Introduction to Bayesian Data Analysis
Foundational Ideas (Chapter 1 of book)

Prof. Dr. Shravan Vasishth
Professor, Linguistics
Cognitive Science / Linguistics, Uni Potsdam, Germany

Random variables

A random variable X is a function $X : S \rightarrow \mathbb{R}$ that associates to each outcome $\omega \in S$ 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 random variable: the number of coin tosses till H

- $X : \omega \rightarrow x$
- ω: H, TH, TTH,. . . (infinite)
- $X(H) = 1, X(TH) = 2, X(TTH) = 3, \dots$
- $x = 1, 2, \dots; x \in S_X$

Random variables

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

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

Random variables: PMFs and PDFs

- 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:

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

defined by

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

Random variables: PMFs and PDFs

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 $X \leq x$.

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

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

Random variables: Demo code

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 0 0 0 0 0 1 0 0 0
```

The probability mass function: Bernoulli

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

where x can have values 0, 1.

Random variables: Demo code

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.

Random variables: Demo code

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

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

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

## [1] 1
```

Random variables: Demo code

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

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

## [1] 0.5
```

The main points of this lecture

- The definition of a random variable
- An example of a discrete random variable (the Bernoulli)
- Generating random data using **rbern**.
- Using the PMF to compute the probabilities of different possible events (the **dbern** function)
- Computing the cumulative distribution function using **pbern**.

You can start reading chapter 1 of the textbook!

Another example of a discrete distribution: 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.

Bernoulli PMF

$$\theta^x (1-\theta)^{1-x} \text{ where } S_x = \{0, 1\}$$

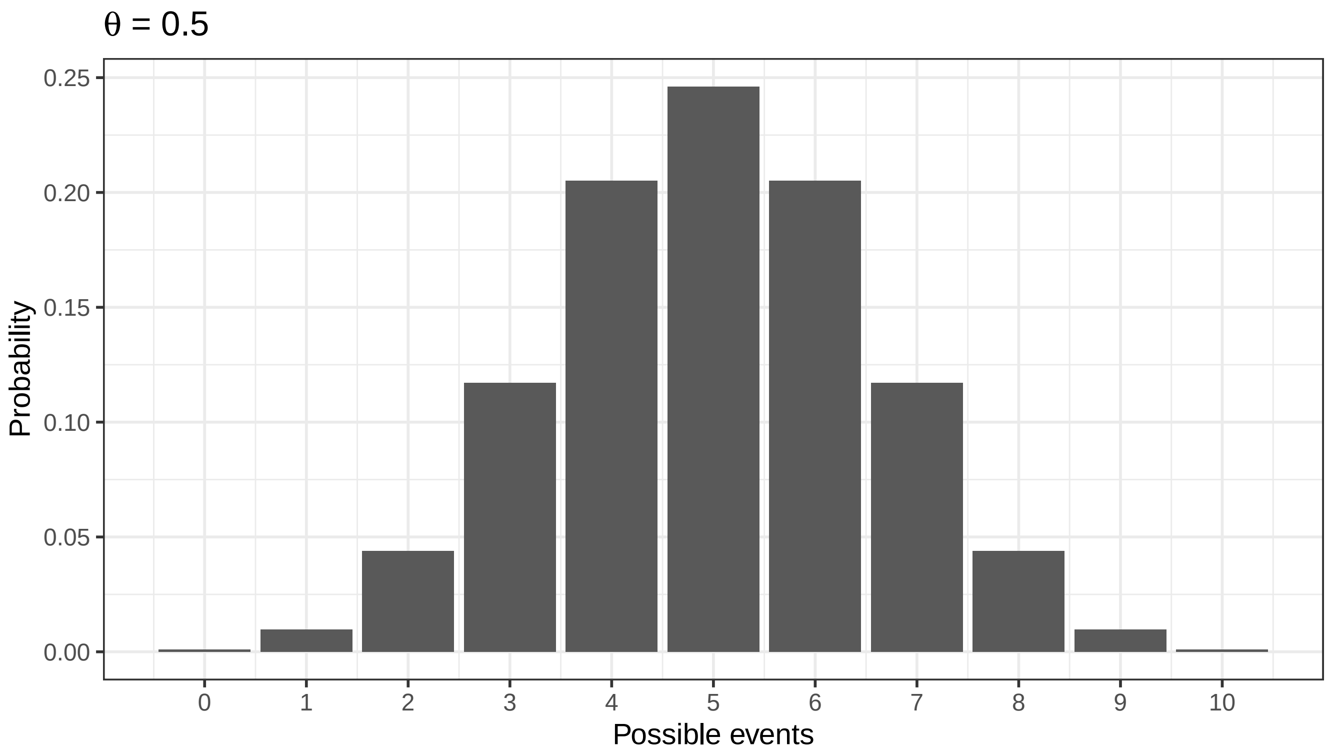
Binomial PMF

$$\binom{n}{x} \theta^x (1-\theta)^{n-x} \text{ where } S_x = \{0, 1, \dots, n\}$$

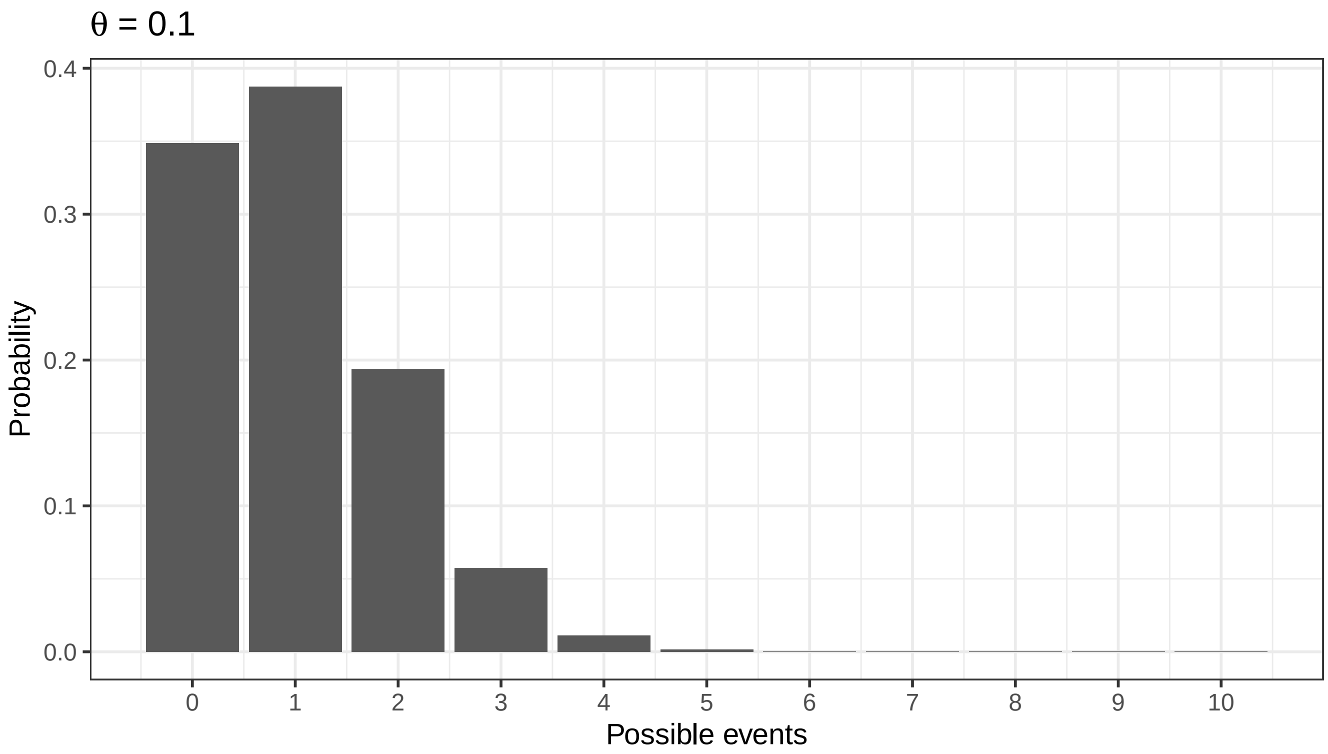
- n is the number of times the coin was tossed (the number of trials; size in R).
- $\binom{n}{x}$ is the number of ways that you can get x successes in n trials.

```
# how many ways can you get two successes in 10 trials?
choose(10, 2)
## [1] 45
```

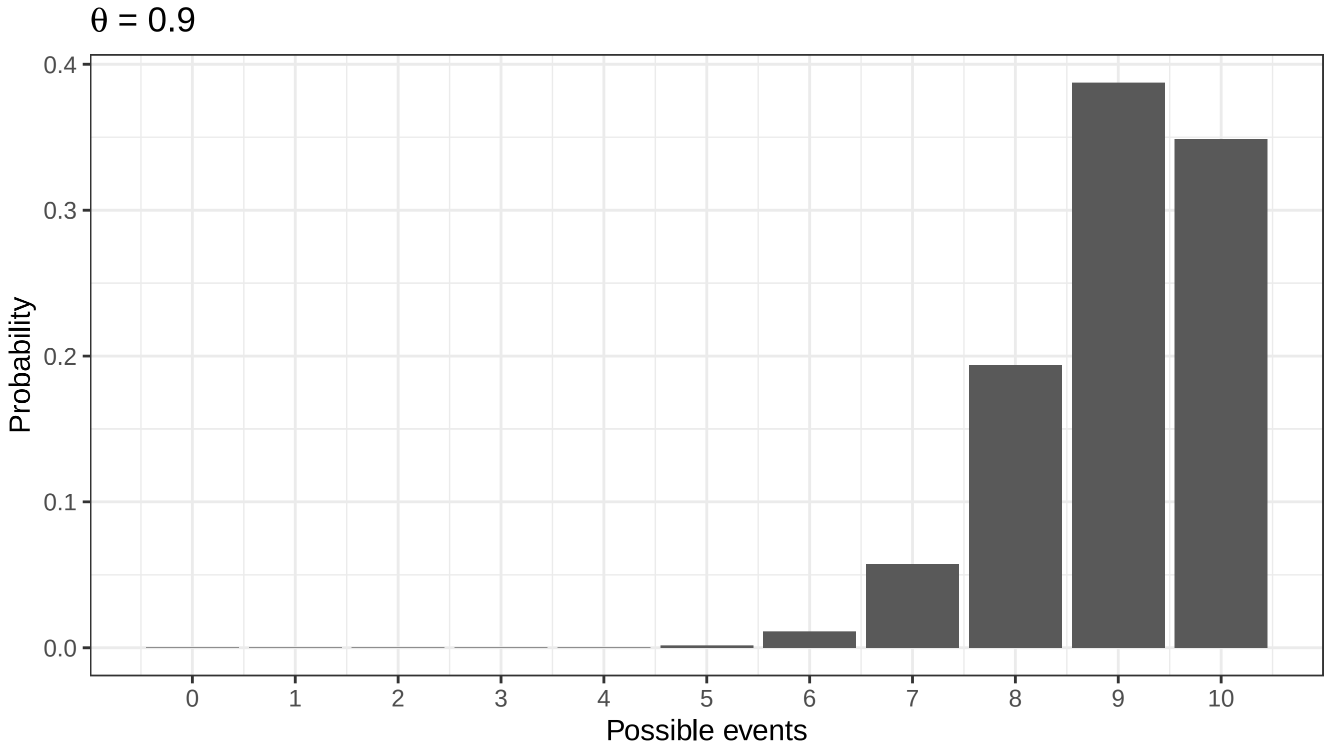
Another example of a discrete distribution: The binomial



Another example of a discrete distribution: The binomial



Another example of a discrete distribution: The binomial



Another example of a discrete distribution: The binomial

Four critical R functions (the d-p-q-r family): rbinom, dbinom, pbinom, and qbinom.

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] 0 0 0 0 0 0 1 0 0 0

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

Compare:

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

##  [1] 4 6 2 2 5 4 4 5 6 7
```

Another example of a discrete distribution: The binomial

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

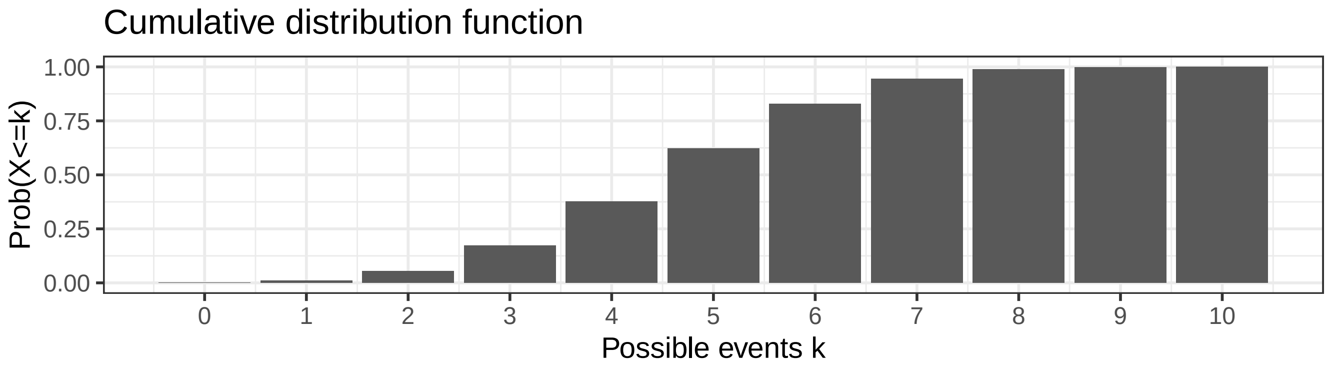
```
probs <- round(dbinom(0:10, size = 10, prob = 0.5), 3)
x <- 0:10
```

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

Another example of a discrete distribution: The binomial

Compute cumulative probabilities using the CDF $Prob(X \leq x)$

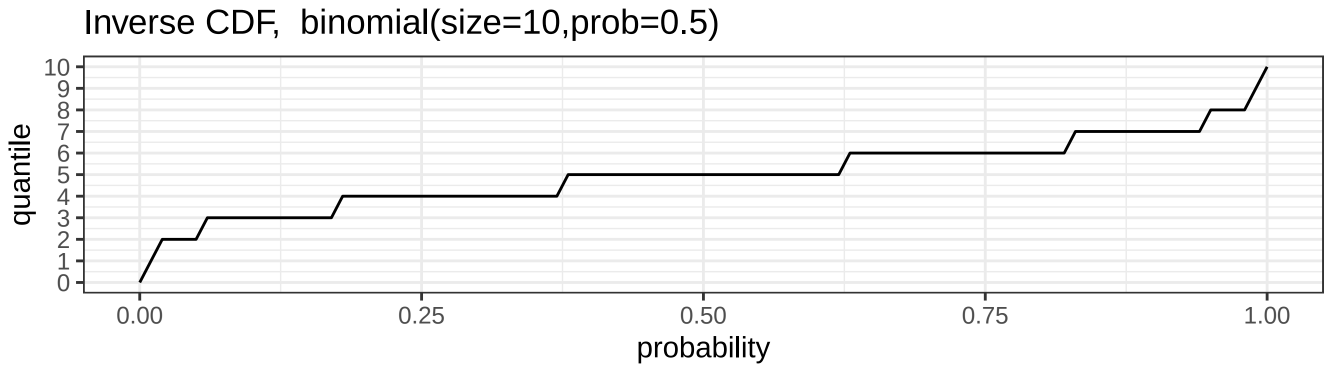
```
pbinom(0:10, size = 10, prob = 0.5)

##  [1] 0.0009765625 0.0107421875 0.0546875000 0.1718750000 0.3769531250
##  [6] 0.6230468750 0.8281250000 0.9453125000 0.9892578125 0.9990234375
## [11] 1.0000000000
```



Another example of a discrete distribution: The binomial

Compute quantiles using the inverse of the CDF: What is the quantile q such that $Prob(X \leq q) = p$?

```
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
```



Summary and the next step

The d-p-q-r family of functions:

- rbinom
- dbinom
- pbinom
- qbinom

Reminder: Discrete random variables

A random variable X is a function $X : S \rightarrow \mathbb{R}$ that associates to each outcome $\omega \in S$ 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 $X \in S_X$.

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

$$p_X : S_X \rightarrow [0,1] \tag{3}$$

defined by

$$p_X(x) = P(X(\omega) = x), x \in S_X \tag{4}$$

The cumulative distribution function (CDF): $\sum p_X(x)$

Reminder: Discrete random variables

Key R functions (Bernoulli random variable):

- rbern: generate random data
- dbern: the probability of a particular event
- pbern: the CDF.

Key R functions (Binomial random variable):

- rbinom: generate random data
- dbinom: the probability of a particular event
- pbinom: the CDF
- qbinom: the inverse of the CDF.

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

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

Continuous random variables: The normal random variable

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

$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \text{ where } -\infty < x < +\infty \tag{7}$$

- 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)

Continuous random variables: The normal random variable

- 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 Prob$(X = x) = 0$.
- In any continuous distribution, we can compute probabilities like 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.

Continuous random variables: The normal random variable

The cumulative distribution function (CDF) is

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

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

Continuous random variables: The normal random variable

Summary

- We saw a first example of a continuous random variable, the normal distribution.
- Next, I will unpack some of the important properties of this important distribution.

Reminder: Continuous random variables: The normal random variable

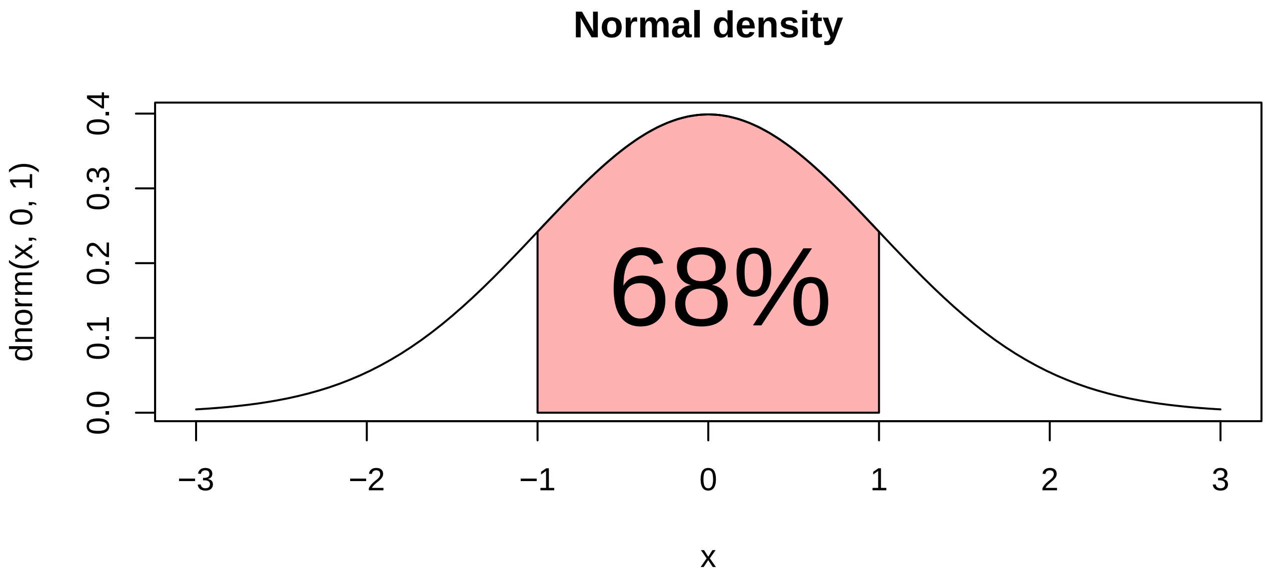
The main points of the previous lecture (1-3):

- We saw an example of a continuous distribution: the normal distribution
- The probability of a point value in a continuous distribution, Prob$(X = x_1)$ is **always** 0.
- The CDF of a continuous distribution allows us to work out probabilities like Prob$(x_1 < X < x_2)$ where $x_1 < x_2$.

In this lecture, I unpack some important properties of the Normal distribution.

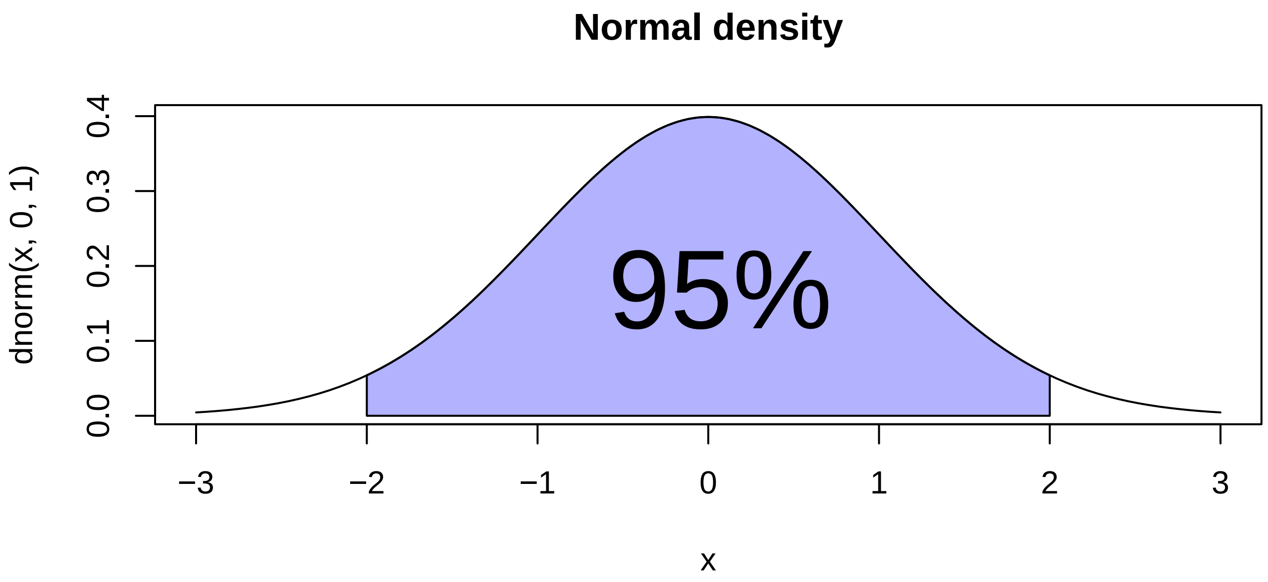
Continuous random variables: The normal random variable

Consider a special instance of the normal distribution, the **standard normal distribution**: $Normal(\mu = 0, \sigma = 1)$.
In the $Normal(\mu = 0, \sigma = 1)$, Prob$(-1 < X < +1) = 0.68$.



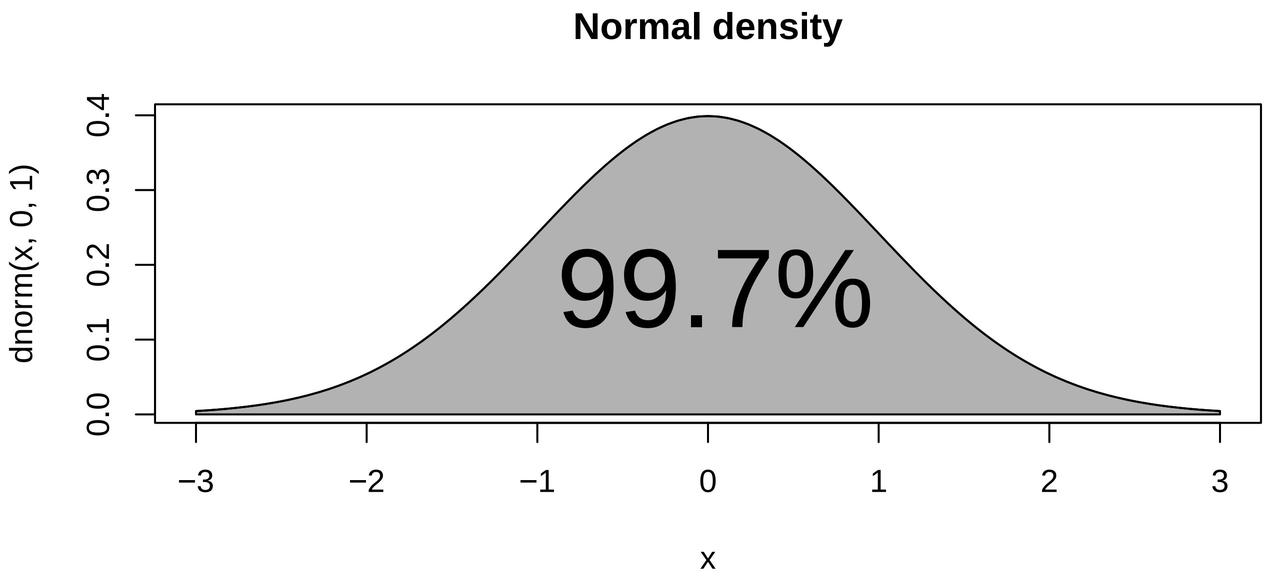
Continuous random variables: The normal random variable

In the $Normal(\mu = 0, \sigma = 1)$, Prob$(-2 < X < +2) = 0.95$ (more accurately: Prob$(-1.96 < X < +1.96) = 0.95$).



Continuous random variables: The normal random variable

In the $Normal(\mu = 0, \sigma = 1)$, Prob$(-3 < X < +3) = 0.997$.



Continuous random variables: The normal random variable

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 $Normal(\mu, \sigma)$,

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

The continuous values on the x-axis (here, $\pm 1, 2, 3$) are called **quantiles**.

Continuous random variables: The normal random variable

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:

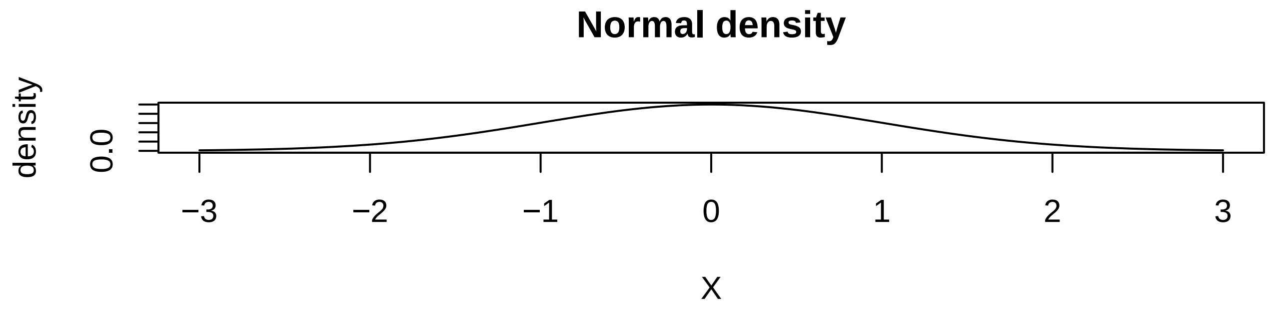
Sum up the area under the curve $\int g(x)\, dx$:

```
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
```

Continuous random variables: The normal random variable

The shape doesn't change of course:

```
x <- seq(-10, 10, by = 0.01)

plot(function(x) normkernel(x), -3, 3,
  main = "Normal density", ylim = c(0, 1),
  ylab = "density", xlab = "X"
)
```



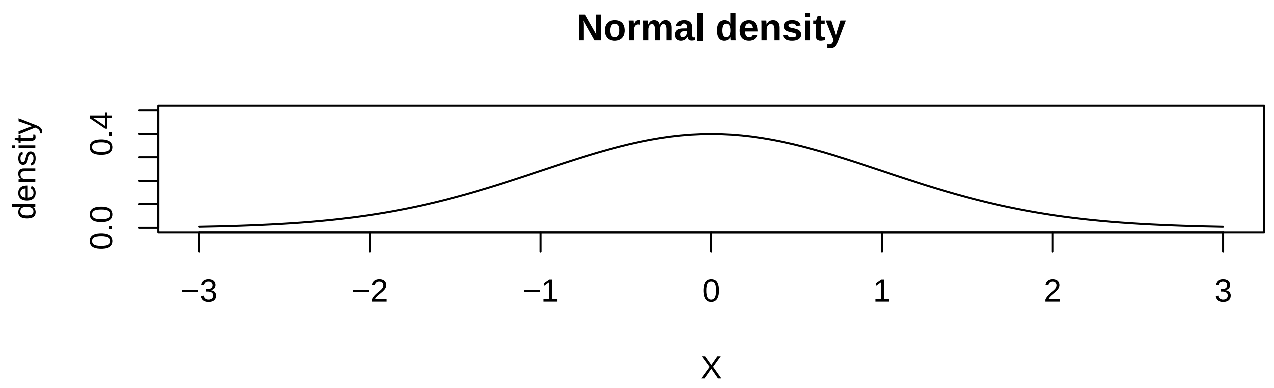
Continuous random variables: The normal random variable

Adding the **normalizing constant**, $\frac{1}{\sqrt{2\pi\sigma^2}}$, makes the above kernel density a PDF.

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

integrate(norm, lower = -Inf, upper = +Inf)

## 1 with absolute error < 9.4e-05
```

Continuous random variables: The normal random variable

```
x <- seq(-10, 10, by = 0.01)

plot(function(x) dnorm(x), -3, 3,
  main = "Normal density", ylim = c(0, 0.5),
  ylab = "density", xlab = "X"
)
```



Continuous random variables: The normal random variable

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.

Continuous random variables: The normal random variable

Summary

We saw some important properties of the Normal distribution

- The area under the curve gives the probability of observing values between x_1 and x_2, Prob$(x_1 < X < x_2)$, where $x_1 < x_2$.
- The Normal PDF has two parts, the normalizing constant and the kernel.
- The normalizing constant can be worked out given the kernel because in a proper PDF the area under the curve must sum to 1.

Next, we will look at some key functions (the d-p-q-r family) in R that relate to the Normal distribution.

Reminder: Continuous random variables: The normal random variable

We saw some important properties of the Normal distribution

- The area under the curve gives the probability of observing values between x_1 and x_2, Prob$(x_1 < X < x_2)$, where $x_1 < x_2$.
- The Normal PDF has two parts, the normalizing constant and the kernel.
- The normalizing constant can be worked out given the kernel because in a proper PDF the area under the curve must sum to 1.

Next, let's look at some key functions in R that relate to the Normal distribution.

Continuous random variables: The normal random variable

Reminder:

Recall these key functions:

Bernoulli:

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

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

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
```

Continuous random variables: The normal random variable

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:

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

## [1]  0.2796176  2.5796858 -1.0757202 -1.2920775 -0.5946934
```

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

```
rnorm(5)

## [1]  0.5726697 -2.2484905 -0.1656357  2.3950025 -1.4808078
```

Continuous random variables: The normal random variable

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
```

Continuous random variables: The normal random variable

3. Compute quantiles corresponding to probabilities using the inverse of the CDF: qnorm

```
qnorm(0.9772499)

## [1] 2.000001
```

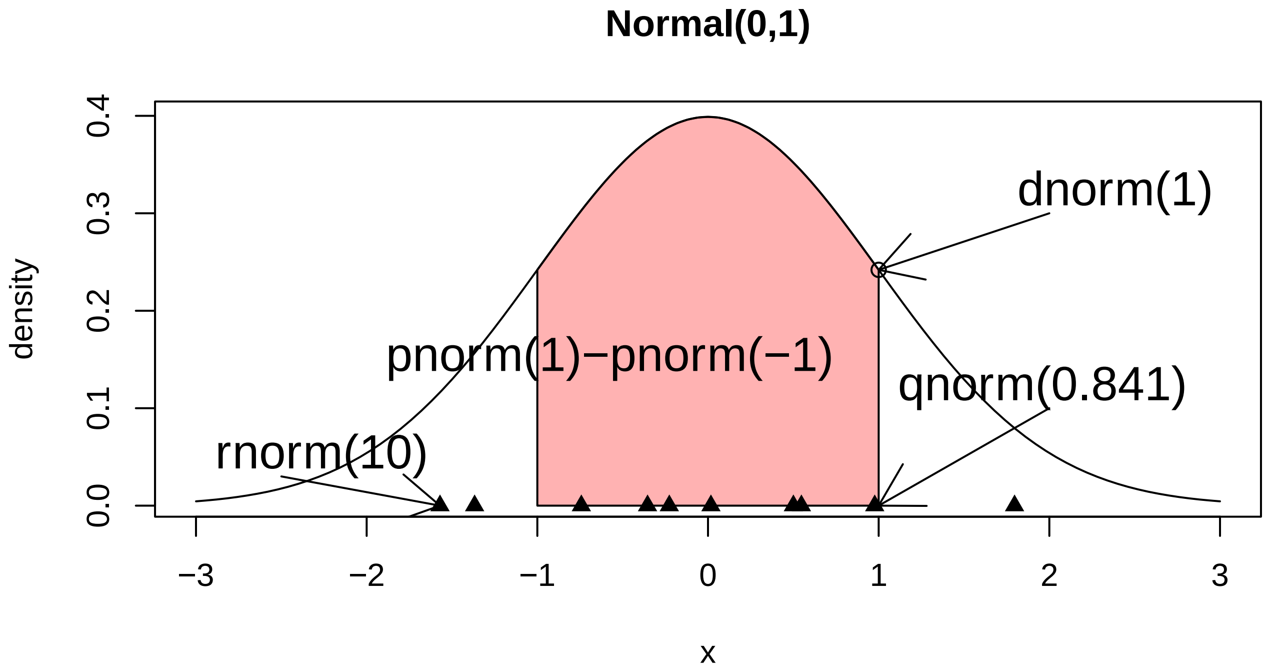
Continuous random variables: The normal random variable

4. Compute the probability density for a quantile value: dnorm

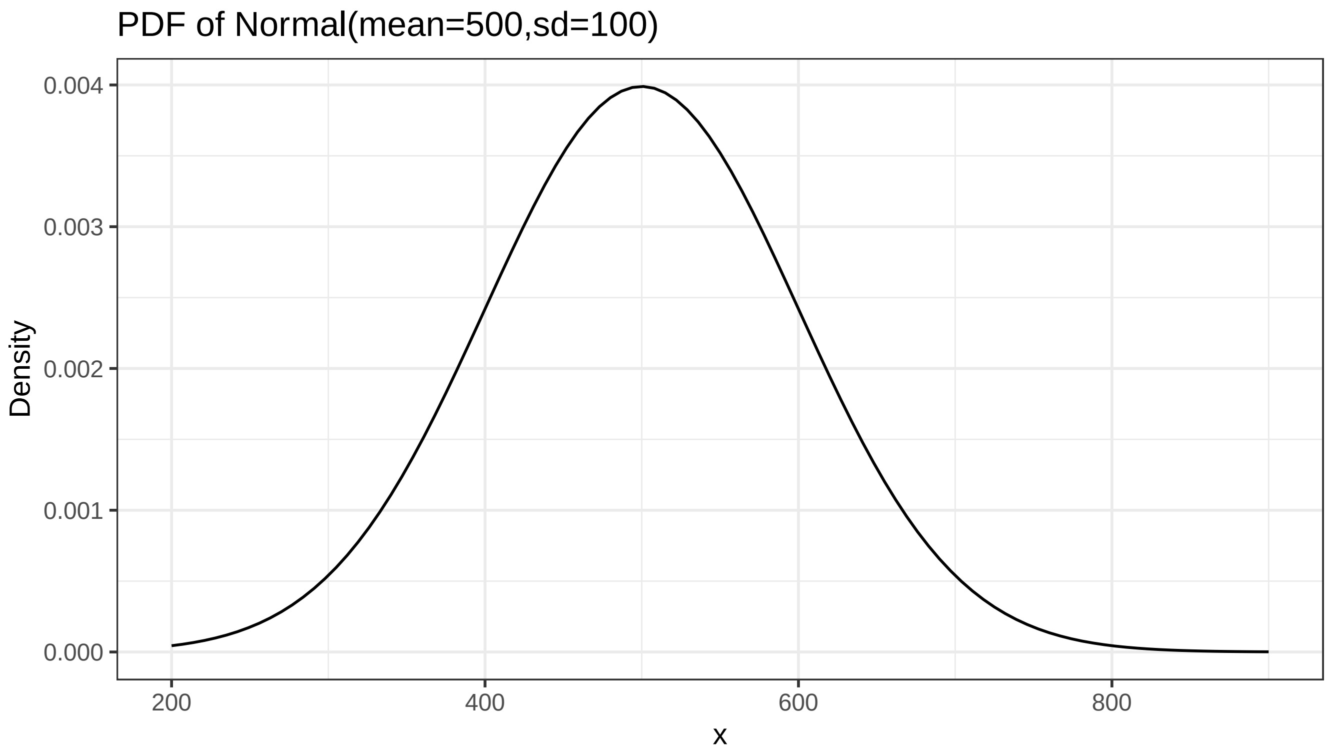
```
dnorm(2)

## [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 point values x.

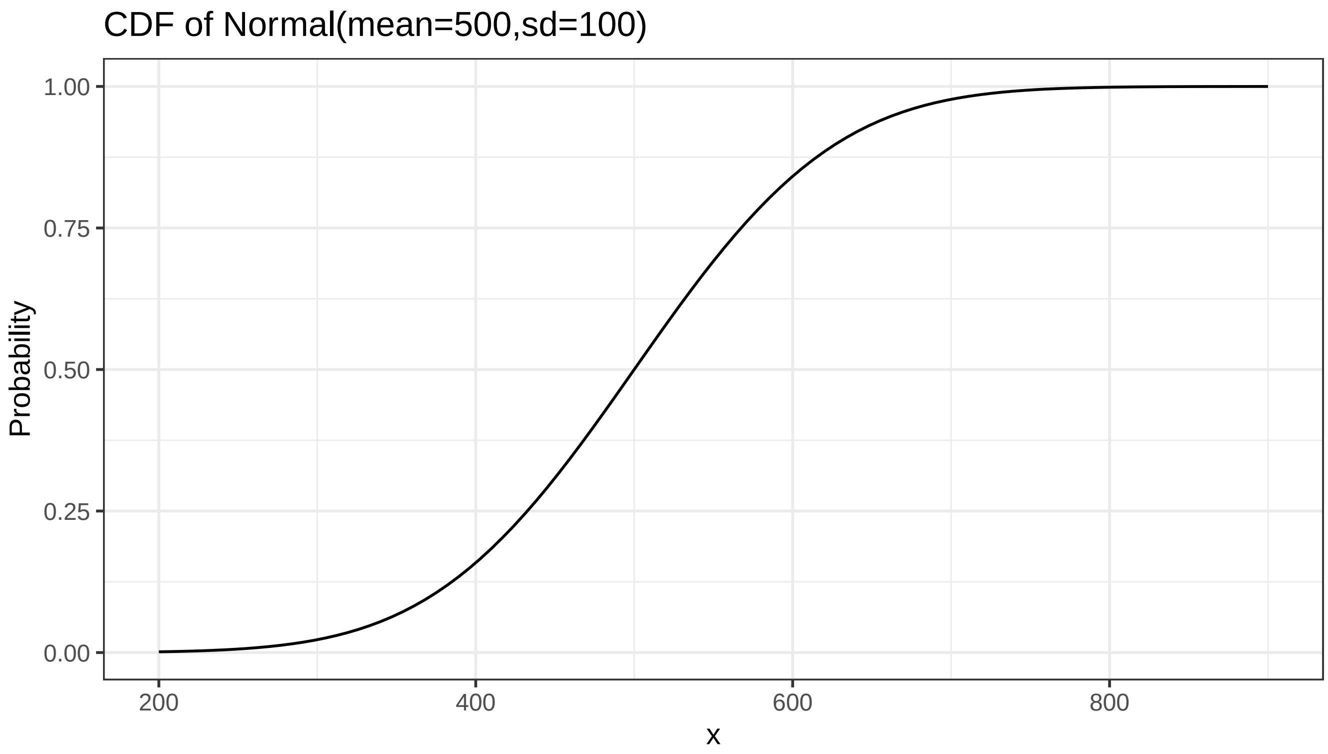
Continuous random variables: The normal random variable



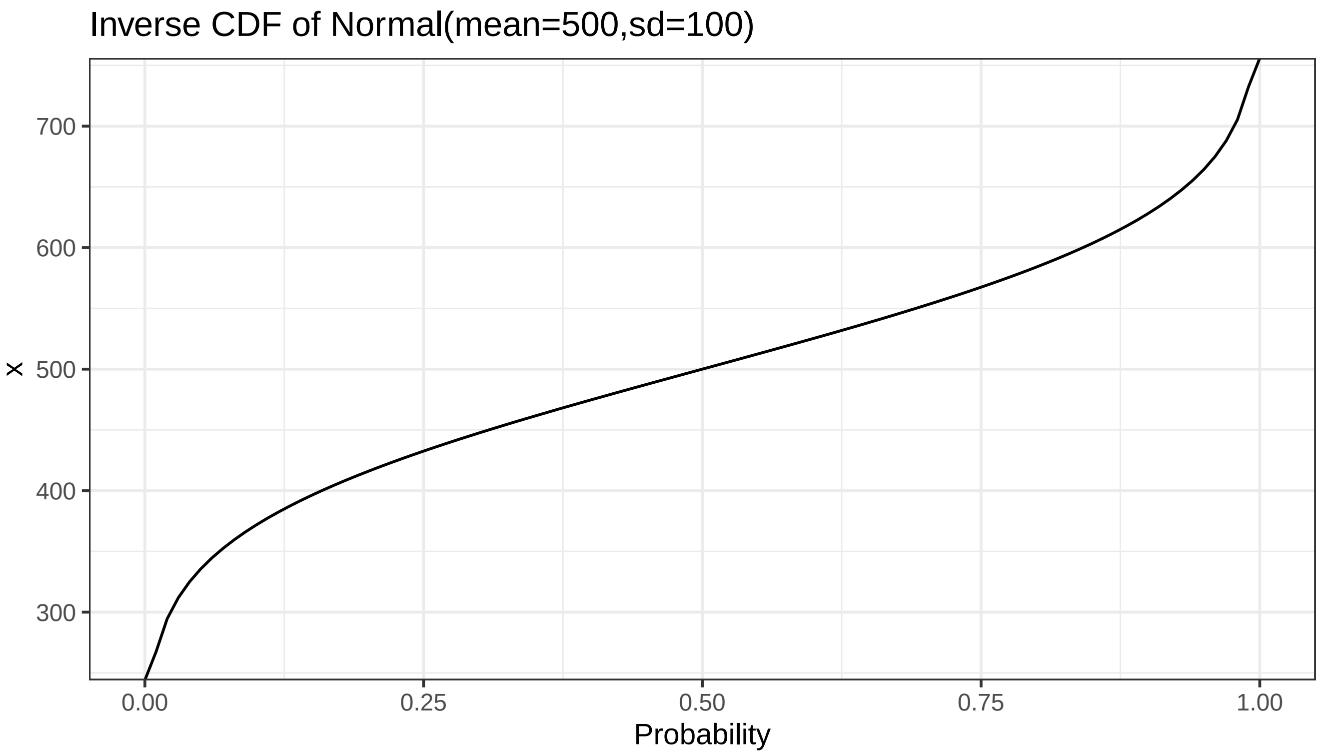
Continuous random variables: The normal random variable



Continuous random variables: The normal random variable



Continuous random variables: The normal random variable



Continuous random variables: The normal random variable

Summary

- Random variables and PMFs and PDFs
- Examples of discrete and continuous random variables
- The d-p-q-r family of functions.

Up next: Maximum likelihood estimates.

The expectation and variance (discrete case)

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:

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

The expectation (discrete case)

- The expectation has the interpretation that if we were to repeatedly do the experiment with a larger and larger number of trials and calculate the sample mean of the observations, we would approach the expected value 0.5.
- Another way to look at the expectation is that the expectation gives us the weighted mean of the possible events, weighted by the respective probabilities of each event.

The expectation (discrete and continuous examples)

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 $\hat{\sigma}^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}$

All these results can be derived analytically; e.g., see my lecture notes on linear models: https://github.com/vasishth/LM

Maximum likelihood estimates: binomial example

- From the observed data, we can compute the estimate of θ, $\hat{\theta} = k/n$.
- The quantity $\hat{\theta}$ is the observed proportion of successes, and is called the **maximum likelihood estimate** of the true (but unknown) parameter θ.
- Once we have estimated θ in this way, we can also obtain an estimate of the variance ($n\hat{\theta}(1-\hat{\theta})$).
- These estimates are then used for statistical inference.

Maximum likelihood estimates: binomial example

What does the term "maximum likelihood estimate" mean?

- In order to understand this term, it is necessary to first understand what a likelihood function is.
- Recall that in the binomial example, the PMF $p(k|n,\theta)$ assumes that θ and n are fixed, and k will vary from 0 to 10 when the experiment is repeated multiple times.
- The **likelihood function** refers to the PMF $p(k|n,\theta)$, treated as a function of θ.
- The likelihood function is written $\mathcal{L}(\theta|k,n)$, or just $\mathcal{L}(\theta)$.

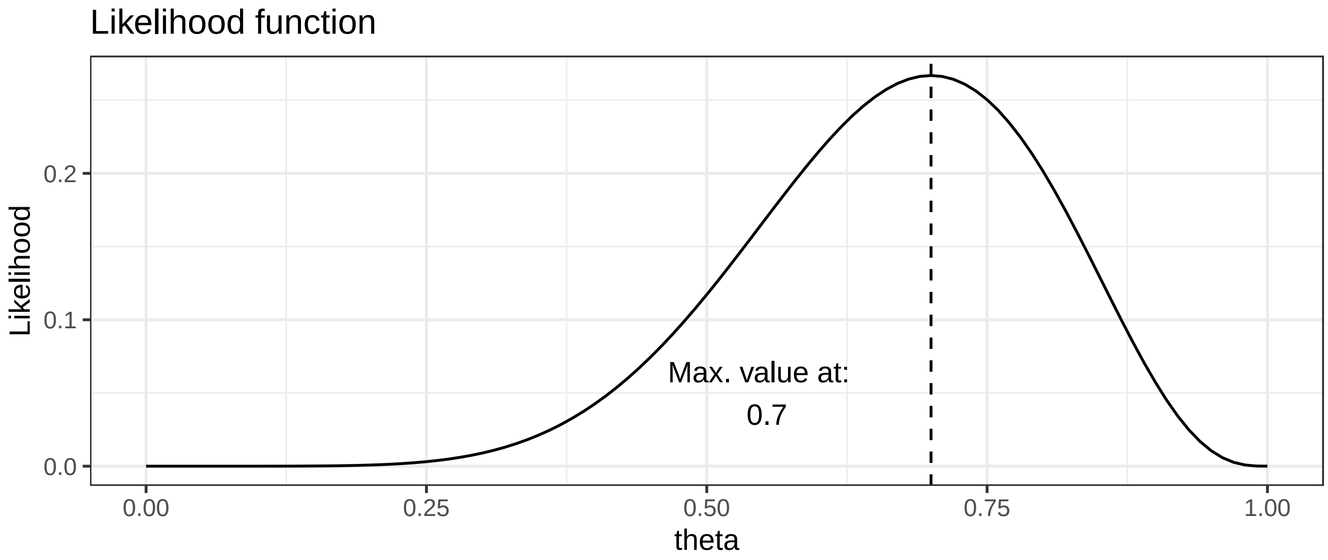
Maximum likelihood estimates: binomial example

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

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

Maximum likelihood estimates

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



Maximum likelihood estimates

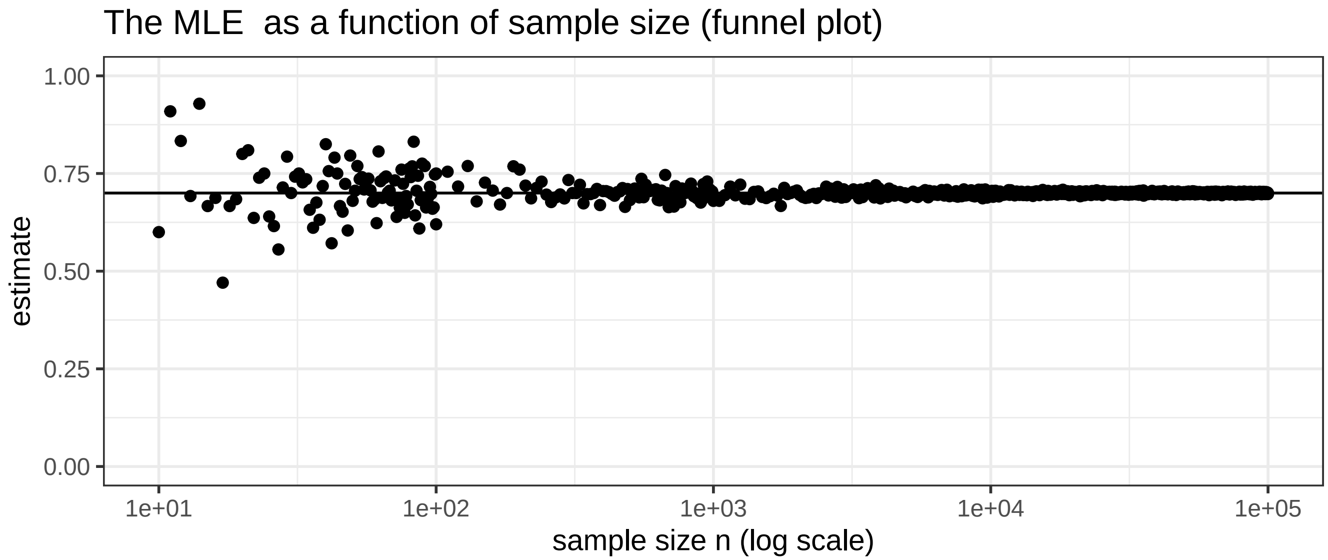
- Thus, the maximum likelihood estimate (MLE) gives us the most likely value that the parameter θ has, given the data.
- In the binomial, the proportion of successes k/n can be shown to be the maximum likelihood estimate of the parameter θ.
- In the normal distribution, the MLEs of μ and σ are the usual formulas we are familiar with.

More details in this free online book (optional reading):

> Introduction to Probability and Statistics Using R, by Kerns.
> https://www.nongnu.org/ipsur/

Maximum likelihood estimates: binomial example

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



Summary

- Given the PMF/PDF we have seen, the expectation and the variance of the relevant parameters:
  - Discrete example: Binomial, $E[Y] = \theta, Var(Y) = n\theta(1-\theta)$.
  - Continuous example: Normal, $E[Y] = \mu, Var(Y) = \sigma^2$
- In the Binomial, we can estimate θ by computing $\frac{k}{n}$.
- In the Normal, we can estimate μ and σ using the formulas shown earlier.
- These estimates can be shown to be MLEs.

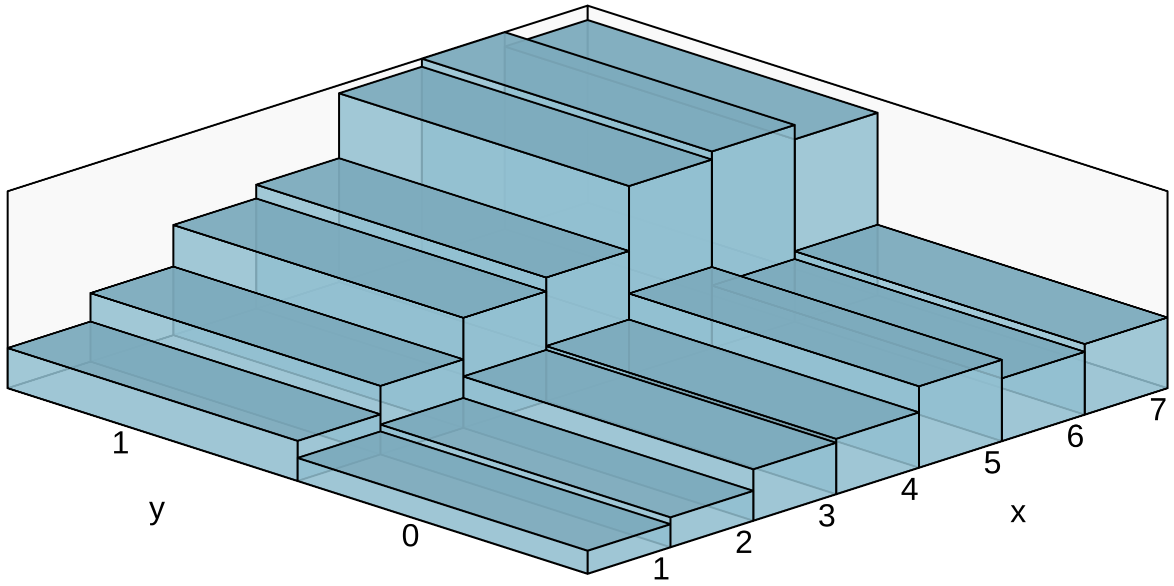
Next, we turn to bivariate/multivariate distributions.

Bivariate distributions: Discrete case

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.



Bivariate distributions: Discrete case

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

Table: 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

Bivariate distributions: Discrete case

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**)**:

$$p_X(x) = \sum_{y \in S_Y} p_{X,Y}(x,y). \tag{16}$$

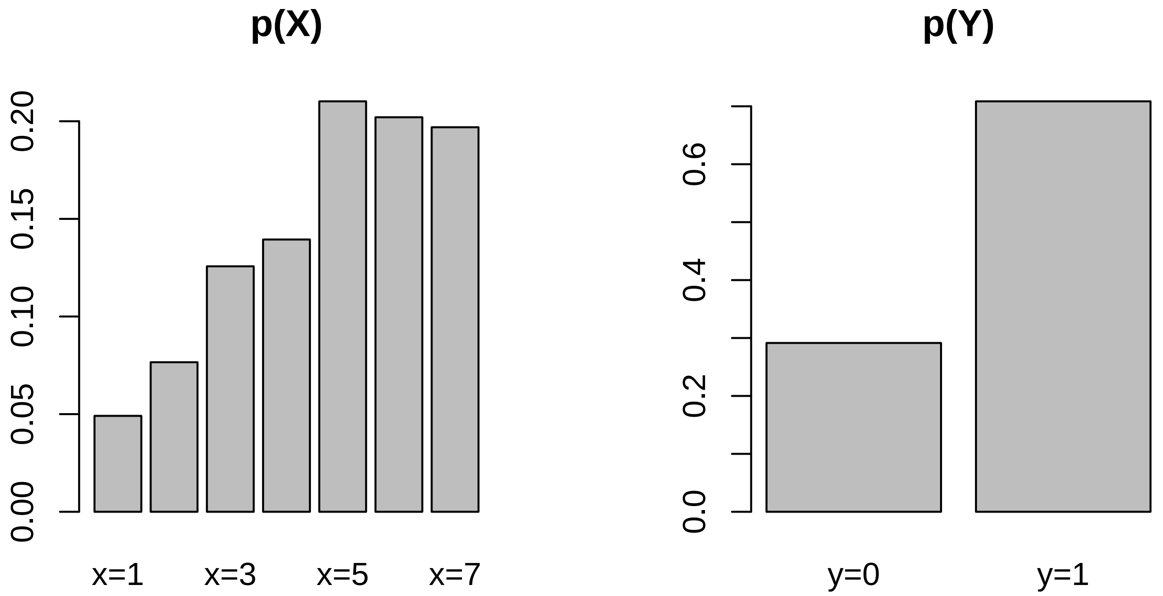
$$p_Y(y) = \sum_{x \in S_X} p_{X,Y}(x,y). \tag{17}$$

Bivariate distributions: Discrete case

Table: 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	

Bivariate distributions: Discrete case



Bivariate distributions: Discrete case

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

$$p_{X|Y}(x \mid y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} \tag{18}$$

and

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

Bivariate distributions: Discrete case

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

Table: 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	

$$
\begin{aligned}
p_{X|Y}(1 \mid 0) =& \frac{p_{X,Y}(1,0)}{p_Y(0)}\\
=& \frac{0.018}{0.291}\\
=& 0.062
\end{aligned}
\tag{20}
$$

Bivariate distributions: Discrete case

Table: A table for listing conditional distributions of X given Y.

	$x = 1$	$x = 2$	$x = 3$	$x = 4$	$x = 5$	$x = 6$	$x = 7$
$p_{X\|Y(x\|y=0)}$	0.062						
$p_{X\|Y(x\|y=1)}$							

Fill in the rest of this table!

Next topic

Next, we turn to continuous bivariate/multivariate distributions.

Bivariate distributions: Continuous case

- Consider two normal random variables X and Y, each of which coming from, for example, a $Normal(0,1)$ distribution, with some correlation $\rho_{X,Y}$ between the two random variables.
- Such a bivariate distribution is expressed in terms of the means and standard deviations of each of the two distributions, and the correlation ρ_{XY} between them.
- The standard deviations and correlation are expressed in a special form of a 2×2 matrix called a variance-covariance matrix Σ.

Bivariate distributions: Continuous case

- If ρ_{XY} is the correlation between the two random variables, and σ_X and σ_Y the respective standard deviations, the variance-covariance matrix is written as:

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \rho_{XY}\sigma_X\sigma_Y \\ \rho_{XY}\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \tag{21}$$

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

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

Bivariate distributions: Continuous case

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

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma\right) \tag{22}$$

- The joint PDF is written with reference to the two variables $f_{X,Y}(x,y)$.
- The joint PDF has the property that the volume under the curve sums to 1---this sum-to-1 property is the same idea that we encountered in the univariate cases (the binomial and normal distributions), except that we are talking about a bivariate distribution here.

Bivariate distributions: Continuous case

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).

$$\iint_{S_{X,Y}} f_{X,Y}(x,y)\,dxdy = 1 \tag{23}$$

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 $v' < v$).

$$\begin{aligned} F_{X,Y}(u,v) &= \text{Prob}(X < u, Y < v) \\ &= \int_{-\infty}^{u}\int_{-\infty}^{v} f_{X,Y}(x,y)\,dydx \text{ for } (x,y) \in \mathbb{R}^2 \end{aligned} \tag{24}$$

Bivariate distributions: Continuous case

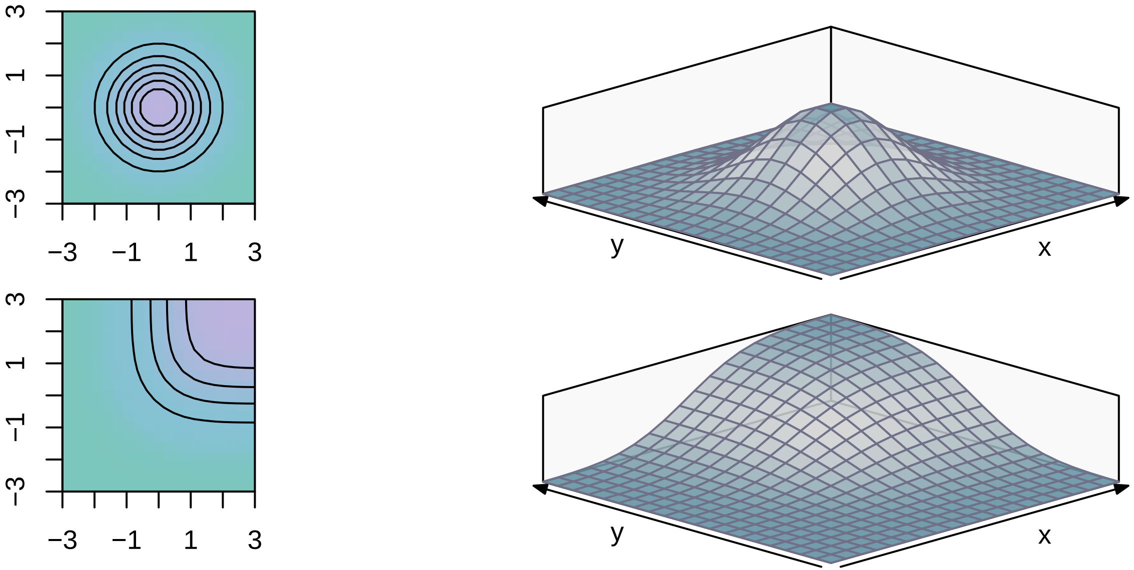
Just as in the discrete case, the marginal distributions can be derived by marginalizing out the other random variable:

$$f_X(x) = \int_{S_Y} f_{X,Y}(x,y)\,dy \quad f_Y(y) = \int_{S_X} f_{X,Y}(x,y)\,dx \tag{25}$$

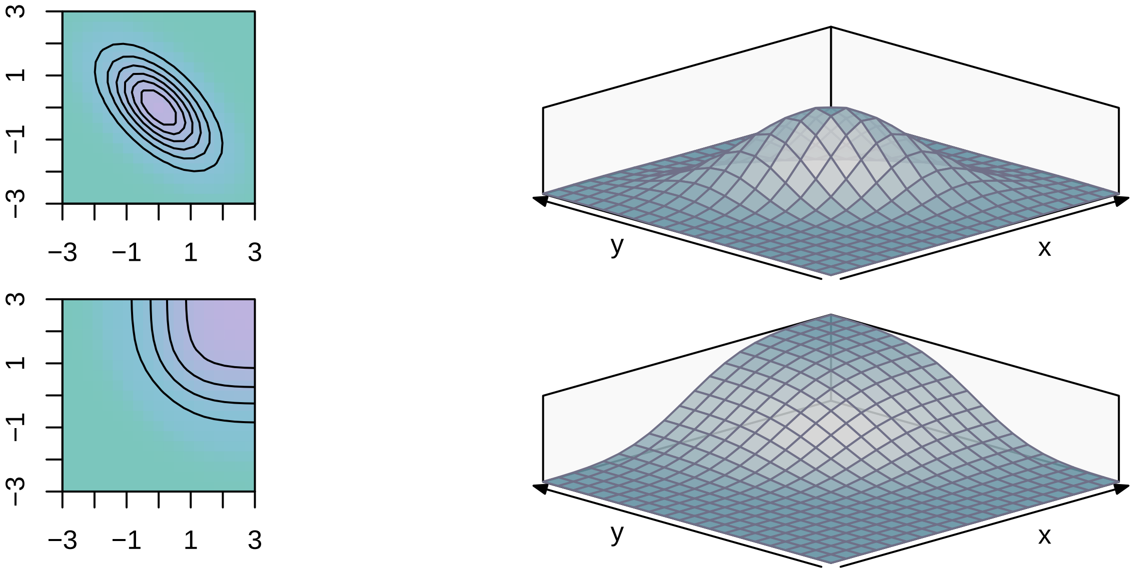
Here, S_X and S_Y are the respective supports.

This is exactly like the discrete example earlier, except that we are working in continuous space, so $\sum$ is replaced with $\int$.

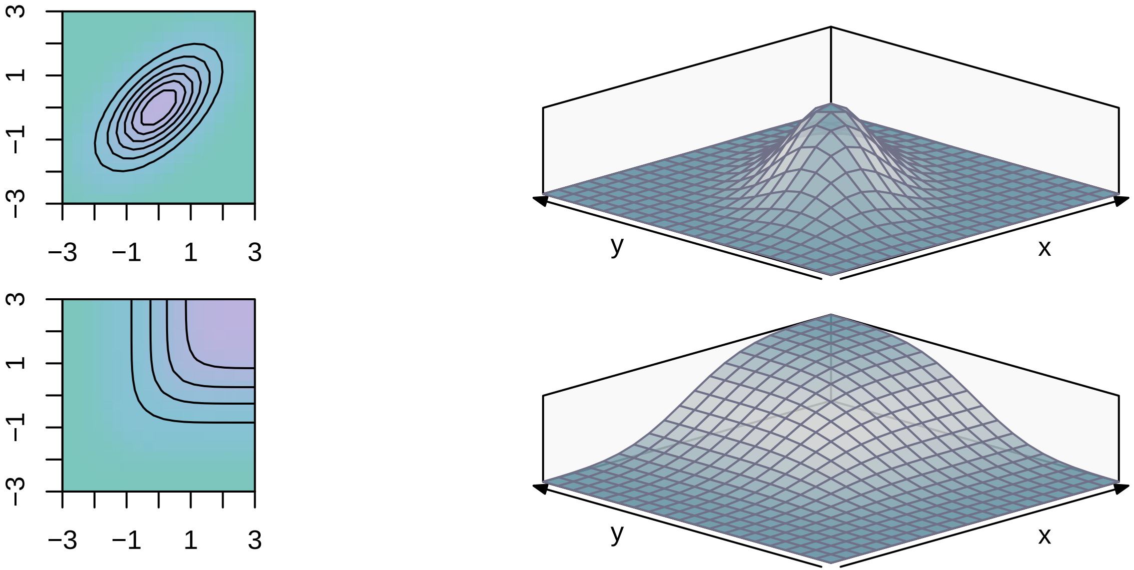
Bivariate distributions: Continuous case ($\rho = 0$)



Bivariate distributions: Continuous case ($\rho = -0.6$)



Bivariate distributions: Continuous case ($\rho = 0.6$)



Bayesian Data Analysis

Shravan Vasishth
vasishth.github.io

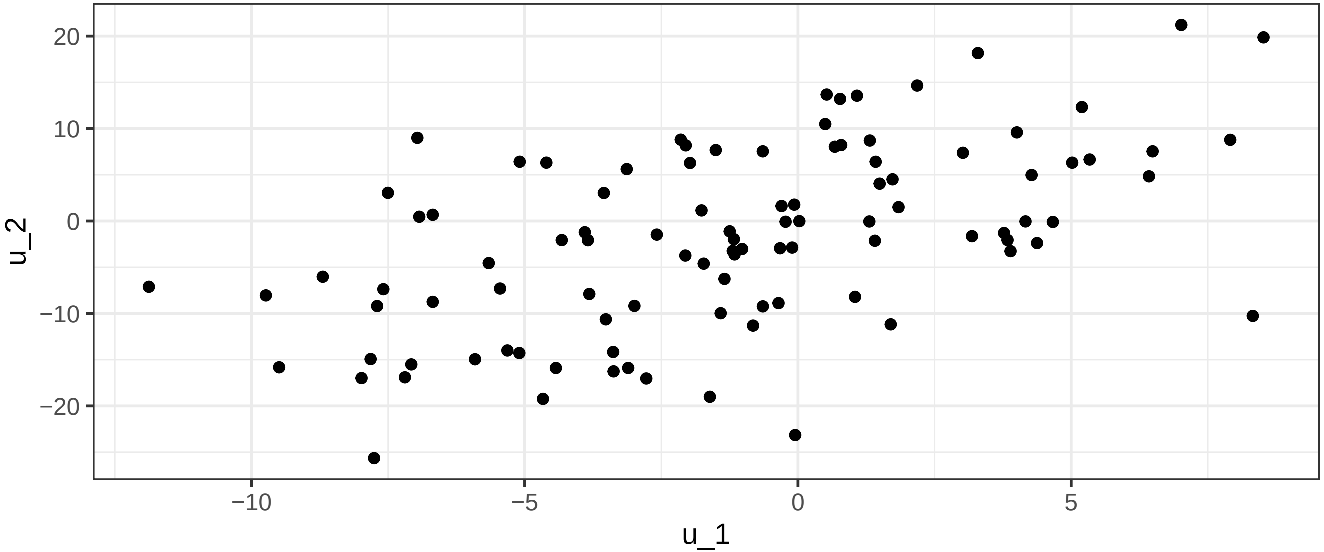
Bivariate distributions: Continuous case

Generate simulated bivariate (multivariate) data

```
## define a variance-covariance matrix:
Sigma <- matrix(c(5^2, 5 * 10 * .6, 5 * 10 * .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,] -3.381358 -14.16423
## [2,]  2.181265  14.65221
## [3,] -7.986015 -16.99031
```

Bivariate distributions: Continuous case



Stepping back

- Why did I waste your time on distributions? Why not just get into data analysis right away?
- These foundational ideas are necessary for understanding the remaining part of this course.
- The minimum you should conceptually understand before doing Bayes:
  - PMFs and PDFs and their properties
  - The d-p-q-r family of functions
  - Likelihood functions and Maximum likelihood estimation
  - Joint, marginal, and conditional distributions
- A shaky foundation will lead to shaky data analysis :)

We will turn next to Bayesian modeling!