MONASH University

MONASH BUSINESS SCHOOL

Statistical Thinking (ETC2420/ETC5242)

Associate Professor Catherine Forbes

Week 6: Distributional models and maximum likelihood

Learning Goals for Week 6

- Apply elementary probability and conditional probability rules
- Identify common discrete and continuous univariate distributions
- Develop distributional models for i.i.d data and estimate them using maximum likelihood methods
- Use CLT- and Bootstrap-based confidence intervals to characterise uncertainty in MLEs

Assigned reading for Week 6:

- Appendix A in ISRS

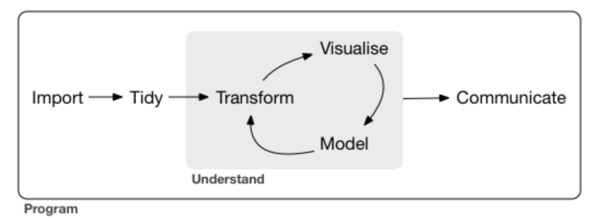


Figure 1: (Grolemond & Wickham, 2017)

So far. . .

- **Week 1**: Introduction to R and RStudio
  - Rmarkdown, Reproducibility, Tidyverse
- **Week 2**: Introduction to data, visualisation and wrangling
  - ggplot2, dplyr, tidyr
- **Week 3**: Randomisation and simulation for testing proportions
  - Permutation test for binary outcomes, sampling distributions
- **Weeks 4 + 5**: Resampling techniques for assessing variability in means
  - CLT, t-tests, confidence intervals, bootstrap
  - paired and independent samples

- We want to now move on to more advanced **modelling of populations**
  - Not just population means
- For this we need to build up our skills relating to **probability**

Uses of probability?

- A tool to describe (and understand) apparent **randomness**
- A way to **characterise uncertainty**
- To model data (for understanding, prediction, learning. . .)
- Think in terms of a **random process**, leading to an outcome
  - may be a single event (H or T)
  - may be a sequence of events (HHTHTTHHTH. . .)

How to derive probabilities in general?

1. Mathematically
   - “Counting rules” (e.g. permutations, and other tricks)
   - These can be difficult!
2. Simulation on a computer
   - Often *much* easier!

Sample spaces and Events

We are most often interested in the outcomes of **experiments**

- An experiment is any activity that produces or observes an outcome.

The **sample space** is the collection of all possible outcomes

- this may be a finite collection
- or an infinitely large set

Events are subsets of outcomes

- including the full sample space
- combinations of individual outcomes
- single outcomes
- the empty set (no outcomes at all)

Need to be able to work out probabilities (somehow) for complex events

- There are probability rules to use!

Axioms of probability

There are three fundamental rules of probability.

1. If $\Pr(A)$ is the probability associated with event A, then $0 \leq \Pr(A) \leq 1$
2. The total probability of all outcomes in the sample space is 1
3. If $A_1, A_2, \ldots$ is a sequence of **mutually exclusive** events, then

$$\Pr(A_1 \cup A_2 \cup \ldots) = \Pr(A_1) + \Pr(A_2) + \ldots$$

Disjoint events

Mutually exclusive events are sometimes referred to as **disjoint** events.

These are events that cannot happen simultaneously (their intersection is empty)

- From the third axiom, if events A_1 and A_2 are **mutually exclusive**
- Then $\Pr(A_1 \text{ or } A_2) = \Pr(A_1) + \Pr(A_2)$

Non-disjoint events

- Outcomes that overlap are called **non-disjoint** events
  - We need a more general rule for working out their probabilities

Example: Consider events $(X > 2)$ and $(X < 4)$

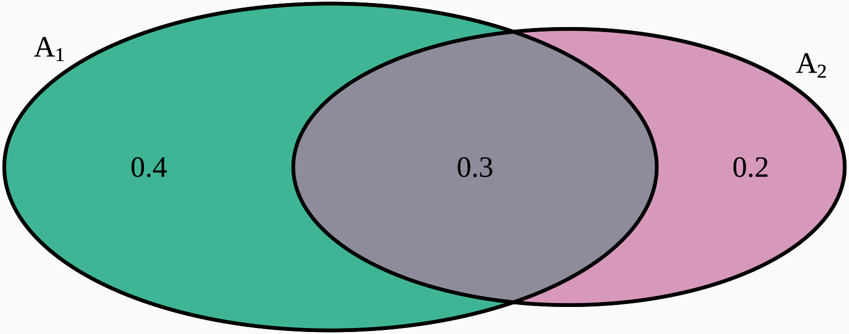
- These are NOT disjoint!
- No "double counting" of probability allowed!!
- Need to take out the "double counted" (overlap) part:

$$\Pr(X > 2 \cup X < 4) = \Pr(X > 2) + \Pr(X < 4) - \Pr(X > 2 \cap X < 4)$$

Example: Venn diagram

Suppose $\Pr(A_1) = 0.7$ and $\Pr(A_2) = 0.5$ and $\Pr(A_1 \cap A_2) = 0.3$

Then $\Pr(A_1 \cup A_2) = 0.7 + 0.5 - 0.3 = 0.9$



Complementary events

The **complement of event** A is the event denoted by A^c

- A^c represents **all outcomes not in** A
- The probabilities for events A and A^c are related:

$$\begin{aligned} P(A) + P(A^c) &= 1 \\ P(A) &= 1 - P(A^c) \\ P(A^c) &= 1 - P(A) \end{aligned}$$

Sometimes it is easier to work out the probability for a complementary event

Random variables

- A random process or variable with a **numerical** outcome

Example of a random process (but not a random variable)

- For $i = 1$ and $i = 2$

	$X_i = x$	$\Pr(X_i = x)$
x	Head	0.5
	Tail	0.5

Example of a random variable (that is also a random process)

- For $i = 1$ and $i = 2$
- Let $X_i =$ the **number of heads** on toss i

	$X_i = x$	$\Pr(X_i = x)$
x	1	0.5
	0	0.5

Random variables may be characterised as being either **discrete** or **continuous**

Discrete random variable over a finite (or countable) sample space

Example

$X = x$	$\Pr(X = x)$
$X = 1$	1/2
$X = 2$	1/8
$X = 3$	1/4
$X = 4$	1/8

Find probabilities for given events:

1. $\Pr(X = 2)$
2. $\Pr(X \leq 2)$
3. $\Pr(X \text{ is even})$
4. $\Pr(X < 4)$
5. $\Pr(X > 2 \text{ and } X < 3) = \Pr(X > 2 \cap X < 3)$
6. $\Pr(X > 2 \text{ or } X < 3) = \Pr(X > 2 \cup X < 3)$

Example

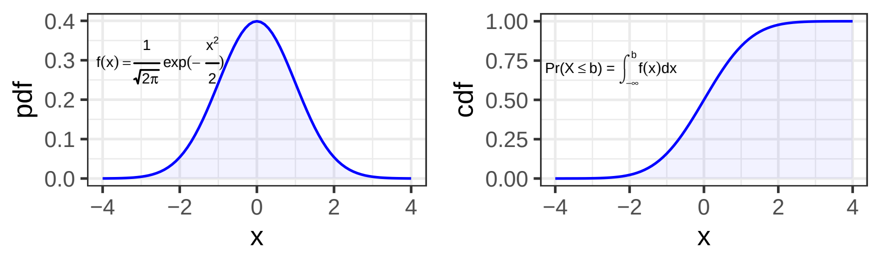
Let's work them out...

$X = x$	$\Pr(X = x)$
$X = 1$	1/2
$X = 2$	1/8
$X = 3$	1/4
$X = 4$	1/8

1. $\Pr(X = 2) = 1/8$
2. $\Pr(X \leq 2) = \Pr(x = 1 \text{ or } X = 2) = 1/2 + 1/8 = 5/8$
3. $\Pr(X \text{ is even}) = \Pr(x = 2 \text{ or } X = 4) = 1/8 + 1/8 = 1/4$
4. $\Pr(X < 4) = 1 - \Pr(X = 4) = 1 - 1/8 = 7/8$
5. $\Pr(X > 2 \text{ and } X < 3) = \Pr(X > 2 \cap X < 3) = 0$ (two events, cannot both occur)
6. $\Pr(X > 2 \text{ or } X < 3) = \Pr(X > 2 \cup X < 3) = 1$ (two events, either one can occur)

Example: Continuous random variable over an infinite sample space

If $X \sim N(0,1)$



Find

1. $\Pr(X = 1)$
2. $\Pr(X < 1)$
3. $\Pr(X \text{ is even})$
4. $\Pr(X < -\frac{1}{2})$
5. $\Pr(X > 2 \text{ and } X < 3) = \Pr(X > 2 \cap X < 3)$
6. $\Pr(X > 2 \text{ or } X < 3) = \Pr(X > 2 \cup X < 3)$

Example: Continuous random variable over an infinite sample space

We can work out these probabilities using the cumulative distribution function (cdf)

- equivalent to corresponding area under the probability density function (pdf)

If $X \sim N(0, 1)$

1. $\Pr(X = 1) = 0$ since there is no area above a single point
2. $\Pr(X < 1) = \int_{-\infty}^{1} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{x^2}{2}\}dx$ = `pnorm(1, mean=0, sd=1)` = 0.8413
3. $\Pr(X \text{ is even}) = 0$ since no area above a countable number of points
4. $\Pr(X < -\frac{1}{2}) = \int_{-\infty}^{-0.5} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{x^2}{2}\}dx$ = `pnorm(-0.5, mean=0, sd=1)` = 0.3085
5. $\Pr(X > 2 \text{ and } X < 3) = \Pr(X > 2 \cap X < 3) = \int_{2}^{3} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{x^2}{2}\}dx$
   = `pnorm(3, mean=0, sd=1) - pnorm(2, mean=0, sd=1)` = 0.0214
6. $\Pr(X > 2 \text{ or } X < 3) = \Pr(X > 2 \cup X < 3)$
   $= \Pr(X > 2) + \Pr(X < 3) - \Pr(X > 2 \cap X < 3)$
   $= [1 - \Pr(X < 2)] + \Pr(X < 3) - [\Pr(X < 3) - \Pr(X < 2)]$
   $= 1 - \Pr(X < 2) + \Pr(X < 3) - \Pr(X < 3) + \Pr(X < 2)$
   $= 1$

Independence

- **Two** processes are **independent** if knowing the outcome of one provides no useful information about the outcome of the other

Multiplication Rule for independent processes

- If A and B are simple events from two **different** and **independent** processes
  - two compound processes but “simple” relationship between them due to assumed independence
- Then the event that **both** A and B occur corresponds to an **intersection**
  - the joint probability can calculated as the product of the individual probabilities:

$$\Pr(A \text{ and } B) = \Pr(A) \times \Pr(B)$$

- Similarly, if there are k simple events $A_1, A_2, \ldots, A_k$ from k independent processes
  - Then the probability that **all events** will occur is given by

$$\Pr(A_1) \times \Pr(A_2) \times \cdots \times \Pr(A_k)$$

Example: Two independent coin tosses

- For $i = 1$ and $i = 2$

	$X_i = x$	$\Pr(X_i = x)$
x	Head	0.5
	Tail	0.5

- If two fair coin tosses are **independent**, then any **joint probability** about **each outcome** will be the **product of the two marginal probabilities** about each outcome.

1. *What is the probability of a "Head" on the **first** toss and a "Tail" on the **second** toss?*

$$\begin{aligned} \Pr(X_1 &= \text{Head and } X_2 = \text{Tail}) \\ &= \Pr(X_1 = \text{Head}) \times \Pr(X_2 = \text{Tail}) \\ &= (0.5) \times (0.5) \\ &= 0.25 \end{aligned}$$

Example: Two independent coin tosses

Note: The calculation above is **NOT** the same calculation as for...

2. *What is the probability of getting a "Head" and a "Tail" on two independent coin tosses?*

- In this case the event in question is more complex, since it includes two possibilities:
  - We could have either $(X_1 = \text{Head and } X_2 = \text{Tail})$ or $(X_1 = \text{Tail and } X_2 = \text{Head})$
- $\Rightarrow$ Here we have to calculate two separate probabilities, each being the product of probabilities for X_1 and X_2

$$
\begin{aligned}
\Pr(X_1 &= \text{Head and } X_2 = \text{Tail OR } X_1 = \text{Tail and } X_2 = \text{Head}) \\
&= \Pr(X_1 = \text{Head and } X_2 = \text{Tail}) + \Pr(X_1 = \text{Tail and } X_2 = \text{Head}) \\
&= \Pr(X_1 = \text{Head}) \times \Pr(X_2 = \text{Tail}) + \Pr(X_1 = \text{Tail}) \times \Pr(X_2 = \text{Head}) \\
&= (0.5) \times (0.5) + (0.5) \times (0.5) \\
&= 2 \times (0.25) \\
&= 0.5
\end{aligned}
$$

Multiplication Rule for independent processes

Example: Left-handedness

- About 9% of people in the population are left-handed
- Suppose 2 people are selected at random from the Australian population
  - (Assume population is so large that the outcomes for the two selected are independent)

1 What is the probability that both people selected are left-handed?

(0.09)(0.09) = 0.0081

8 What is the probability that both people selected are right-handed?

- $(1-0.09)(1-0.09) = (0.91)^2 = 0.8281$

9 What is the probability that one person selected is left-handed, and the other is right-handed?

- Note probabilities of all events must sum to one
- $1-(0.0081+0.8281) = 0.1638$

Probability table (raw counts)

Example: Travel survey data

- Random sample survey of 100 people with particular credit card
- *Are you planning to travel abroad next year?*
- Take these as proportional to “true probabilities”

		Age group			
		25 or less	26-40	41 or more	Total
Response	Yes	2	12	15	29
	Undecided	5	10	16	31
	No	10	15	15	40
	Total	17	37	46	100

1. $\Pr($Card holder intends to travel over next 12 months$)$?
2. $\Pr($Card holder intends to travel over next 12 months OR is undecided$)$?
3. $\Pr($Card holder intends to travel over next 12 months AND is 25 years old or less$)$?

Probability table (relative frequencies)

Example: Travel survey data

Convert to probabilites (divide all by 100 to make sum to 1.0)

- Take relative frequencies as "true" probabilities

		Age group			
		25 or less	26-40	41 or more	Total
	Yes	0.02	0.12	0.15	0.29
Response	Undecided	0.05	0.10	0.16	0.31
	No	0.10	0.15	0.15	0.40
	Total	0.17	0.37	0.46	1.0

$\Pr(\text{Card holder intends to travel over next 12 months}) = 0.29$

$\Pr(\text{Card holder intends to travel over next 12 months OR is undecided}) = 0.29 + 0.31 = 0.60$

$\Pr(\text{Card holder intends to travel over next 12 months AND is 25 years old or less}) = 0.02$

Joint, Marginal and Conditional distributions

- Joint probability
  - probability of outcomes for two or more variables or processes
- Marginal probability
  - probability of outcomes for a single variable or process
- Conditional probability
  - probability of outcomes for a single variable or process **given information about a second variable or process**

Joint probabilities

Example: Travel survey data (revisited)

Probability for all possible pairs

Age group AND Response combination	Prob
Yes response AND (25 or less)	0.02
Yes response AND (26-40)	0.12
Yes response AND (41 or more)	0.15
Undecided response AND (25 or less)	0.05
Undecided response AND (26-40)	0.10
Undecided response AND (41 or more)	0.16
No response AND (25 or less)	0.10
No response AND (26-40)	0.15
No response AND (41 or more)	0.15
	1.0

This information was obtained directly from the original probability table, but is expressed here more formally for compound events

Marginal probabilities

Age group

Sum across all rows to get column totals for each age group

Age group	Prob
25 or less	0.17
26-40	0.37
1 or more	0.46
	1.0

Are you planning to travel abroad next year?

Sum across all columns to get row totals for each response

Response	Prob
Yes	0.29
Undecided	0.31
No	0.40
	1.0

Conditional probability

The conditional probability for a single **outcome of interest** A, given **conditioned on an event** B, is defined as

$$\Pr(A \mid B) = \frac{\Pr(A \text{ and } B)}{\Pr(B)}$$

A **conditional probability distribution** concerns a list of possible outcomes, with their corresponding conditional probabilities satisfying three rules:

1. All outcomes listed in the sample space must be disjoint
2. Each conditional probability must be between 0 and 1 (inclusive)
3. Conditional probabilities must sum to 1

Conditional probability distributions (We'll list 6 of them here!)

Example: # 1

Response	$\Pr(\text{Response}\mid\text{25 years or less})$
Yes	0.02/0.17 = 0.1176
Undecided	0.05/0.17 = 0.2941
No	0.10/0.17 = 0.5882
	0.17/0.17 = 1.0

Example: # 2

Response	$\Pr(\text{Response}\mid\text{26 - 40 years})$
Yes	0.12/0.37 = 0.3243
Undecided	0.10/0.37 = 0.2703
No	0.15/0.37 = 0.4054
	0.37/0.37 = 1.0

Conditional probability distributions (We'll list 6 of them here!)

Example: # 3

Response	$\Pr(\text{Response}\|\text{40 years or more})$
Yes	0.15/0.46 = 0.3261
Undecided	0.16/0.46 = 0.3478
No	0.15/0.46 = 0.3261
	0.46/0.46 = 1.0

Example: # 4

Age group	$\Pr(\text{Age group}\|\text{Yes response})$
25 years or less	0.02/0.29 = 0.0690
26 - 40 years	0.12/0.29 = 0.4138
40 years or more	0.15/0.29 = 0.5172
	0.29/0.29 = 1.0

Conditional probability distributions (We'll list 6 of them here!)

Example: # 5

Age group	$\Pr($Age group$\|$Undecided response$)$
25 years or less	0.05/0.31 = 0.1613
26 - 40 years	0.10/0.31 = 0.3226
40 years or more	0.16/0.31 = 0.5161
	0.31/0.31 = 1.0

Example: # 6

Age group	$\Pr($Age group$\|$No response$)$
25 years or less	0.10/0.40 = 0.250
26 - 40 years	0.15/0.40 = 0.375
40 years or more	0.15/0.40 = 0.375
	0.40/0.40 = 1.0

General Multiplication Rule

- If A and B represent two outcomes or events, then

$$\Pr(A \text{ and } B) = \Pr(A \mid B) \times \Pr(B)$$

- Here A is the outcome of interest, and B is the event being conditioned upon
- Alternatively,

$$\Pr(A \text{ and } B) = \Pr(B \mid A) \times \Pr(A)$$

- Here B is the outcome of interest, and A is the event being conditioned upon

Tree Diagram

Work out joint probabilities using the product of marginal and conditional probabilities

Example: Travel survey data (revisited)

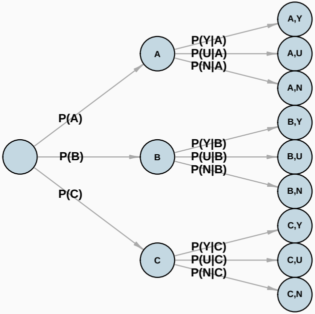
Let:

- "A" = 25 years or less
- "B" = 26 - 40 years
- "C" = 41 years or more

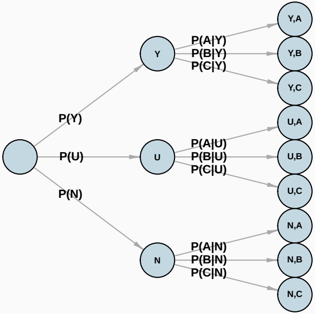
and Let:

- "Y" = Yes response
- "U" = Undecided response
- "N" = No response

Tree diagram 1



Tree diagram 2



Bayes' theorem: inverting probabilities

- Consider two variables, V_1 and V_2, and suppose
  - Outcome A relates to V_1
  - Outcome B relates to V_2

Then

$$\Pr(A \mid B\ ) = \frac{\Pr(B \cap A)}{\Pr(B)} = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B \mid A)\Pr(A) + \Pr(B \mid A^c)\Pr(A^c)}$$

Now suppose $A_1, A_2, A_3, \ldots, A_k$ represent all possible outcomes of V_1. Then

$$\Pr(A_1 \mid B\ ) = \frac{\Pr(B \cap A_1)}{\Pr(B)} = \frac{\Pr(B \mid A_1)\Pr(A_1)}{\Pr(B \mid A_1)\Pr(A_1) + \cdots + \Pr(B \mid A_k)\Pr(A_k)}$$

Expected value of a discrete random variable

- If X takes outcomes $x_1, \dots, x_k$ with probabilities $\Pr(X = x_1), \dots, \Pr(X = x_k)$, respectively, then the **expected value** of X is

$$E[X] = x_1 \Pr(X = x_1) + \cdots + x_k \Pr(X = x_k) = \sum_{i=1}^{k} x_i \Pr(X = x_i)$$

- We often denote $\mu = E[X]$
  - μ is an **an attribute** of the probability distribution for X
  - μ is not random

Variance of a discrete random variable

- If X takes outcomes $x_1, \ldots, x_k$ with probabilities $\Pr(X = x_1), \ldots, \Pr(X = x_k)$, respectively, then the **variance** of X is

$$Var(X) = (x_1 - \mu)^2 \Pr(X = x_1) + \cdots + (x_k - \mu)^2 \Pr(X = x_k) = \sum_{i=1}^{k} (x_i - \mu)^2 \Pr(X = x_i)$$

- We often denote $\sigma^2 = Var(X)$
- The **standard deviation** of X is given by $\sigma = \sqrt{\sigma^2}$.
  - σ^2 and σ are features of the probability distribution
  - σ^2 and σ are not random

Some commonly used discrete distributions

- Bernoulli
- Binomial
- Negative Binomial
- Geometric
- Uniform (discrete)
- Poisson

Bernoulli

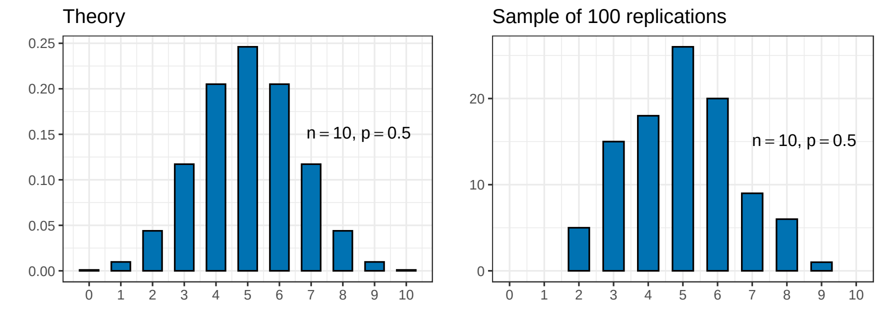
$$P(X = x \mid p) = p^x (1 - p)^{1-x} \quad \text{for } x \in \{0, 1\}, \quad \text{given } 0 < p < 1$$

- We have seen this before
- $X = 1$ if a "heads" appears, $X = 0$ otherwise
- $E[X] = p$ and $Var(X) = p(1 - p)$

Binomial

$$P(X = x \mid n, p) = \begin{pmatrix} n \\ x \end{pmatrix} p^x (1-p)^{n-x} \quad \text{for } x \in \{0, 1, 2, ..., n\}$$

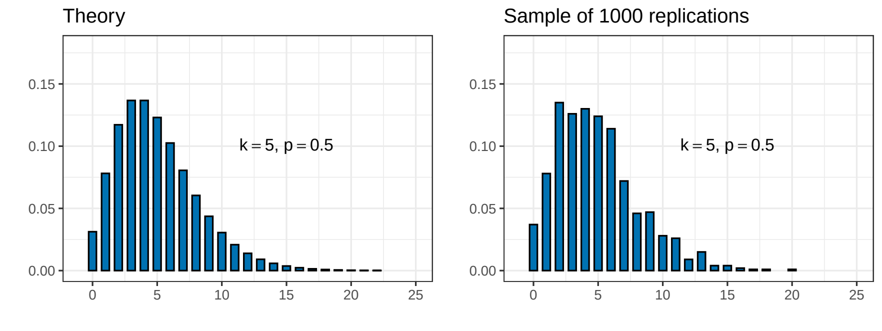
- Discrete, unimodal, right- or left-skewed or unimodal depending on p
- Arises from counting the number of successes from n independent Bernouilli trials, e.g. the number of heads in 10 coin flips
- $E[X] = np$ and $Var(X) = np(1-p)$



Negative Binomial

$$P(X = x \mid k, p) = \frac{\Gamma(x+k)}{\Gamma(k)x!} p^k (1-p)^x, \quad \text{for } x \in \{0, 1, 2, \ldots\}, \quad \text{given } 0 < p < 1$$

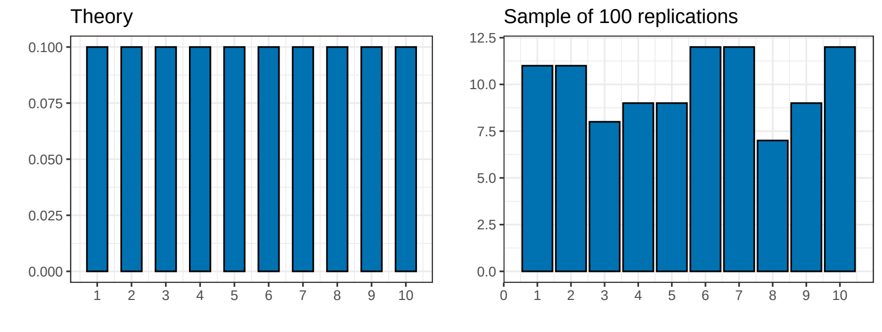
- Discrete, unimodal, right- or left-skewed or unimodal depending on p
- Arises from counting the number of 'failures' that occur in a sequence of independent Bernoulli trials until the targeted k^{th} success occurs
- $E[X] = \frac{k(1-p)}{p}$ and $Var(X) = \frac{k(1-p)}{p^2}$
- Called the **geometric distribution** when $k = 1$.



Uniform (discrete)

$$P(X = x \mid a, b) = \frac{1}{b - a + 1} \quad \text{for integers } a, b, \ \text{with } x \in \{a, a+1, ..., b\}$$

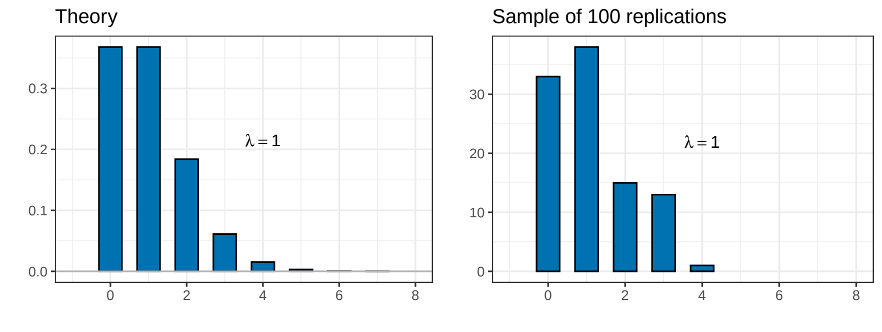
- Discrete, symmetric, unimodal over values $\{a, a+1, ..., b\}$
- Arises from equally likely outcomes
- $E[X] = \frac{b-a}{2}$ and $Var(X) = \frac{(b-a+1)^2-1}{12}$



Poisson distribution

$$P(X = x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \quad \text{for } x \in \{0, 1, 2, ...\}, \quad \text{given } \lambda > 0$$

- Discrete, right-skewed, unimodal
- Arises when counting number of times event occurs in an interval of time, e.g. the number of patients arriving in an emergency room between 11 and 12 pm
- $E[X] = \lambda$ and $Var(X) = \lambda$



Continuous random variables

- X is a continuous random variable taking outcomes over the real line, it has a probability density function (pdf) given by $f(x)$, for all x
- Then

1. $f(x) \geq 0$ for all $x \in \mathcal{R}$
2. The probability associated X associated with an (open) interval $A = (L_A, U_A)$ is given by

$$\Pr(X \subset A) = \int_{L_A}^{U_A} f(x)dx$$

3. The probability associated with all possible outcomes is given b

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

Expected value and variance of a continuous random variable

- X is a continuous random variable taking outcomes over the real line, having pdf $f(x)$, then the **expected value** and **variance* of X are given by

$$E[X] = \mu = \int_{-\infty}^{\infty} x\, f(x) dx$$

and

$$Var(X) = E[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2\, f(x) dx$$

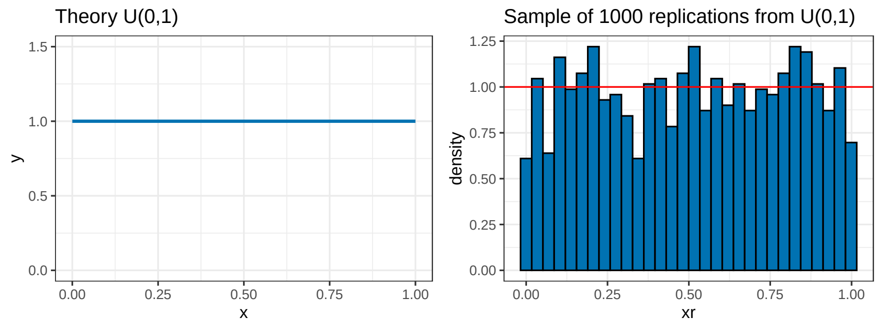
Some commonly used continuous distributions

- Uniform
- Normal
- Exponential
- Gamma
- Pareto
- Weibull
- Lognormal
- Beta

Uniform distribution (continuous)

$$p(x \mid a,\ b) = \frac{1}{(b-a)} \quad \text{for } x \in (a,\ b),\ \text{for } a < b$$

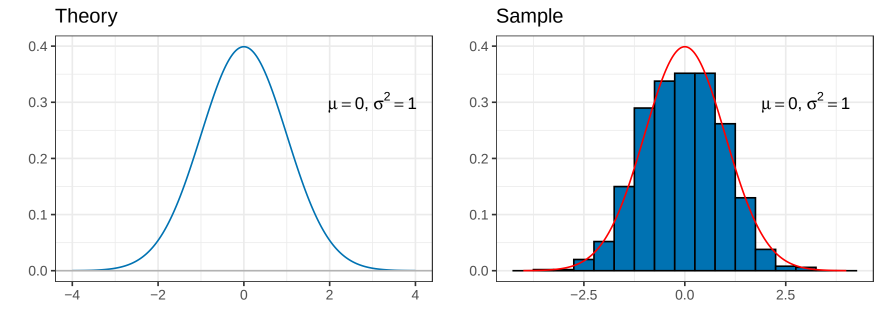
- continuous, symmetric, unimodal
- $E[X] = \frac{a+b}{2}$ and $Var(X) = \frac{(b-a)^2}{12}$



Normal distribution $N(\mu,\ \sigma^2)$

$$f(x \mid \mu,\ \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty,\ \text{for } -\infty < \mu < \infty, \sigma^2 > 0$$

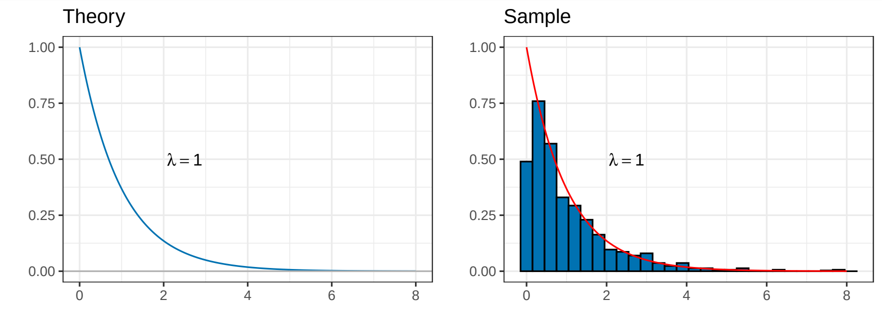
- Gaussian, bell-shaped
- symmetric, unimodal
- $E[X] = \mu$ and $Var(X) = \sigma^2$



Exponential distribution

$$f(x \mid \lambda) = \lambda e^{-\lambda x} \quad x \geq 0, \text{ for } \lambda > 0$$

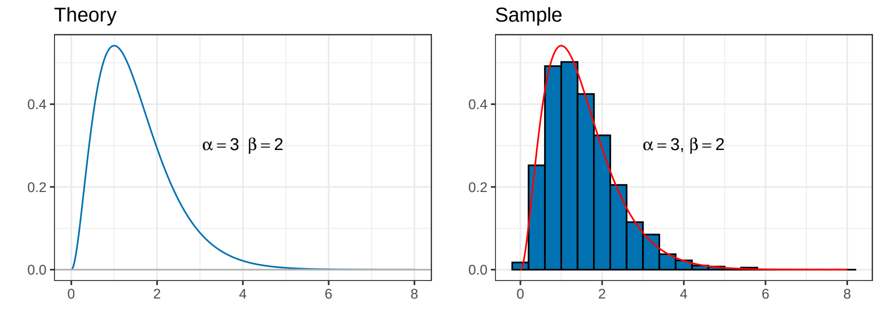
- right-skewed, unimodal
- Arises in time between or duration of events, e.g. time between successive failures of a machine, duration of a phone call to a help center
- λ is a **rate** parameter ($\beta = 1/\lambda$) is a **scale** parameter
- $E[X] = \frac{1}{\lambda}$, $Var(X) = \frac{1}{\lambda^2} = \beta^2$



Gamma distribution

$$f(x \mid \alpha,\ \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-x\beta} \quad x \geq 0,\ \text{for } \alpha > 1, \beta > 0$$

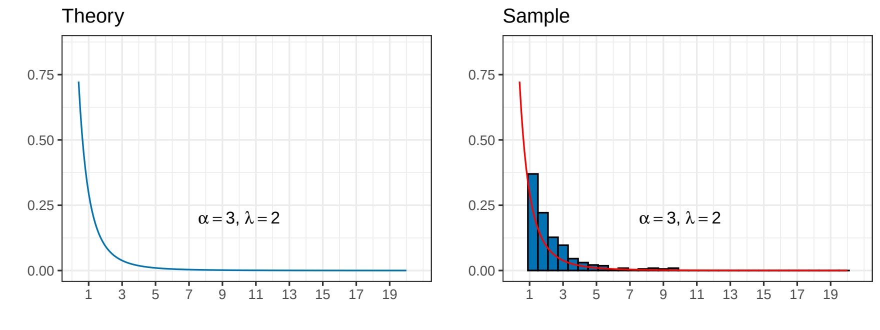
- right-skewed, unimodal
- α changes shape substantially
- β is a **rate** parameter ($b = 1/\beta$) is a **scale** parameter
- Special case is χ^2_v when $\alpha = \frac{v}{2}$ and $\beta = \frac{1}{2}$
- $E[X] = \frac{\alpha}{\beta} = \alpha b$, $Var(X) = \frac{\alpha}{\beta^2} = \alpha b^2$



Pareto

$$f(x \mid \alpha, \lambda) = \frac{\alpha\lambda^\alpha}{(\lambda + x)^{\alpha+1}} \quad x > 0, \text{ for } \alpha > 0, \lambda > 0$$

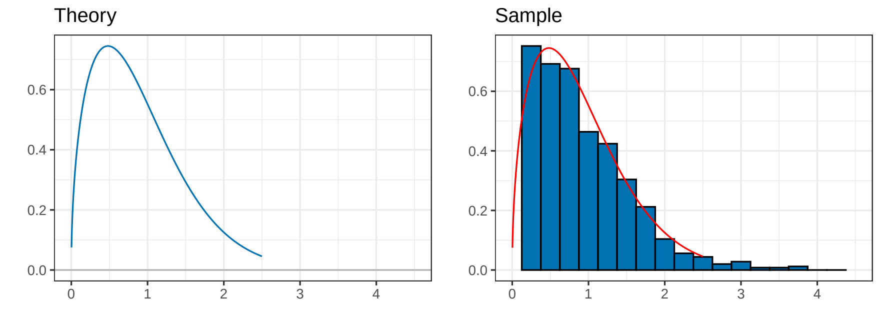
- Used to describe allocation of wealth, sizes of human settlement
- Heavier tailed than exponential distribution
- $E[X] = \frac{\lambda}{\alpha-1}$, for $\alpha > 1$, and $Var(X) = \frac{\alpha\lambda^2}{(\alpha-1)^2(\alpha-2)}$, for $\alpha > 2$



Weibull

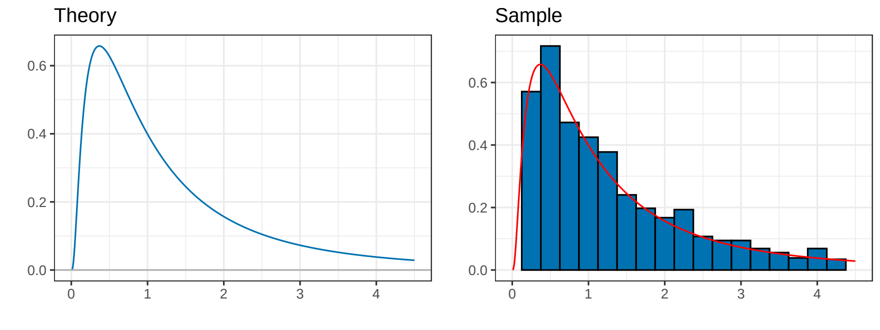
$$f(x \mid \lambda, k) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}, \quad x > 0, \text{ for } \lambda > 0, k > 0$$

- used for particle size distribution, failure analysis, delivery time, extreme value theory
- shape changes considerably with different k
- $E[X] = \lambda\Gamma\left(1+\frac{1}{k}\right)$ and $Var(X) = \lambda^2\left[\Gamma\left(1+\frac{2}{k}\right) - \left(\Gamma\left(1+\frac{1}{k}\right)\right)^2\right]$



Lognormal

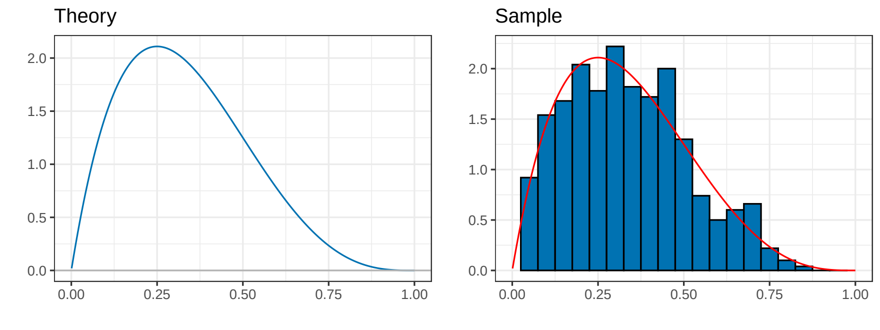
- Also called Galton's distribution
- Generated when $Y \sim N(\mu, \sigma^2)$, and study $X = exp(Y)$
- used for modeling length of comments posted in internet discussion forums, users' dwell time on the online articles, size of living tissue, highly communicable epidemics
- $E[X] = \exp\{\mu + \frac{\sigma^2}{2}\}$ and $Var(X) = \exp\{2\mu + \sigma^2\}\left(\exp\{\sigma^2\} - 1\right)$



Beta

$$f(x \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad x \in (0,1) \text{ for } \alpha > 0, \beta > 0$$

- Parameters $\alpha > 0$ and $\beta > 0$
- Generalisation of a continuous uniform on (0,1)
  - Same as a continuous uniform when $\alpha = \beta = 1$
- $E[X] = \frac{\alpha}{\alpha+\beta}$ and $Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$



Linear Combinations of random variables

- If X and Y are random variables, with mean values μ_X and μ_Y, respectively, and
- a and b are non-random constants, then
- the linear combination of X and Y, denoted by Z, is given by

$$Z = aX + bY$$

- The expected value of Z is given by

$$E[Z] = E[aX + bY] = a\ \mu_X + b\ \mu_Y$$

Linear Combinations of random variables

- If X and Y are random variables, with variances σ_X^2 and σ_Y^2, respectively, and
- a and b are non-random constants
- then a linear combination of X and Y given by

$$Z = aX + bY$$

- has **variance** given by

$$Var(Z) = a^2\ \sigma_X^2 + b^2\ \sigma_Y^2 + 2ab\ Cov(X, Y)$$

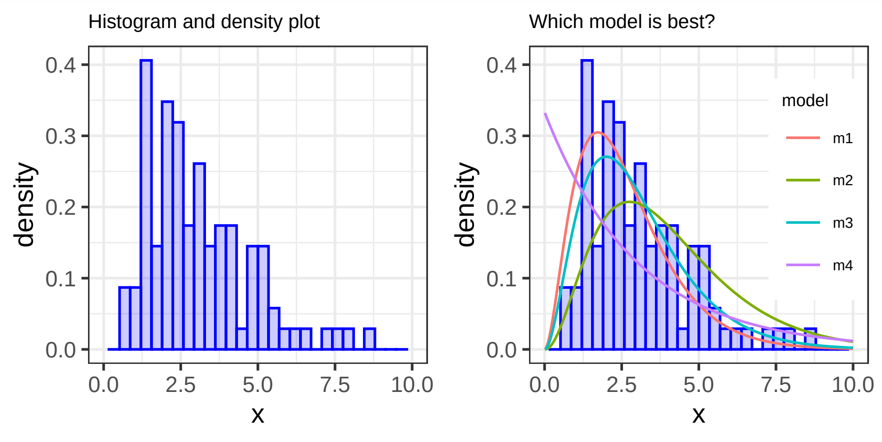
- Where the **covariance** between X and Y is given by

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E[X\ Y] - \mu_X\ \mu_Y$$

- If X and Y are independent, then $Cov(X, Y) = 0$.
    - The converse is not true

Modelling a single population (MLE)

- Which distributions might fit this data?
  - A normal distribution? An exponential? A gamma distribution? Something else?



Which model and fit?

- Assuming the data are a random sample, we need to **choose a model** $F_X(x \mid \theta)$
  - We fit models using the sample and well-established distributional families
- Once we choose a model, we'll need to **estimate** the parameter θ
  - use the **maximum likelihood estimation** (MLE) method
- A fitted model will imply an estimate of the population mean
  - and other features

The MLE

- Start with a given a **population model** $F_X(x \mid \theta)$
- Given **sample data** $x_1, x_2, \ldots, x_n$
  - assume data are i.i.d. (this can be relaxed)
  - With fixed sample size n
- How do we estimate the parameter, θ?
- $\Rightarrow$ we use the **Maximum Likelihood Estimator (MLE)**, denoted by $\hat{\theta}_{MLE}$
  - We find the MLE by maximising the likelihood function

Likelihood Function

If $x_1, x_2, ..., x_n \overset{i.i.d.}{\sim} F_X(x \mid \theta)$, then the likelihood function is

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f_X(x_i \mid \theta)$$

And the **MLE** for θ is

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} \mathcal{L}_n(\theta)$$

Notes about the MLE

- Under assumed population model with cdf F_X
- For **continuous**-valued x, $f_X(x|\theta)$ is a pdf
- For **discrete**-valued x, $f_X(x|\theta)$ is a probability (mass) function
- $\mathcal{L}_n(\theta)$ **is viewed as a function of the parameter** θ, for $\theta \in \Theta$
- **With the data fixed** at their observed values, $x_1, x_2, \ldots, x_n$

Optimising the likelihood function

- It is often easier to maximise the **log-likelihood function**

$$\ell_n(\theta) = \ln \mathcal{L}_n(\theta) = \left[\sum_{i=1}^{n} \ln f_X(x_i|\theta) \right]$$

- The **same** $\hat{\theta}_{MLE}$ maximises $\mathcal{L}_n(\theta)$ and $\ell_n(\theta)$
- In simple cases we can solve for $\hat{\theta}_{MLE}$ through differentiation
  - set first derivative of $\ell_n(\theta)$ equal to zero and solve
  - then check the second derivative of $\ell_n(\theta)$ is negative at $\hat{\theta}_{MLE}$
- More generally MLE is found using numerical optimisation on a computer

A Central Limit Theorem for MLE

- Under some regularity conditions, an MLE has an asymptotic Normal distribution:

$$\sqrt{n}\left(\hat{\theta}_{MLE} - \theta\right) \xrightarrow{D} N\left(0, V\right)$$

Some details

- This CLT comes from the **repeated sampling** behaviour of $\hat{\theta}_{MLE}$
- From the usual **Frequentist perspective**:
  - θ is **fixed**, but unknown
  - Estimator $\hat{\theta}_{MLE}$ is **random**
- If θ is a vector, then V is a matrix (related to the "*Fisher information*")
- V is usually unknown, and may depend on θ $\Rightarrow$ use $\hat{V}^{-1} = -\frac{1}{n} \left.\frac{\partial^2 \ell_n(\theta)}{\partial\theta\, \partial\theta'}\right|_{\hat{\theta}_{MLE}}$

$\Rightarrow \hat{\theta}_{MLE} \overset{approx}{\sim} N\left(\theta, \dfrac{\hat{V}}{n}\right)$ Here $\sqrt{\dfrac{\hat{V}}{n}}$ is the estimated standard error *SE*

MLE + CLT-based confidence intervals

- Approximate 95% confidence interval for scalar θ

$$\left(\hat{\theta}_{MLE} + z_{0.025}\sqrt{\frac{\hat{V}}{n}},\ \hat{\theta}_{MLE} + z_{0.975}\sqrt{\frac{\hat{V}}{n}}\right)$$

- Approximate 95% confidence interval for element $\theta[j]$

$$\left(\hat{\theta}[j] + z_{0.025}\ \sqrt{\frac{\hat{V}[j,j]}{n}},\ \hat{\theta}[j] + z_{0.975}\ \sqrt{\frac{\hat{V}[j,j]}{n}}\right)$$

- $\hat{V}[j,j]$ is j^{th} diagonal element of $\hat{V}$
- Use the standard normal quantiles: $z_{0.025} = -1.96$ and $z_{0.975} = 1.96$
  - not t_{n-1} quantiles

MLE in R with MASS::fitdistr()

- **Example**: Fit a *Gamma*(α, β) distribution to data in “x”

```r
fit <- fitdistr(x, "gamma")
fit
```

```
    shape       rate
  3.4697    1.1235
 (0.4690) (0.1634)
```

```r
fit %>% tidy() %>% kable() %>% kable_styling()
```

term	estimate	std.error
shape	3.470	0.4690
rate	1.123	0.1634

- Elements in “fit”: **estimate**, **sd**, **vcov**, **n** and **loglik**

MLE + Bootstrap-based approximate confidence Intervals for θ

We can also use a Bootstrap approach!

The Bootstrap CI for $\hat{\theta}_{MLE}$

1. Generate a Bootstrap sample of B potential $\hat{\theta}$ values
   - For each b in $1 : B$
     - resample n draws from the observed data values, with replacement
     - label these values as $\{x_1^{[b]}, x_2^{[b]}, \ldots, x_n^{[b]}\}$
     - compute the MLE $\hat{\theta}^{[b]}$ by maximising $\mathcal{L}_n^{[b]}(\theta)$, constructed from the bootstrap sample
   - Bootstrap sample: $\{\hat{\theta}^{[1]}, \hat{\theta}^{[2]}, \ldots, \hat{\theta}^{[B]}\}$
2. Use the empirical distribution from this Bootstrap sample to approximate the sampling distribution of $\hat{\theta}_{MLE}$
3. Construct an approximate 95% confidence interval by selecting interval from 2. with (empirical) probability (at least) 95%
   - (lower) 2.5% quantile to 97.5% quantile

For vector θ

- Do steps 1. and 2. Then do 3. for each component of θ

Why Bootstrap?

Relative advantages of the Bootstrap are:

- Relies only on the actual sample observed
- Approximates the sampling distribution of $\hat{\theta}_{MLE}$ for **finite n**
- With a large pool of potential bootstrap samples
  - we can get reasonably accurate CIs
  - (so long as your original sample is representative of the true population!)

Note this is called a PARAMETRIC bootstrap here

- Because we are using the **parametric model** assumption with MLE to get the estimator

Assessing model fit

Both CLT-based confidence intervals **and** Bootstrap-based confidence intervals

- Constructed from the output of an ML procedure
- Implicitly assume the selected “model” for ML is “correct” for the data

If the model doesn’t match the data well

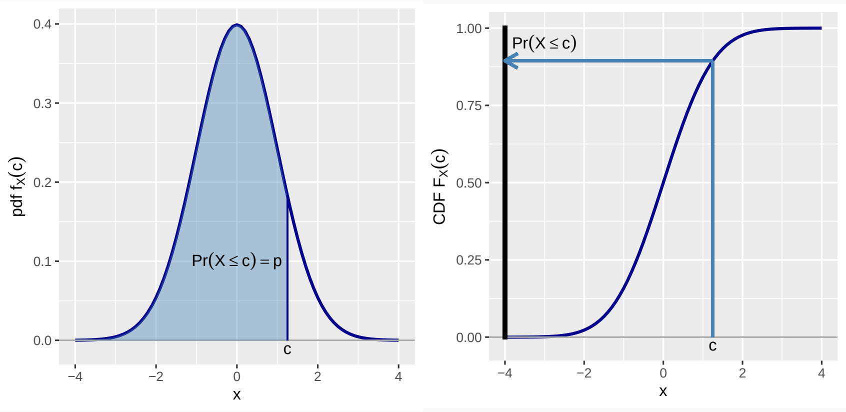
- $\Rightarrow$ parameter estimate and confidence interval(s) will not be very useful!

We need a way to assess the MODEL itself

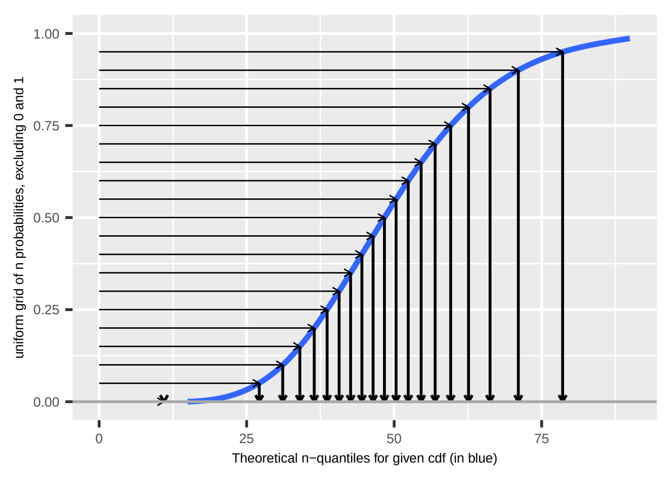
- Is the fitted model suitable for the data?
- Use QQplots, which are based on pairs that match:
  - **theoretical *n*-quantiles** (obtained by inverting the model’s cdf) with
  - **empirical *n*-quantiles** (i.e. the sorted sample data values)
- If these pairs “match” then the model is a good fit to the data!

Relationship between quantiles (percentiles), the pdf and the cdf

- The cdf of X, denoted $F_X(c)$, returns a value $p \in [0, 1]$
- This is equal to the area under the pdf of X, denoted $f_X(c)$, between $(-\infty, c]$



Inversion of a cdf



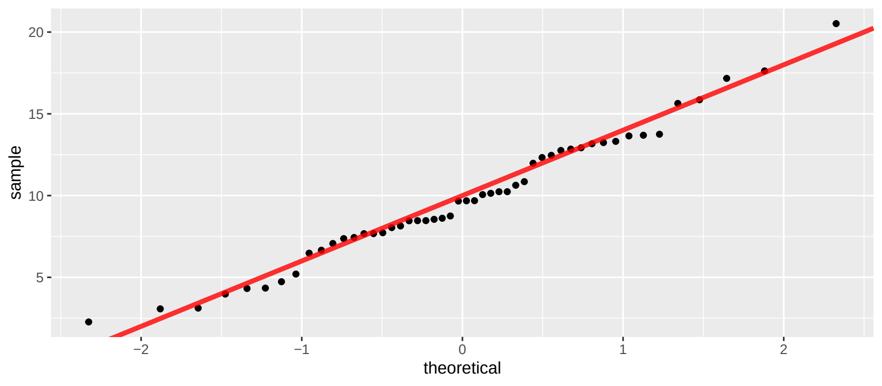
- Avoid potential inversion of cdf at 0 or 1 if range of distribution reaches $-\infty$ or ∞
  - e.g. set (n+1)-quantiles for $p_i = \frac{i}{n} - \frac{1}{2n},\ i = 1, 2, \ldots, n$

Quantile-Quantile Plot (QQplot)

- A graphical tool (subjective visual check) to help assess if plausible that data came from specified distribution
  - e.g. a distribution from MLE fit
- Create scatterplot
  - ordered data (y-axis) against theoretical quantiles (x-axis), or
  - ordered sample data against ordered simulated data
- If both sets of quantiles from same distribution $\Rightarrow$ points should lie on a straight line
  - if not straight, may get an idea of where data doesn't fit
- Often useful to add a line to QQplot
  - 45° line (perfect alignment)
  - line connecting specified quantiles (e.g. 25th- and 75th-%iles)

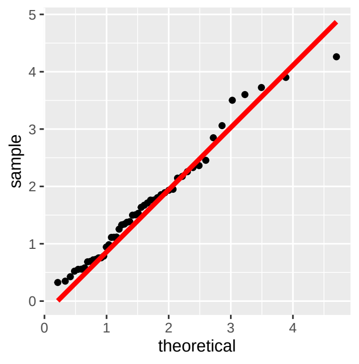
Example 1: N(μ, σ^2) against N(0,1) quantiles

```
n <- 50
df <- tibble(x = rnorm(n, 10, 4))
p <- df %>% ggplot() + geom_qq(aes(sample = x))
p <- p + geom_abline(intercept = 10, slope = 4, color = "red",
    size = 1.5, alpha = 0.8)
p
```



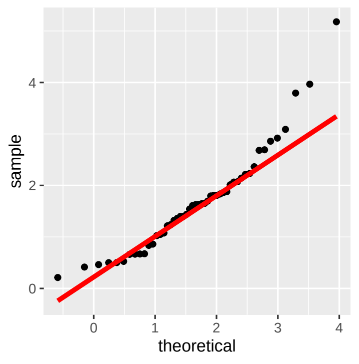
Example 2: stat_qq() for different distributions

```
df <- tibble(mydata = rgamma(n = 50, shape = 3, rate = 2))
fit <- fitdistr(df$mydata, "gamma")
params <- fit$estimate
ggplot(df, aes(sample = mydata)) + stat_qq(distribution = qgamma,
    dparams = params) + stat_qq_line(distribution = qgamma, dparams = params,
    color = "red", size = 1.5) + theme(aspect.ratio = 1)
```



Example 3

```
df <- tibble(mydata = rgamma(n = 50, shape = 3, rate = 2))
fit <- fitdistr(df$mydata, "normal")
params <- fit$estimate
ggplot(df, aes(sample = mydata)) + stat_qq(distribution = qnorm,
    dparams = params) + stat_qq_line(distribution = qnorm, dparams = params,
    color = "red", size = 1.5) + theme(aspect.ratio = 1)
```



About QQplots

- Can we test?
- H_0: data comes from the specified model vs. H_1 data does not come from the specified model
- In most cases, fit will not be perfect

Various approaches available for informal test:

- Use a ‘thick-marker’ judgment approach
- Use a bootstrap technique to obtain “confidence set”
- Embed QQplot from among many QQplots from data simulated from the model