StatMeth超浓缩整合攻略(期中篇)

`statmeth`

Introduction to statistics

Collecting Sample data

Different sampling methods:

Voluntary response sample

subjects decide themselves to be included in sample.
(very biased)

Random sample

each member of population has equal probability of being selected.

Simple random sample

each sample of size n has equal probability of being chosen.

Systematic sampling

after starting point, select every k-th member.

Stratified sampling

divide population into subgroups such that subjects within groups have same characteristics, then draw a (simple) random sample from each group.

Cluster sampling

divide population into clusters, then randomly select some of these clusters.

Convenience sampling

easily available results.

Different variables:

Variable

varying quantity.

Response (dependent) variable

representing the effect to study

Explanatory (independent) variable

possibly causing that effect

Confounding

mixing influence of several explanatory variables on response

`Example`

Independent variable -> alcohol consumption
Dependent variable -> mortality
Confounding variables -> age, gender, education ...

Different types of study:

Observational study

characteristics of subjects are observed; subjects are not modified.

- Retrospective (case-control) : data from past
- Cross-sectional : data from one point in time
- Prospective (longitudinal) : data are to be collected

Experiment

some subject treatment.

- Sometimes control and treatment group; single-blind or double-blind（设置对照组；单盲：被测试者 | 双盲：被测者和测试者）
- To measure placebo effect or experimenter effect. (安慰剂效应和观察者效应)

Types of data

> Differ in sample size

Parameter

numerical measurement describing a **population's** characteristic.

Notation: typically Greek symbols, e.g. μ, σ.

Statistic

numerical measurement describing a **samples's** characteristic.
Notation: small letters, e.g. $\overline{x}, s$.

> Differ in data type

Qualitative (categorical)

names or labels represent counts or measurements
Examples : good/bad/fair

Quantitative (numerical)

numbers represent counts or measurements

- **Discrete** : the set of possible values is countable (e.g. number of siblings)
- **Continous** : the set of possible values is uncountable (e.g. weight of oldest sibling)

> Based on the level of measurement

Qualitative data:

- **Nominal** : names, labels, categories (no ordering). No computation possible. (e.g. gender, eye colour)
- **Ordinal** : categories with ordering, but no meaningful differenes. (e.g. grades(A-F), opinions (totally disagree/agree))

Quantitative data:

- **Interval** : ordering possible and meaningful differences, but no natural zero starting point. (e.g. year of birth, temperature)
- **Ratio** : ordering possible and meaningful differences & natural starting point. (e.g. body lenth, marathon times.)

Summarising and graphing data

Describe data distribution :

Graphical :

- **Frequency distribution** (table) : count occurences of category
- **Bar chart**
- **Pareto bar chart** : categories ordered w.r.t. frequency, required data of nominal meansurement level!
- **Pie chart** : pie piece sized determined by relative frequency of category.(Mainly : qualitative data)
- **Histogram** : bar areas are proportional to frequency in respective interval.
- **Time series** : visualization of time-varying quantity(e.g.yearly number of sunspots).

Descriptive :

- Qualitative : describe shape, location and dispersion
- Quantitative : numerical summaries of location and variation

> Qualitative description:

Shape

make smooth approximation of histogram.

- Symmetrical
- Skewed (right-skewed, left skewed)
- Uniform

Location

position on x axis.

Dispertion (spread/variation)

measure of variation with dataset.

> Numerical summaries:

Measure of center

value at the center or middle of a data set.

- **mean** : the "average". Every data value used.
  Not robust: strongly affected by extreme values.
  Sample mean : $\overline{x} = (\sum_{i=1}^{n} x_i)/n$
  Population mean: $\mu = (\sum_{i=1}^{N} x_i)/N$
- **median** : the "middle" value of the data set (after sorting).

Robust : not much affected by extreme values.

- **mode** : the value that occurs with highest frequency.
  Hardly used for numerical data, but applicable to nominal data.
  Dataset with unique mode : **unimodal, bimodal/multimodal** (graphs with different peaks).

Measure of variation

- **sample standard deviation** : common measure of variation. Measures how much the values deviate from the sample mean.
  $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} = \sqrt{\frac{n\sum_{i=1}^{n} x_i - (\sum_{i=1}^{n} x_i)^2}{n(n-1)}}$
- **sample variance** : the square of standard deviation.
  $s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$
- **population standard deviation** : σ
- **population variance** : σ^2
- **Range** : maximum - minimum

Measure of relative standing and boxplots :

Percentiles P_i :

i% of data values is smaller than P_i and $(100-i)$% is larger than P_i.
Special percentiles : **quartiles Q_1, Q_2, Q_3**.

- $Q_1 = P_25$: first quartile
- $Q_2 = P_50$ = median : first quartile
- $Q_3 = P_75$: third quartile

5-number summary :

1. Minimum
2. First quartile, Q_1
3. Median, Q_2
4. Third quartile, Q_3
5. Maximum

Interquartile range (IQR):

IQR = $Q_3 - Q_1$

Boxplots :

provide information about distribution

- **Whiskers** : lines extending from the box. Not exceed $1.5 * IQR$
- **Outliers** : all points not included

Probability

Basic concepts of probability

Probability experiment :

Production of (random outcome).
E.g. die roll, coin toss.

Sample space Ω :

Set of all possible outcomes.
E.g. $\Omega =$ {1,2,3,4,5,6}

Event A, B, ... :

Collection of outcomes.
E.g. A = {even number is thrown} = {2,4,6}

Simple event :

Consists 1 outcome.
E.g. {1}.

Probability measure :

Function $P(\cdot)$ assigning values between 0 and 1 to events.
E.g. $P(A)$ = $P($ {2,4,6} $) = \frac{1}{2}$.

Interpretation of probabilities :

- $P(A) = 0$: occurence of A is impossible.
  e.g. $P(\emptyset) = 0$.($\emptyset =$ empty event : nothing happens)
- $P(A) = 1$: occurence of A is certain.
  e.g. $P(\Omega) = 1$.
- Event A is unlikely when $P(A)$ is small, e.g. < 0.05

Law of Large numbers (LLN) :

> Suppose a procedure is repeated again and again and outcomes are independent. Then the relative frequency probability of an event A tends towards true $P(A)$.

`Notice` -> Special case

Three ways to determine probability $P(A)$ of event A:

1. Estimate with **relative frequency** :

$$P(A) = \frac{number\ of\ times\ A\ occurred}{number\ of\ times\ the\ procedure\ was\ repeated}$$

   `Many trials` -> relative frequency $\approx$ real (true) value of $P(A)$ (Supported by Law of Large numbers)

2. **Classical (theoretical) approach** :
   Make probability model (outcome space, probability measure, etc.) and compute $P(A)$ using properties of P.
   E.g: rolling dice, card games...

3. **Subjective approach** :
   Estimate $P(A)$, based on intuition and/or experience.

> **Example of classical approach** : Throw a fair (unbiased) coin 3 times. What is the probability of 1 time Heads?

- Sample space Ω has $2 * 2 * 2 = 8$ outcomes.
  Ω = {HHH,HHT, HTH, HTT, THH, THT, TTH, TTT}.
- Interestion event A = {1 H} -> A = {HTT, THT, TTH}.
- The outcomes are equally like, hence:

$$P(A) = \frac{number\ of\ times\ A\ occurred}{total\ number\ of\ different\ simple\ events} = \frac{3}{8}$$

Counting principle :

> Suppose two probability experiments are performed. If

- experiment 1 has $a \geq 0$ possible outcomes
- experiment 2 has $b \geq 0$ possible outcomes
  Then the experiments combined have $a * b$ possible outcomes.
  This principle extends to any number of experiments.

> **Example of counting principle** : First throw coin, then roll die.

⇒ total number of outcomes of both experiments: 2 * 6 = 12

General probability measure for finite/countable sample space Ω

In general it is not necessarily true that all outcomes are equally likely. E.g.: biased die.

In all cases of **discrete sample spaces** (finite/countable):

- Each outcome $\omega \in \Omega$ has a probability, and
  $P(\omega) \geq 0$（任何事件的概率一定是正数）
  $\sum_{\omega \in \Omega} P(\omega) = 1$（sample space中所有单独项概率的和等于1）
- The probability of an event A is defined by

$$P(A) = \sum_{\omega \in A} P(\omega)$$

> **Example : biased die.**
> What is the probability of throwing an even number?

Ω = {1,2,3,4,5,6}.

Outcomes not equally likely :
$P(6) = \frac{2}{7}$ and $P(1) = P(2) = \ldots = P(5) =$

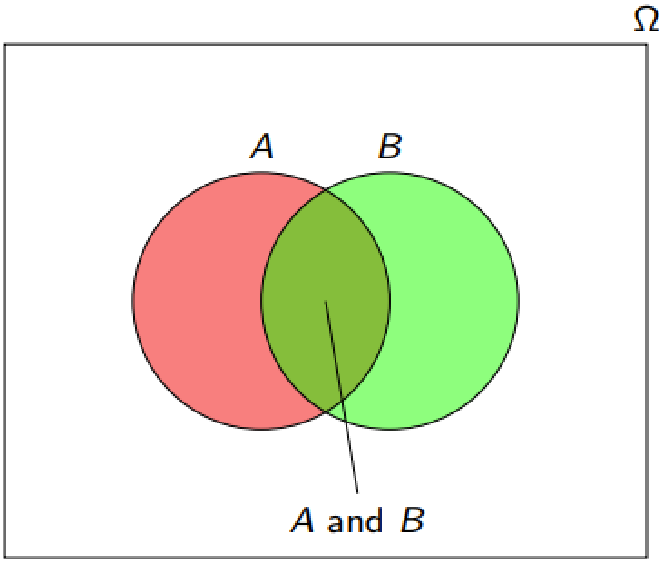
A = {even number} = {2, 4, 6}

⇒ $P(A) = P(\{2,4,6\}) = P(2) + P(4) + P(6) = \frac{1}{7} + \frac{1}{7} + \frac{2}{7} = \frac{4}{7}$

Addition rule

Idea : every outcome is counted only once.

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$



`Notation:`

$A \cup B = A \text{ or } B$:

union, set of outcomes which are in A **or** B (both allowed!)

$A \cap B = A \text{ and } B$:

intersection, set of outcomes which are both in A **and** B.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

> **Example : three coin tosses (unbiased coin)**
> Compute the probability of the event "Tails twice or Heads in first throw".

Ω = {HHH, HHT, HTH, HTT, THH, THT, TTH}.

A = {Tails twice} = {HTT, THT, TTH}, so $P(A) = \frac{3}{8}$.

B = {Heads in first throw} = {HTT, HHT, HTH, HHH}, so $P(B) = \frac{4}{8} = \frac{1}{2}$.

$A \cap B$ = {Tails twice and heads in first throw} = {HTT}, so $P(A \cap B) = \frac{1}{8}$.

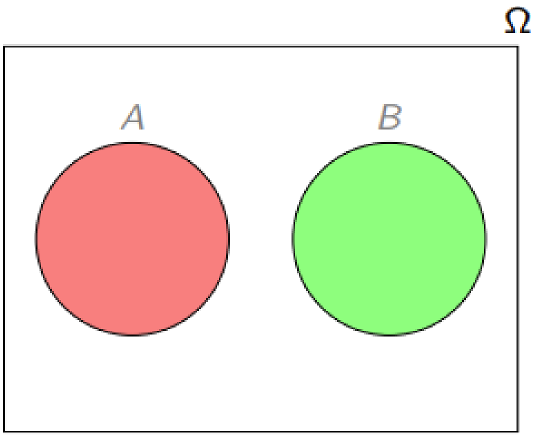
$\Rightarrow P$(Tails twice or Heads in first throw)
$= P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{3}{8} + \frac{1}{2} - \frac{1}{8} = \frac{3}{4}$

Addition rule for two **disjoint events** :

A and B are **disjoint** if they exclude each other, i.e. $A \cap B = \emptyset$.

If A and B are disjoint then:

$$P(A \cup B) = P(A) + P(B)$$



`Notice` : This is **different from independence**!!

> **Example : Roll a fair die once.**
> What is the probability you throw an even number or 3?

Ω = {1, 2, 3, 4, 5, 6}.

A = {even number} = {2, 4, 6}, so $P(A) = \frac{3}{6} = \frac{1}{2}$.

B = { 3 }, so $P(B) = \frac{1}{6}$.

Furthermore, $A \cap B = \emptyset$, so A and B disjoint. Hence,

$$P(A \cup B) = P(A) + P(B) = \frac{1}{2} + \frac{1}{6} = \frac{2}{3}$$

> General addition rule for **disjoint events** :

Let $A_1, \ldots, A_m$ be disjoint, i.e. $A_i \cap A_j = \emptyset$ for $i \neq j$. Then :

$$P(A_1 \cup \ldots \cup \ldots A_m) = \sum_{i=1}^{m} P(A_i)$$

> **Example : rolling two fair dice**
> What is the probability of "sum equals 4, 8, 9"?

$\Omega =${(1,1), ... , (1,6), (2,1), ... , (6,6)} contains 6 x 6 = 36 outcomes, which are all equally likely.

A = {Sum is 4} = {(1,3), (2,2), (3,1)},

B = {Sum is 8} = {(2,6), (3,5), (4,4), (5,3), (6,2)},

C = {Sum is 9} = {(3,6), (4,5), (5,4), (6,3)}.

$$P(sum\ is\ 4,8,9) = P(A) + P(B) + P(C) = \frac{3}{36} + \frac{5}{36} + \frac{4}{36} = \frac{1}{3}$$

Complement rule :

$\overline{A}$ (or A^c) : complement of A; outcomes which are not in A.

$$P(\overline{A}) = 1 - P(A)$$

> **Example : three fair coin tosses**
> What is the probability of at least one Heads?

A = {at least 1 Heads} $\Rightarrow \overline{A} =$ {no Heads}.

$$P(A) = 1 - P(\overline{A}) = 1 - P(no\ Heads) = 1 - P(TTT) = 1 - \frac{1}{8} = \frac{7}{8}$$

> Complement of **at least one** is no occurence of ...

Multiplication rule :

$P(B|A)$: **conditional** probability that B occurs **given** that A has occurred.

If $P(A) > 0$, then :

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

- If A has occurred, B only happens if outcome is in both A and B. Hence, in $A \cap B$.
- The sample space is reduced to A.
- Hence, given A has occured, compute $P(A \cap B)$ relative to $P(A)$.

`Notice`: $P(B|A) \neq P(A|B)$ in general.

> **Example : 2 fair coin tosses**
> What is the conditional probability of "twice Heads" given that
> 1. the first flip is Heads?
> 2. there is at least one Heads?

(1) : (Sample space 和 event 陈述省略)

$$P(B|A_1) = \frac{P(A_1 \cap B)}{P(A_1)} = \frac{P(HH)}{P(HH, HT)} = \frac{1/4}{1/2} = \frac{1}{2}$$

(2) :

$$P(B|A_2) = \frac{P(A_2 \cap B)}{P(A_2)} = \frac{P(HH)}{P(HH, HT, TH)} = \frac{1/4}{3/4} = \frac{1}{3}$$

> The formula can also be written as :

$$P(A \cap B) = P(A) \cdot P(B|A)$$

> **Example : Draw balls from vase**
> Vase with ball 1 to 9.
> Draw two balls, after each other.
> What is the probability of first is 1 and then 2 ?

$$P((1,2)) = P(first\ 1, then\ 2) = P(first\ 1) \cdot P(draw\ ball\ 2|ball\ 1\ is\ drawn) = \frac{1}{9} \cdot \frac{1}{8} = \frac{1}{72}$$

Independence :

Two events A and B are **independent** if

$$P(A \cap B) = P(A) \cdot P(B)$$

Thus $P(B) = P(B|A)$ when A and B are independent.

`Notice` : Independence $\neq$ disjointness !

Independence depend on the sampling methods:

- sampling with replacement : selections are independent events
- sampling without replacement : selections are dependent events

`However`, to simplify calculations :

Small sample rule :

> When drawing a small sample from a large population, we treat the selections as independent events.

Law of Total Probability and Baye's Theorem

Baye's Theorem

> Addition rule for disjoint events ($B \cap A$ & $B \cap \overline{A}$) :

$$P(B) = P(B \cap A) + P(B \cap \overline{A})$$

Then, by the multiplication rule :

$$P(B) = P(B \cap A) + P(B \cap \overline{A}) = P(B|A) \cdot P(A) + P(B|\overline{A}) \cdot P(\overline{A})$$

Simple law of total probability :

Let A and B be events. Then

$$P(B) = P(B \cap A) + P(B \cap \overline{A}) = P(B|A) \cdot P(A) + P(B|\overline{A}) \cdot P(\overline{A})$$

> **Baye's Theorem :**

Let A and B be events, then:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|\overline{A}) \cdot P(\overline{A})}$$

`Notice` :

$P(B|A) + P(\overline{B}|A) = 1$

but in general :

$P(B|A) + P(B|\overline{A}) \neq 1$

> **Example : medical test for certain disease**
> Suppose 0.1% of population has the disease.
> Medical test : if someone
> - has the disease $\Rightarrow$ positive test result with probability 0.98.
> - does not have the disease $\Rightarrow$ negative test result with probability 0.99.
> Suppose Dennis conducts the test: the result is positive.
> What is the probability that Dennis has the disease **given the positive test** outcome?

Let B = {positive} and A = {disease}. Compute P(A|B).

$P(B|A) = 0.98 \Rightarrow$ use Bayes's theorem :

First, compute $P(B|\overline{A})$, $P(A)$ and $P(\overline{A})$.

$\overline{A}$ = {does not have disease}

We know : $P(B|\overline{A}) = 0.01$, $P(A) = 0,001$ and $P(\overline{A}) = 1 - 0.001 = 0.999$.

$\Rightarrow P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|\overline{A}) \cdot P(\overline{A})} = \frac{0.98 * 0.001}{0.98 * 0.001 + 0.01 * 0.999} \approx 0.089$

The probability that Dennis has the disease is 8.9%.

Partition :

Events $A_1, \ldots, A_m$ are called a **partition** if

- pairwise disjoint : $A_i \cap A_j = \emptyset$,if $i \neq j$
- union is entire sample space : $A_1 \cup A_2 \cup \ldots \cup A_m = \Omega$

Let $A_1, \ldots, A_m$ be a partition, then also $B \cap A_1, \ldots B \cap A_m$ disjoint. Then :

$P(B) = P(B \cap \Omega) = P(B \cap (A_1 \cup A_2 \cup \ldots \cup A_m))$

$= P((B \cap A_1) \cup (B \cap A_2) \cup \ldots \cup (B \cap A_m))$

$= \sum_{i=1}^{m} P(B \cap A_i)$ `(general addition rule for disjoint event)`

$= \sum_{i=1}^{m} P(B|A_i) \cdot P(A_i)$ `(multiplication rule)`

> **Law of Total Probability :**

Let $A_1, \ldots, A_m$ be a partition, then :

$$P(B) = \sum_{i=1}^{m} P(B \cap A_i) = \sum_{i=1}^{m} P(B|A_i) \cdot P(A_i)$$

> **Example : defective products in a factory**
> Machines 1, 2 and 3 produce 30%, 45% and 25% of all products.
> Respectively 2%, 3% and 2% thereof are defective.
> A randomly selected product is defective.
> What is the probability that it came from machine 2?

A_i = {machine i made product}, B = {product defective},

so interested in $P(A_2|B)$.

We have $P(A_1) = 0.30$, $P(A_2) = 0.45$, $P(A_3) = 0.25$.

$P(B|A_1) = 0.02$, $P(B|A_2) = 0.03$ and $P(B|A_3) = 0.02$. Hence,

$P(A_2|B) = \frac{P(B|A_2) \cdot P(A_2)}{P(B|A_1) \cdot P(A_1) + P(B|A_2) \cdot P(A_2) + P(B|A_3) \cdot A_3} = \frac{0.0135}{0.0245} \approx 0.55$

Probability Distributions

Random Variable:

A random variable is a variable that assigns a numerical value to each outcome of a

probability experiment.
`Notation` : $X, Y, \dots$
x -> value of random variable

> **Example : two coin tosses**
> Throw a fair coin twice. Let the random variable X be the number of heads.

Sampe space : Ω = {HH, HT, TH, TT} .

Values of X for those outcomes :
X(HH) = 2, *X(HT)* = 1, *X(TH)* = 1, *X(TT)* = 0.

So, X takes values 0, 1, 2.

A probability distribution :

determines all probabilities of possible values of a random variable. Given by a table, formula or graph.

A discrete random variable :

has finite (or countably) many different values.

- Its probability distribution is the collection of all their individual probabilities.
- The total sum of these probabilities is 1.

A continous random variable :

has uncountably many different values.

- Its probability distribution is given by **probability density function**.
- Probabilities can be computed by area under this function.
- The total area is 1.

Discrete random variable

> **Recipe to find probability distribution of discrete random variable.**

- Determine the sample space of the underlying probability experiment and the probabilities of the outcomes ω.

- List the values $X(\omega)$ for all ω in Ω.
- For each value x of X, find all simple events { \omega} with value x. They form the event $\{X = x\} = \{\omega : X(\omega) = x\}$.
- Probabilities $P(\{\omega\})$ determine the probability of $\{X = x\}$:

$P(X = x) = P(\{\omega : X(\omega) = x\}) = \sum_{\omega:X(\omega)=x} P(\{\omega\})$

- Make a table : left column with all values x of X, right column with probabilities $P(X = x)$.

> **Example : two coin tosses (fair)**
> Random variable X: number of heads.

⇒ $X(HH) = 2$, $X(HT) = 1$, *X(TH) = 1, $X(TT) = 0$.

$P(X = 0) = P(\{TT\}) = \frac{1}{4}$,
$P(X = 1) = P(\{TH, HT\}) = \frac{2}{4} = \frac{1}{2}$,
$P(X = 2) = P(\{TT\}) = \frac{1}{4}$,

x	$P(X = x)$	num. $P(X = x)$
0	1/4	0.25
1	1/2	0.50
2	1/4	0.25

Check : $P(X = 0) + P(X = 1) + P(X = 1) = \frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1$.

Expected value (expectation / mean) :

of a discrete random variable X with possible values $x_1, \ldots, x_k$ is the weighted average of all possible values of X:

$$\mu = E(X) = \sum_{i=1}^{k} \cdot P(X = x_i)$$

> **Example : X = maximum of two fair dice**
> What is $E(X)$?

x	$P(X = x)$	num. $P(X = x)$	$x \cdot P(X = x)$

x	*P(X = x)*	num. *P(X = x)*	*x • P(X = x)*
1	1/36	0.028	0.028
2	1/12	0.083	0.167
3	5/36	0.139	0.417
4	7/36	0.194	0.778
5	1/4	0.250	1.250
6	11/36	0.306	1.833

Thus :

$$E(X) = \sum_{i=1}^{6} \cdot P(X = x_i) \approx 4.472$$

Variance :

of a discrete random variable X with values $x_1, \ldots, x_k$ is

$$\sigma^2 = Var(X) = \sum_{i=1}^{k} [(x_i - \mu)^2 P(X = x_i)]$$

Standard deviation of X:

$$\sigma = SD(X) = \sqrt{Var(X)} = \sqrt{\sum_{i=1}^{k} [(x_i - \mu)^2 P(X = x_i)]}$$

`NB`: convenient manual computation

$Var(x) = \sum_{i=1}^{k} [x_i^2 P(X = x_i)] - \mu^2$

> **Example : X = maximum of two fair dice**
> What is $SD(X)$?

Probability distribution + weighted averages :

x	*P(X = x)*	num. *P(X = x)*	*x • P(X = x)*	x^2*•P(X = x)*
1	1/36	0.028	0.028	0.028
2	1/12	0.083	0.167	0.333
3	5/36	0.139	0.417	1.250
4	7/36	0.194	0.778	3.110
5	1/4	0.250	1.250	6.250
6	11/36	0.306	1.833	11.000

Thus $\sum_{i=1}^{6} i^2 \cdot P(x=i) \approx$. Hence,

$$\sigma^2 = Var(x) = \sum_{i=1}^{6}[i^2 P(X=i)] - \mu^2 \approx 21.972 - 20.000 = 1.972$$

Finally,

$$\sigma = \sqrt{Var(X)} \approx \sqrt{1.972} \approx 1.404$$

Law of Large Numbers Theorem :

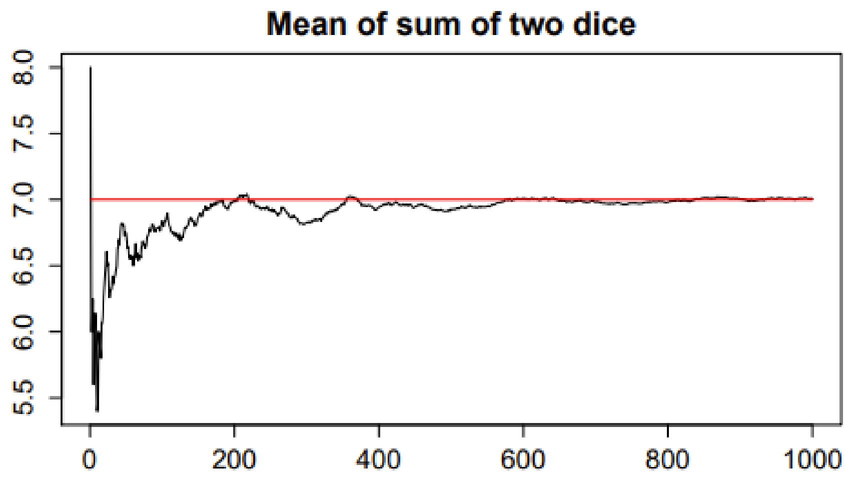
> Let $X_1, \dots, X_n$ be *n* **independent** versions of random variable *X*, where X has expected value μ. Then their mean $\frac{1}{n}(X_1 + \dots + X_n)$ tends to approach μ.

`Notice`

This is a special version of the LLN in basic Probability section :
random variable $X_i = 1$ if *A* occurs, $X_i = 0$ if *A* does not occur.

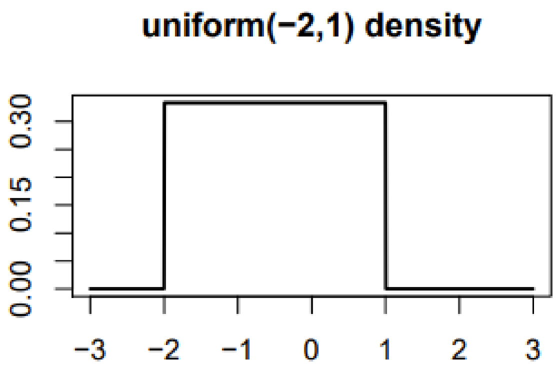
> Example : *X* = sum of two fair dice

We can find that $E(X) = 7$. Behaviour of mean of X_i's after $n(\to \infty)$ double rolls.



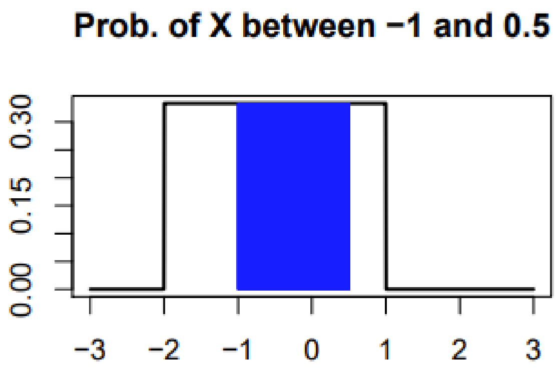
Continuous random variables

> **Example : choose point in interval**
> Let X denote a random point between -2 and 1.
> What is the probability distribution of X?



The probability density function is :

$p(x) = \frac{1}{3}$ for $x \in [-2, 1]$.



(长 x 宽 = 长方形面积)

$P(-1 \leq X \leq \frac{1}{2}$ = blue area = $(\frac{1}{2}-(-1)) \cdot \frac{1}{3} = \frac{3}{2} \cdot \frac{1}{3} = \frac{1}{2}$

Standard normal distribution

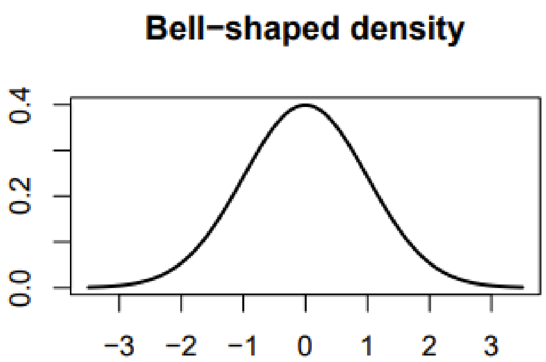
Probability density function :

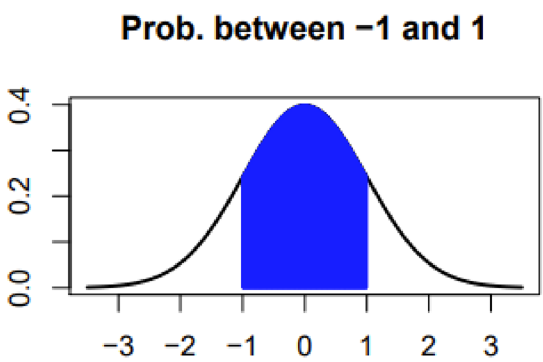
a curve *p(x)* such that

- $p(x) \geq 0$ for all x,
- total area under curve = 1.

The **probability** that *X* takes values between *a* and *b*, i.e. $P(a \leq X \leq b)$ equals the **area** under the curve $p(x)$ between *a* and *b*.

> **Example : bell-shaped density**





Normal distribution :

A random variable X has a normal distribution if it has probability density

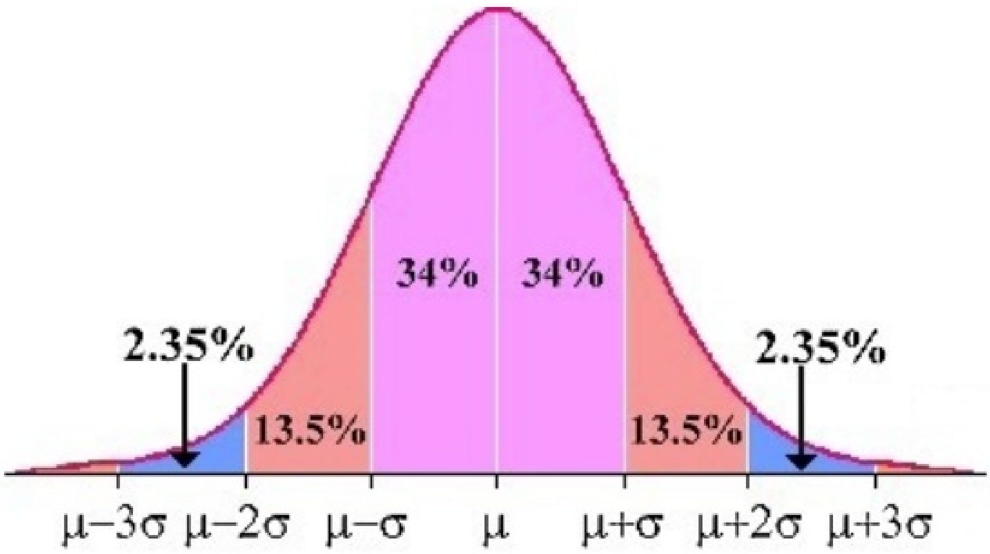
$$p(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

This density is continuous, **bell-shaped** and **symmetric**.

We write $X \sim N(\mu, \sigma^2)$ and for X normally distributed with mean μ and variance σ^2.

he standard normal distribution has mean 0 and variance 1 : $N(0, 1)$.

> **Rule of thumb for $N(\mu, \sigma^2)$**



- 68% of probability mass lies between $\mu - \sigma$ and $\mu + \sigma$
- 95% of probability mass lies between $\mu - 2\sigma$ and $\mu + 2\sigma$
- 99.7% of probability mass lies between $\mu - 3\sigma$ and $\mu + 3\sigma$

> **Determine probabilities of a normally distributed random variable**

$P(X \leq Z)$ = area under density to the left of z

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a)$$

$$P(X \geq b) = 1 - P(X \leq b)$$

- In case of $N(0, 1)$: Use Table 2 of book (p.786-787)
- For $N(\mu, \sigma^2)$: compute z-scores and use Table 2.

> **Example : Probabilities of standard normal distribution**
> Let $Z \sim N(0,1)$.

1. $P(Z \leq 0.5) = 0.6915$ (cumulative area to the left of 0.5)
2. $P(Z \geq -1.33) = 1 - P(Z \leq -1.33) = 1 - 0.0918 = 0.9082$
3. $P(Z \in [-1.33, 0.5]) = P(-1.33 \leq Z \leq 0.5)$
   $= P(Z \leq 0.5) - P(Z \leq -1.33) = 0.6915 - 0.09$

Applications of normal distributions

> Relationship $N(\mu, \sigma^2)$ versus $N(0,1)$

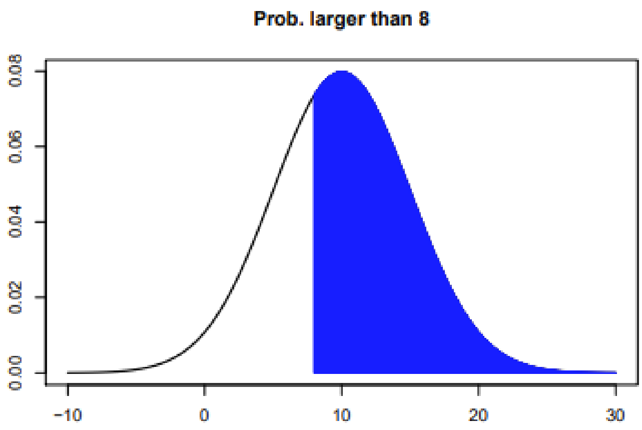
If random variable $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$.

Z-score of value x:

Let x be a (data) value of interest, related to a population distribution with mean μ and standard deviation σ. The z-score of x is $z = \frac{x-\mu}{\sigma}$.

`Interpretation` : number of standard deviations away from the mean.

> **Exampe : $X \sim N(10, 25)$.**
> What is $P(X \geq 8)$?



$X \sim N(10, 25)$ so $\mu = 10$ and $\sigma = 5$.

Since $Z = \frac{X-10}{5} \sim N(0,1)$,

$$\begin{aligned} P(X \geq 8) &= P(\frac{X-10}{5} \geq \frac{8-10}{5}) \\ &= P(Z \geq -0.4) \\ &= 1 - 0.3446 \\ &= 0.6554 \end{aligned} \tag{8}$$

> **Example : X = "random test score"**
> X is appoximately $N(500, 10000)$-distributed.
> What is the probability that random participant scores are between 550 and 700?

Compute z-scores of 550 and 700 :
$x = 550 \rightarrow z = \frac{550-500}{100} = 0.5$,

$x = 8700 \rightarrow z = \frac{700-500}{100} = 2.0$

Hence, $P(500 \leq X \leq 700) = 0.9772 - 0.6915 = 0.2825$

The Central Limit Theorem

The Central Limit Theorem (CLT) :

> Independently draw a sample of size *n > 30* from a population with mean μ and standard deviation σ.
> Then $\overline{X}_n$ has **approxmately** a $N(\mu, \frac{\sigma^2}{n})$-distribution
> (hence, standard deviation $\frac{\sigma}{\sqrt{n}}$).

`Notice` : the population can have any distribution!

> Special case :

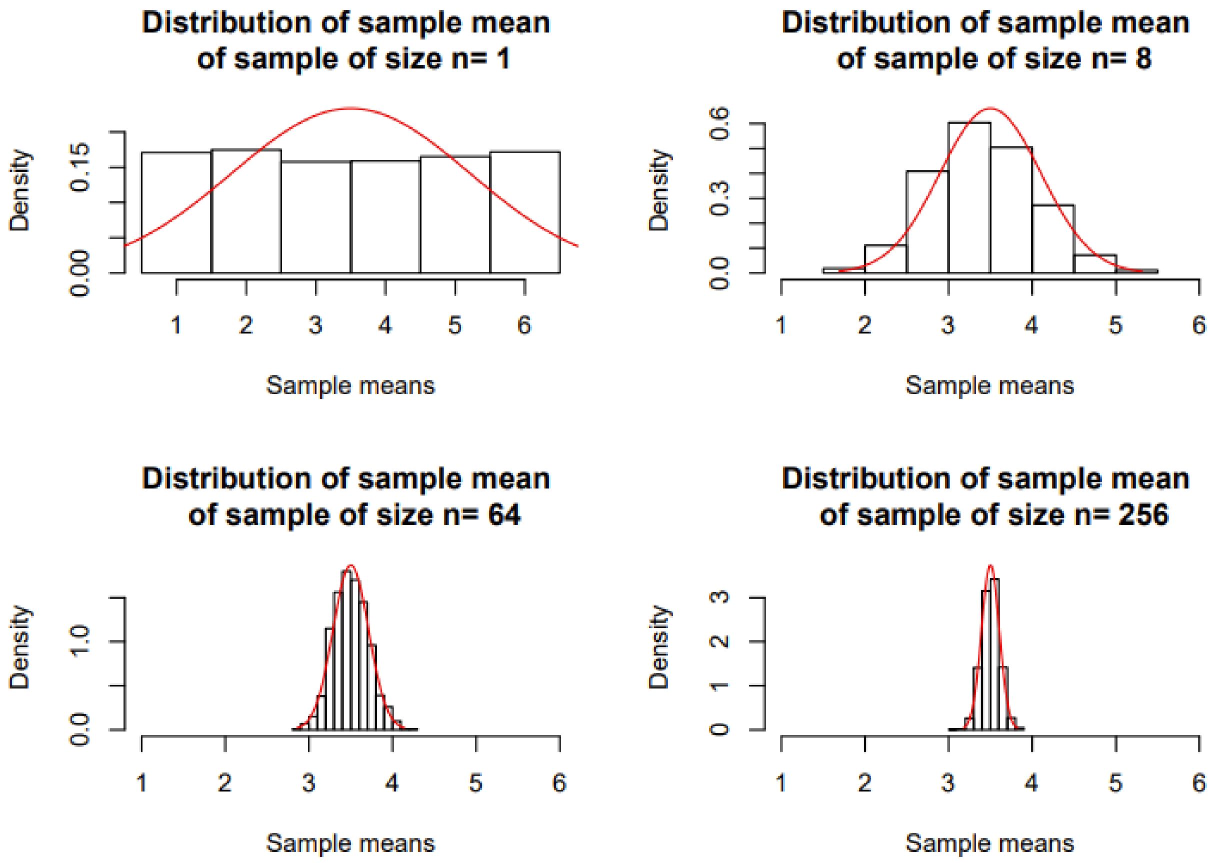
Independently draw a sample of size *n* from a **normal** population with mean μ and standard deviation σ.
Then $\overline{X}_n \sim N(\mu, \frac{\sigma^2}{n})$.

`Notice` : *n* can be any number.

Example : illustration of CLT for sample mean of a fair die.

Histograms : distribution of 1000 sample means of n = 1, 8, 64, and 256 die rolls.

Red line: normal distribution according to CLT, i.e. $N(3.5, \frac{2.92}{n}$



Example application of CLT : test scores
Test scores are approximately $N(500, 10000)$-distributed.
1.Alice scores 475. What perentage of students performs better?
2.A school of 100 students has an everage score of 475. What percentage of schools performs better?

1.The z-score of x = 475 is $\frac{475-500}{100}$ = -0.25.

Table 2 : 1 - 0.4013 = 0.5987,
so ca. 60% of students performs better.

2.CLT - >
Distribution of mean score of a school of 100 students is $N(500, \frac{10000}{100})$,

so mean $\mu = 500$ and standard deviation $\sigma = \frac{100}{\sqrt{100}} = 10$. Hence, z-score of x = 475 is $\frac{475-500}{10} = -2.5$.

Table 2 : 1 - 0.0062 = 0.9938,
so 99.38% of comparable schools perform better.

> Is the sample mean normally distributed?

Consider a population distribution with mean μ and standard deviation σ
Take a sample of size n from this population.

The sample mean $\overline{X}$ has a normal distribution if

- **Sample size n > 30.** Then CLT applies and $\overline{X}$ has approximately a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.
- **The population distribution is a normal distribution.** Then, $\overline{X}$ has a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ for any n.

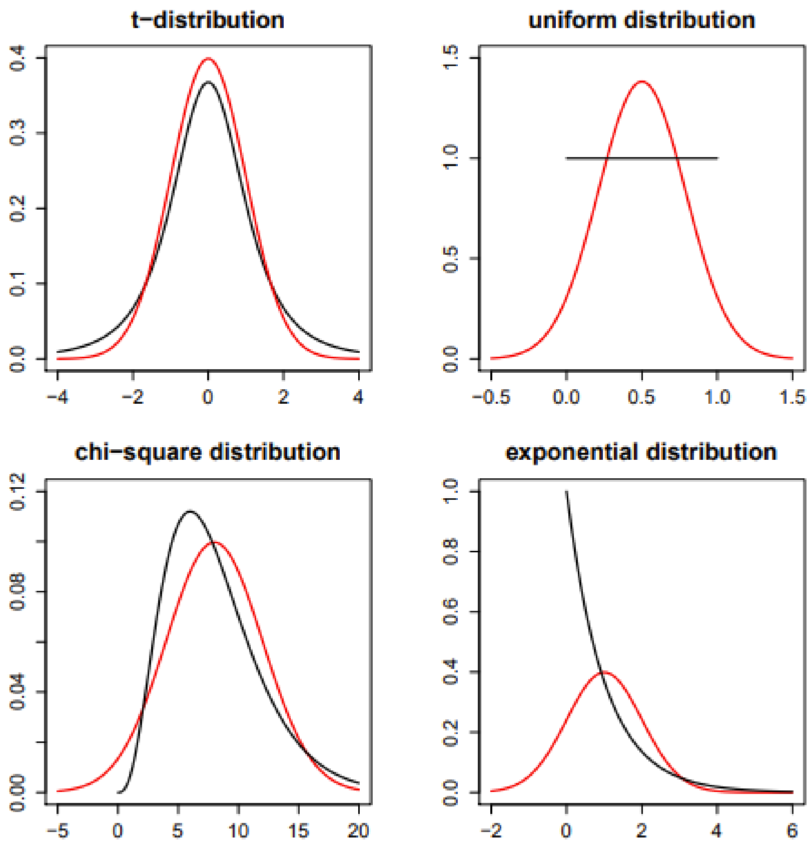
Normally assumption for X is reasonable if

- X is a mean of many independent measurements.(CLT applies)
- The dataset's shape suggests normality:
  - histogram bell-shaped curve
  - Normal QQ-plot approximately straight line

Assessing normality

> **Example of non-normal distributions**

A normal distribution with the same mean and standard deviation is plotted in red.



A model distribution :

> is a (theoretical) probability distribution for describing the **unknown** true population distribution.

`Examples (continous variables)` : normal, uniform, *t*, χ^2, exponential.

If a model distribution is used, we say:

The variable <`...`> is modelled as a random variable
having a <`model distribution`>
with <`relevant parameters`>.

> Example

The variable "Birth date - Due date" is a random variable
having a normal distribution
with mean 0 and standard deviation 10.

Normal QQ - Plot :

QQ-plot = quantile-quantile plot
consider the dataset $x_1, \ldots, x_n$.

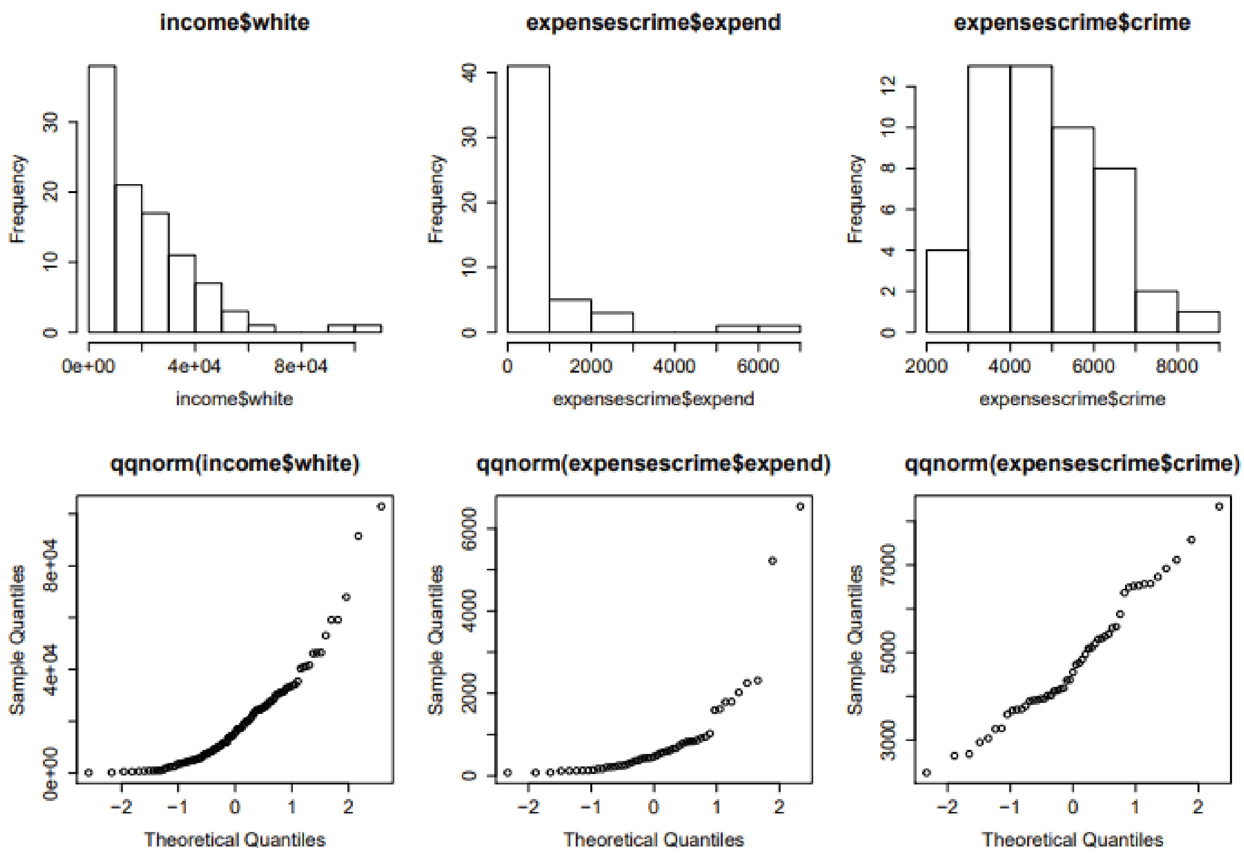
- Ordered values $x_(1), \ldots, x_(n)$ are plotted against theoretical quantiles $z_{a_1}, \ldots, z_{a_n}$ of $N(0,1)$.
- If points approximately follow a straight line, then $N(\mu, \sigma^2)$ is a reasonable model distribution.
- If the straight line is $y = a + bx$, then $\mu \sim a$ (line's intercept) and $\sigma \sim b$ (line's slope).

`Notice` : There are QQ-plots other than "normal QQ-plots", those use theoretical quantiles of other continuous distribution.

> Sample size matters :

Small n : more variation $\Rightarrow$ histogram and QQ-plot can deviate a lot from bell shape and straight line
respectively even if data come from $N(\mu, \sigma^2)$.
Large n : the histogram and QQ-plot are more reliable.

> Example : normal QQ-plots



Left & middle : no straight line at all, obviously not from normal distribution.

Right: approximately straight line *y = 5000 + 1000x*,
so $N(5000, 1000000)$ is a reasonable model distribution.

A location-scale family :

> is a family of probability distributions such that each family member is obtained from another by
>
> - shifting (change in location) and/or
> - stretching/squeezing (change in scale).

In short : by a linear transformation, *Y = a + bX* , for some *a* and *b > 0*.

Normal distributions form a location-scale family.
(If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$)

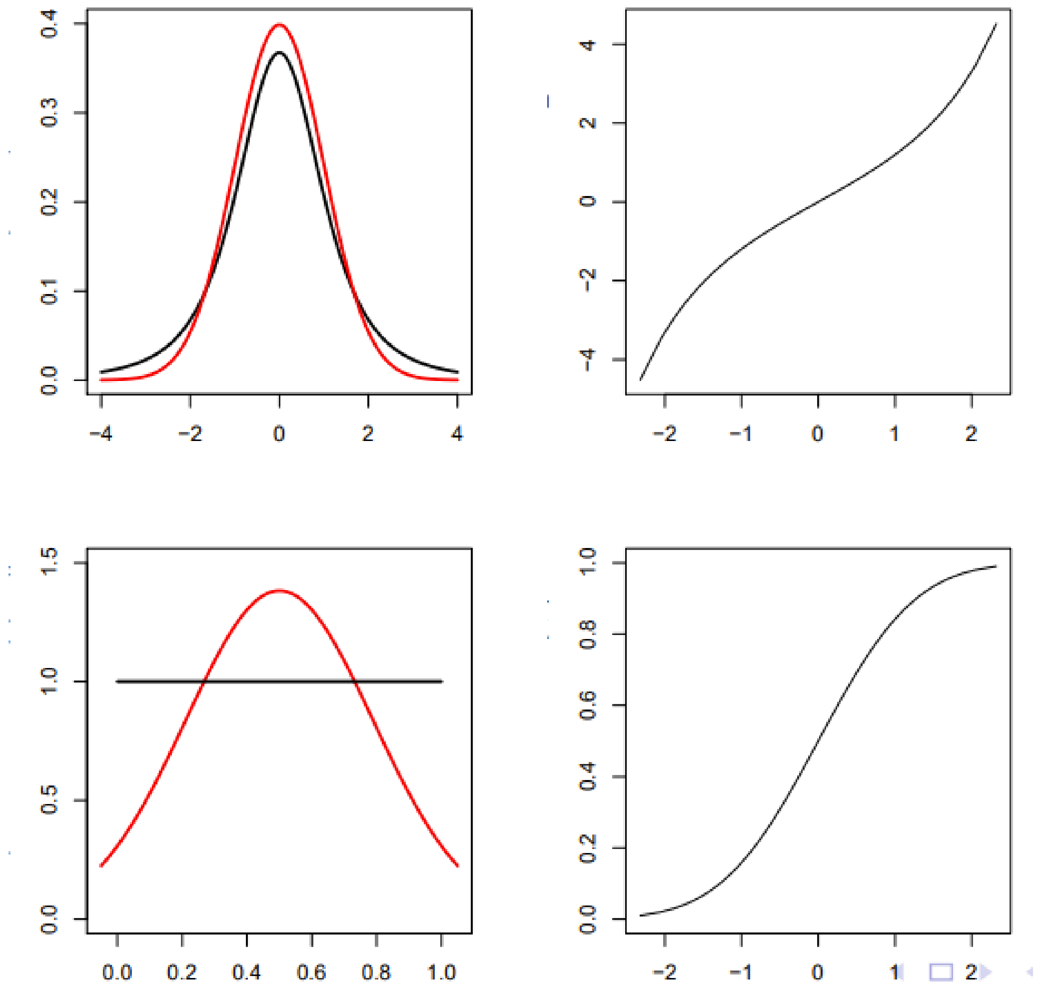
Stochasts(随机) *X* and *Y* have probability distributions that are in the same location-scale family $\Longleftrightarrow$ the QQ-Plot shows a straight line *Y = a + bX*.

> **There are three types of QQ-Plots:**

1. x-axis: theoretical quantiles of a probability distribution.
   y-axis: sample quantiles of a dataset.
   Used to **assess whether the particular distribution could be used as model distribution**.
2. x-axis: theoretical quantiles of a probability distribution.
   y-axis: theoretical quantiles of another probability distribution.
   Used to **compare the shape of two probability distributions**, for instance to verify whether they belong to the same location-scale family.
3. x-axis: sample quantiles of a dataset.
   y-axis: sample quantiles of another dataset.
   Used to **compare the shape** of the two data distributions and assess whether they could possibly **originate from two model distributions belonging to the same location-scale family**.

> Example : theoretical QQ-plots

Top: t-distribution with 3 degrees of freedom,
Bottom: uniform(0,1) distribution,
vs. a normal distribution with the same mean and standard deviation.

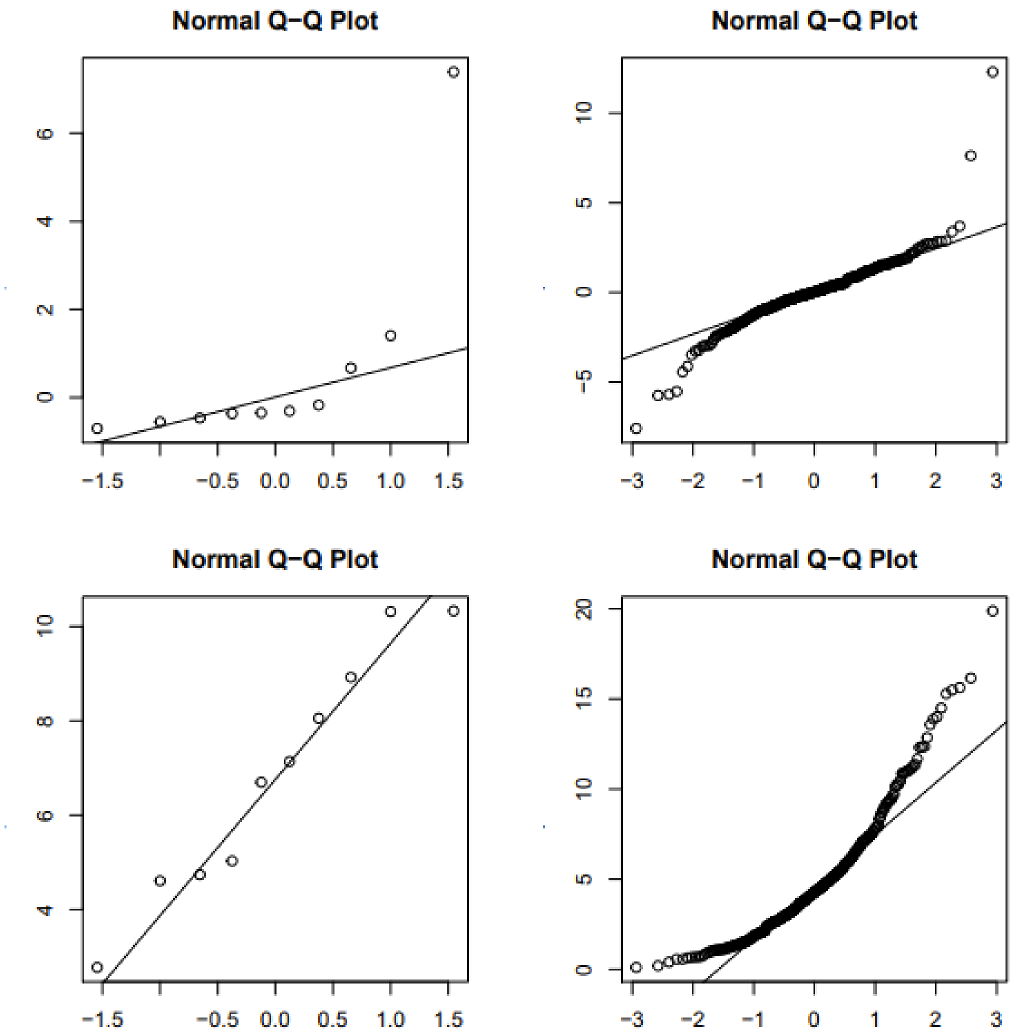


How to interpret QQ-plots

Draw (imaginary) straight line through middle of the QQ-plot.

- Points on left side below straight line?
  $\Rightarrow$ left tail of sample is heavier than left tail of N(0, 1).
- Points on left side above straight line?
  $\Rightarrow$ left tail of N(0, 1) is heavier than left tail of sample.
- Points on right side above straight line?
  $\Rightarrow$ right tail of sample is heavier than right tail of N(0, 1).
- Points on right side below straight line?
  $\Rightarrow$ right tail of N(0, 1) is heavier than right tail of sample.

Example : interpreting normal QQ-plots



How to assess normality of data with QQ-plot

- Make a normal QQ-plot (`qqnorm()`).
- If points follow approximately a straight line $y = a + bx$ (with slope $b > 0$), then $N(a, b^2)$ is reasonable as model distribution.
- If points don't follow a straight line, then the sample is most likely not from a normal distribution.

In latter case: the sample is most likely from a location-scale family with lighter or heavier tails than those of the normal distribution, depending on the shape of the QQ-plot.