Random Variables and Distributions

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.

$\longrightarrow$ 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.

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

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
$\overline{H}$ = mean of the 20 student heights
S_H = sample SD of heights
T_i = number of teeth of student i
$\overline{T}$ = average number of teeth

Random variables are . . .

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

Example: the 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).

Example: height, weight, mass, some 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).

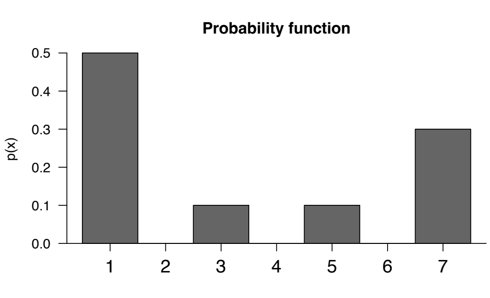
Probability function

Consider a *discrete* random variable, X.

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

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

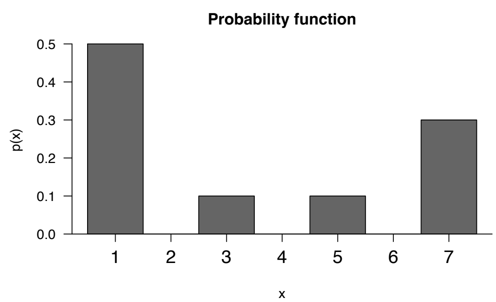
Note that $p(x) \geq 0$ for all x and $\sum p(x) = 1$.



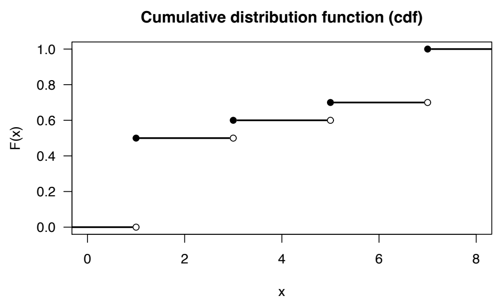
x	p(x)
1	0.5
3	0.1
5	0.1
7	0.3

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.
$\rightarrow$ 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.

That is, the number of red balls in n draws with replacement from an urn for which the proportion of red balls is p.

$\longrightarrow$ What is its probability function?

Example: Let $X \sim$ Binomial(n=9,p=0.2).

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

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

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

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

Binomial distribution

$$
\begin{aligned}
p(1) &= \Pr(X = 1) = \Pr(\text{exactly one red ball}) \\
&= \Pr(\text{RBBBBBBBB or BRBBBBBBB or} \ldots \text{or BBBBBBBBR}) \\
&= \Pr(\text{RBBBBBBBB}) + \Pr(\text{BRBBBBBBB}) + \Pr(\text{BBRBBBBBB}) \\
&\quad + \Pr(\text{BBBRBBBBB}) + \Pr(\text{BBBBRBBBB}) \\
&\quad + \Pr(\text{BBBBBRBBB}) + \Pr(\text{BBBBBBRBB}) \\
&\quad + \Pr(\text{BBBBBBBRB}) + \Pr(\text{BBBBBBBBR}) \\
&= p(1-p)^8 + p(1-p)^8 + \ldots p(1-p)^8 = 9p(1-p)^8 \approx 30\%.
\end{aligned}
$$

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

How many outcomes have 2 red balls among the 9 balls drawn?

$\longrightarrow$ 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?

$9 \times 8 / 2 = 36.$

The binomial coefficient

The number of possible samples of size k selected from a population of size n :

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

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

$\longrightarrow$ $0! = 1$

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.

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

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

The world is entropy driven

Assume we are flipping a fair coin (independently) ten times. Let X be the random variable that describes the number of heads H in the experiment.

Pr(TTTTTTTTTT) = Pr(HTTHHHTHTH) = $(1/2)^{10}$

$\longrightarrow$ There is only one possible outcome with zero heads.

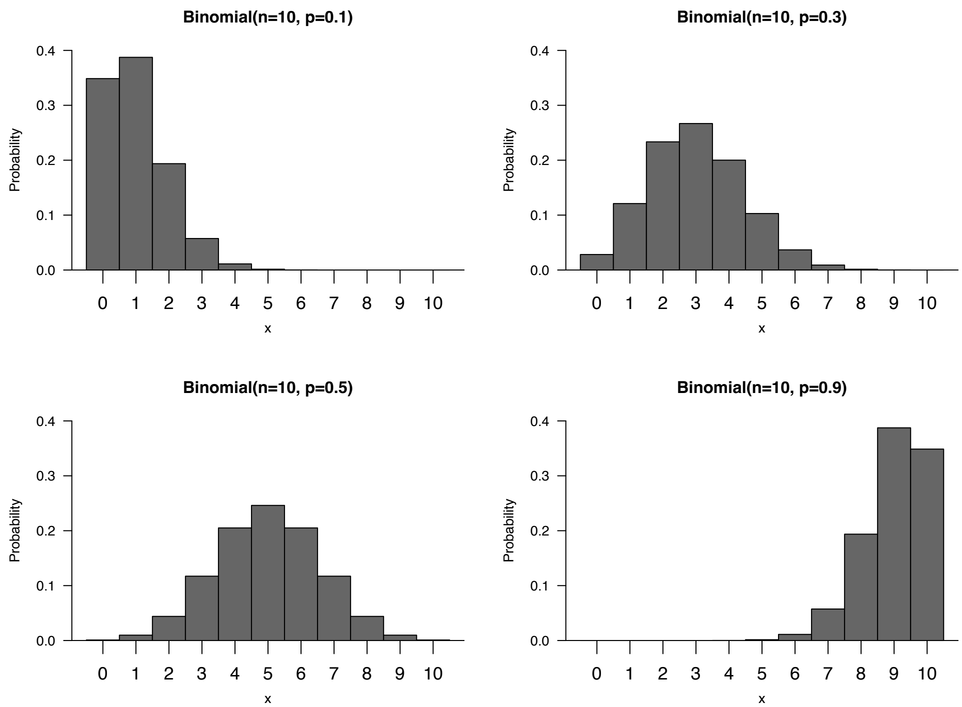
$\longrightarrow$ There are 210 possibilities for outcomes with six heads.

Thus,

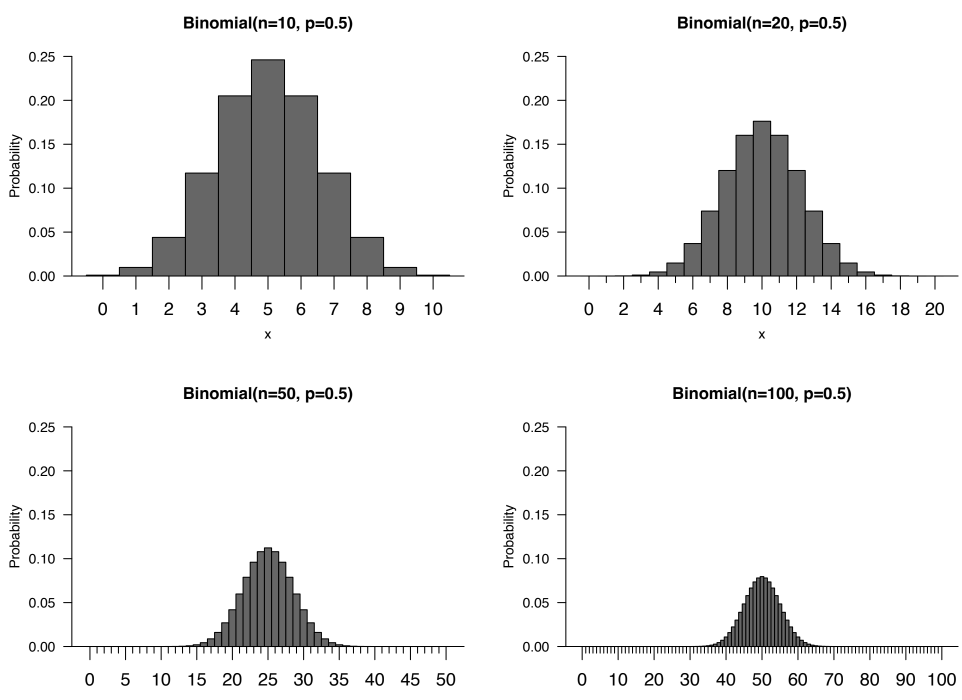
$\longrightarrow$ $\Pr(X = 0) = (1/2)^{10} \approx 0.1\%$.

$\longrightarrow$ $\Pr(X = 6) = 210 \times (1/2)^{10} \approx 20.5\%$.

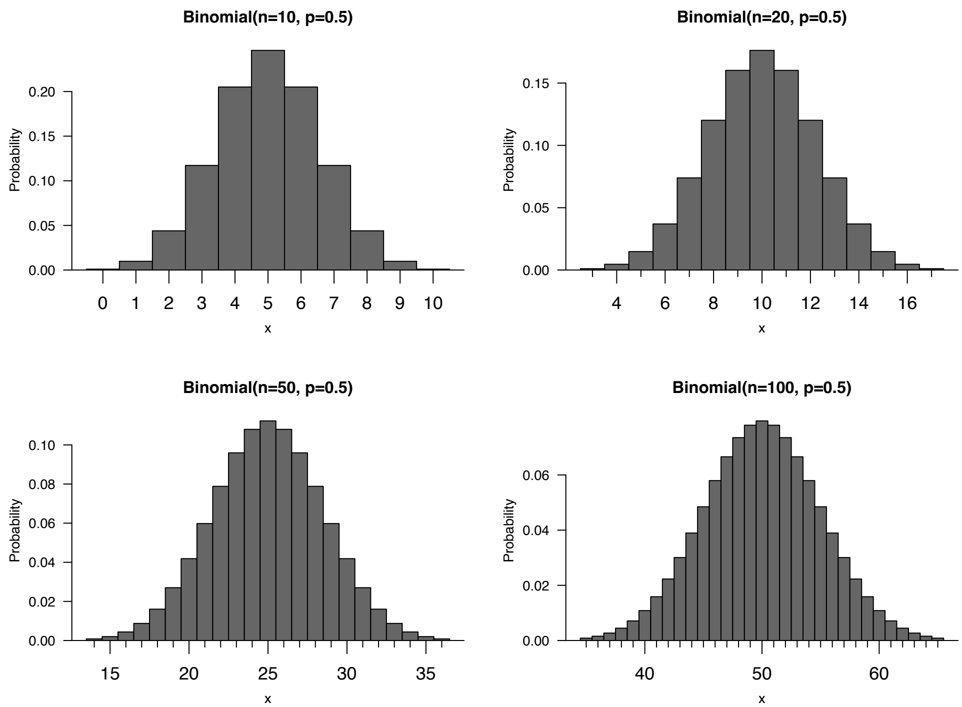
Binomial distributions



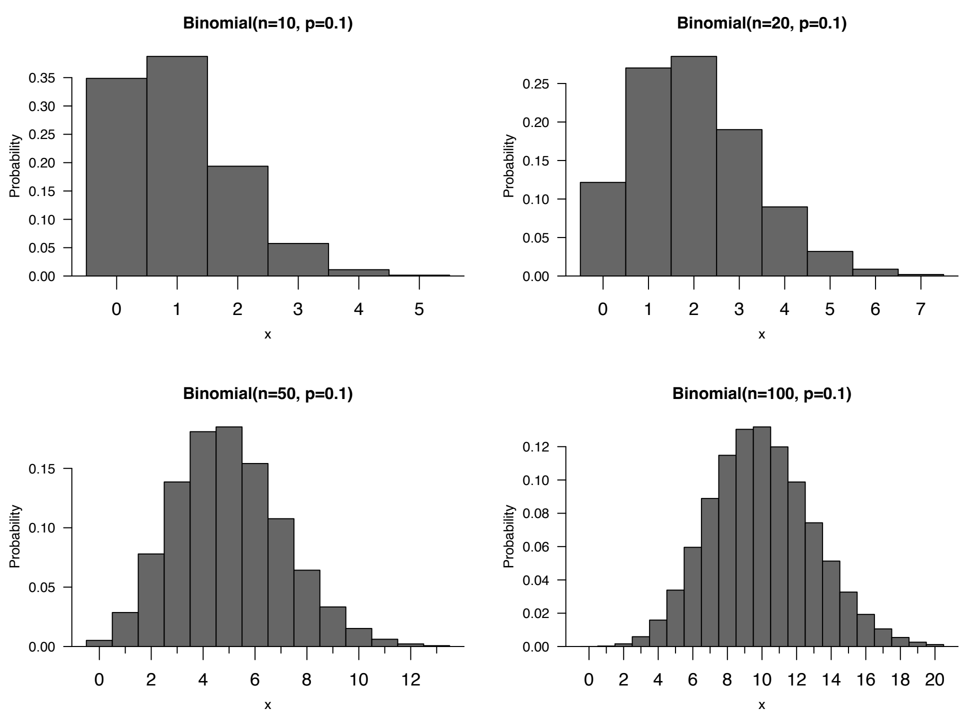
Binomial distributions



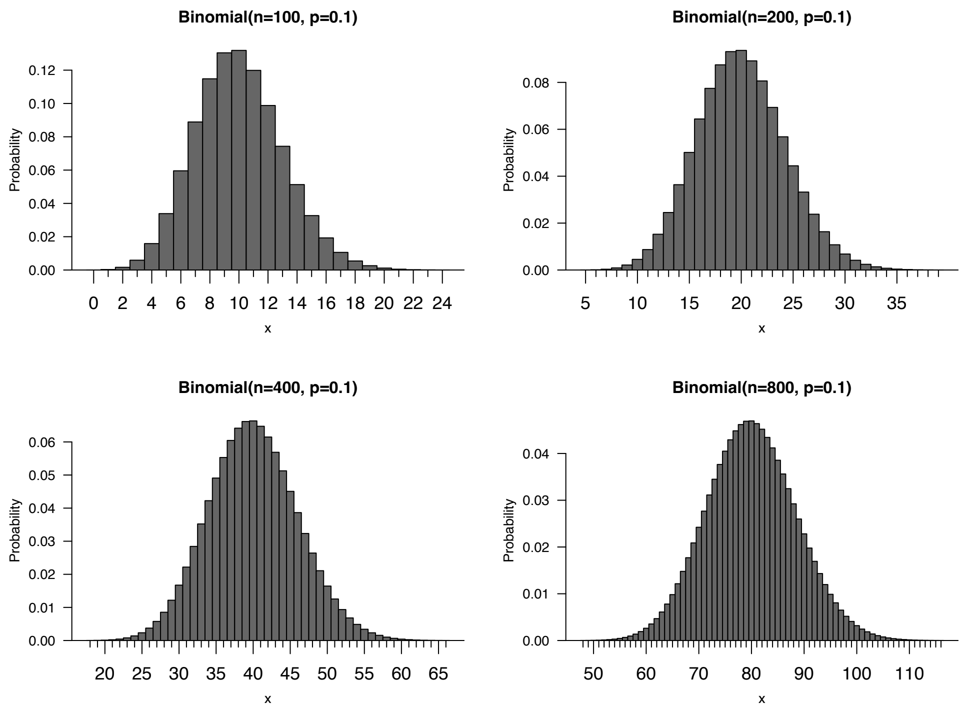
Binomial distributions



Binomial distributions



Binomial distributions



Expected value and standard deviation

⟶ The expected value (or mean) of a discrete random variable X with probability function p(x) is

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

⟶ The variance of a discrete random variable X with probability function 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 RVs

If $X \sim$ Binomial(n,p), then

$$E(X) = n\,p$$

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

$\longrightarrow$ 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 in R

$\longrightarrow$ Simulate binomial random variables

```
rbinom(m, size, prob)
```

$\longrightarrow$ The binomial probability function: $Pr(X = x)$

```
dbinom(x, size, prob)
```

$\longrightarrow$ The binomial CDF: $Pr(X \le q)$

```
pbinom(q, size, prob)
```

$\longrightarrow$ The inverse CDF: the smallest q such that $Pr(X \le q) \ge p$

```
qbinom(p, size, prob)
```

Binomial random variable

Number of successes in n trials where:

$\longrightarrow$ Trials independent

$\longrightarrow$ p = Pr(success) is constant

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

Deviations from the binomial:

$\longrightarrow$ Varying p

$\longrightarrow$ Clumping or repulsion (non-independence)

Examples

Suppose survival differs between genders:

Pr(survive | male) = 10%   but   Pr(survive | female) = 80%.

$\longrightarrow$ Pick 4 male mice and 6 female mice.

The number of survivors is not binomial.

$\longrightarrow$ Pick 10 random mice (with Pr(mouse is male) = 40%).

The number of survivors is binomial.

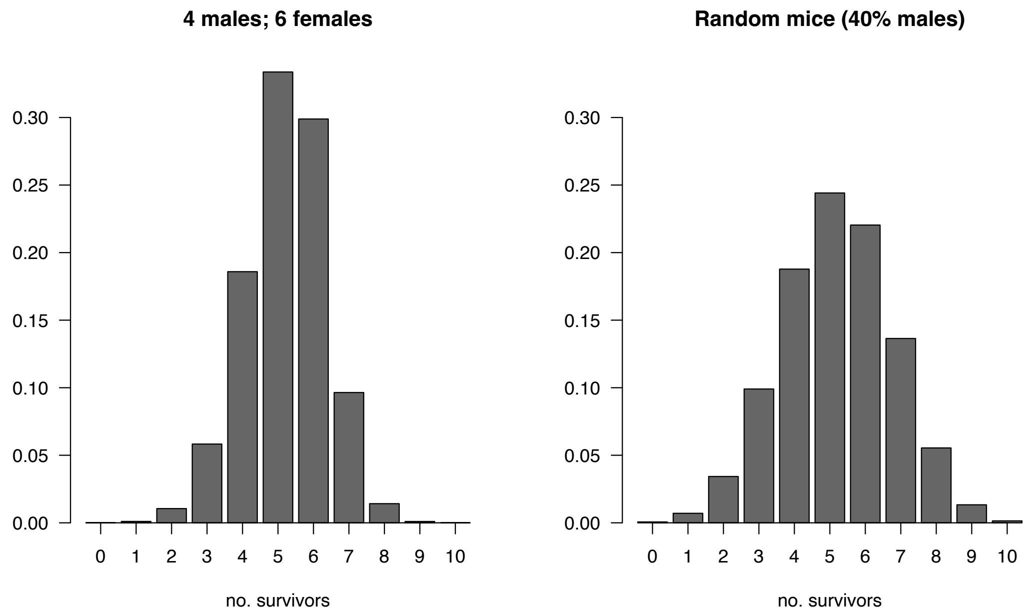
$p = 0.4 \times 0.1 + 0.6 \times 0.8 = 0.52.$

Pr(survive) =

Pr(survive and male) + Pr(survive and female) =

Pr(male) $\times$ Pr(survive | male) + Pr(female) $\times$ Pr(survive | female)

Examples



Examples

Consider Mendel's pea experiments.

Purple or white flowers, purple dominant to white:

F_0 genotypes are *PP* and *ww*, F_1 genotypes are *Pw*.

$\longrightarrow$ 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.

$\longrightarrow$ Pick 10 random F_2's. Self each and take a 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.

Pr(a progeny has a purple flower) =

Pr(purple and {F2 is PP}) + Pr(purple and {F2 is Pw}) + Pr(purple and {F2 is ww}) =

Pr(F2 is PP) $\times$ Pr(purple | F2 is PP) + Pr(F2 is Pw) $\times$ Pr(purple | F2 is Pw) + Pr(F2 is ww) $\times$ Pr(purple | F2 is ww)

Multinomial distribution

- Imagine an urn with k types of balls.
- Let p_i denote the proportion of type i.
- Draw n balls with replacement.
- Outcome: $(n_1, n_2, \ldots, n_k)$, with $\sum_i n_i = n$, where n_i is the no. balls drawn that were of type i.

$\longrightarrow$ $$P(X_1{=}n_1, \ldots, X_k{=}n_k) = \frac{n!}{n_1! \times \cdots \times n_k!} \, p_1^{n_1} \times \cdots \times p_k^{n_k}$$

if $\quad 0 \le n_i \le n\,, \quad \sum_i n_i = n \quad$ and $\quad 0 < p_i < 1, \quad \sum_i p_i = 1.$

Otherwise $\quad P(X_1{=}n_1, \ldots, X_k{=}n_k) = 0.$

Example

AA	AB	BB
35	43	22

$\longrightarrow$ Do these data correspond reasonably to the proportions 1:2:1?

Let $\quad (p_1, p_2, p_3) = (0.25, 0.50, 0.25)$ and $n = 100$.

$$P(X_1{=}35, X_2{=}43, X_3{=}22) = \frac{100!}{35!\;43!\;22!} \, 0.25^{35}\; 0.50^{43}\; 0.25^{22}$$

$$\approx 7.3 \times 10^{-4}$$

Poisson distribution

Consider a Binomial(n,p) where

$\longrightarrow$ n is really large

$\longrightarrow$ 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.

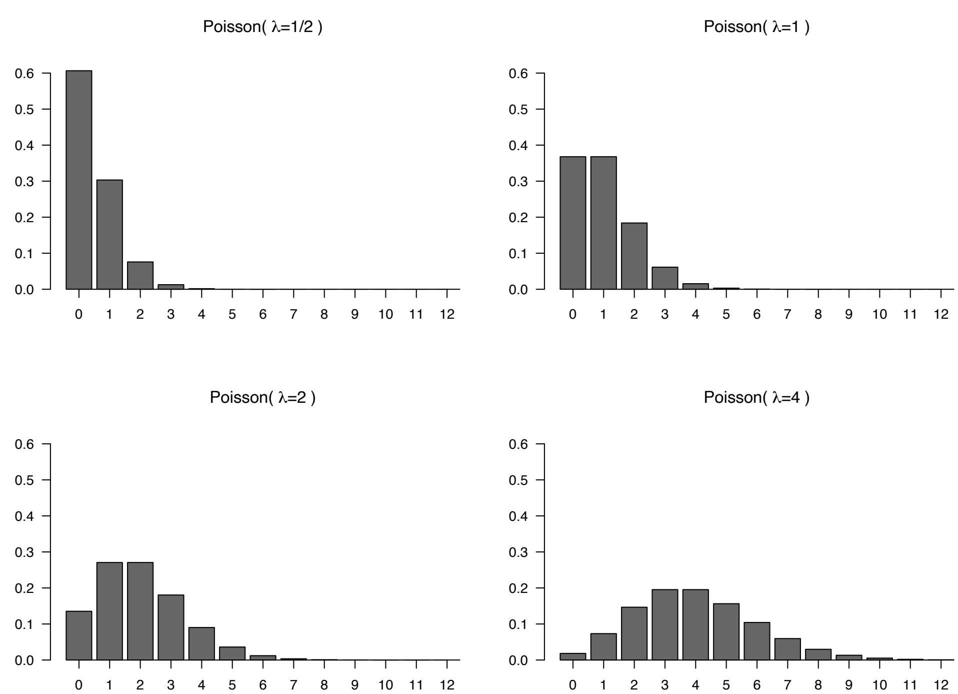
$\longrightarrow$ In this case, X follows a Poisson distribution approximately.

Let λ = n p = E(X).

$\longrightarrow$ p(x) = Pr(X = x) = $e^{-\lambda}\lambda^x/x!$

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

Poisson distribution



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.

- $\longrightarrow$ $X \sim$ Poisson(λ = 1.25).
- $\longrightarrow$ E(X) = 1.25
- $\longrightarrow$ 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) $\times$ $(1.25)^2$ / 2 $\approx$ 22%.

Calculations in R

- $\longrightarrow$ Simulate poisson random variables
  `rpois(m, lambda)`
- $\longrightarrow$ The poisson probability function: Pr(X = x)
  `dpois(m, lambda)`
- $\longrightarrow$ The poisson CDF: Pr(X $\leq$ q)
  `ppois(m, lambda)`
- $\longrightarrow$ The inverse CDF: the smallest q such that Pr(X $\leq$ q) $\geq$ p
  `qpois(m, lambda)`

$Y = a + b\ X$

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

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

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

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

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

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

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

$\longrightarrow$ $E(Z) = 0$

$\longrightarrow$ $SD(Z) = 1$

Example

Suppose $X \sim$ Binomial(n,p) $\rightarrow$ number of successes

$\longrightarrow$ $E(X) = n\ p$

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

Let $P = X / n$ $\rightarrow$ proportion of successes

$\longrightarrow$ $E(P) = E(X / n) = E(X) / n = p$

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

Continuous random variables

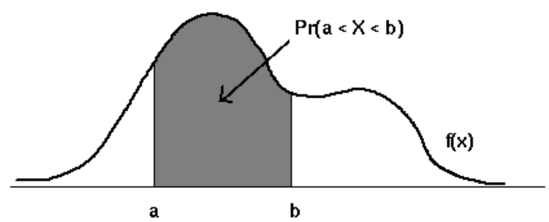
Suppose X is a continuous random variable.

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

$\longrightarrow$ f(x) $\geq$ 0

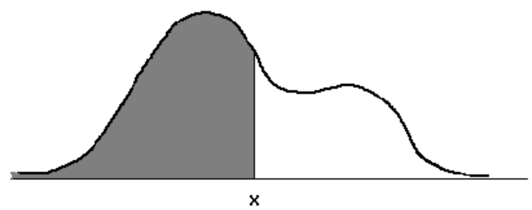
$\longrightarrow$ $\int_{-\infty}^{\infty}$ f(x) d(x) $= 1$

$\longrightarrow$ Areas under curve = probabilities



Cumulative distr. function:

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



Means and standard deviations

Expected value:

$\longrightarrow$ Discrete RV: E(X) = $\sum_{\mathrm{x}}$ x p(x)

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

Standard deviation:

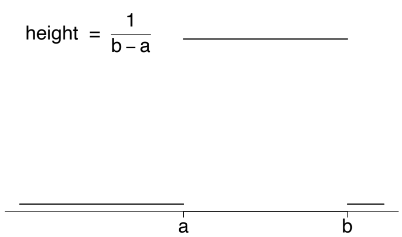
$\longrightarrow$ Discrete RV: SD(X) = $\sqrt{\sum_{\mathrm{x}} [\mathrm{x} - \mathrm{E(X)}]^2 \, \mathrm{p(x)}}$

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

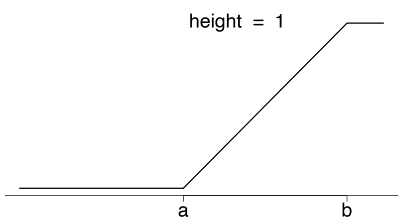
Uniform distribution

$X \sim$ Uniform(a, b)

$\longrightarrow$ Draw a number at random from the interval (a, b).



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



$\longrightarrow$ E(X) = (a + b) / 2

$\longrightarrow$ SD(X) = (b – a) / $\sqrt{12}$

$\approx 0.29 \times$ (b – a)

Normal distribution

By far the most important distribution:
The normal distribution (also called the Gaussian distribution).

If $X \sim N(\mu, \sigma)$, then the pdf of X is

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

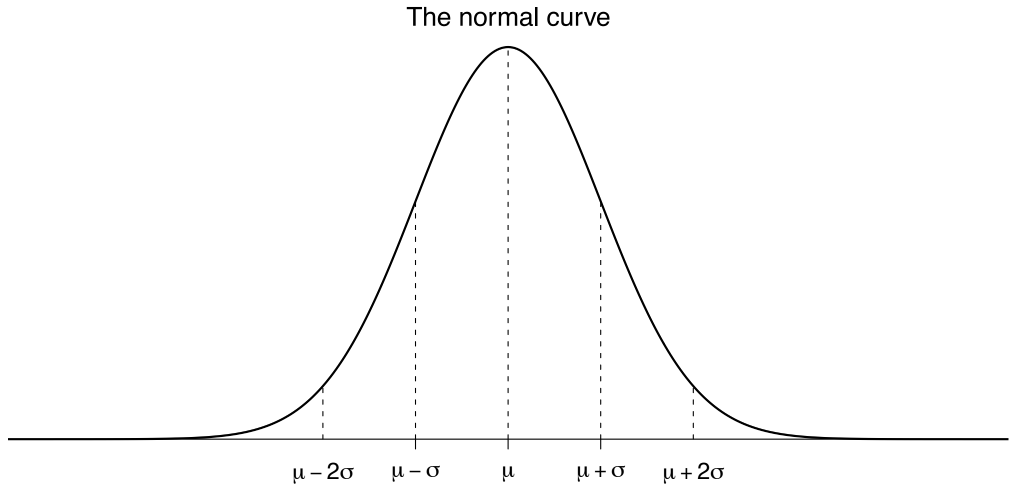
Note: E(X) = μ and SD(X) = σ.

Of great importance:

$\longrightarrow$ If $X \sim N(\mu,\sigma)$ and $Z = (X - \mu) / \sigma$, then $Z \sim N(0, 1)$.

This is the standard normal distribution.

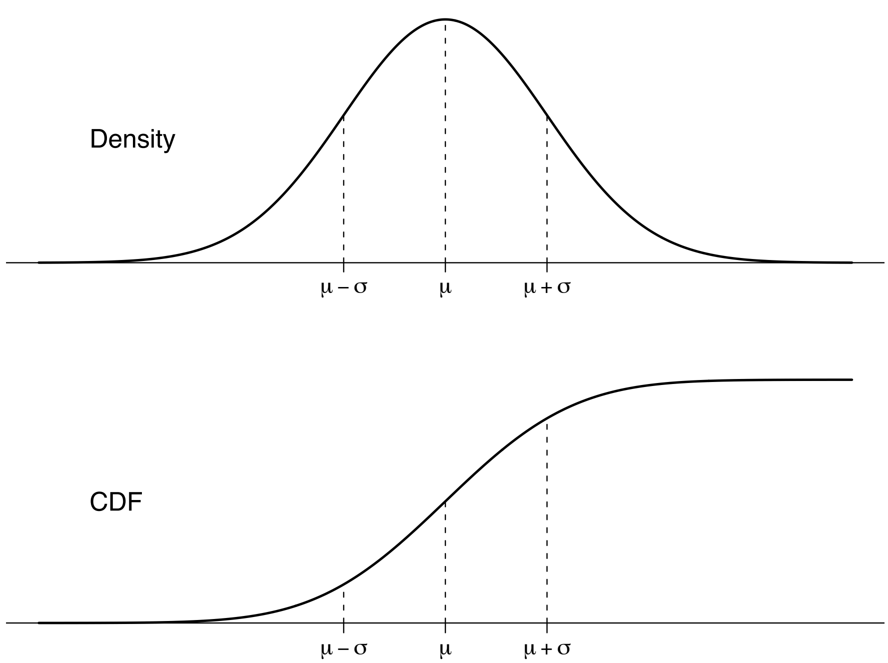
Normal distribution



$\longrightarrow$ Remember:

Pr($\mu - \sigma \leq X \leq \mu + \sigma$) ≈ 68%   and   Pr($\mu - 2\sigma \leq X \leq \mu + 2\sigma$) ≈ 95%.

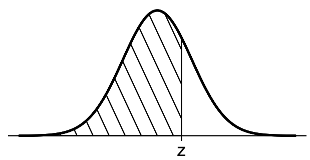
The normal CDF



Calculations with the normal curve in R

- Convert to a statement involving the cdf.
- Use the function `pnorm()`.

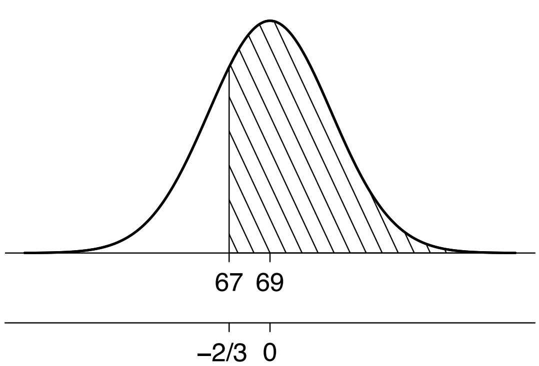
$\longrightarrow$ Draw a picture!



Examples

Suppose the heights of adult males in the U.S. are approximately normal distributed, with mean = 69 in and SD = 3 in.

$\longrightarrow$ What proportion of men are taller than 5’7”?



$X \sim$ N(μ=69, σ=3)

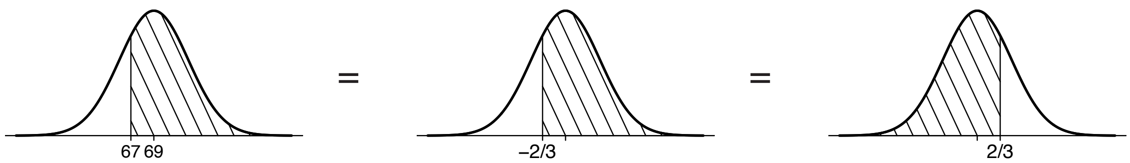
Z = (X – 69)/3 $\sim$ N(0,1)

Pr($X \geq$ 67) =

Pr($Z \geq$ (67 – 69)/3) =

Pr($Z \geq$ – 2/3)

R



Use either of the following three:

$\longrightarrow$ `pnorm(67, 69, 3, lower=FALSE)`

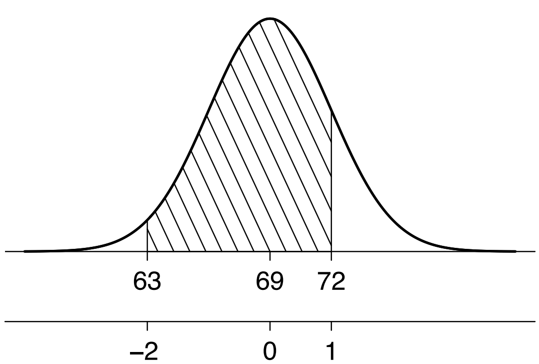
$\longrightarrow$ `1 - pnorm(67, 69, 3)`

$\longrightarrow$ `pnorm(2/3)`

The answer: 75%.

Another calculation

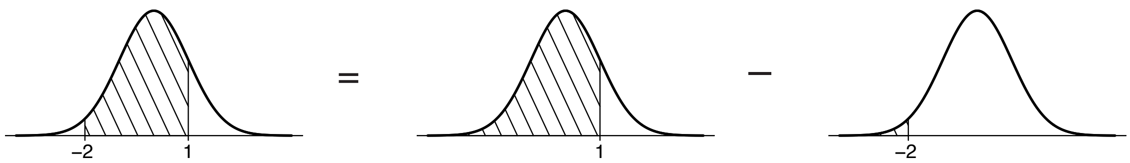
$\longrightarrow$ What proportion of men are between 5’3” and 6’?



$\Pr(63 \leq X \leq 72) =$

$\Pr(-2 \leq Z \leq 1)$

R



Use either of the following:

→ `pnorm(72,69,3) - pnorm(63,69,3)`

→ `pnorm(1) - pnorm(-2)`

The answer: 82%.