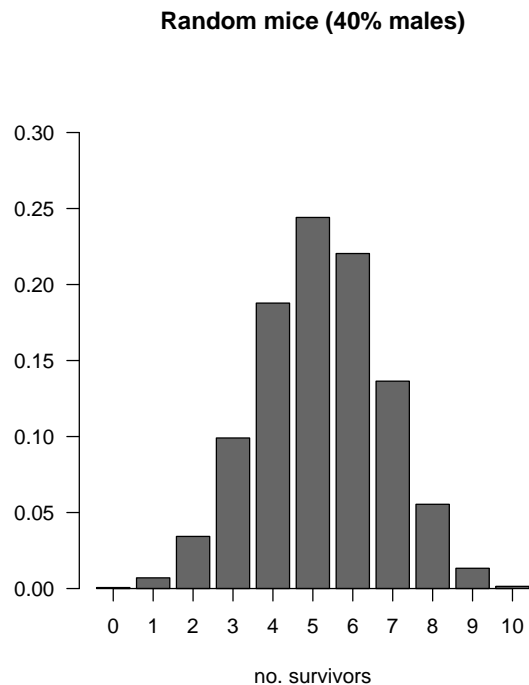
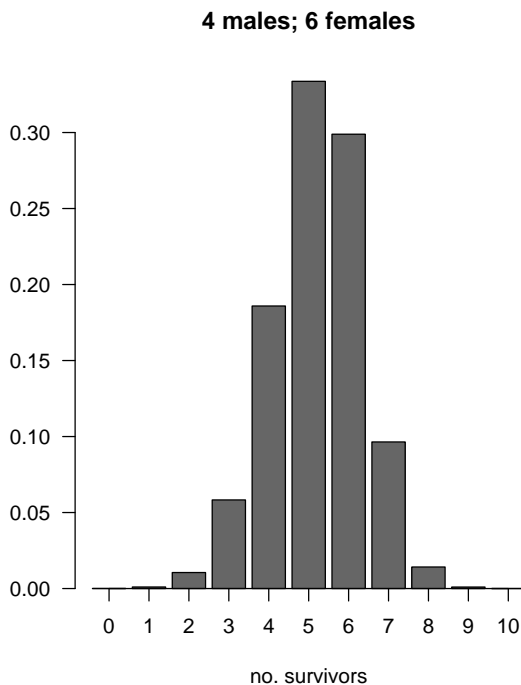
Examples

Consider Mendel's pea experiments. (Purple or white flowers; purple dominant to white.)

- Pick a random F_2 . Self it and acquire 10 progeny. The number of progeny with purple flowers is *not* binomial (unless we condition on the genotype of the F_2 plant).
- Pick 10 random F_2 's. Self each and take one child from each. The number of progeny with purple flowers *is* binomial. ($p = (1/4) \times 1 + (1/2) \times (3/4) + (1/4) \times 0 = 5/8$.)

Suppose $\Pr(\text{survive} \mid \text{male}) = 10\%$ but $\Pr(\text{survive} \mid \text{female}) = 80\%$.

- Pick 4 male mice and 6 female mice. The number of survivors is *not* binomial.
- Pick 10 random mice (with $\Pr(\text{mouse is male}) = 40\%$). The number of survivors *is* binomial.



Poisson distribution

Consider a binomial(n, p) where

- n is really large
- p is really small

For example, suppose each well in a microtiter plate contains 50,000 T cells, and that 1/100,000 cells respond to a particular antigen.

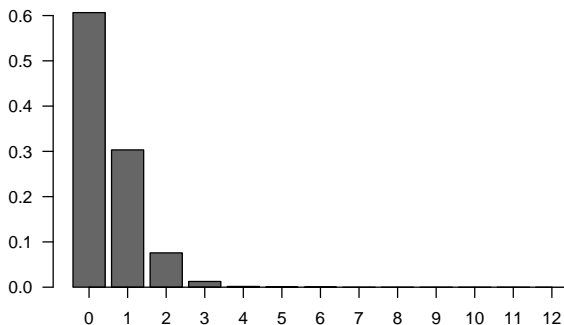
Let X be the number of responding cells in a well.

In this case, X follows a **Poisson** distribution.

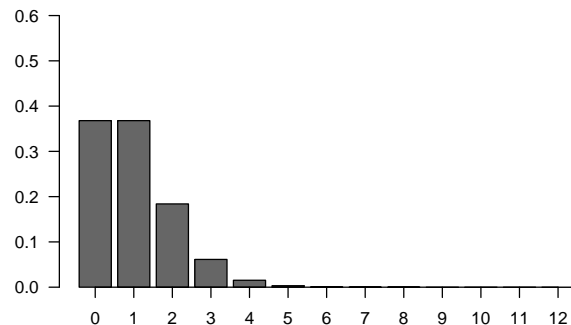
Let $\lambda = n \times p = E(X)$. Then $p(x) = \Pr(X = x) = e^{-\lambda} \lambda^x / x!$

Note that $SD(X) = \sqrt{\lambda}$.

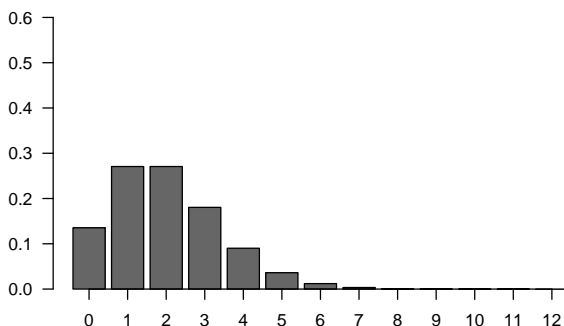
Poisson($\lambda=1/2$)



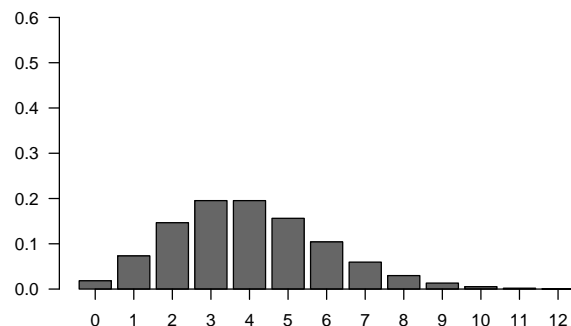
Poisson($\lambda=1$)



Poisson($\lambda=2$)



Poisson($\lambda=4$)



Example

Suppose there are 100,000 T cells in each well of a microtiter plate. Suppose that 1/80,000 T cells respond to a particular antigen.

Let X = number of responding T cells in a well.

$X \sim \text{Poisson}(\lambda = 1.25)$.

$$E(X) = 1.25; \text{SD}(X) = \sqrt{1.25} \approx 1.12.$$

$$\Pr(X = 0) = \exp(-1.25) \approx 29\%.$$

$$\Pr(X > 0) = 1 - \exp(-1.25) \approx 71\%.$$

$$\Pr(X = 2) = \exp(-1.25) (1.25)^2/2 \approx 22\%.$$

In R

The following functions act just like `rbinom`, `dbinom`, etc., for the binomial distribution:

```
rpois(m, lambda)
```

```
dpois(x, lambda)
```

```
ppois(q, lambda)
```

```
qpois(p, lambda)
```

$$Y = a + b X$$

Suppose X is a discrete random variable with probability function p , so that $p(x) = \Pr(X = x)$.

Expected value (mean): $E(X) = \sum_x x p(x)$

Standard deviation (SD): $SD(X) = \sqrt{\sum_x [x - E(X)]^2 p(x)}$

Let $Y = a + b X$ where a and b are numbers. Then Y is a random variable (like X), and

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

$$SD(Y) = |b| SD(X)$$

In particular, if $\mu = E(X)$, $\sigma = SD(X)$, and $Z = (X - \mu) / \sigma$, then

$$E(Z) = 0 \text{ and } SD(Z) = 1$$

Example

Suppose $X \sim \text{binomial}(n, p)$.

(The **number** of successes in n independent trials where $p = \Pr(\text{success})$.)

$$\text{Then } E(X) = n p \text{ and } SD(X) = \sqrt{n p (1 - p)}$$

Let $P = X / n = \text{proportion}$ of successes.

$$E(P) = E(X / n) = E(X) / n = p.$$

$$SD(P) = SD(X / n) = SD(X) / n = \dots = \sqrt{p (1 - p) / n}$$

Continuous random variables

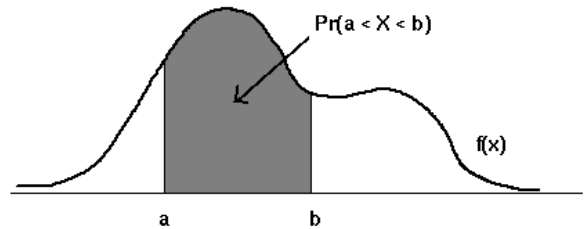
Suppose X is a **continuous** random variable.

Instead of a probability function, X has a **probability density function** (pdf), sometimes called just the **density** of X .

$$f(x) \geq 0$$

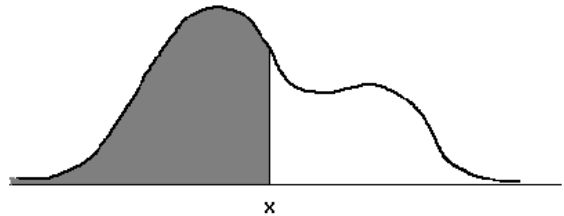
$$\int_{-\infty}^{\infty} f(x) dx = 1$$

Areas under curve = probabilities



Cumulative distr'n func'n (cdf):

$$F(x) = \Pr(X \leq x) =$$



Means and SDs

Expected value (mean):

$$\text{Discrete RV: } E(X) = \sum_x x p(x)$$

$$\text{Continuous RV: } E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

Standard deviation (SD):

$$\text{Discrete RV: } SD(X) = \sqrt{\sum_x [x - E(X)]^2 p(x)}$$

$$\text{Continuous RV: } SD(X) = \sqrt{\int_{-\infty}^{\infty} [x - E(X)]^2 f(x) dx}$$

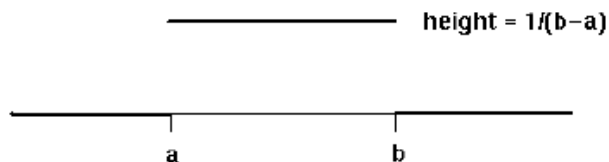
Example: Uniform distribution

$X \sim \text{Uniform}(a, b)$

i.e., draw a number at random from the interval (a, b) .

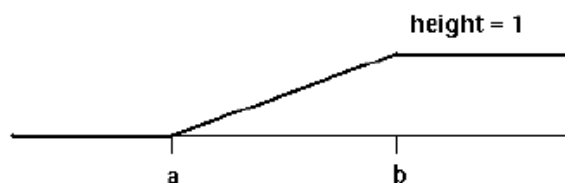
Density function:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}$$



$$E(X) = (b+a)/2 \quad SD(X) = (b-a)/\sqrt{12} \approx 0.29 \times (b-a)$$

Cumulative dist'n fdn (cdf):



The normal distribution

By far the most important distribution:

The Normal distribution

(also called the Gaussian distribution)

If $X \sim N(\mu, \sigma)$, then

$$\text{The pdf of } X \text{ is } f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Also $E(X) = \mu$ and $SD(X) = \sigma$.

Of great importance: If $X \sim N(\mu, \sigma)$ and $Z = (X - \mu) / \sigma$,

Then $Z \sim N(0, 1)$.

This is the "Standard normal distribution."

The normal curve

