Statistical Methods: Lecture 4

Lecture Overview

Standard normal distribution

General normal distributions

Central Limit Theorem

Assessing normality and QQ plots

Continuous Random Variables

Recall from beginning: definition of continuous random variable:

Definition (Continuous random variable)

- uncountably many different values.
- probability distribution given by probability density function;
- probabilities computed by area under this function.

Total area: 1.

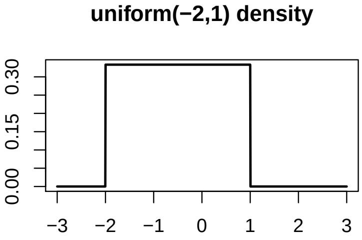
Let X be a continuous random variable.
$P(X = x) = 0$...

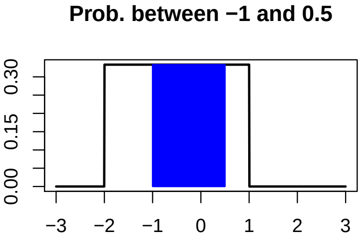
Instead, consider $P(\text{values of } X \text{ lie in } I)$ (I some interval)
$=$ area under probability density function restricted to I.

Continuous Random Variables

Example: choose point in interval

Random variable X: random point between -2 and 1.
Probability distribution of X? Density function $p(x) = \frac{1}{3}$ for $x \in [-2, 1]$.





$P(X \in [-1, \frac{1}{2}]) = \text{blue area} = (\frac{1}{2} - (-1)) \cdot \frac{1}{3} = \frac{3}{2} \cdot \frac{1}{3} = \frac{1}{2}$.

5.2 Probability density function

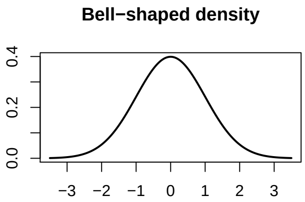
Definition (probability density function)

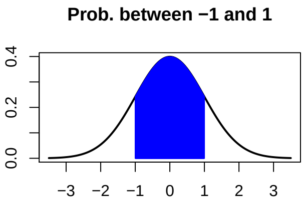
Probability density function is a function $p(x)$ such that

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

$P(X \in [a, b])$ = area under the curve $p(x)$ between a and b.

Example of a bell-shaped density





Here: $P(X \in [-1, 1]) \approx 0.68$.

5.2 The standard normal distribution

Definition (normal distribution)

Random variable X has a normal distribution with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$ if its density is

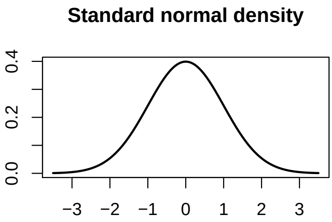
$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

Notice that $p(x)$ is continuous, bell-shaped and symmetric, $E(X) = \mu$, $Var(X) = \sigma^2$.

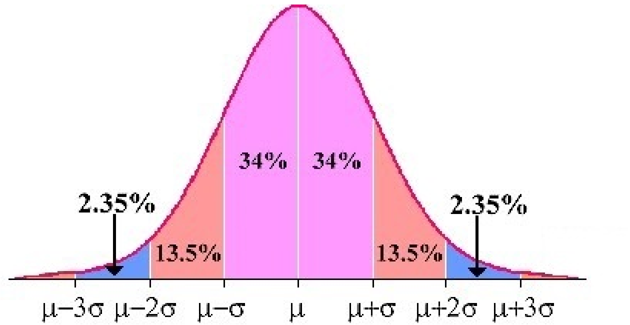
Notation: $N(\mu, \sigma^2)$ for normal distribution with mean μ, variance σ^2.

Standard normal distribution: $N(0, 1)$.

In R: `dnorm(x,mean=0,sd=1)`.



5.2 The standard normal distribution



- 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 $\mu + 3\sigma$

5.2 The standard normal distribution

Determine probabilities of normal distribution

$$\begin{aligned} P(X \leq z) &= \text{area under density to the left of } z \\ P(X \in [a, b]) &= P(X \leq b) - P(X \leq a) \\ P(X \geq b) &= 1 - P(X \leq b) \end{aligned}$$

- In R: `pnorm(x)` computes the probability $P(X \leq x)$ for $X \sim N(0, 1)$.
- In case of $N(0, 1)$: Table 2 of book (p. 786-787); shows cumulative area under density to the left of z, i.e., the probability $P(X \leq x)$ for $X \sim N(0, 1)$.
- For $N(\mu, \sigma^2)$ we can compute probabilities by using $N(0, 1)$ (later).

5.2 The standard normal distribution

Example: Probabilities of standard normal distribution

Let $X \sim N(0,1)$.

1. $P(X \leq 0.6) = ?$
2. $P(X \geq -1.45) = ?$
3. $P(X \in [-1.45, 0.6]) = P(-1.45 \leq X \leq 0.6) = ?$

1. Use Table 2: cumulative area to the left of 0.6 is $0.7257 = P(X \leq 0.6)$.
2. $P(X \geq -1.45) = 1 - P(X \leq -1.45)$.
   Table 2 with $z = -1.45$: $P(X \leq -1.45) = 0.0735$
   Hence, $P(X \geq -1.45) = 1 - 0.0735 = 0.9265$.
3. $P(-1.45 \leq X \leq 0.6) = P(X \leq 0.6) - P(X \leq -1.45) = 0.7057 - 0.0735 = 0.6322$.

5.2 The standard normal distribution: recap

Probability density function

A probability density is a function $p(x)$ such that $p(x) \geq 0$ and the total area under the curve is 1.
The probability that X takes a value between a and b, i.e. $P(X \in [a, b])$, can be obtained by determining the area under the curve $p(x)$ between a and b.

Definition (Normal distribution)

A random variable X has a normal distribution if its probability density $p(x)$ is continuous, bell-shaped and symmetric.
Notation: $N(\mu, \sigma^2)$ for a normal distribution with mean μ and variance σ^2.
The standard normal distribution has mean 0 and standard deviation 1: $N(0, 1)$.

Determine probabilities of standard normal distribution

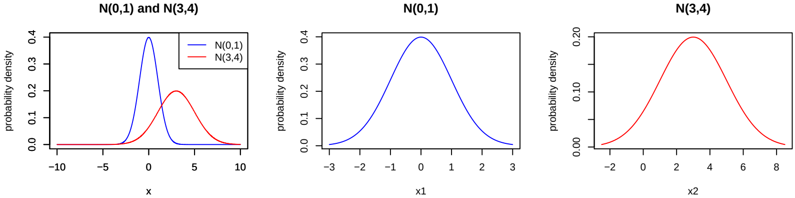
Let X has $N(0, 1)$ distribution.
R: `pnorm(x)` gives probability $P(X \leq x)$.
Manual: use Table 2 of book (p. 786-787), which shows the cumulative area under the curve to the left of a z-score, $P(X \leq z)$.

5.3 Applications of normal distributions

Relating $N(\mu, \sigma^2)$ to $N(0, 1)$

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

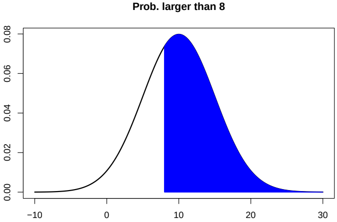
Example: Normal distributions have similar shape



5.3 Applications of normal distributions

Example: $X \sim N(10, 25)$.

So, $\mu = 10, \sigma = 5$. What is $P(X \geq 8)$?



Since $Z = \frac{X-10}{5} \sim N(0,1)$, $P(X \geq 8) = P\left(\frac{X-10}{5} \geq \frac{8-10}{5}\right) = P(Z \geq -0.4)$.
Table 2: 0.3446 of the area is to the left of -0.4, so $P(X \geq 8) = 1 - 0.3446 = 0.6554$.

5.3 Applications of normal distributions

Definition: 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.

Let $X \sim N(\mu, \sigma^2)$. Since $P(X \leq x) = P(Z \leq z)$, where $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$, use Table 2!

Example: X = random SAT test score of math section

approximately $N(500, 10000)$-distributed.
Probability that random participant scores between 550 and 700?

Compute z scores of 550 and 700:

$$x = 550 \rightarrow z = \frac{550 - 500}{100} = 0.5, \quad x = 700 \rightarrow z = \frac{700 - 500}{100} = 2.0.$$

Table 2: 0.6915 of the area is to the left of $z = 0.5$ and 0.9772 is to the left of $z = 2.0$.
Hence, $P(550 \leq X \leq 700) = 0.9772 - 0.6915 = 0.2858$.

5.5 The Central Limit Theorem

Rolling (fair) dice n times

- Let X_i = outcome of i-th roll.
- Behaviour of mean $\overline{X}_n = \frac{1}{n}(X_1 + \ldots + X_n)$ for large n?
- Recall Law of Large Numbers (LLN): $\overline{X}_n = \frac{1}{n}(X_1 + \ldots + X_n)$ approximates $E(X_1) = 3.5$.
- $\overline{X}_n$ is a random variable; which probability distribution?
- For fixed n: determine possible values x of $\overline{X}_n$ and probabilities $P(\overline{X}_n = x)$.
- Doable for $n = 1$ and $n = 2$, but practically impossible for larger n.
- Solution: Central Limit Theorem.

5.5 The Central Limit Theorem

The Central Limit Theorem (CLT)

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

NB: the population can have any (non-degenerate) distribution.

The Central Limit Theorem (CLT) for normal population (special case)

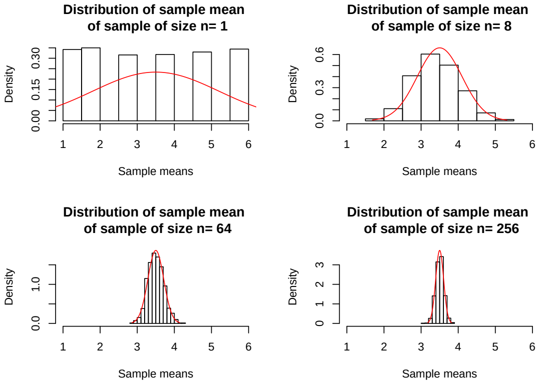
Take a sample of size n from a normal population with mean μ and standard deviation σ. Then $\overline{X}_n \sim N(\mu, \frac{\sigma^2}{n})$.

NB: n can be any number.

5.5 The Central Limit Theorem

Example: CLT for sample mean of dice throws

Histograms: distribution of 1000 sample means of 1, 8, 64, and 256 die rolls.
Red line: normal distribution according to CLT, i.e., $N(3.5, 2.92/n)$



5.5 The Central Limit Theorem

Example application of CLT

SAT test scores of math section approximately $N(500, 10000)$-distributed.

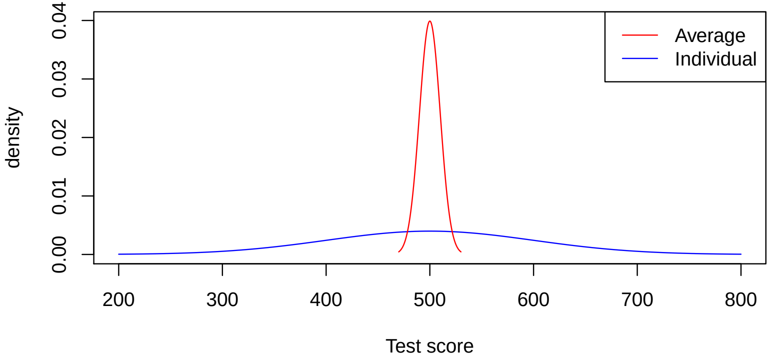
1. Alice scores 475. What percentage of students performs better?
2. A school of 100 students has an average score of 475.
   What percentage of schools performs better?

1. z score of $x = 475$ is: $\frac{475-500}{100} = -0.25$.
   Table 2: $P(Z > -0.25) = 1 - P(Z \leq -0.25) = 1 - 0.4013 = 0.5987$,
   so ca. 60% of students performs better.
2. CLT applies ($n > 30$).
   Distribution of mean SAT score of a school of 100 students is approx. $N(500, \frac{10000}{100})$,
   so $\mu = 500$ and $\sigma = \frac{100}{\sqrt{100}} = 10$.
   Hence, z score of $x = 475$: $\frac{475-500}{10} = -2.5$.
   Table 2: $P(\overline{X}_n > 475) = P(Z > -2.5) = 1 - 0.0062 = 0.9938$,
   so 99.38% of comparable schools (i.e. of 100 students) perform better.

5.5 The Central Limit Theorem

Example: application of CLT

Difference in distributions: individual vs. average scores.



5.5 The Central Limit Theorem: recap

Sample mean normally distributed?

Consider a population distribution with mean μ and st. 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.

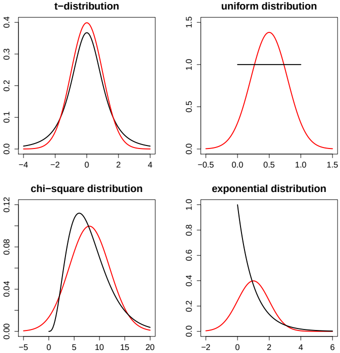
Normality assumption for X reasonable if

- X is a mean of many independent measurements. (CLT applies.)
- Dataset shape suggests normality:
  histogram bell-shaped curve?
  Normal Q-Q plot approximately straight line? Treated later.

5.6 Assessing normality

More extensive than in the book.

Examples of distributions different than normal



Normal distribution with same mean and standard deviation as distribution in black.

5.6 Assessing normality

More extensive than in the book.

Definition (model distribution)

Theoretical probability distribution for describing the unknown true population distribution.

Examples (continuous variables): normal, uniform, t, χ^2, exponential.

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

Example: The variable 'Date of birth - Due date' is a random variable
having a normal distribution
with mean 0 and standard deviation 10.

5.6 Assessing normality

Assessing normality

Consider dataset $x_1, \dots, x_n$. When is model distribution $N(\mu, \sigma^2)$ reasonable?

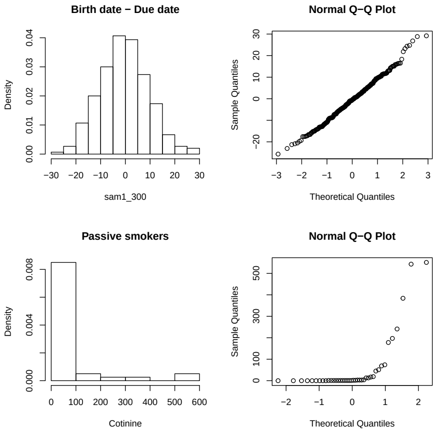
- Shape of histogram
  Strong deviation from bell shape? Then $N(\mu, \sigma^2)$ unlikely.
- Boxplot (not always helpful).
- Normal QQ plot.

5.6 Assessing normality

Top: Birth date - due date.
Bottom: cotinine passive smokers.

Top right: approx. straight line $y = 10x$, so $N(0, 100)$ reasonable model distribution.

Bottom right: no straight line at all, so obviously not from normal distribution.



5.6 Assessing normality

What is a Normal QQ plot?

Consider dataset $x_1, \ldots, x_n$.

- ordered values $x_{(1)}, .., x_{(n)}$ plotted vs. theoretical quantiles $z_{a_1}, .., z_{a_n}$ of $N(0,1)$. Here, z_{a_i} is the z-score with $\frac{2i-1}{2n}$ ($\approx \frac{i}{n}$) of the N(0,1) area to the left.
- If points follow approx. straight line, then $N(\mu, \sigma^2)$ possible model distribution.
- If straight line $y = a + bx$, then $\mu \approx a$ (line's intercept) and $\sigma \approx b$ (line's slope).
- In R: `qqnorm()`

What is a QQ plot?

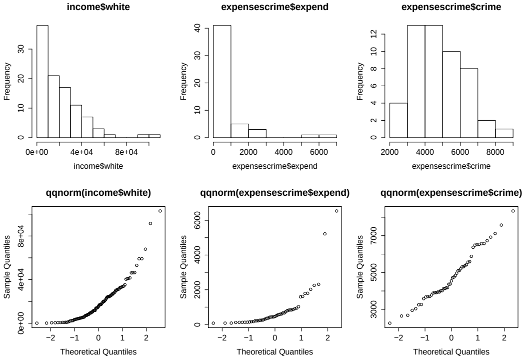
There are QQ plots other than "normal QQ plots":
use theoretical quantiles of other continuous distributions.

Sample size

Small n: more variation $\Rightarrow$ histogram / QQ plot could deviate (from bell shape / straight line), even if $N(\mu, \sigma^2)$ true. Large n: histogram and QQ plot: more reliable.

5.6 Assessing normality

Example: normal QQ plots



Left and middle: no straight line at all, obviously not from normal distribution. Right: approx. straight line $y = 5000 + 1000x$, $N(5000, 10^6)$ reasonable model distribution.

5.6 Assessing normality

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

Definition: a location-scale family of probability distributions

Each member obtained from another by

- $\triangleright$ shifting (change in location) and/or
- $\triangleright$ stretching/squeezing (change in scale).

Random variables X and Y have probability distributions that are in the same location-scale family $\iff$ the QQ-plot shows a straight line $Y = a + bX$.

5.6 Assessing normality

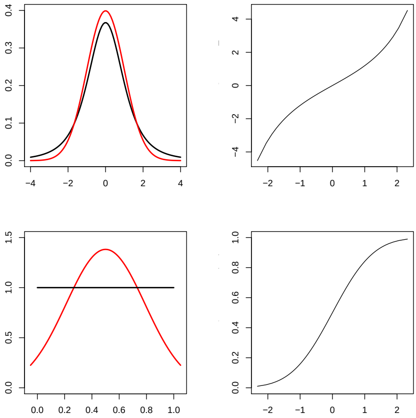
Three types of QQ-plots

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.

5.6 Assessing normality

Example: theoretical QQ plots

Top: t-distribution with 3 degrees of freedom
Bottom: uniform(0,1) distribution.



5.6 Assessing normality

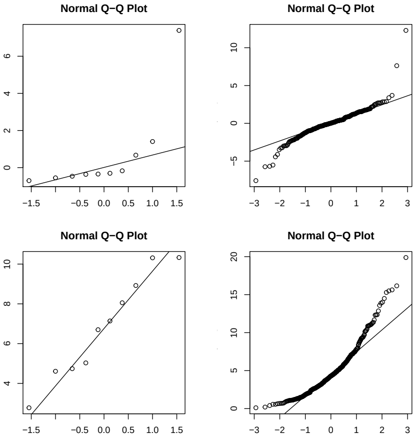
How to interpret QQ plots

Draw (imaginary) straight line through middle of 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.

5.6 Assessing normality

Example: interpreting normal QQ plots



qqline(): straight line through first and third quartiles.

Which tails of which distributions are heavier?

5.6 Assessing normality

How to assess normality of data with QQ plot

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

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