Chapter 1: Foundations

Shravan Vasishth

https://bruno.nicenboim.me/bayescogsci/

Contents

Textbook

Introduction to Bayesian Data Analysis for Cognitive Science

Nicenboim, Schad, Vasishth

https://bruno.nicenboim.me/bayescogsci/

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., $x \in S_X$. We can also sloppily write $X \in S_X$.

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)
- $X(H) = 1, X(TH) = 2, X(TTH) = 3, \dots$
- $x = 1, 2, \dots; x \in S_X$

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

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

Every discrete (continuous) random variable X has associated with it a **probability mass (density) function (PMD, 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) $F(X \leq x)$ gives the cumulative probability of observing all the events $X \leq x$.

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

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

Simulate tossing a coin ten times, with probability (which I call θ below) of heads 0.5:

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

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

```
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: The binomial

- Consider tossing a coin 10 times (number of trials, `size` in R).
- When 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

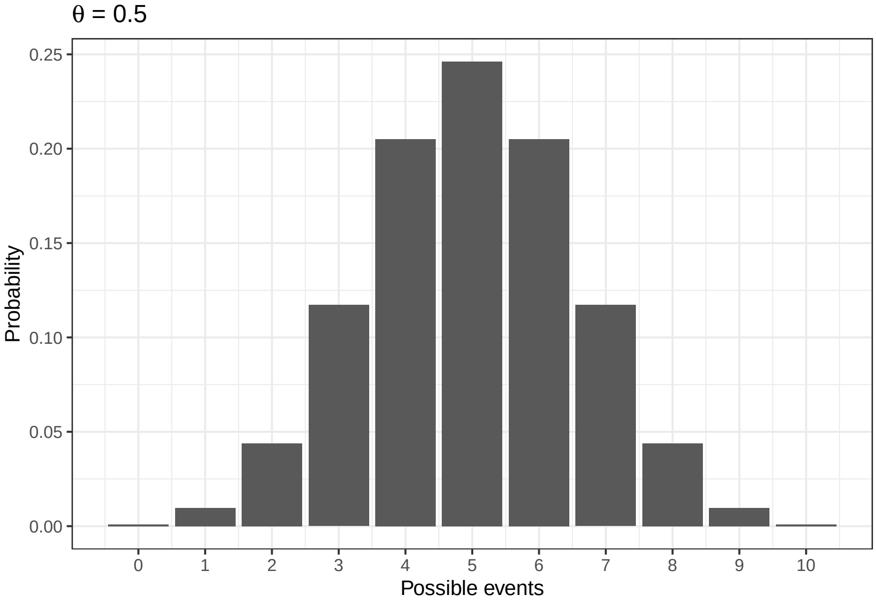
$$S_x = \{0, 1, \ldots, n\}$$

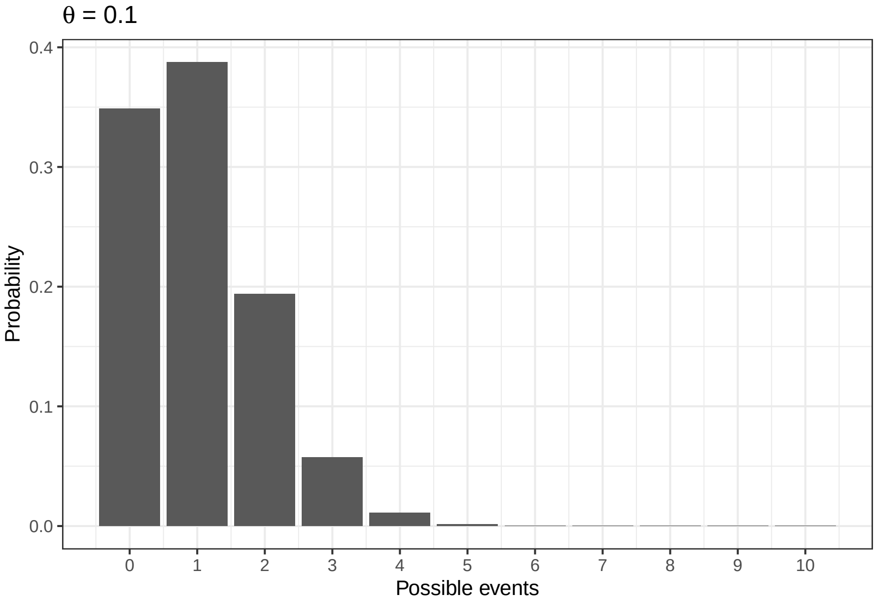
- n is the number of times the coin was tossed (the number of trials; size in R).
- $\binom{n}{x}$ is the number of ways that you can get x successes in n trials.

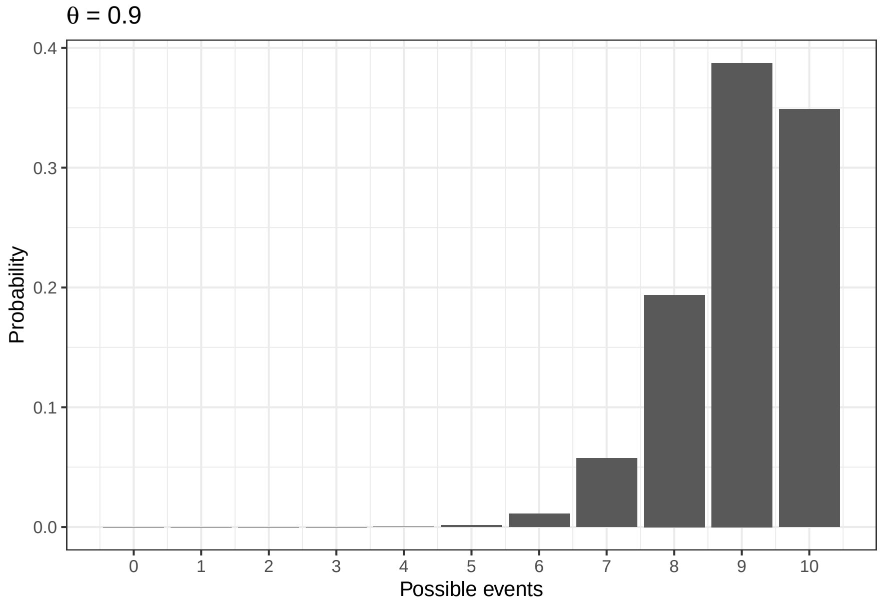
```
choose(10, 2)
```

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

- θ is the probability of success in n trials.







Four critical R functions

1. Generate random data

n: number of experiments done (**Note**: we used n for trials above)

size: the number of times the coin was tossed in each experiment

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

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

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

Compare:

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

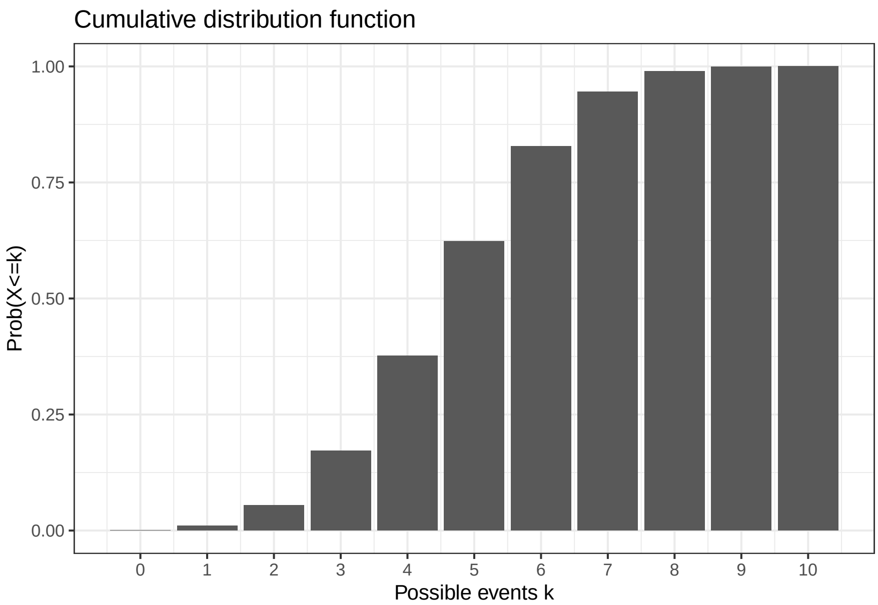
```
##  [1] 6 4 8 6 4 4 3 7 4 6
```

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

```
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
```

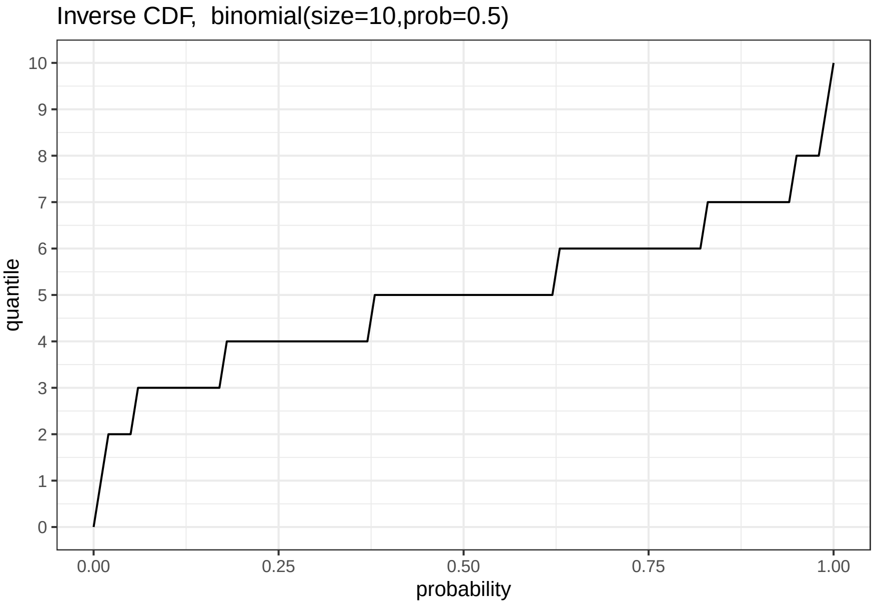
3. Compute cumulative probabilities



4. Compute quantiles using the inverse of the CDF

```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
```



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 has 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 $-\infty < x < +\infty$

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 $+\infty$ (we can **truncate** the normal to have finite values—this comes later)
- 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 Prob($x_1 < X < x_2$) =?, we need the cumulative distribution function.

The cumulative distribution function (CDF) is

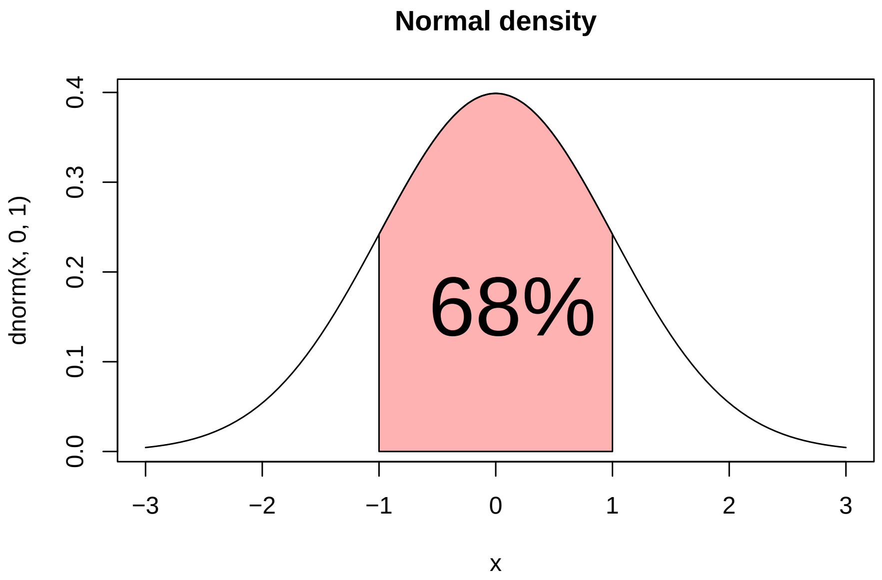
$$P(X < u) = F(X < u) = \int_{-\infty}^{u} f(x)\,dx \quad (8)$$

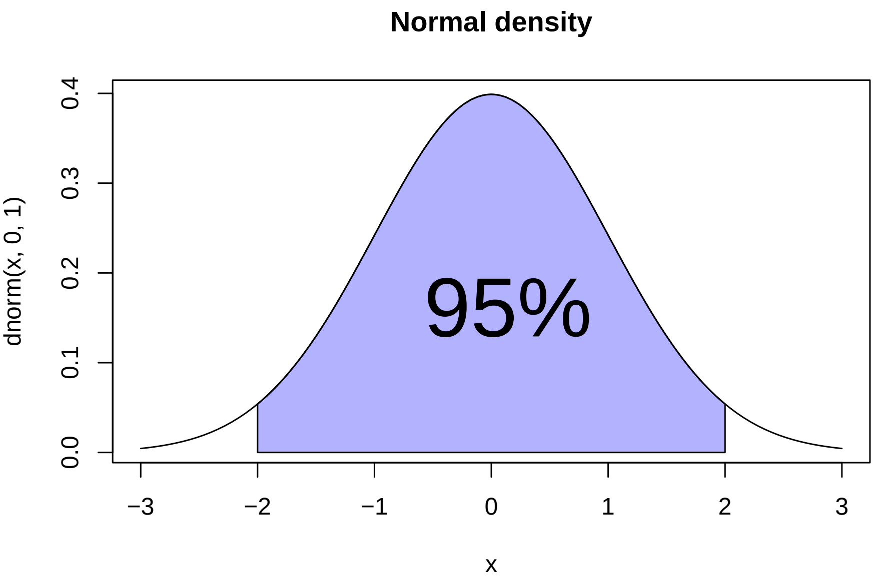
- The integral sign $\int$ is just the summation symbol in continuous space.
- Recall the summation in the CDF of the Bernoulli!

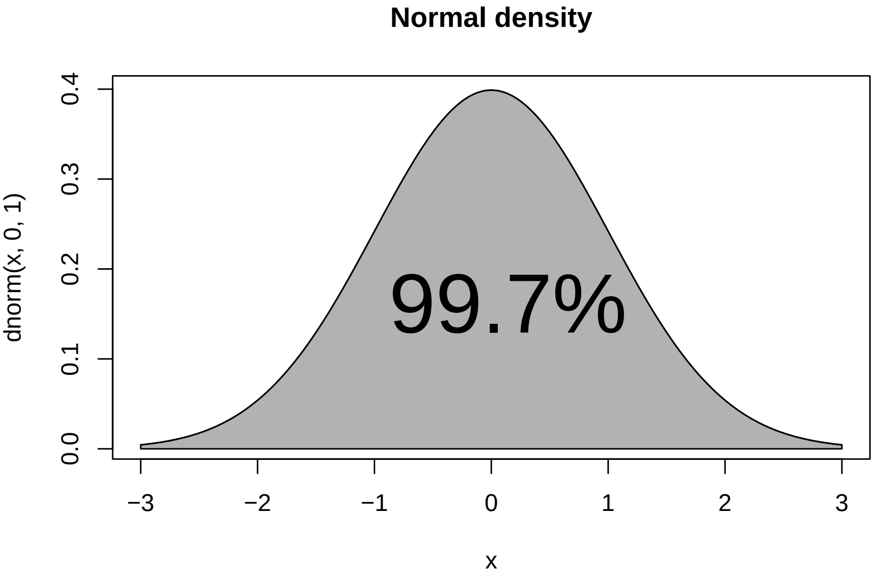
The standard normal distribution

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

- Prob($-1 < X < +1$) = 0.68
- Prob($-2 < X < +2$) = 0.95
- 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

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

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

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

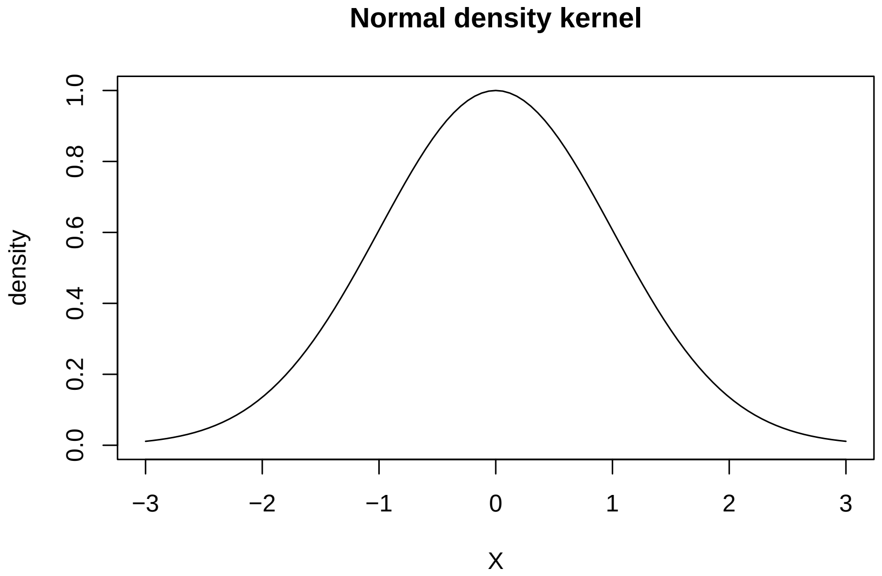
Sum up the area under the curve $\int g(x)\,dx$:

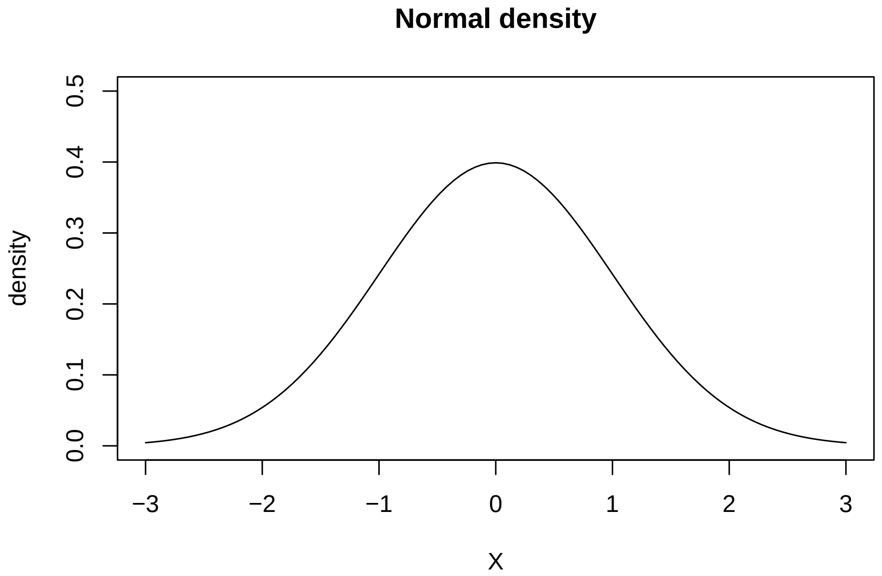
```
normkernel <- function(x, mu = 0, sigma = 1) {
  exp((-(x - mu)^2 / (2 * (sigma^2))))
}

integrate(normkernel, lower = -Inf, upper = +Inf)
```

```
## 2.506628 with absolute error < 0.00023
```

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:

$$k \int g(x)\, dx = 1 \tag{10}$$

Solving for k just amounts to computing:

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

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

The four key functions for the normal distribution

Recall these key functions for the Bernoulli:

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

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

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

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

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

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

```
extraDistr::qbern(p = 1, prob = 0.5)
```

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

For the binomial:

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

```
## [1] 3
```

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

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

```r
pbinom(q = 2, size = 10, prob = 0.5)
```

```
## [1] 0.0546875
```

```r
qbinom(p = 0.0546875, size = 10, prob = 0.5)
```

```
## [1] 2
```

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

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

```
## [1] -0.348 -0.967 -0.961 -1.909 -0.300
```

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

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

```
## [1] -2.057 -0.411  0.207  0.058 -1.171
```

2. Compute probabilities using CDF: pnorm

Some examples of usage:

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

```r
pnorm(2)
```

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

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

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

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

3. Compute quantiles: qnorm

```r
qnorm(0.9772499)
```

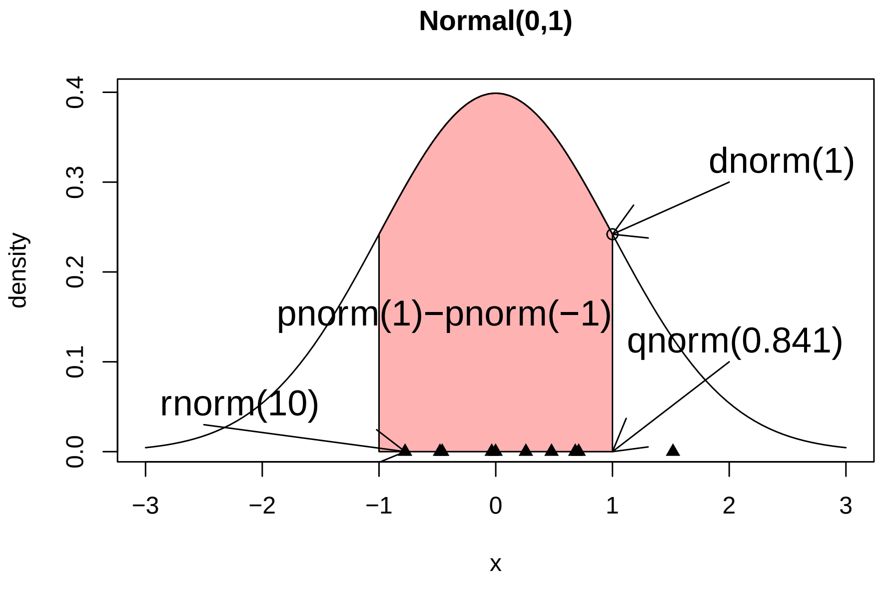
```
## [1] 2.000001
```

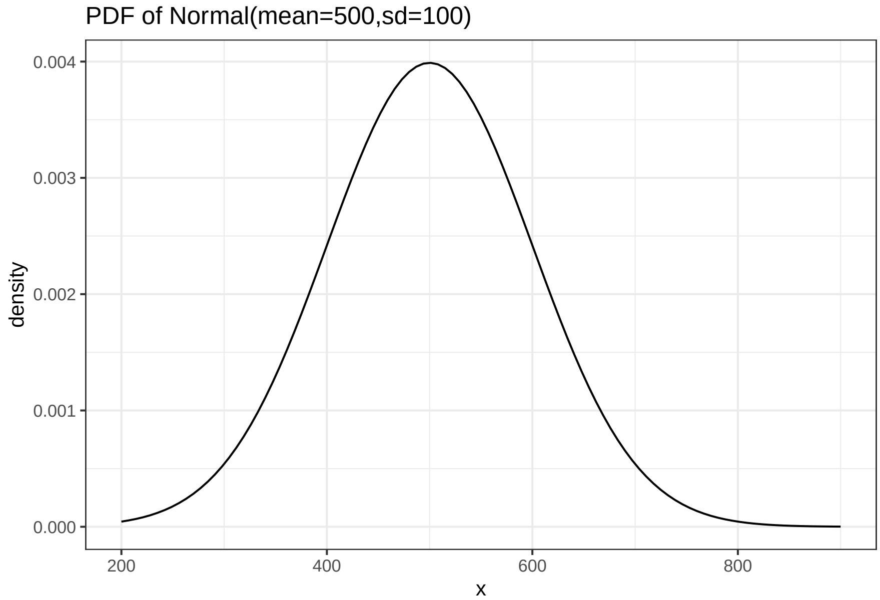
4. Compute the probability density: dnorm

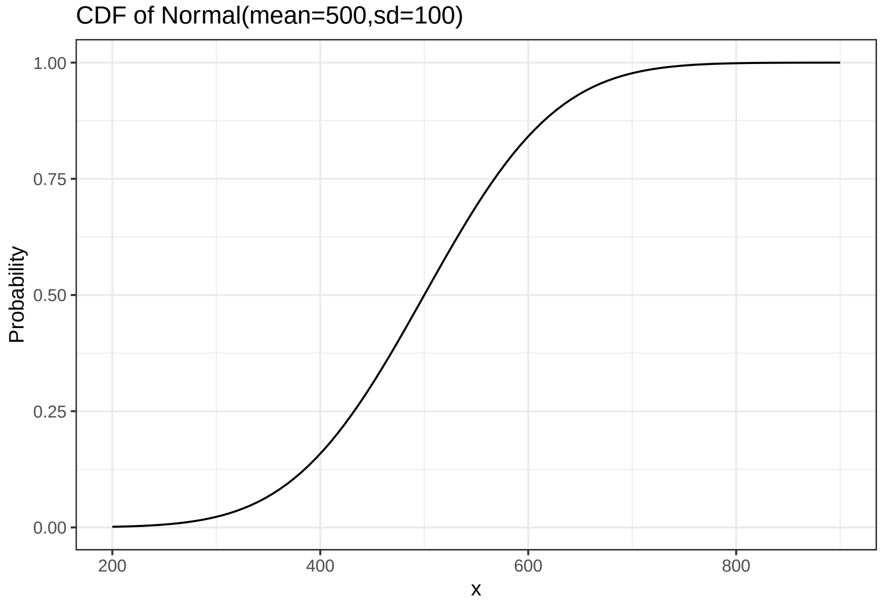
```r
dnorm(2)
```

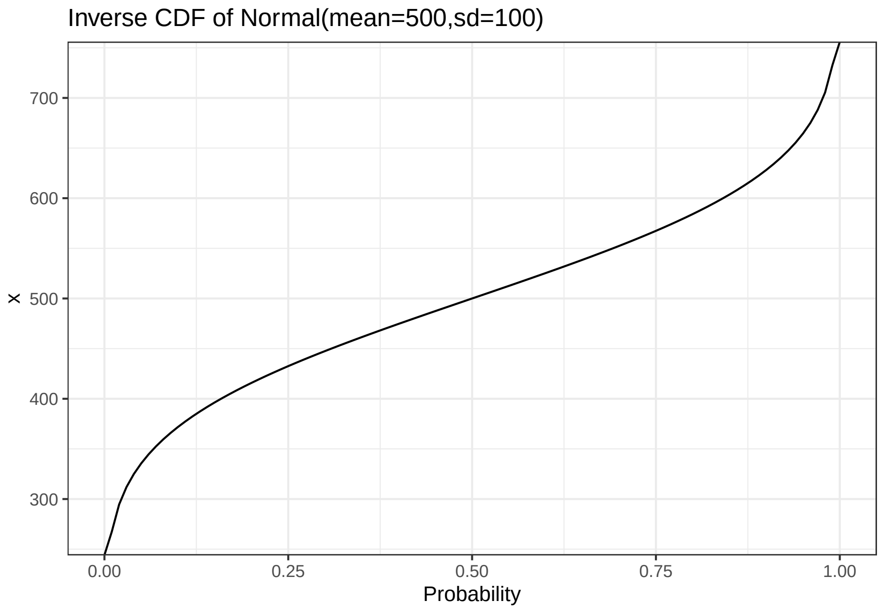
```
## [1] 0.05399097
```

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.




PDF of Normal(mean=500,sd=100)
density
0.004
0.003
0.002
0.001
0.000
200
400
600
800
x


CDF of Normal(mean=500,sd=100)
Probability
1.00
0.75
0.50
0.25
0.00
200
400
600
800
x


Inverse CDF of Normal(mean=500,sd=100)
x
700
600
500
400
300
0.00
0.25
0.50
0.75
1.00
Probability

The expectation and variance of an RV

The expectation of a discrete random variable Y with probability mass function f(y), is defined as

$$E[Y] = \sum_y y \cdot f(y) \tag{12}$$

Example: Toss a fair coin once. The possible events are Tails (represented as 0) and Heads (represented as 1), each with equal probability, 0.5. The expectation is:

$$E[Y] = \sum_y y \cdot f(y) = 0 \cdot 0.5 + 1 \cdot 0.5 = 0.5 \tag{13}$$

The variance is defined as:

$$Var(Y) = E[Y^2] - E[Y]^2 \tag{14}$$

In the **binomial**, $Y \sim \binom{n}{k}\theta^k(1-\theta)^{n-k}$:

- The expectation: $E[Y] = n\theta$
  - Estimated by $\hat{\theta} = \frac{k}{n}$
- The variance: $Var(Y) = n\theta(1-\theta)$
  - Estimated by $Var(y) = n\hat{\theta}(1-\hat{\theta})$

In the **normal**, $Y \sim Normal(\mu, \sigma)$:

- The expectation: $E[Y] = \int y f(y)\, dy = \mu$
  - Estimated by $\hat{\mu} = \bar{y} = \frac{\sum y}{n}$
- The variance: $Var(Y) = \sigma^2$
  - Estimated by $\hat{\sigma}^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}$

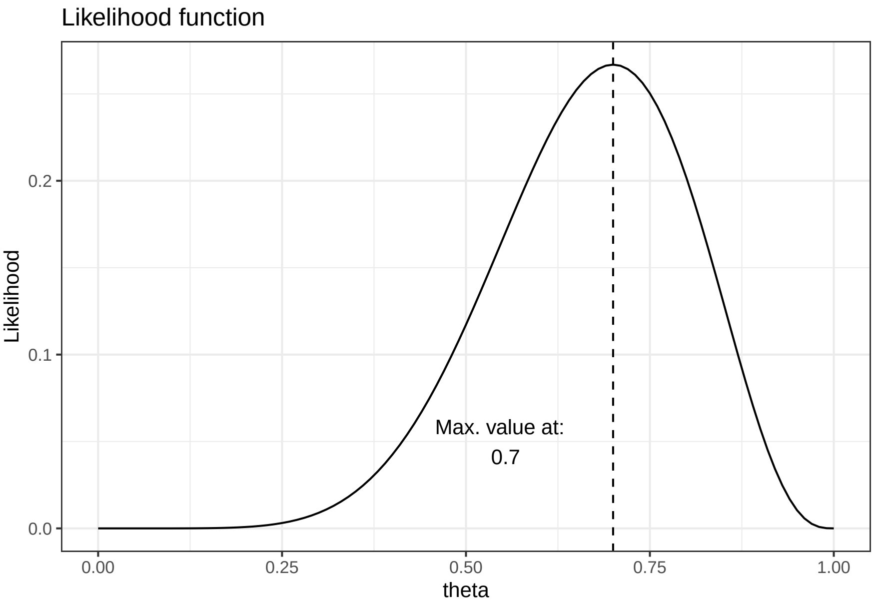
The likelihood function (Binomial)

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

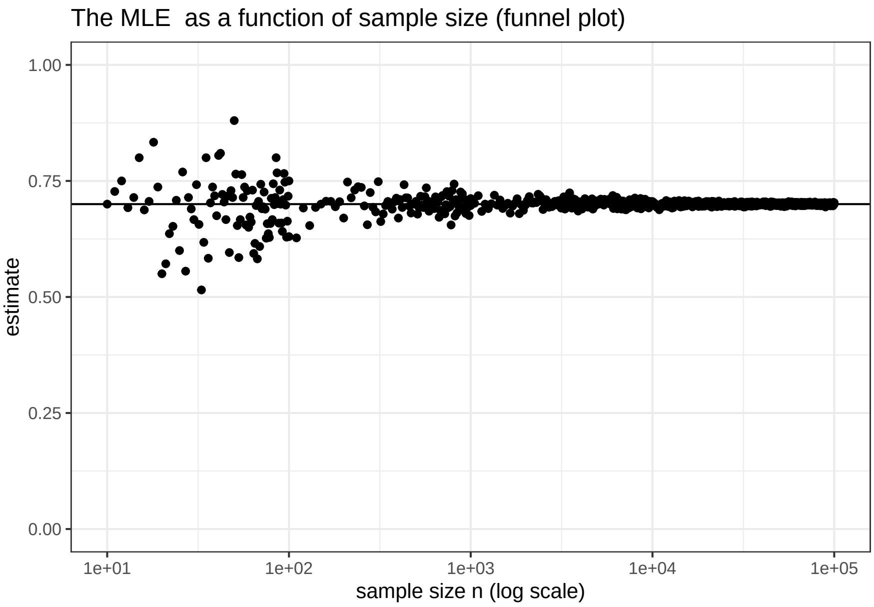
For example, suppose that we record $n = 10$ trials, and observe $k = 7$ successes. The likelihood function is:

$$\mathcal{L}(\theta|k=7,n=10) = \binom{10}{7}\theta^7(1-\theta)^{10-7} \quad (15)$$

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.



The likelihood function (Normal)

$$\mathcal{L}(\mu, \sigma | x) = Normal(x, \mu, \sigma) \qquad (16)$$

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

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

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

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

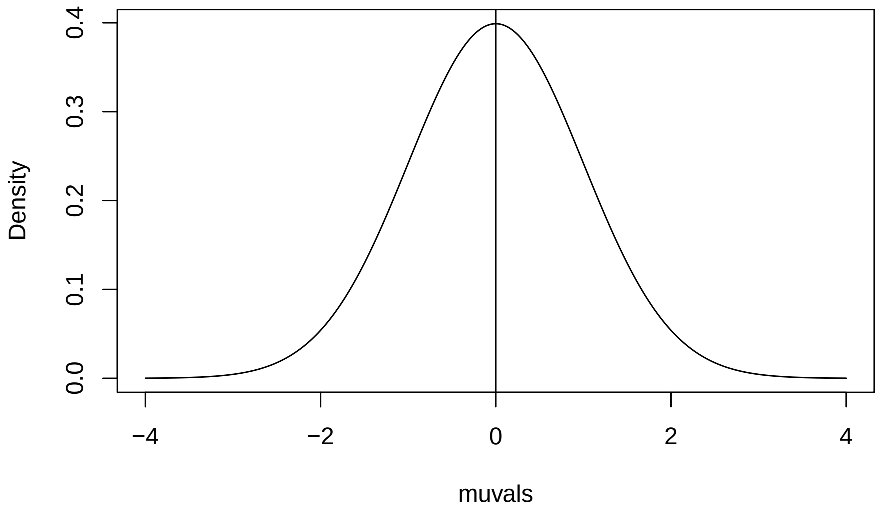
```
dnorm(x,mean=0,sd=10)
```

```
## [1] 0.03989423
```

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

```
## [1] 0.02419707
```

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



If we have two independent data points, the joint likelihood for two arbitrary values of μ and σ:

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

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

```
## log likelihood:
dnorm(x1,mean=0,sd=1,log=TRUE) +
  dnorm(x2,mean=0,sd=1,log=TRUE)
```

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

```
## 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:

```
## 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
```

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

$$\mathcal{L}(\mu, \sigma|x) = \prod_{i=1}^{n} Normal(x_i, \mu, \sigma) \tag{17}$$

or

$$\ell(\mu, \sigma|x) = \sum_{i=1}^{n} log(Normal(x_i, \mu, \sigma)) \tag{18}$$

Bivariate/multivariate 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.

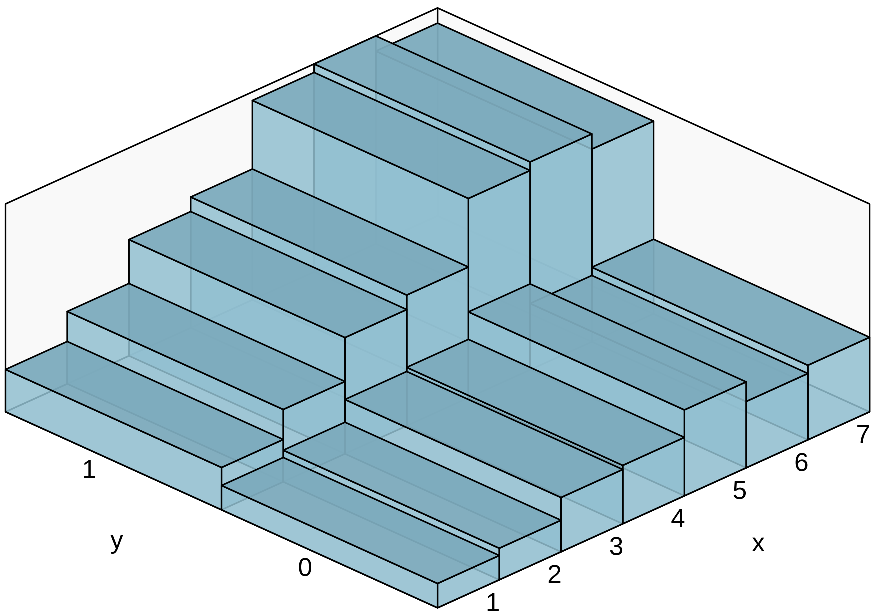


Table 1: The joint PMF for two random variables X and Y.

	$x = 1$	$x = 2$	$x = 3$	$x = 4$	$x = 5$	$x = 6$	$x = 7$
$y = 0$	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

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** $p_{X,Y}(x,y)$.

Two useful quantities that we can compute:

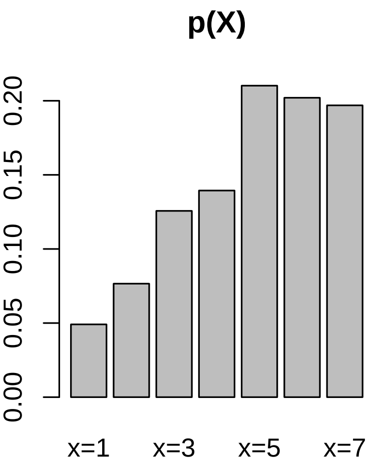
The marginal distributions (p_X and p_Y):

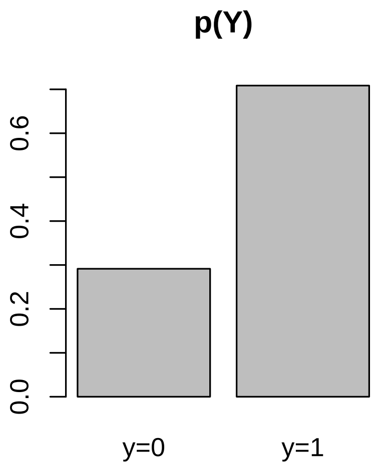
$$p_X(x) = \sum_{y \in S_Y} p_{X,Y}(x,y). \tag{19}$$

$$p_Y(y) = \sum_{x \in S_X} p_{X,Y}(x,y). \tag{20}$$

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 $p_{Y|X}$)

$$p_{X|Y}(x \mid y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} \tag{21}$$

and

$$p_{Y|X}(y \mid x) = \frac{p_{X,Y}(x,y)}{p_X(x)} \tag{22}$$

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	

$$\begin{aligned} p_{X|Y}(1 \mid 0) =& \frac{p_{X,Y}(1,0)}{p_Y(0)} \\ =& \frac{0.018}{0.291} \\ =& 0.062 \end{aligned} \tag{23}$$

Next, we turn to continuous bivariate/multivariate distributions.

The variance-covariance matrix:

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \rho_{XY}\sigma_X\sigma_Y \\ \rho_{XY}\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \tag{24}$$

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:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma \right) \tag{25}$$

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

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

$$\iint_{S_{X,Y}} f_{X,Y}(x,y)\,dxdy = 1 \tag{26}$$

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

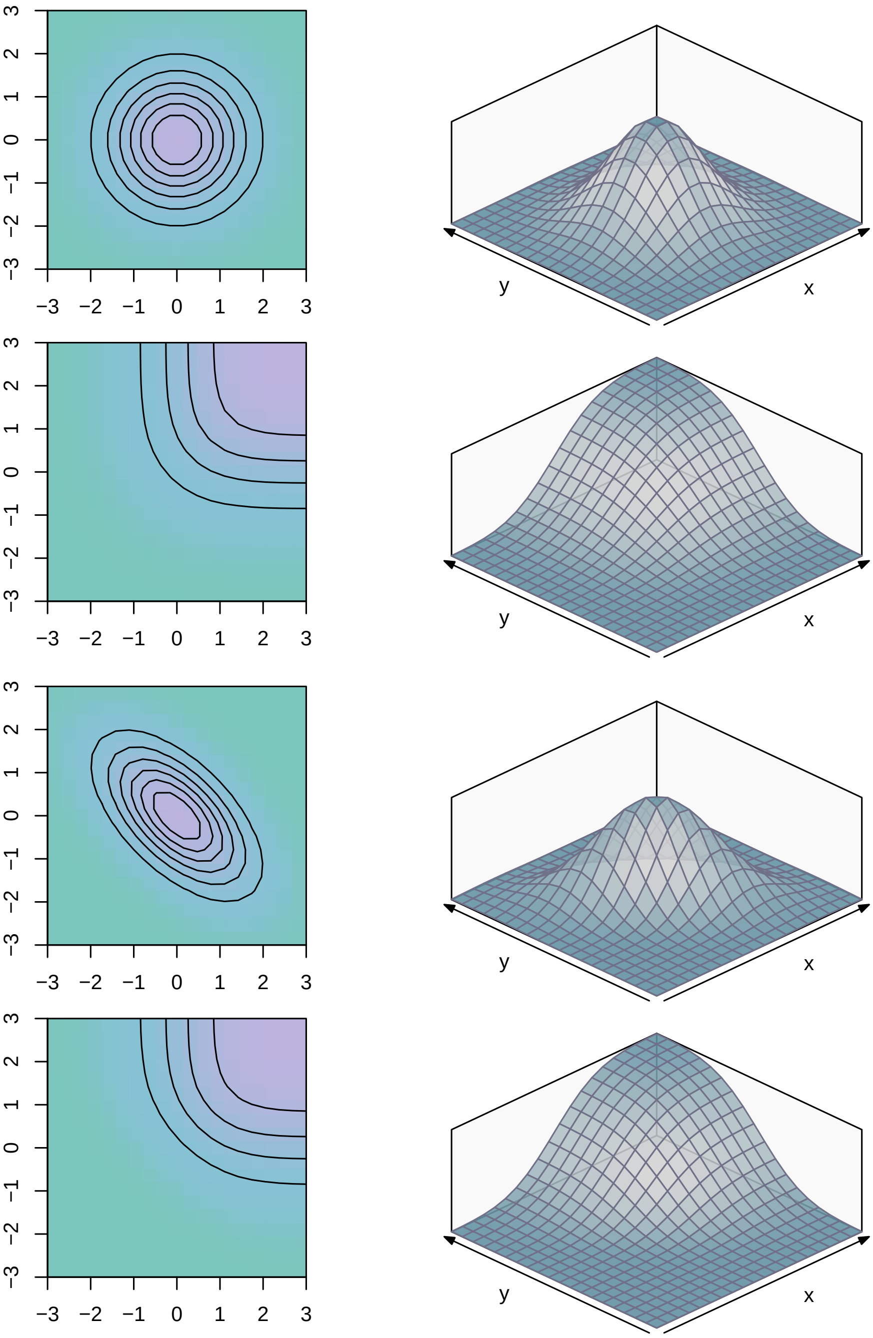
The joint CDF would be written as follows. The equation below gives us the probability of observing a value like (u,v) or some value smaller than that (i.e., some (u',v'), such that $u' < u$ and $v' < v$).

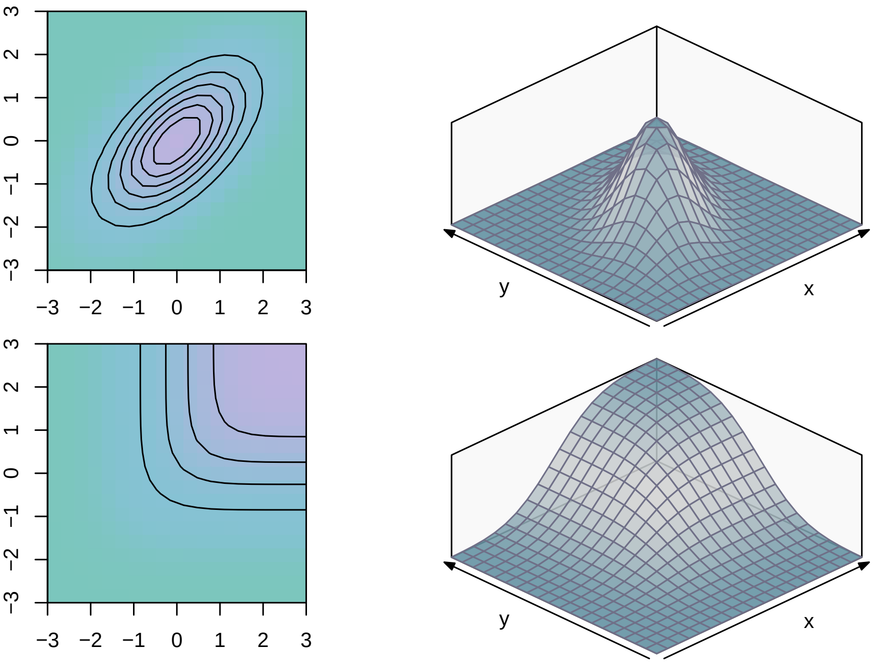
$$\begin{aligned} F_{X,Y}(u,v) =& \text{Prob}(X < u, Y < v) \\ =& \int_{-\infty}^{u} \int_{-\infty}^{v} f_{X,Y}(x,y)\,dydx \\ & \text{for } (x,y) \in \mathbb{R}^2 \end{aligned} \tag{27}$$

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

$$f_X(x) = \int_{S_Y} f_{X,Y}(x,y)\,dy \quad f_Y(y) = \int_{S_X} f_{X,Y}(x,y)\,dx \tag{28}$$

Here, S_X and S_Y are the respective supports.

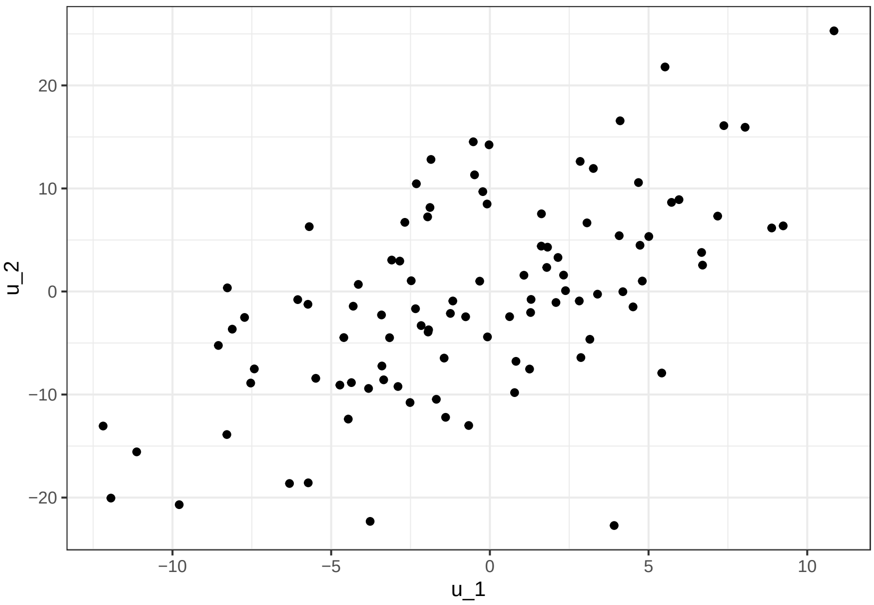




Generate simulated bivariate (multivariate) data

```
## 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,] -11.940225 -20.047789
## [2,]  -3.160155  -4.485012
## [3,]  -5.723035 -18.566348
```



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