Chapter 1: Foundations

Shravan Vasishth (vasishth.github.io)

June 2025

Contents

Textbook

Introduction to Bayesian Data Analysis for Cognitive Science

Nicenboim, Schad, Vasishth

- Online version: https://bruno.nicenboim.me/bayescogsci/
- Source code: https://github.com/bnicenboim/bayescogsci
- Physical book: here

Be sure to read the textbook's chapter 1 in addition to watching this lecture.

Introduction: Motivation for this lecture

- Whenever we collect data, an implicit assumption is that the data are being generated from a **random variable**.
- Understanding the basic properties of random variables is of key importance when learning statistical modeling.
- The ideas and concepts in this lecture are often not taught in statistics courses in linguistics and psychology.
- The commonly used cookbook approach to teaching statistics leads to all kinds of misunderstandings that have a snowball effect and are a big part of the cause for the replication crisis and other problems in inference that we see so often in empirical work in linguistics and psychology.

It only takes about a day to understand these materials, but the content here will positively impact your ability to carry out statistical modeling and data analysis.

Discrete random variables

A random variable X is a function $X : \Omega \to \mathbb{R}$ that associates to each **outcome** $\omega \in \Omega$ exactly one number $X(\omega) = x$.

S_X is all the x's (all the possible values of X, the **support of X**). I.e., $x \in S_X$.

An example of a discrete RV

An example of a discrete random variable: keep tossing a coin again and again until you get a Heads.

- $X : \omega \to x$
- ω: H, TH, TTH,…(infinite)
- $X(H) = 1, X(TH) = 2, X(TTH) = 3, \ldots$
- $x = 1, 2, \ldots; x \in S_X$

A second example of a discrete random variable: tossing a coin once.

- $X : \omega \to x$
- ω: H, T
- $X(T) = 0, X(H) = 1$
- $x = 0, 1; x \in S_X$

The probability mass function (PMF)

Every discrete (continuous) random variable X has associated with it a **probability mass (density) function (PMF, PDF)**.

- PMF is used for discrete distributions and PDF for continuous.
- (Some books use PDF for both discrete and continuous distributions.)

Thinking just about discrete random variables for now:

$$p_X : S_X \to [0, 1] \tag{1}$$

defined by

$$p_X(x) = \text{Prob}(X(\omega) = x), x \in S_X \tag{2}$$

Example of a PMF: a random variable X representing tossing a coin once.

- In the case of a fair coin, x can be 0 or 1, and the probability of each possible event (each event is a subset of the set of possible outcomes) is 0.5.
- Formally: $p_X(x) = \text{Prob}(X(\omega) = x), x \in S_X$
- The probability mass function defines the probability of each event: $p_X(0) = p_X(1) = 0.5$.

The cumulative distribution function (CDF)

The **cumulative distribution function** (CDF) $F(X \leq x)$ gives the cumulative probability of observing all the events $X \leq x$.

$$
\begin{aligned}
F(x=1) &= \text{Prob}(X \leq 1) \\
&= \sum_{x=0}^{1} p_X(x) \\
&= p_X(x=0) + p_X(x=1) \\
&= 1
\end{aligned}
\tag{3}
$$

$$
\begin{aligned}
F(x=0) &= \text{Prob}(X \leq 0) \\
&= \sum_{x=0}^{0} p_X(x) \\
&= p_X(x=0) \\
&= 0.5
\end{aligned}
\tag{4}
$$

Do 10 coin-tossing experiments, each with one trial. The probability (which I call θ below) of heads 0.5:

```
extraDistr::rbern(n = 10, prob = 0.5)
```

```
##  [1] 0 0 1 1 0 0 0 0 0 0
```

The probability mass function: Bernoulli

$$p_X(x) = \theta^x (1-\theta)^{(1-x)}$$

where x can have values 0, 1.

What's the probability of a tails/heads? The d-family of functions:

```
extraDistr::dbern(0, prob = 0.5)
```

```
## [1] 0.5
```

```r
extraDistr::dbern(1, prob = 0.5)
```

```
## [1] 0.5
```

Notice that these probabilities sum to 1.

The cumulative probability distribution function: the p-family of functions:

$$F(x=1) = Prob(X \leq 1) = \sum_{x=0}^{1} p_X(x) = 1$$

```r
extraDistr::pbern(1, prob = 0.5)
```

```
## [1] 1
```

$$F(x=0) = Prob(X \leq 0) = \sum_{x=0}^{0} p_X(x) = 0.5$$

```r
extraDistr::pbern(0, prob = 0.5)
```

```
## [1] 0.5
```

Another example of a discrete random variable: The binomial

- Consider carrying out a single experiment where you toss a coin 10 times (the number of trials, `size` in R).
- When the number of trials (size) is 1, we have a Bernoulli; when we have size greater than 1, we have a Binomial.

$$\theta^x (1-\theta)^{1-x}$$

where

$$S_X = \{0, 1\}$$

Binomial PMF

$$\binom{n}{x}\theta^x(1-\theta)^{n-x}$$

where

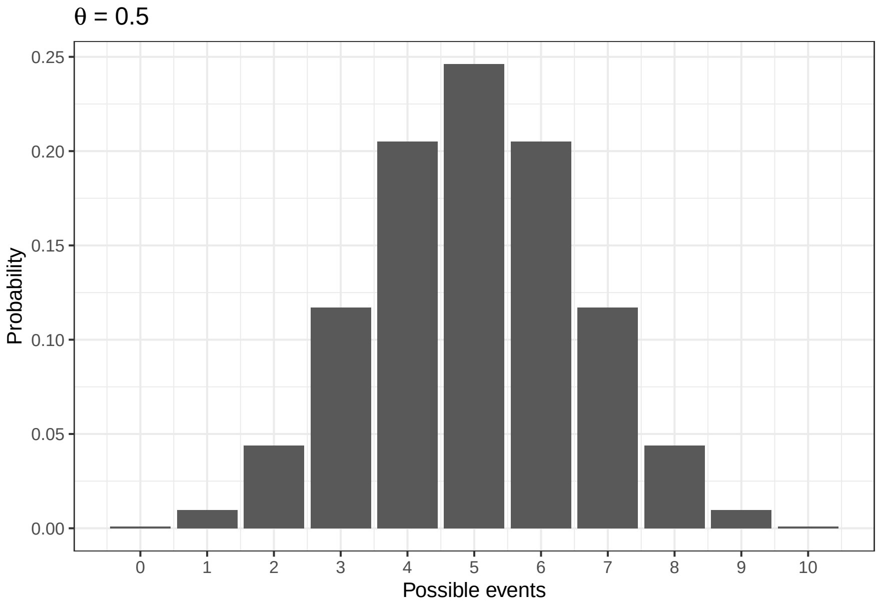
$$S_X = \{0, 1, \ldots, n\}$$

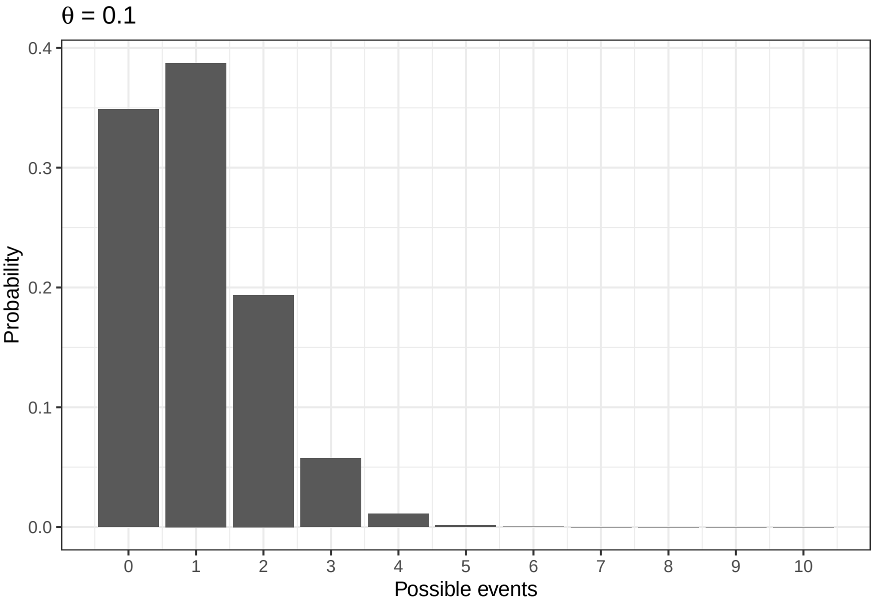
- n is the number of times the coin was tossed (the number of trials; size in R).
- $\binom{n}{x}$ is the number of ways that you can get x successes in n trials.

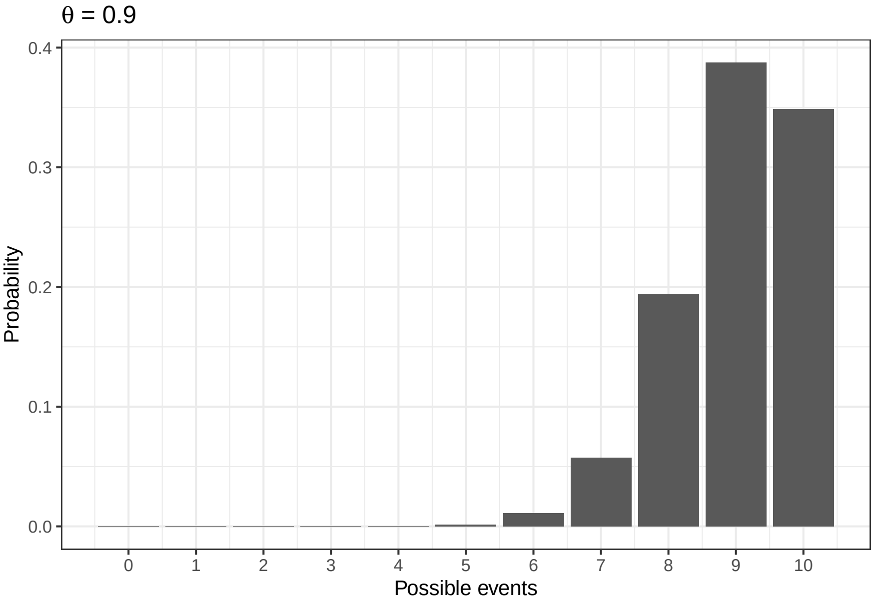
```
choose(10, 2)
```

```
## [1] 45
```

- θ is the probability of success in n trials.







Four critical R functions for the binomial RV

1. Generate random data: rbinom

- n: number of experiments done (**Note**: in the binomial pdf, n stands for the number of trials). In R, n is called the number of observations.
- size: the number of times the coin was tossed in each experiment (the number of trials)

Example: 10 separate experiments, each with 1 trial:

```r
rbinom(n = 10, size = 1, prob = 0.5)
```

```
##  [1] 0 1 0 1 0 1 1 1 0 1
```

```r
## equivalent to: rbern(10,0.5)
```

Example: 10 separate experiments, each with 10 trials:

```r
rbinom(n = 10, size = 10, prob = 0.5)
```

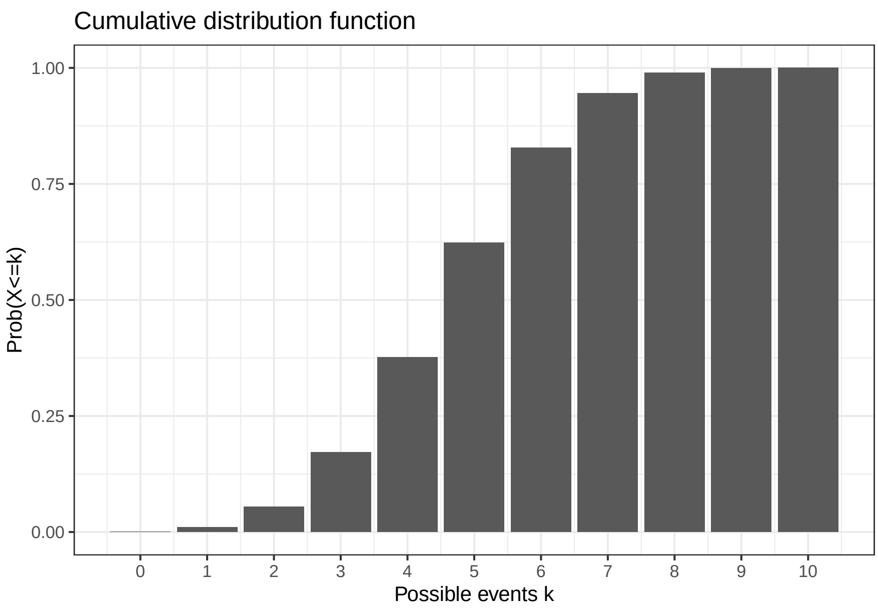
```
##  [1] 7 5 9 5 6 4 8 6 6 4
```

2. Compute probabilities of particular events (0,1,...,10 successes when n=10): dbinom

```r
probs <- round(dbinom(0:10, size = 10,
                      prob = 0.5), 3)
x <- 0:10
```

```
##     x probs
## 1   0 0.001
## 2   1 0.010
## 3   2 0.044
## 4   3 0.117
## 5   4 0.205
## 6   5 0.246
## 7   6 0.205
## 8   7 0.117
## 9   8 0.044
## 10  9 0.010
## 11 10 0.001
```

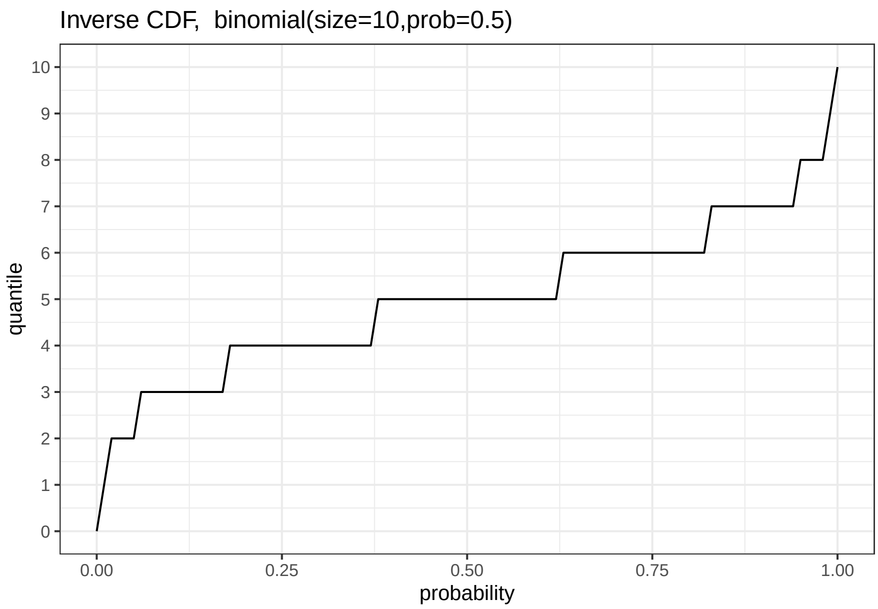
3. Compute cumulative probabilities: pbinom



4. Compute quantiles using the inverse of the CDF: qbinom

```r
probs <- pbinom(0:10, size = 10, prob = 0.5)
qbinom(probs, size = 10, prob = 0.5)
```

```
##  [1]  0  1  2  3  4  5  6  7  8  9 10
```



These four functions are the d-p-q-r family of functions, and are available for all the distributions available in R (e.g., Poisson, geometric, normal, beta, uniform, gamma, exponential, Cauchy, etc.).

Continuous random variables

In coin tosses, H and T are discrete possible outcomes.

- By contrast, variables like reading times range from 0 milliseconds up–these are **continuous variables**.
- Continuous random variables have a probability **density** function (PDF) $f(\cdot)$ associated with them. (cf. PMF in discrete RVs)
- The expression

$$X \sim f(\cdot) \tag{5}$$

means that the random variable X is assumed to have PDF $f(\cdot)$.

For example, if we say that $X \sim Normal(\mu, \sigma)$, we are assuming that the PDF is

$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{6}$$

where $-\infty < x < +\infty$

We can **truncate** the normal distribution such that S_X is bounded between some lower bound and/or upper bound–this comes later.

The normal random variable

The PDF below is associated with the normal distribution that you are probably familiar with:

$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \quad (7)$$

where $-\infty < x < +\infty$.

- The support of X, i.e., the elements of S_X, has values ranging from $-\infty$ to $+\infty$
- μ is the location parameter (here, mean)
- σ is the scale parameter (here, standard deviation)

In the discrete RV case, we could compute the probability of a **particular** event occurring:

```
extraDistr::dbern(x = 1, prob = 0.5)
```

```
## [1] 0.5
```

```
dbinom(x = 2, size = 10, prob = 0.5)
```

```
## [1] 0.04394531
```

- In a continuous distribution, probability is defined as the **area under the curve**.
- As a consequence, for any particular **point** value x, where $X \sim Normal(\mu, \sigma)$, it is always the case that $\text{Prob}(X = x) = 0$.
- In any continuous distribution, we can compute probabilities like $\text{Prob}(x_1 < X < x_2) =?$, where $x_1 < x_2$, by summing up the **area under the curve**.
- To compute probabilities like $\text{Prob}(x_1 < X < x_2) =?$, we need the cumulative distribution function.

The cumulative distribution function (CDF) is

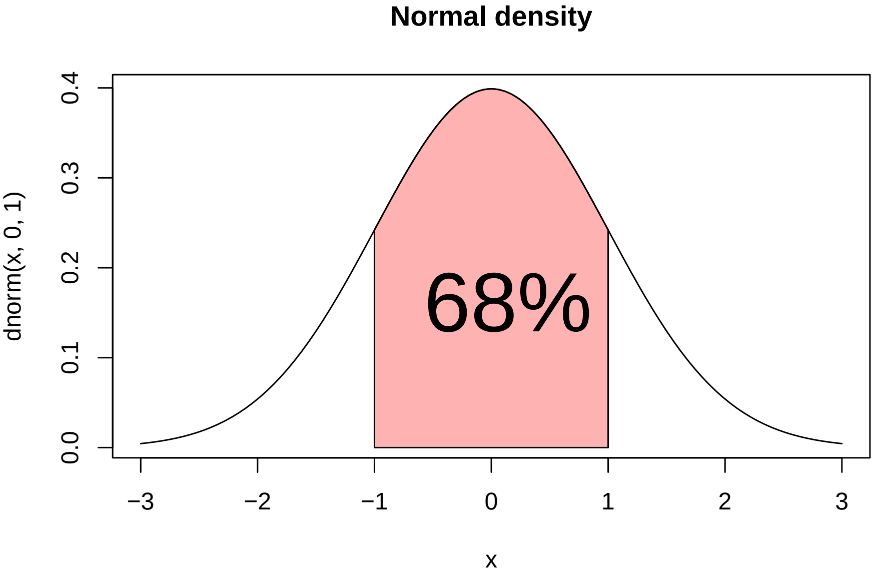
$$P(X < u) = F(X < u) = \int_{-\infty}^{u} f(x)\,dx \quad (8)$$

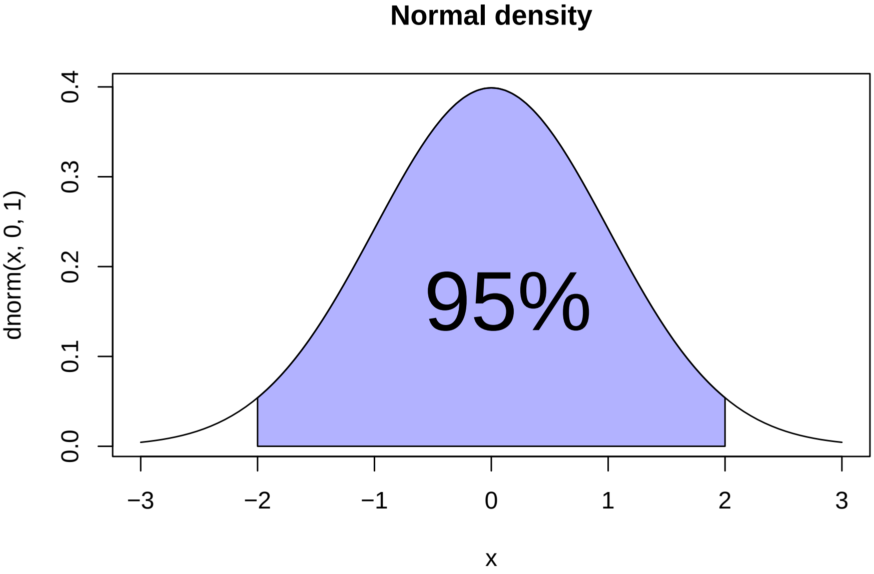
- The integral sign $\int$ is just the summation symbol in continuous space.
- Recall the summation in the CDF of the Bernoulli!

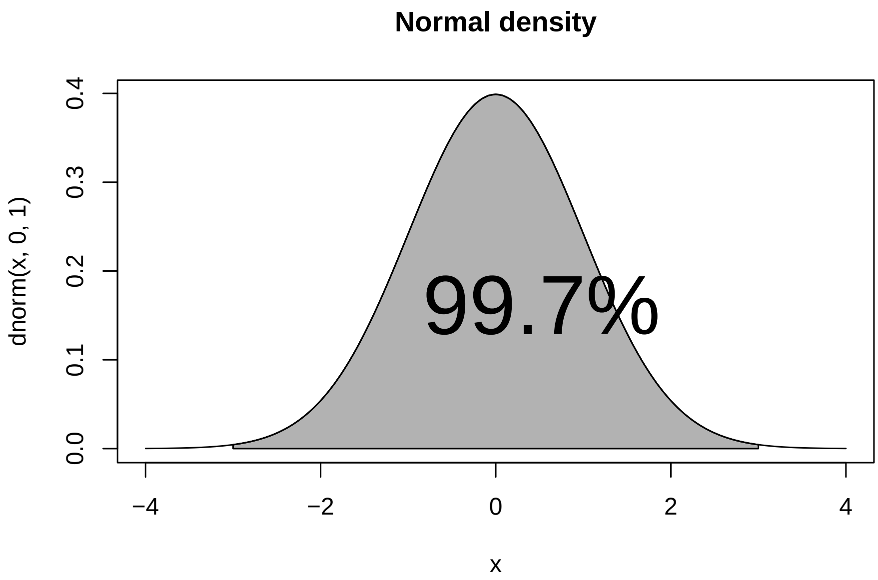
The standard normal distribution

In the $Normal(\mu = 0, \sigma = 1)$,

- $\text{Prob}(-1 < X < +1) = 0.68$
- $\text{Prob}(-2 < X < +2) = 0.95$
- $\text{Prob}(-3 < X < +3) = 0.997$







More generally, for any $Normal(\mu, \sigma)$,

- $\text{Prob}(-1 \times \sigma < X < +1 \times \sigma) = 0.68$
- $\text{Prob}(-2 \times \sigma < X < +2 \times \sigma) = 0.95$
- $\text{Prob}(-3 \times \sigma < X < +3 \times \sigma) = 0.997$

The normalizing constant and the kernel

The PDF of the normal again:

$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{9}$$

This part of $f(x \mid \mu, \sigma)$ (call it $g(x)$) is the "kernel'' of the normal PDF:

$$g(x \mid \mu, \sigma) = \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{10}$$

For the above function, the area under the curve doesn't sum to 1:

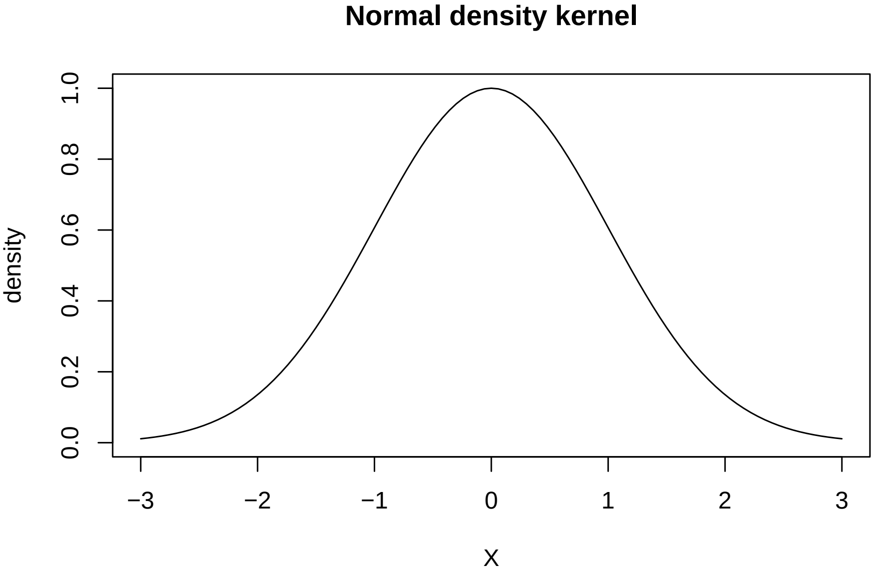
Sum up the area under the curve $\int g(x)\, dx$:

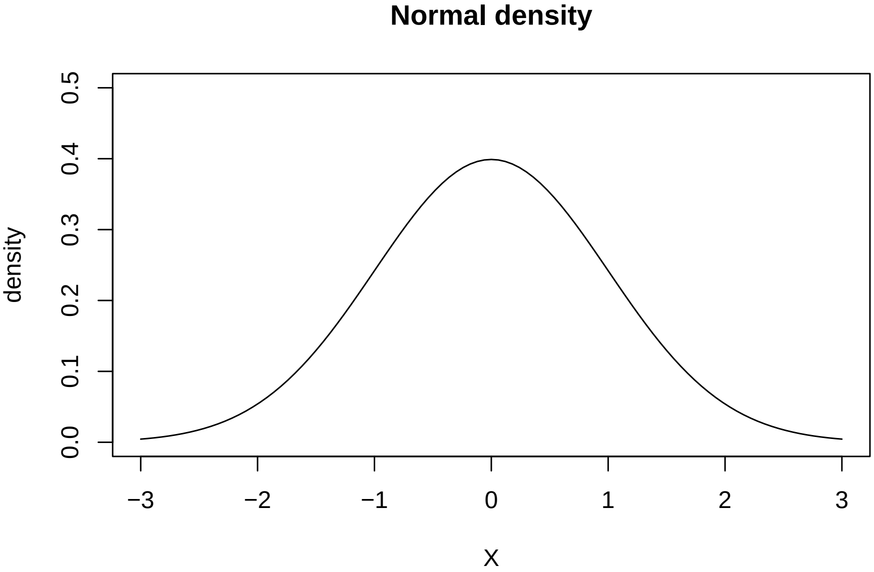
```r
g <- function(x, mu = 0, sigma = 1) {
  exp((-(x - mu)^2 / (2 * (sigma^2))))
}

integrate(g, lower = -Inf, upper = +Inf)$value
```

```
## [1] 2.506628
```

The shape doesn't change of course:





In simple examples like the one shown here, given the kernel of some PDF like $g(x)$, we can figure out the normalizing constant by solving for k in:

$$k \int g(x)\, dx = 1 \tag{11}$$

Solving for k just amounts to computing:

$$k = \frac{1}{\int g(x)\, dx} \tag{12}$$

So, in our example above,

```
(k<-1/integrate(g, lower = -Inf, upper = +Inf)$value)
```

```
## [1] 0.3989423
```

The above number is just $\frac{1}{\sqrt{2\pi\sigma^2}}$, where $\sigma = 1$:

```
1/(sqrt(2*pi*1))
```

```
## [1] 0.3989423
```

Once we include the normalizing constant, the area under the curve in $g(x)$ sums to 1:

```
k * integrate(g, lower = -Inf, upper = +Inf)$value
```

```
## [1] 1
```

We will see the practical implication of this when we move on to chapter 2 of the textbook.

The d-p-q-r functions for the normal distribution

In the continuous case, we also have this family of d-p-q-r functions. In the normal distribution:

1. Generate random data using rnorm

```
round(rnorm(5, mean = 0, sd = 1),3)
```

```
## [1]  1.537 -1.240 -0.878  0.441  0.305
```

For the standard normal, mean=0, and sd=1 can be omitted (these are the default values in R).

```
round(rnorm(5),3)
```

```
## [1] -0.354  0.439  1.408  0.162 -1.478
```

2. Compute probabilities using CDF: pnorm

Some examples of usage:

- Prob($X < 2$) (e.g., in $X \sim Normal(0, 1)$)

```
pnorm(2)
```

```
## [1] 0.9772499
```

- Prob($X > 2$) (e.g., in $X \sim Normal(0, 1)$)

```
pnorm(2, lower.tail = FALSE)
```

```
## [1] 0.02275013
```

3. Compute quantiles: qnorm

```
qnorm(0.9772499)
```

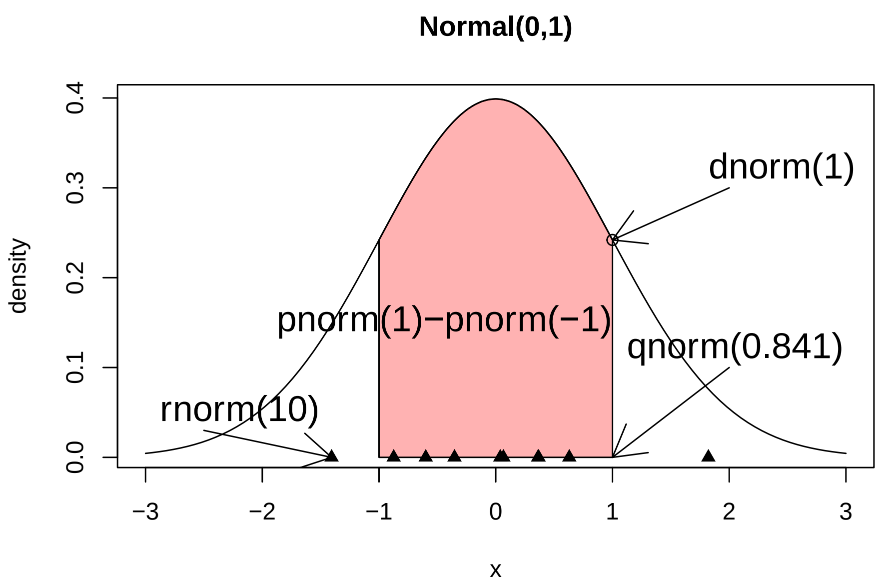
```
## [1] 2.000001
```

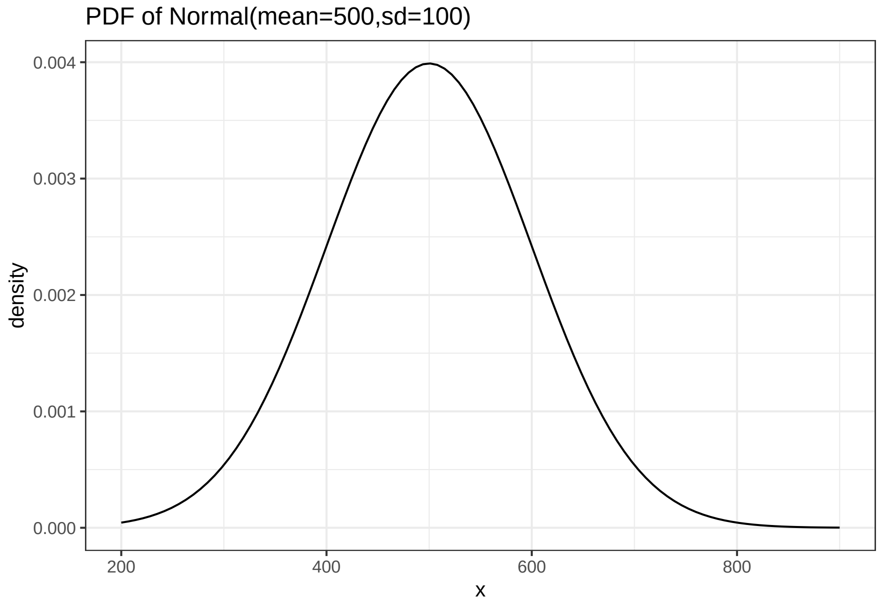
4. Compute the probability density: dnorm

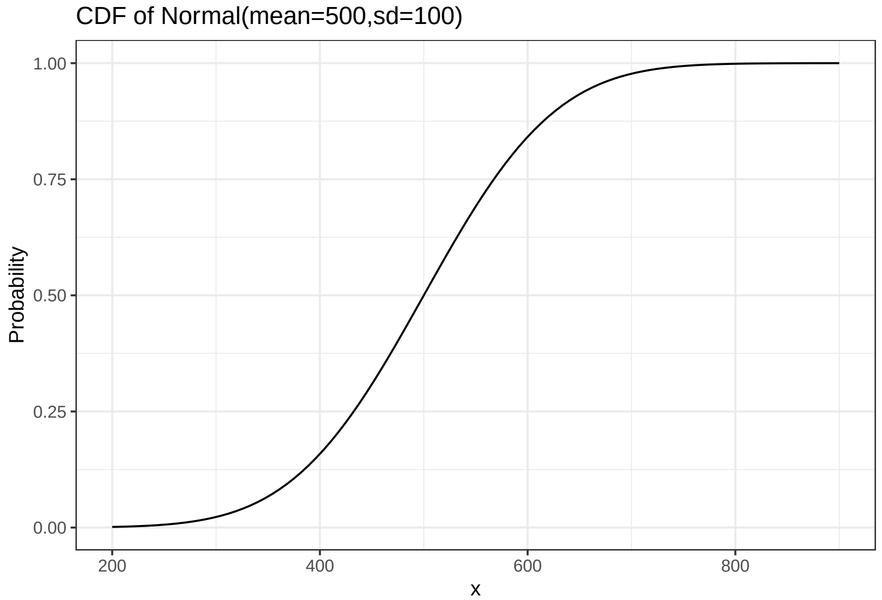
```
dnorm(2)
```

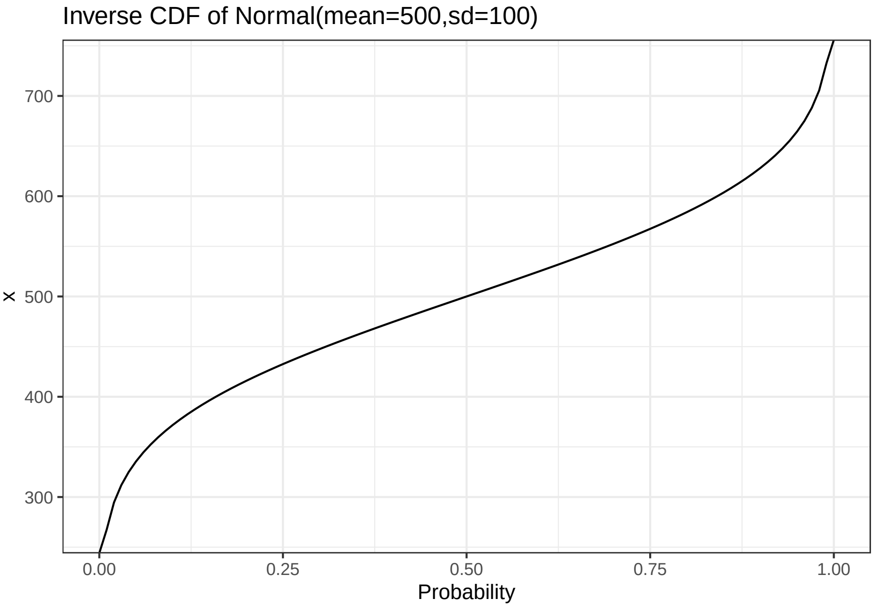
```
## [1] 0.05399097
```

Note: In the continuous case, this is a **density**, the value $f(x)$, not a probability. Cf. the discrete examples dbern and dbinom, which give probabilities of a point value x.


Normal(0,1)
dnorm(1)
pnorm(1)−pnorm(−1)
qnorm(0.841)
rnorm(10)
density
x


PDF of Normal(mean=500,sd=100)
density
x


CDF of Normal(mean=500,sd=100)
Probability
x



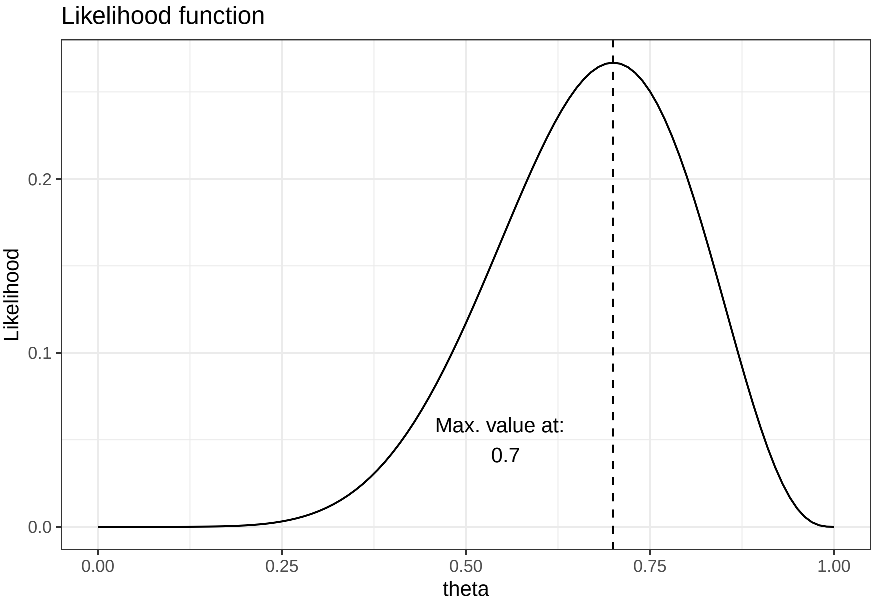
The likelihood function (Binomial)

The **likelihood function** refers to the PMF $p(k|n,\theta)$, treated as a function of θ.

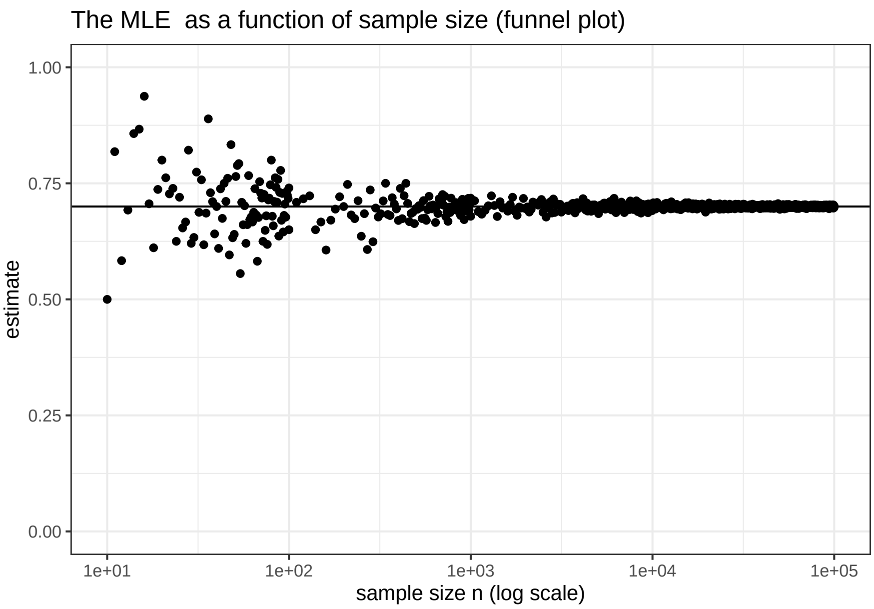
For example, suppose that we record $n = 10$ trials, and observe $k = 7$ successes. The likelihood function is:

$$\mathcal{L}(\theta|k=7,n=10) = \binom{10}{7}\theta^7(1-\theta)^{10-7} \quad (13)$$

If we now plot the likelihood function for all possible values of θ ranging from 0 to 1, we get the plot shown below.



The MLE (**from a particular sample** of data need not invariably give us an accurate estimate of θ.



Sample size is key here: as $n \to \infty$, we approach the true value of the parameter (here, θ).

The likelihood function (Normal)

$$\mathcal{L}(\mu, \sigma | x) = Normal(x, \mu, \sigma) \qquad (14)$$

Below, assume that $\sigma = 1$.

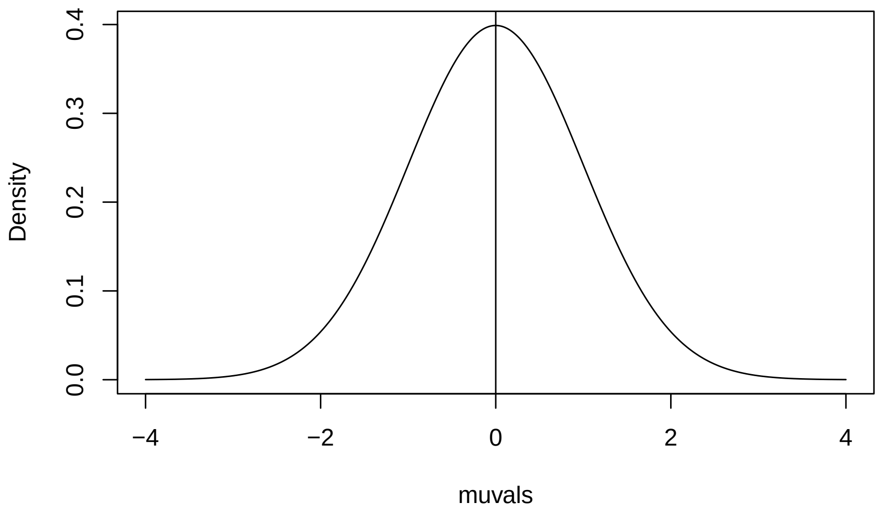
```r
## the data:
x<-0
## the likelihood under different values
## of mu:
dnorm(x,mean=0,sd=1)
```

```
## [1] 0.3989423
```

```r
dnorm(x,mean=10,sd=1)
```

```
## [1] 7.694599e-23
```

Assuming that $\sigma = 1$, the likelihood function of μ:



If we have two **independent** data points, the joint likelihood given the data of μ, assuming $\sigma = 1$:

```r
x1<-0
x2<-1.5
dnorm(x1,mean=0,sd=1) *
  dnorm(x2,mean=0,sd=1)
```

```
## [1] 0.05167004
```

```r
## log likelihood:
dnorm(x1,mean=0,sd=1,log=TRUE) +
```

```r
  dnorm(x2,mean=0,sd=1,log=TRUE)
```

```
## [1] -2.962877
```

```r
## more compactly:
x<-c(x1,x2)
sum(dnorm(x,mean=0,sd=1,log=TRUE))
```

```
## [1] -2.962877
```

One practical implication: one can use the log likelihood to compare competing models' fit:

```r
## Model 1:
sum(dnorm(x,mean=0,sd=1,log=TRUE))
```

```
## [1] -2.962877
```

```r
## Model 2:
sum(dnorm(x,mean=10,sd=1,log=TRUE))
```

```
## [1] -87.96288
```

Model 1 has higher likelihood than Model 2, so we'd prefer to assume that the data are better characterized by Model 1 than 2 (neither may be the true model!).

More generally, for independent and identically distributed data $x = x_1, \dots, x_n$:

$$\mathcal{L}(\mu, \sigma | x) = \prod_{i=1}^{n} Normal(x_i, \mu, \sigma) \tag{15}$$

or

$$\ell(\mu, \sigma | x) = \sum_{i=1}^{n} log(Normal(x_i, \mu, \sigma)) \tag{16}$$

The expectation and variance of an RV

Read section 1.4.1 of chapter 1 of the textbook, and (optionally) chapter 2 of the linear modeling lecture notes here:

https://github.com/vasishth/LM

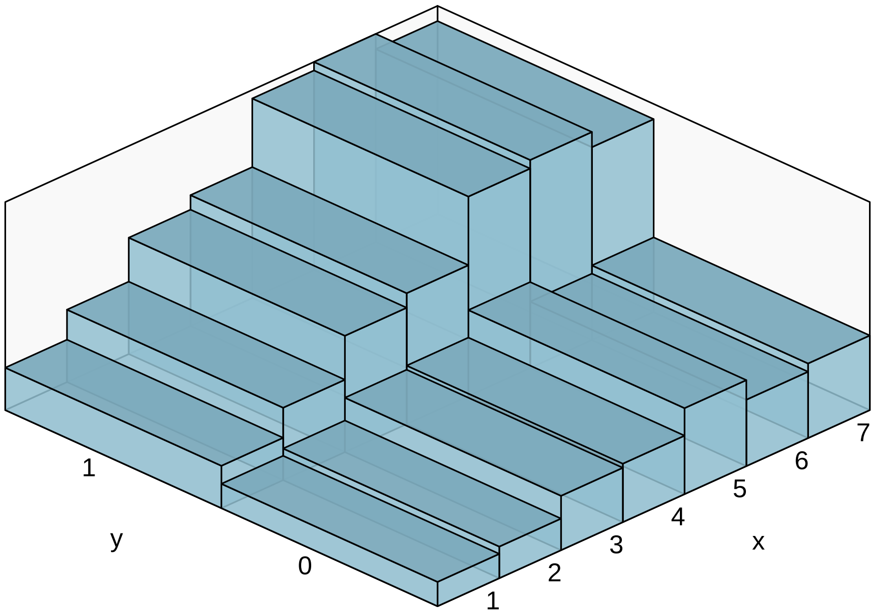
Bivariate/multivariate distributions

Discrete bivariate distributions

Data from: Laurinavichyute, A. (2020). Similarity-based interference and faulty encoding accounts of sentence processing. dissertation, University of Potsdam.

X: Likert ratings 1-7.

Y: 0, 1 accuracy responses.



The joint PMF: $p_{X,Y}(x, y)$

For each possible pair of values of X and Y, we have a **joint probability mass function** $p_{X,Y}(x, y)$.

Table 1: The joint PMF for two random variables X and Y.

	$x=1$	$x=2$	$x=3$	$x=4$	$x=5$	$x=6$	$x=7$
$y=0$	0.018	0.023	0.04	0.043	0.063	0.049	0.055
$y=1$	0.031	0.053	0.086	0.096	0.147	0.153	0.142

Two useful quantities that we can compute:

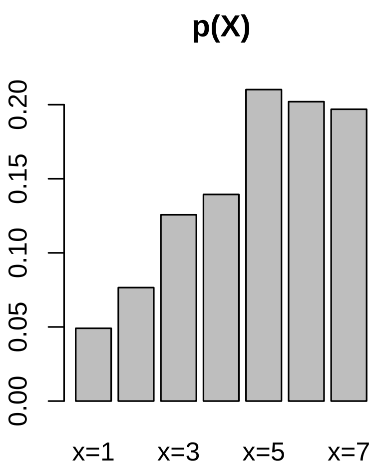
The marginal distributions (p_X and p_Y)

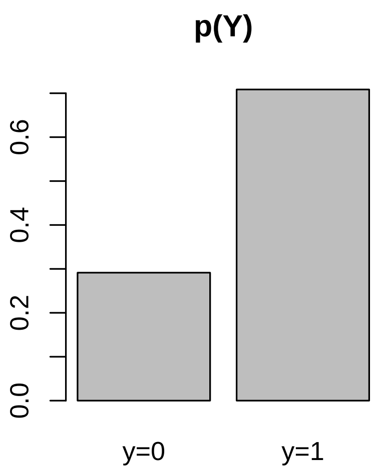
$$p_X(x) = \sum_{y \in S_Y} p_{X,Y}(x,y). \tag{17}$$

$$p_Y(y) = \sum_{x \in S_X} p_{X,Y}(x,y). \tag{18}$$

Table 2: The joint PMF for two random variables X and Y, along with the marginal distributions of X and Y.

	$x=$ 1	$x=$ 2	$x=$ 3	$x=$ 4	$x=$ 5	$x=$ 6	$x=$ 7	$p(Y)$
$y=$ 0	0.018	0.023	0.04	0.043	0.063	0.049	0.055	0.291
$y=$ 1	0.031	0.053	0.086	0.096	0.147	0.153	0.142	0.709
$p(X)$	0.049	0.077	0.126	0.139	0.21	0.202	0.197	





The conditional distributions ($p_{X|Y}$ and $p_{Y|X}$)

$$p_{X|Y}(x \mid y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} \tag{19}$$

and

$$p_{Y|X}(y \mid x) = \frac{p_{X,Y}(x,y)}{p_X(x)} \tag{20}$$

Let's do the calculation for $p_{X|Y}(x \mid y = 0)$.

Table 3: The joint PMF for two random variables X and Y, along with the marginal distributions of X and Y.

	$x = 1$	$x = 2$	$x = 3$	$x = 4$	$x = 5$	$x = 6$	$x = 7$	$p(Y)$
$y = 0$	0.018	0.023	0.04	0.043	0.063	0.049	0.055	0.291
$y = 1$	0.031	0.053	0.086	0.096	0.147	0.153	0.142	0.709
$p(X)$	0.049	0.077	0.126	0.139	0.21	0.202	0.197	

$$\begin{aligned} p_{X|Y}(1 \mid 0) =& \frac{p_{X,Y}(1,0)}{p_Y(0)} \\ =& \frac{0.018}{0.291} \\ =& 0.062 \end{aligned} \tag{21}$$

As an exercise, figure out the conditional distribution of X given Y, and the conditional distribution of Y given X.

Continuous bivariate distributions

Next, we turn to continuous bivariate/multivariate distributions.

The variance-covariance matrix:

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \rho_{XY}\sigma_X\sigma_Y \\ \rho_{XY}\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \tag{22}$$

The off-diagonals of this matrix contain the covariance between X and Y:

$$Cov(X, Y) = \rho_{XY}\sigma_X\sigma_Y$$

The joint distribution of X and Y is defined as follows:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma \right) \tag{23}$$

The joint PDF has the property that the volume under the surface sums to 1.

Formally, we would write the volume under the surface as a double integral: we are summing up the volume under the surface for both X and Y (hence the two integrals).

$$\iint_{S_{X,Y}} f_{X,Y}(x, y)\, dxdy = 1 \tag{24}$$

Here, the terms dx and dy express the fact that we are computing the volume under the surface along the X axis and the Y axis.

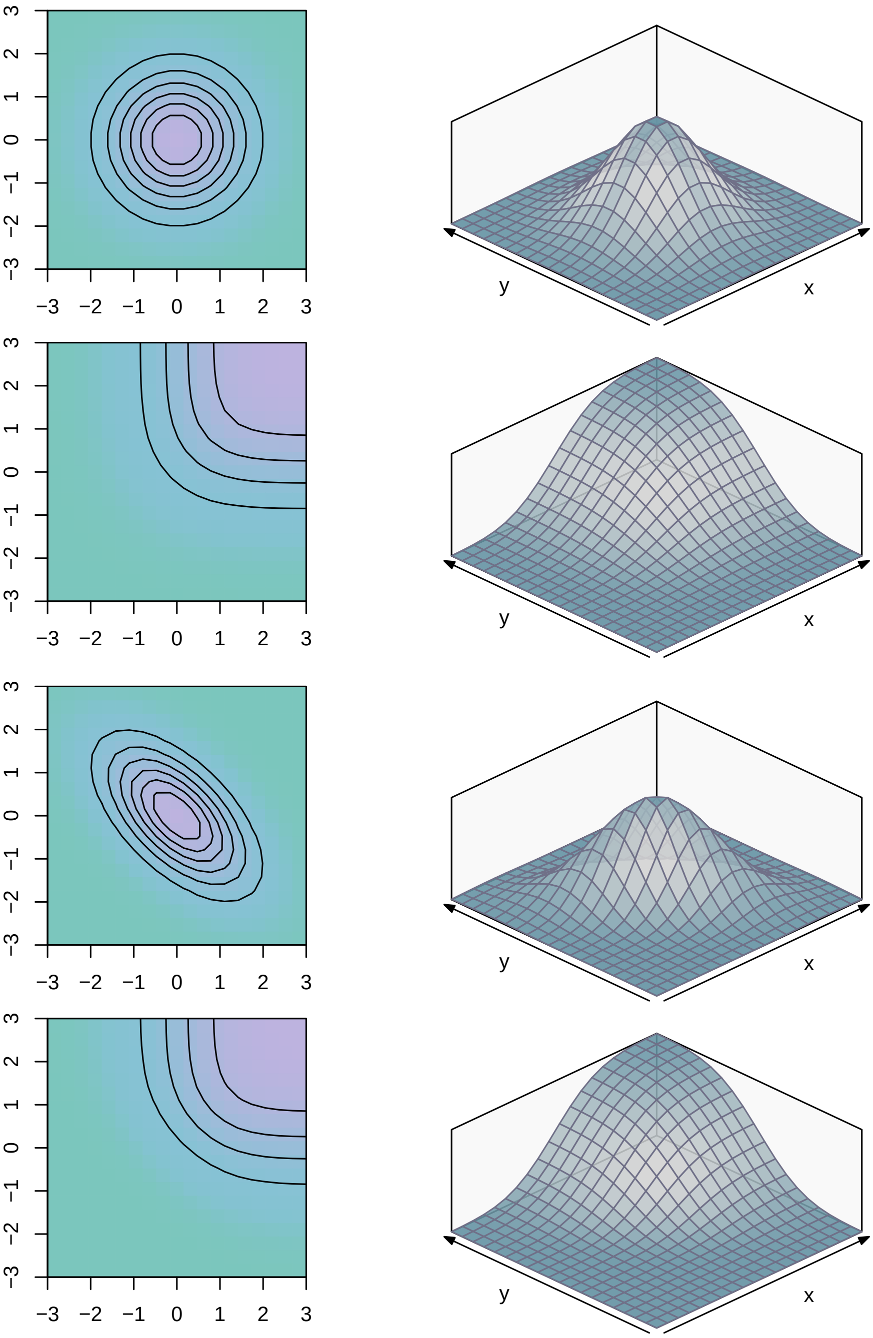
The joint CDF would be written as follows. The equation below gives us the probability of observing a value like (u, v) or some value smaller than that (i.e., some (u', v'), such that $u' < u$ and $v' < v$).

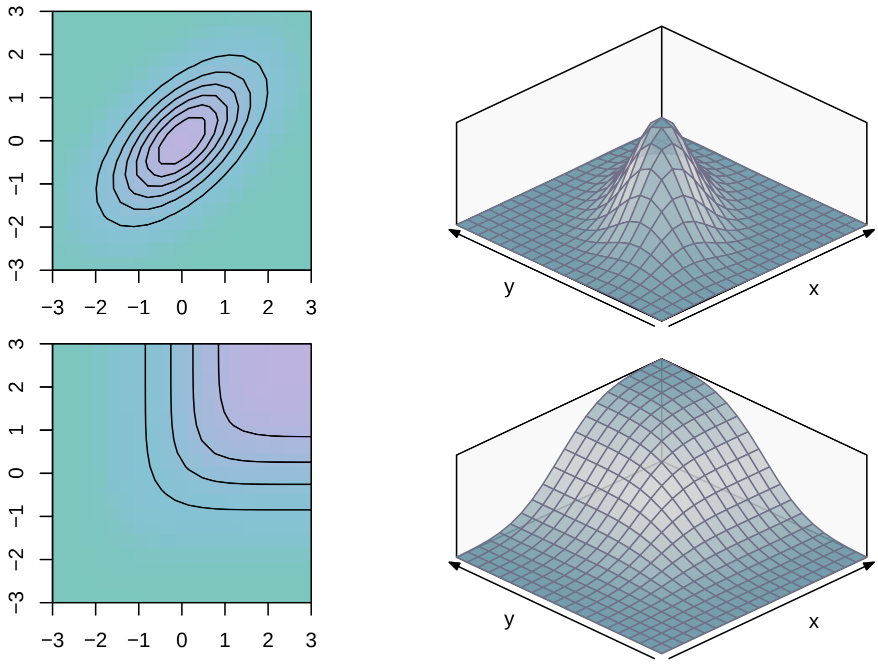
$$\begin{aligned} F_{X,Y}(u, v) =& \text{Prob}(X < u, Y < v) \\ =& \int_{-\infty}^{u} \int_{-\infty}^{v} f_{X,Y}(x, y)\, dydx \\ & \text{for } (x, y) \in \mathbb{R}^2 \end{aligned} \tag{25}$$

Just as in the discrete case, the marginal distributions can be derived by marginalizing out the other random variable:

$$f_X(x) = \int_{S_Y} f_{X,Y}(x,y)\,dy \quad f_Y(y) = \int_{S_X} f_{X,Y}(x,y)\,dx \tag{26}$$

Here, S_X and S_Y are the respective supports.

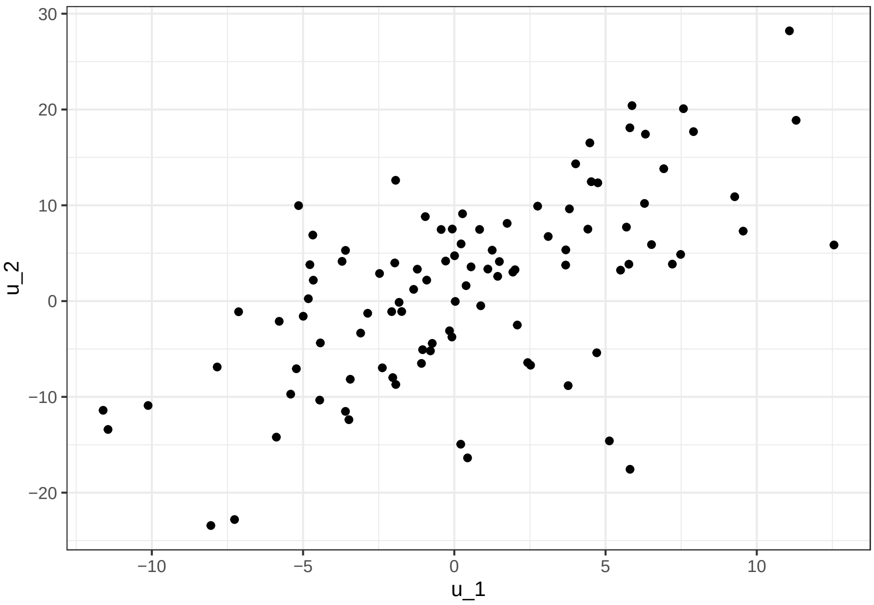




Generate simulated bivariate (multivariate) data

```
## define a variance-covariance matrix:
Sigma <- matrix(c(5^2, 5 * 10 * 0.6,
                  5 * 10 * 0.6, 10^2),
  byrow = FALSE, ncol = 2
)
## generate data:
u <- MASS::mvrnorm(n = 100, mu = c(0, 0),
                   Sigma = Sigma)
head(u, n = 3)
```

```
##            [,1]       [,2]
## [1,] -4.8250945  0.2469200
## [2,]  4.4153751  7.5110094
## [3,]  0.8741429 -0.4878949
```



One practical implication: Such bi/multivariate distributions become critically important to understand when we turn to hierarchical (linear mixed) models.