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 | B) = Pr(A and B) / Pr(B)

A and B are independent if Pr(A 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 University.

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 University.

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).

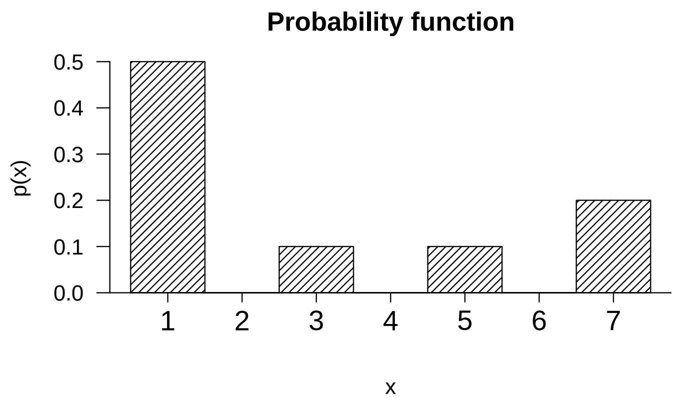
Probability function

Consider a *discrete* random variable, X.

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

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

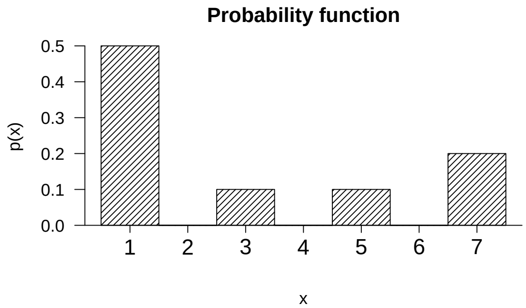
Note that $\{X = x\}$ is an *event*—a set of possible outcomes.
Also $p(x) \geq 0$ for all x and $\sum p(x) = 1$.



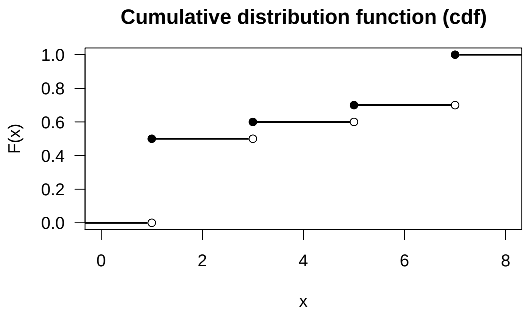
x	p(x)
1	0.5
3	0.1
5	0.1
7	0.3

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	0.3



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(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 $X \sim$ binomial($n = 9$, $p = 0.2$).

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

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

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

p(1) = Pr(X = 1) = Pr(exactly one red ball) = . . .

Binomial distribution

p(1) = Pr(X = 1) = Pr(exactly one red ball)

= Pr(RBBBBBBBB or BRBBBBBBB or . . . or BBBBBBBBR)

= Pr(RBBBBBBBB) + Pr(BRBBBBBBB) + Pr(BBRBBBBBB)
+ Pr(BBBRBBBBB) + Pr(BBBBRBBBB)
+ Pr(BBBBBRBBB) + Pr(BBBBBBRBB)
+ Pr(BBBBBBBRB) + Pr(BBBBBBBBR)

= $p(1-p)^8 + p(1-p)^8 + \ldots p(1-p)^8 = \mathbf{9p(1-p)^8} \approx$ 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	RBRBBBBBB	RBBRBBBBB	RBBBRBBBB
RBBBBRBBB	RBBBBBRBB	RBBBBBBRB	RBBBBBBBR
BRRBBBBBB	BRBRBBBBB	BRBBRBBBB	BRBBBRBBB
BRBBBBRBB	BRBBBBBRB	BRBBBBBBR	BBRRBBBBB
BBRBRBBBB	BBRBBRBBB	BBRBBBRBB	BBRBBBBRB
BBRBBBBBR	BBBRRBBBB	BBBRBRBBB	BBBRBBRBB
BBBRBBBRB	BBBRBBBBR	BBBBRRBBB	BBBBRBRBB
BBBBRBBRB	BBBBRBBBR	BBBBBRRBB	BBBBBRBRB
BBBBBRBBR	BBBBBBRRB	BBBBBBRBR	BBBBBBBRR

How many are there?

9 × 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 $n! = n \times (n-1) \times (n-2) \times \ldots \times 3 \times 2 \times 1$

with $0! = 1$

For a binomial(n,p) random variable,

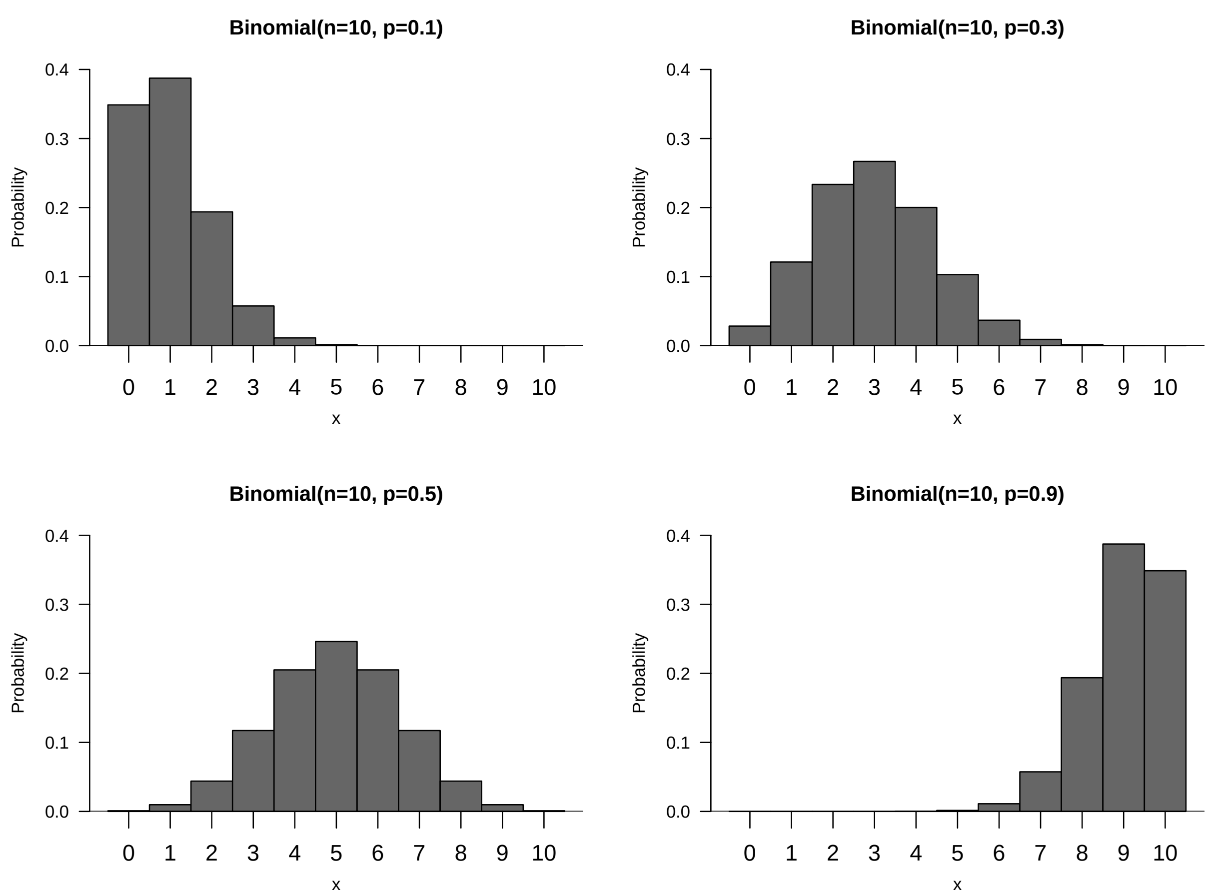
$$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{(n-k)}$$

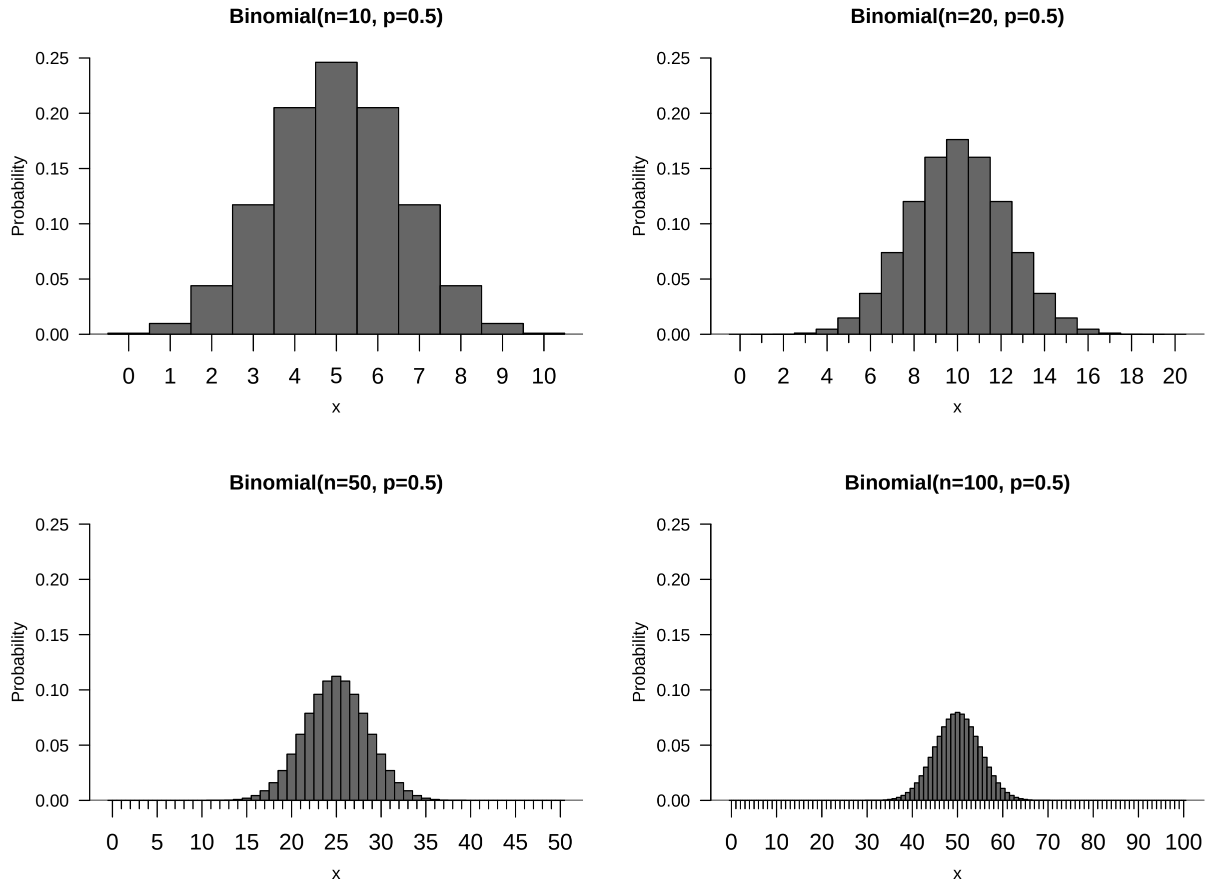
Example

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

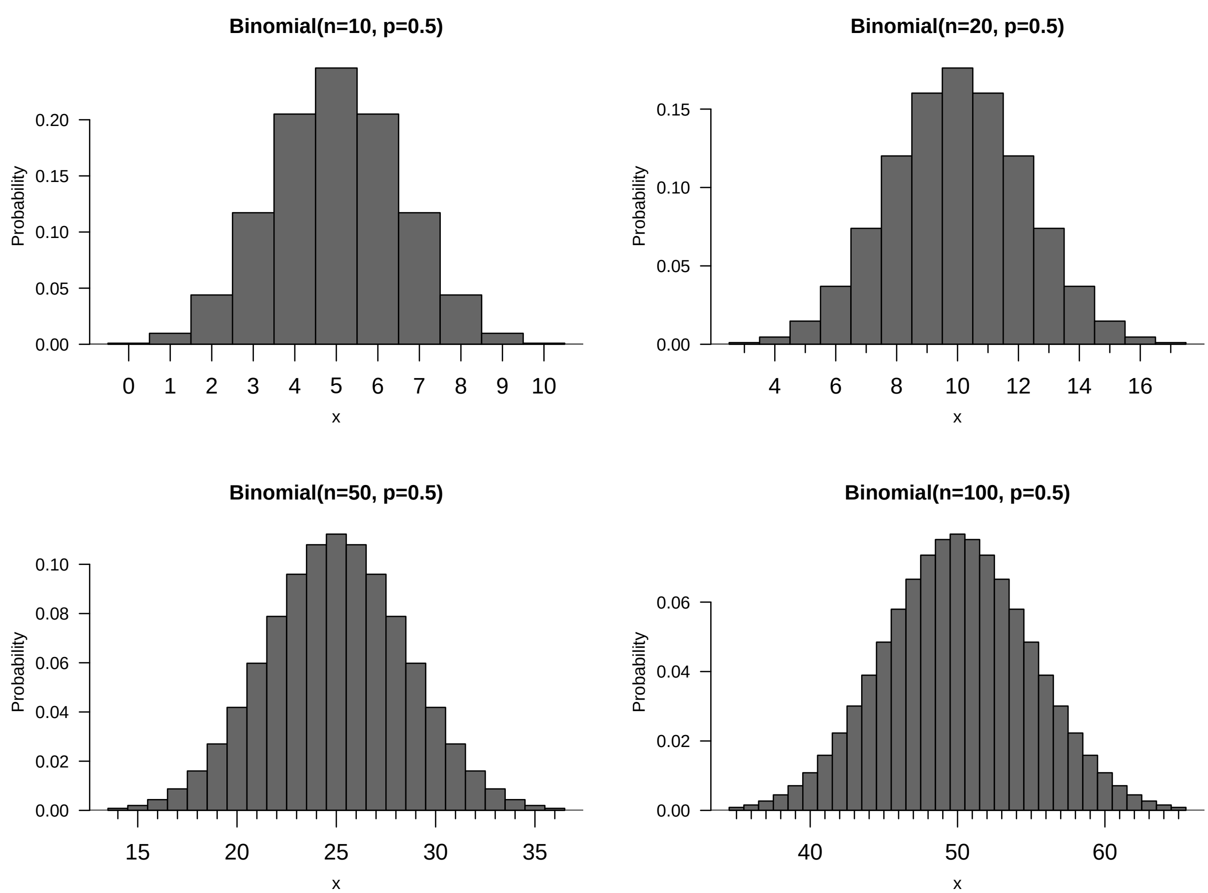
$$
\begin{aligned}
\text{Pr(exactly 7 mice survive)} &= \binom{10}{7}(0.9)^7(0.1)^3 \\
&= \frac{10\times9\times8}{3\times2}\,(0.9)^7(0.1)^3 \\
&= 120\,(0.9)^7(0.1)^3 \\
&\approx 5\%.
\end{aligned}
$$

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

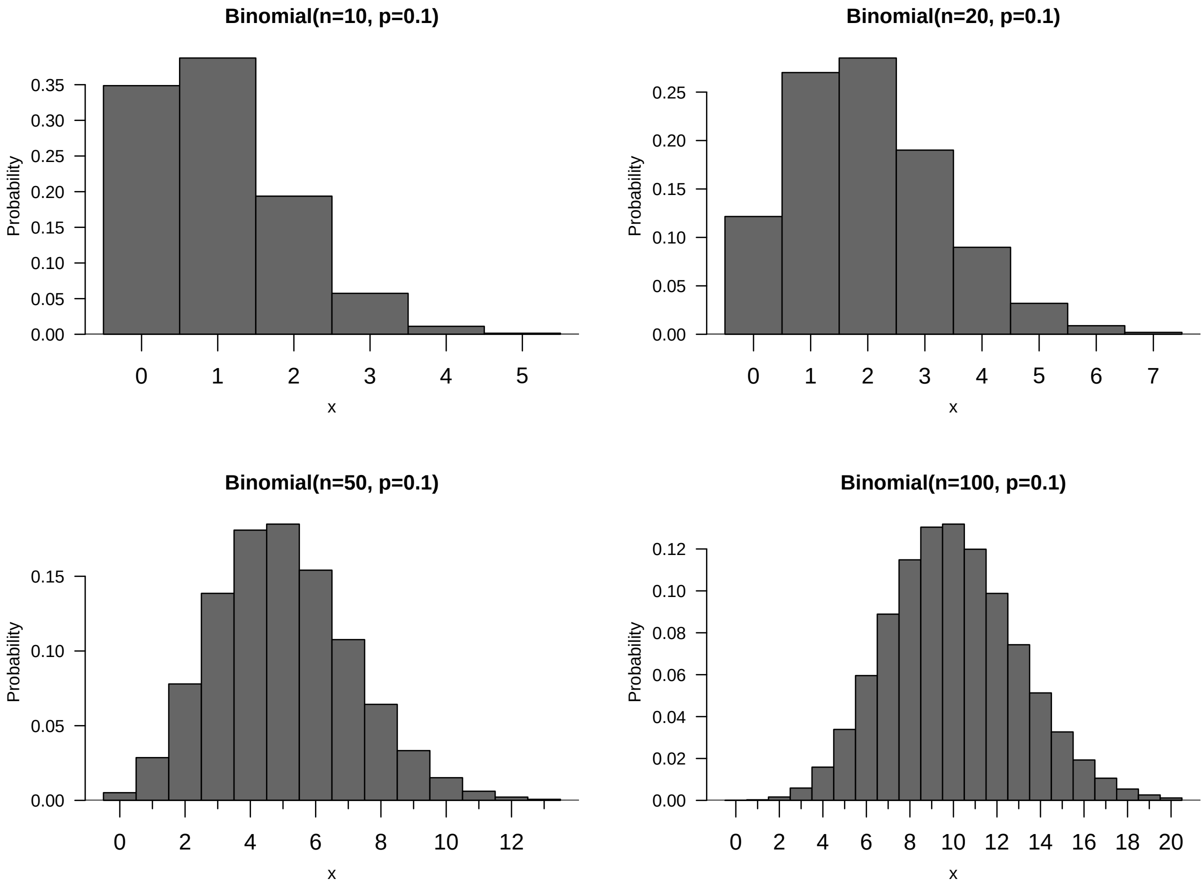




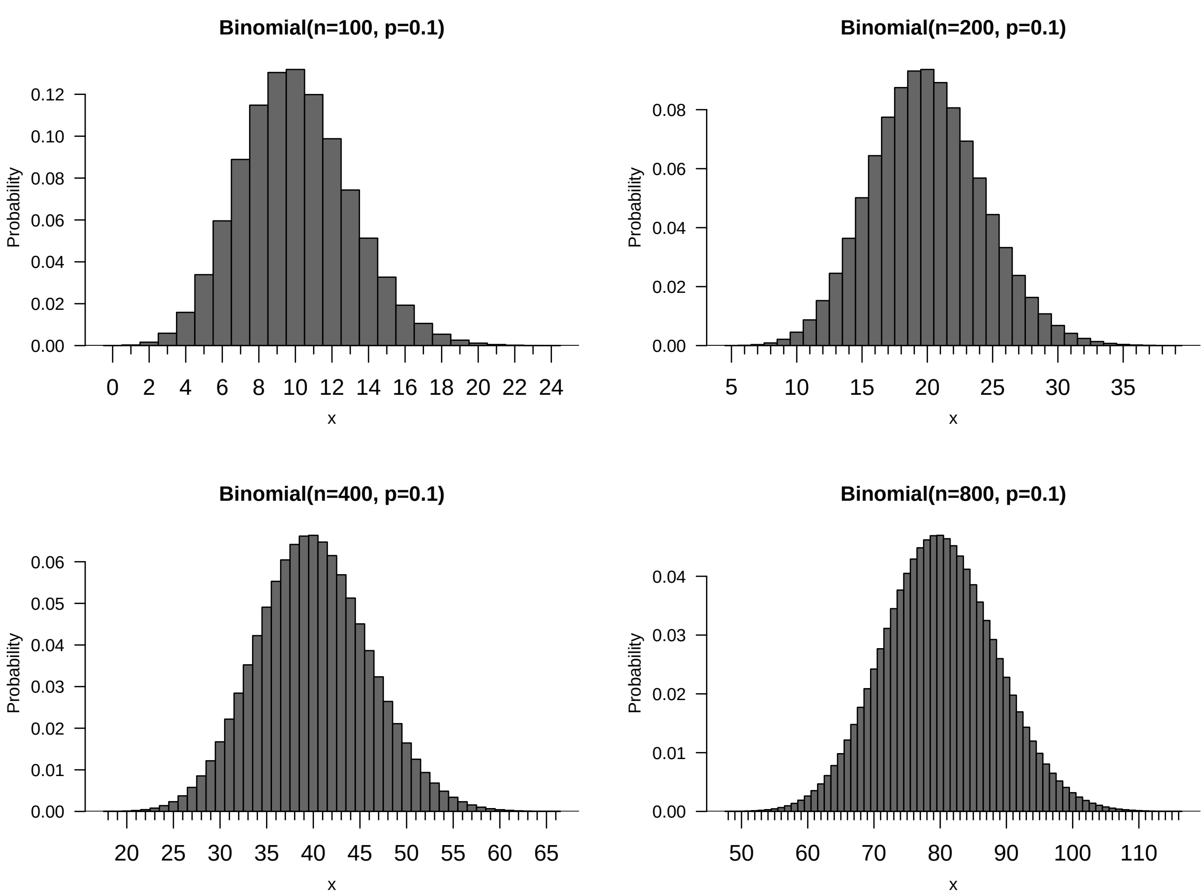
Binomial(n=10, p=0.5)
Binomial(n=20, p=0.5)
Binomial(n=50, p=0.5)
Binomial(n=100, p=0.5)
Probability
x



Binomial(n=10, p=0.5)
Binomial(n=20, p=0.5)
Binomial(n=50, p=0.5)
Binomial(n=100, p=0.5)
Probability
x


Binomial(n=10, p=0.1)
Binomial(n=20, p=0.1)
Binomial(n=50, p=0.1)
Binomial(n=100, p=0.1)
Probability
x

17


Binomial(n=100, p=0.1)
Binomial(n=200, p=0.1)
Binomial(n=400, p=0.1)
Binomial(n=800, p=0.1)
Probability
x

18

Expected value and SD

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

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

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

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

The standard deviation (SD) of X is

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

Mean and SD of binomial RV

If $X \sim$ binomial(n,p),

$$\mathrm{E}(X) = n\,p$$

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

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

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