Mathematical Statistics

MAS 713

Chpater 3.3

Previous lectures

1. Random variables
2. Discrete Random Variables
3. Continuous Random Variables
4. Expectation of a random variable
5. Variance of a random variable
6. Special random variables:
    1. Binomial random variables
    2. Hypergeometric random variables
    3. Poisson random variables
    4. Uniform random variables
    5. Exponential random variables

Questions?

This lecture

- 3.3.1 Normal random variables
- 3.3.2 Checking if the data are normally distributed

Additional reading : Chapter 3 in the textbook

The Normal distribution : introduction

- Last subchapter we introduced several useful continuous probability distributions

- However, the most widely used, and therefore the most important, continuous distribution is undoubtedly the

Normal distribution

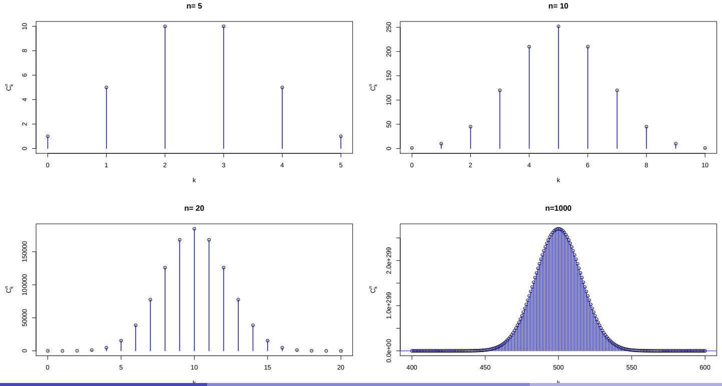
- Its prevalence was first highlighted when it was observed that in many natural processes, random variation among individuals systematically conforms to a particular pattern :

- most of the observations concentrate around one single value (which is the mean)
- the number of observations smoothly decreases, symmetrically on either side, with the deviation from the mean
- it is very unlikely, yet not impossible, to find very extreme values

$\rightsquigarrow$ this yields the famous **bell-shaped** curve

The Normal distribution : introduction

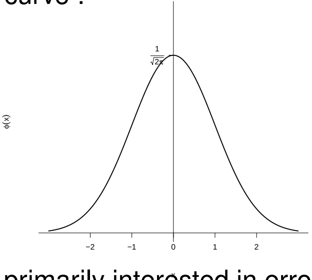
- The bell-shaped curve was first spotted by the French mathematician **Abraham de Moivre** (1667-1754) who in his 1738 book “The Doctrine of Chances” showed that the coefficients $C_k^n = \binom{n}{k}$ in the binomial expansion of $(a+b)^n$ precisely follow the bell shape pattern when n is large



The Normal distribution : introduction

- Later, **Carl-Friedrich Gauss** (1777-1855), a German mathematician (sometimes referred to as the *Princeps mathematicorum*, latin for "the Prince of Mathematicians" or "the foremost of mathematicians") , was the first to write an explicit equation for the bell-shaped curve :

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$



- When deriving his distribution, Gauss was primarily interested in errors of measurement, whose distribution typically follows the bell-shaped curve as well. He called his curve the “normal curve of errors”, which was to become the Normal distribution.
- In honour of Gauss, the Normal distribution is also referred to as the **Gaussian distribution**

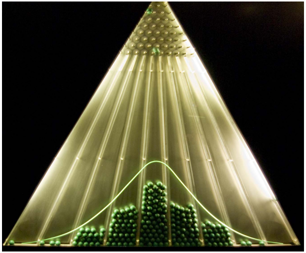
The Normal distribution : introduction

It is important to note that the Normal distribution is not just a convenient mathematical tool, but also occurs in natural phenomena

- For instance, in 1866 **James Maxwell** (1831-1879), a Scottish physicist, determined the distribution of molecular velocity in a gas at equilibrium. As a result of collisions with other molecules, molecular velocity in a given direction is randomly distributed, and from basic assumptions, that distribution can be shown to be the Normal distribution

The Normal distribution : introduction

- Another famous example is the “bean machine”, invented by **Sir Francis Galton** (English scientist, 1822-1911) to demonstrate the Normal distribution. The machine consists of a vertical board with interleaved rows of pins. Balls are dropped from the top, and bounce left and right as they hit the pins. Eventually, they are collected into bins at the bottom. The height of ball columns in the bins approximately follows the bell-shaped curve



The Normal distribution

- A random variable is said to be normally distributed with parameters μ and σ ($\sigma > 0$), i.e.

$$X \sim \mathcal{N}(\mu, \sigma),$$

if its probability density function is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (\rightsquigarrow S_X = \mathbb{R})$$

- Unfortunately, no closed form exists for

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy$$

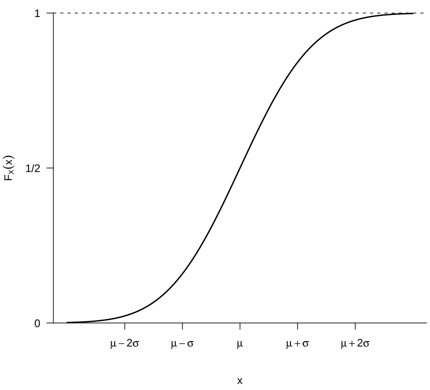
The Normal distribution

Important remark : Be careful! Some (many?) sources use the alternative notation

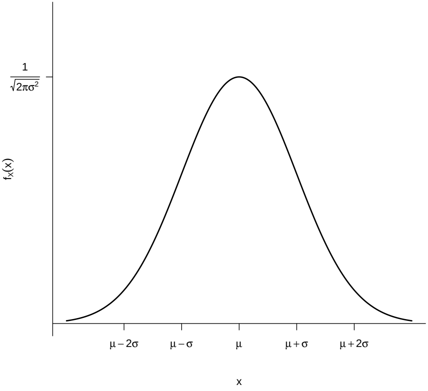
$$X \sim \mathcal{N}(\mu, \sigma^2)$$

$\rightsquigarrow$ in the textbook and in Matlab, the notation $\mathcal{N}(\mu, \sigma)$ is used, so we adopt it in these slides as well

The Normal distribution



cdf $F(x)$



pdf $f(x) = F'(x)$

The Standard Normal distribution

Standard Normal distribution is the Normal distribution with

- $\mu = 0$,
- $\sigma = 1$.

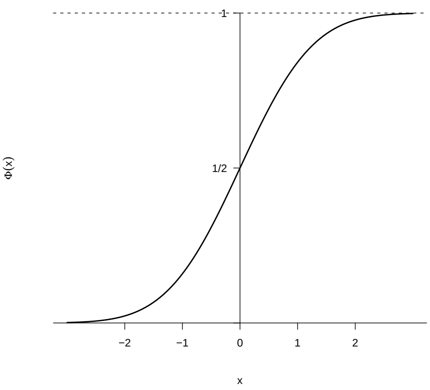
- This yields

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

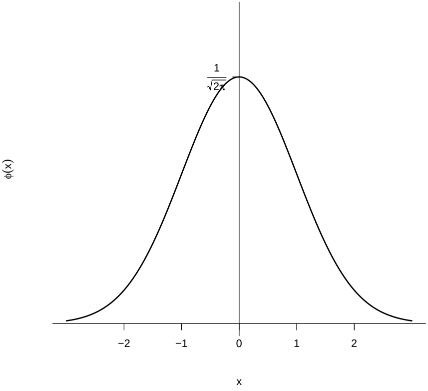
- Usually, in this situation, the following particular notation is used:

$$f(x) \doteq \phi(x) \qquad \text{and} \qquad F(x) \doteq \Phi(x)$$

The Standard Normal distribution



cdf $\Phi(x)$



pdf $\phi(x) = \Phi'(x)$

Normal distribution : properties

- It can be shown that, for any μ and σ,

$$\int_{S_X} f(x)\,dx = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx = 1$$

- Similarly, we can find

$$\mathbb{E}(X) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} x\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx = \mu$$

- and

$$\mathbb{V}\mathrm{ar}(X) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} (x-\mu)^2\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx = \sigma^2$$

Normal distribution : properties

Mean and variance of the Normal distribution

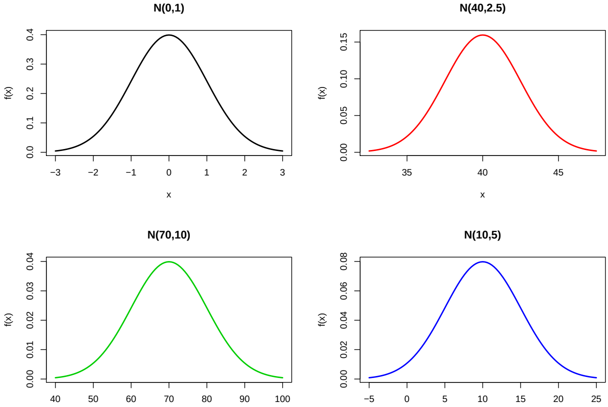
If $X \sim \mathcal{N}(\mu, \sigma)$,

$$\mathbb{E}(X) = \mu \qquad \text{and} \qquad \mathbb{V}\text{ar}(X) = \sigma^2 \qquad (\rightsquigarrow \text{sd}(X) = \sigma)$$

Normal distribution : properties

An important observation is that all normal probability distribution functions have the same bell shape

They only differ in where they are centred (at μ) and in their spread (quantified by σ)



- Let $\mathrm{i} := \sqrt{-1} \in \mathbb{C}$ be the imaginary unit ($\mathrm{i}^2 = -1$)

Definition: **Characteristic Function**

The characteristic function $\varphi_X : \mathbb{R} \to \mathbb{C}$ is defined by

$$\varphi_X(t) := \mathbb{E}[e^{\mathrm{i}t\mathrm{X}}]$$

Theorem:

The characteristic function completely determines the distribution, i.e.

$$X \sim F \quad \Longleftrightarrow \quad \varphi_X(t) = \varphi_F(t) \quad \forall t \in \mathbb{R}.$$

Property:

Let X_1 and X_2 be independent random variables. Then

$$\varphi_{X_1+X_2}(t) = \varphi_{X_1}(t)\varphi_{X_2}(t) \quad \forall t \in \mathbb{R}.$$

Careful:

$\varphi_{X_1+X_2}(t) = \varphi_{X_1}(t)\varphi_{X_2}(t) \quad \forall t \in \mathbb{R}$ does NOT imply X_1, X_2 independent

Characteristic for various distributions

- $X \sim \mathcal{N}(\mu, \sigma) \iff \varphi_X(t) = e^{\mathrm{i}\mu\mathrm{t} - \frac{1}{2}\sigma^2\mathrm{t}^2} \quad \forall t \in \mathbb{R}$.
- $X \sim \mathrm{Ber}(\pi) \iff \varphi_X(t) = (\pi e^{\mathrm{it}} + 1 - \pi) \quad \forall t \in \mathbb{R}$.
- $X \sim \mathrm{Bin}(n, \pi) \iff \varphi_X(t) = (\pi e^{\mathrm{it}} + 1 - \pi)^n \quad \forall t \in \mathbb{R}$.
- $X \sim \mathcal{P}(\lambda) \iff \varphi_X(t) = e^{\lambda(e^{\mathrm{it}} - 1)} \quad \forall t \in \mathbb{R}$.
- $X \sim U_{[\alpha,\beta]} \iff \varphi_X(t) = \frac{e^{\mathrm{i}\beta\mathrm{t}} - e^{\mathrm{i}\alpha\mathrm{t}}}{\mathrm{i}(\beta-\alpha)\mathrm{t}} \quad \forall t \in \mathbb{R} \setminus \{0\}$.
- $X \sim \mathrm{Exp}(\lambda) \iff \varphi_X(t) = \frac{\lambda}{\lambda - \mathrm{it}} \quad \forall t \in \mathbb{R}$.

Normal distribution : standardisation

From the property of the characteristic function
(or alternatively from the expression and the shape of the Normal pdf)
one sees that:

- If $X \sim \mathcal{N}(\mu, \sigma)$, then $Y = aX + b \sim \mathcal{N}(a\mu + b, a\sigma)$

i.e. Y is normally distributed with
mean $\mathbb{E}(Y) = a\mu + b$ and variance $\mathbb{V}\text{ar}(Y) = a^2\sigma^2$

Normal distribution : standardisation

The following result directly follows from the foregoing :

Property: Standardisation

If $X \sim \mathcal{N}(\mu, \sigma)$, then

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

- This linear transformation is called the **standardisation** of the normal random variable X, as it transforms X into a standard normal random variable Z
- For an observed value x of X, the corresponding standardised value $z = \frac{x-\mu}{\sigma}$ is often called z-score
- Standardisation will play a paramount role in the sequel

Normal distribution : properties

- This extremely important fact allows us to deduce any required information for a given Normal distribution $\mathcal{N}(\mu, \sigma)$ from the features of the 'simple' standard normal distribution

- For instance, for the standard pdf $\phi(x)$, it can be found that

$$\int_{-1}^{1} \phi(x)\,dx = \mathbb{P}(-1 < Z < 1) \simeq 0.6827$$
$$\int_{-2}^{2} \phi(x)\,dx = \mathbb{P}(-2 < Z < 2) \simeq 0.9545$$
$$\int_{-3}^{3} \phi(x)\,dx = \mathbb{P}(-3 < Z < 3) \simeq 0.9973$$

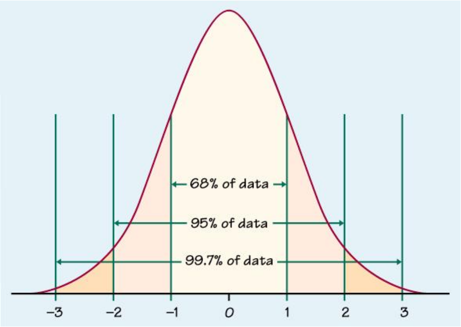
- This automatically translates to the general case $X \sim \mathcal{N}(\mu, \sigma)$:

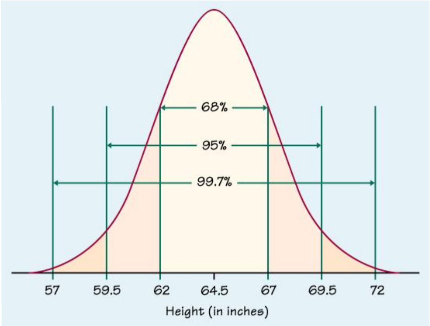
$$\mathbb{P}(\mu - \sigma < X < \mu + \sigma) \simeq 0.6827$$
$$\mathbb{P}(\mu - 2\sigma < X < \mu + 2\sigma) \simeq 0.9545$$
$$\mathbb{P}(\mu - 3\sigma < X < \mu + 3\sigma) \simeq 0.9973$$

- This is known as the **68-95-99 rule** for normal distributions

Normal distribution : properties

For instance, suppose we are told that women's heights in a given population follow a normal distribution with mean $\mu = 64.5$ inches and $\sigma = 2.5$ inches





$\rightsquigarrow$ we expect 68.27 % of women to be between $\mu - \sigma = 64.5 - 2.5 = 62$ inches and $\mu + \sigma = 64.5 + 2.5 = 67$ inches

Normal distribution : remark

- Theoretically, the domain of variation S_X of a normally distributed random variable X is $\mathbb{R} = (-\infty, +\infty)$
- However, there is 99.7% chance to find X between $\mu - 3\sigma$ and $\mu + 3\sigma$
- It almost impossible to find X outside that interval, and virtually impossible to find it very far away from μ

$\rightsquigarrow$ there is in general no problem in modelling the distribution of a positive quantity with a Normal distribution, provided μ is large compared to σ

- This also explains why 6σ is sometimes called the width of the normal distribution

Normal distribution : examples

Example

Suppose that $Z \sim \mathcal{N}(0,1)$. What is $\mathbb{P}(Z \leq 1.25)$?

In principle, this should be directly given by

$$\mathbb{P}(Z \leq 1.25) = \Phi(1.25) = \int_{-\infty}^{1.25} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$$

However, we know that this integral cannot be evaluated analytically

$\rightsquigarrow$ we must use a software (command `normcdf` in Matlab) or the

Standard Normal table

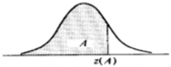
The table gives the 'area under the standard normal curve to the left of z', that is

$$\mathbb{P}(Z \leq z)$$

Here we have just to read : $\mathbb{P}(Z \leq 1.25) \overset{\text{table}}{=} 0.8944$

Standard Normal table

Entry is area A under the standard normal curve from $-\infty$ to $z(A)$



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Normal distribution : examples

Any other kind of probabilities must be written in terms of $\mathbb{P}(Z \leq z)$ and calculated from the value found in the table

Example

Suppose that $Z \sim \mathcal{N}(0, 1)$. What is $\mathbb{P}(Z < 1.25)$? What is $\mathbb{P}(Z > 1.25)$? What is $\mathbb{P}(-0.38 \leq Z < 1.25)$?

- as Z is a continuous random variable, $\mathbb{P}(Z < z) = \mathbb{P}(Z \leq z)$ for any z
  $\leadsto \mathbb{P}(Z < 1.25) = \mathbb{P}(Z \leq 1.25) \overset{\text{table}}{=} 0.8944$
- $\mathbb{P}(Z > 1.25) = 1 - \mathbb{P}(Z \leq 1.25) \overset{\text{table}}{=} 1 - 0.8944 = 0.1056$
- $$\begin{aligned} \mathbb{P}(-0.38 \leq Z < 1.25) &= \mathbb{P}(Z < 1.25) - \mathbb{P}(Z < -0.38) \\ &= \mathbb{P}(Z \leq 1.25) - \mathbb{P}(Z \leq -0.38) \\ &\overset{\text{table}}{=} 0.8944 - 0.3520 \\ &= 0.5424 \end{aligned}$$

Normal distribution : examples

Any other kind of probabilities must be written in terms of $\mathbb{P}(Z \leq z)$ and calculated from the value found in the table

Example

Suppose that $Z \sim \mathcal{N}(0,1)$. What is $\mathbb{P}(Z < 1.25)$? What is $\mathbb{P}(Z > 1.25)$? What is $\mathbb{P}(-0.38 \leq Z < 1.25)$?

- as Z is a continuous random variable, $\mathbb{P}(Z < z) = \mathbb{P}(Z \leq z)$ for any z
  $\leadsto \mathbb{P}(Z < 1.25) = \mathbb{P}(Z \leq 1.25) \overset{\text{table}}{=} 0.8944$
- $\mathbb{P}(Z > 1.25) = 1 - \mathbb{P}(Z \leq 1.25) \overset{\text{table}}{=} 1 - 0.8944 = 0.1056$
- $$\begin{aligned}\mathbb{P}(-0.38 \leq Z < 1.25) &= \mathbb{P}(Z < 1.25) - \mathbb{P}(Z < -0.38) \\ &= \mathbb{P}(Z \leq 1.25) - \mathbb{P}(Z \leq -0.38) \\ &\overset{\text{table}}{=} 0.8944 - 0.3520 \\ &= 0.5424\end{aligned}$$

Normal distribution : examples

Example

The time it takes a driver to react to the brake light on a decelerating vehicle can be modeled with a Normal distribution having parameters $\mu = 1.25$ sec and $\sigma = 0.46$ sec.

In the long run, what proportion of reaction times will be between 1 and 1.75?

- We have $X \sim \mathcal{N}(1.25, 0.46)$ and we desire $\mathbb{P}(1 \leq X \leq 1.75)$.

$$\mathbb{P}(1 \leq X \leq 1.75) = \mathbb{P}(X \leq 1.75) - \mathbb{P}(X \leq 1)$$

but 'unfortunately' we do not have the $\mathcal{N}(1.25, 0.46)$ table

- However, we know $Z = \frac{X-1.25}{0.46} \sim \mathcal{N}(0,1)$ and we do have the $\mathcal{N}(0.1)$ table!

$$\mathbb{P}(X \leq 1.75) = \mathbb{P}\left(\frac{X-1.25}{0.46} \leq \frac{1.75-1.25}{0.46}\right) = \mathbb{P}(Z \leq 1.09) \overset{\text{table}}{=} 0.8621$$

- Similarly, $\mathbb{P}(X \leq 1) = \mathbb{P}(Z \leq -0.54) \overset{\text{table}}{=} 0.2946$, so that

$$\mathbb{P}(1 \leq X \leq 1.75) = 0.8621 - 0.2946 = 0.5675$$

Normal distribution : examples

Example

The time it takes a driver to react to the brake light on a decelerating vehicle can be modeled with a Normal distribution having parameters $\mu = 1.25$ sec and $\sigma = 0.46$ sec.

In the long run, what proportion of reaction times will be between 1 and 1.75?

- We have $X \sim \mathcal{N}(1.25, 0.46)$ and we desire $\mathbb{P}(1 \leq X \leq 1.75)$.

$$\mathbb{P}(1 \leq X \leq 1.75) = \mathbb{P}(X \leq 1.75) - \mathbb{P}(X \leq 1)$$

but ‘unfortunately’ we do not have the $\mathcal{N}(1.25, 0.46)$ table

- However, we know $Z = \frac{X-1.25}{0.46} \sim \mathcal{N}(0, 1)$ and we do have the $\mathcal{N}(0.1)$ table!

$$\mathbb{P}(X \leq 1.75) = \mathbb{P}\left(\frac{X-1.25}{0.46} \leq \frac{1.75-1.25}{0.46}\right) = \mathbb{P}(Z \leq 1.09) \stackrel{\text{table}}{=} 0.8621$$

- Similarly, $\mathbb{P}(X \leq 1) = \mathbb{P}(Z \leq -0.54) \stackrel{\text{table}}{=} 0.2946,$ so that

$$\mathbb{P}(1 \leq X \leq 1.75) = 0.8621 - 0.2946 = 0.5675$$

Normal distribution : examples

Example

The time it takes a driver to react to the brake light on a decelerating vehicle can be modeled with a Normal distribution having parameters $\mu = 1.25$ sec and $\sigma = 0.46$ sec.

In the long run, what proportion of reaction times will be between 1 and 1.75?

- We have $X \sim \mathcal{N}(1.25, 0.46)$ and we desire $\mathbb{P}(1 \leq X \leq 1.75)$.

$$\mathbb{P}(1 \leq X \leq 1.75) = \mathbb{P}(X \leq 1.75) - \mathbb{P}(X \leq 1)$$

but 'unfortunately' we do not have the $\mathcal{N}(1.25, 0.46)$ table

- However, we know $Z = \frac{X-1.25}{0.46} \sim \mathcal{N}(0, 1)$ and we do have the $\mathcal{N}(0.1)$ table!

$$\mathbb{P}(X \leq 1.75) = \mathbb{P}\left(\frac{X-1.25}{0.46} \leq \frac{1.75-1.25}{0.46}\right) = \mathbb{P}(Z \leq 1.09) \overset{\text{table}}{=} 0.8621$$

- Similarly, $\mathbb{P}(X \leq 1) = \mathbb{P}(Z \leq -0.54) \overset{\text{table}}{=} 0.2946$, so that

$$\mathbb{P}(1 \leq X \leq 1.75) = 0.8621 - 0.2946 = 0.5675$$

Normal distribution : examples

Example

The actual amount of instant coffee that a filling machine puts into a "4-ounce" jars may be looked upon as a random variable having a normal distribution with $\sigma = 0.04$ ounce.

If only 2% of the jars are to contain less than 4 ounces, what should be the mean fill of these jars?

- Let X denote the actual amount of coffee put into the jar by the machine
- We have $X \sim \mathcal{N}(\mu, 0.04)$, with μ such that $\mathbb{P}(X \leq 4) = 0.02$

Hence,

$$0.02 = \mathbb{P}(X \leq 4) = \mathbb{P}\left(\frac{X-\mu}{0.04} \leq \frac{4-\mu}{0.04}\right) = \mathbb{P}\left(Z \leq \frac{4-\mu}{0.04}\right)$$

- In the standard normal table, we can find that $\mathbb{P}(Z \leq -2.05) \simeq 0.02$
- We conclude that $\frac{4-\mu}{0.04} = -2.05$, that is,

$$\mu = 4 + 0.04 \times 2.05 = 4.082 \text{ ounces}$$

Normal distribution : examples

Example

The actual amount of instant coffee that a filling machine puts into a "4-ounce" jars may be looked upon as a random variable having a normal distribution with $\sigma = 0.04$ ounce.

If only 2% of the jars are to contain less than 4 ounces, what should be the mean fill of these jars?

- Let X denote the actual amount of coffee put into the jar by the machine
- We have $X \sim \mathcal{N}(\mu, 0.04)$, with μ such that $\mathbb{P}(X \leq 4) = 0.02$

Hence,

$$0.02 = \mathbb{P}(X \leq 4) = \mathbb{P}\left(\frac{X-\mu}{0.04} \leq \frac{4-\mu}{0.04}\right) = \mathbb{P}\left(Z \leq \frac{4-\mu}{0.04}\right)$$

- In the standard normal table, we can find that $\mathbb{P}(Z \leq -2.05) \simeq 0.02$
- We conclude that $\frac{4-\mu}{0.04} = -2.05$, that is,

$$\mu = 4 + 0.04 \times 2.05 = 4.082 \text{ ounces}$$

Normal distribution : quantiles

Normal distribution : quantiles

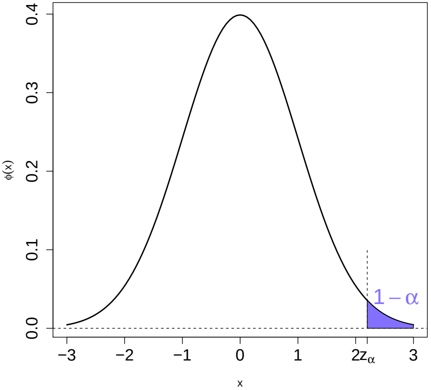
As in the previous example, we are sometimes given a probability and asked to find the corresponding value z

For instance, for any $\alpha \in (0, 1)$, let z_α be such that

$$\mathbb{P}(Z > z_\alpha) = 1 - \alpha.$$

$$\text{i.e.,} \quad \mathbb{P}(Z < z_\alpha) = \alpha,$$

for $Z \sim \mathcal{N}(0, 1)$

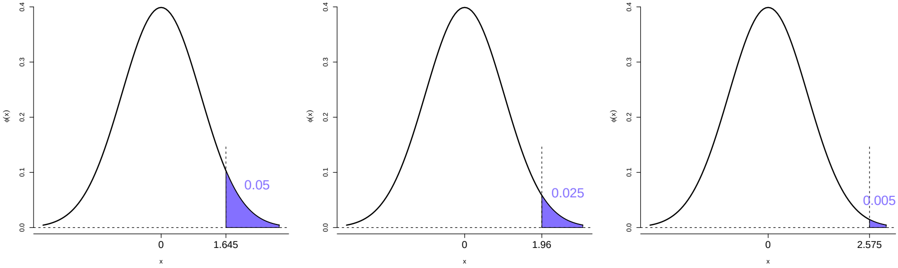


This value z_α is called the **quantile** of level α of the standard normal distribution

Normal distribution : quantiles

Some particular quantiles will be used extensively in subsequent chapters. These are the quantiles of **level 0.95**, **0.975** and **0.995** :

$$\mathbb{P}(Z > 1.645) = 0.05, \quad \mathbb{P}(Z > 1.96) = 0.025, \quad \mathbb{P}(Z > 2.575) = 0.005$$



Note : by symmetry of the normal pdf, it is easy to see that

$$\boxed{z_{1-\alpha} = -z_\alpha}$$

Some further properties of the Normal distribution

Some further properties of the Normal distribution

- We know that if $X \sim \mathcal{N}(\mu, \sigma)$, then $aX + b$ is also normally distributed, for any real values a and b

This generalises further :

Property

Let $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$, $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$ and X_1, X_2 are **independent**. Then, for any real values a and b,

$$aX_1 + bX_2 \sim \mathcal{N}\left(a\mu_1 + b\mu_2, \sqrt{a^2\sigma_1^2 + b^2\sigma_2^2}\right)$$

Some further properties of the Normal distribution

- As a direct application of the preceding property, we have, with $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$,

$$X_1 + X_2 \sim \mathcal{N}\left(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right), \quad X_1 - X_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right)$$

- Besides, the previous property can be readily extended to an arbitrary number of independent normally distributed random variables.

Some further properties of the Normal distribution

Example

Let X_1, X_2, X_3 represent the times necessary to perform three successive repair tasks at a certain service facility. Suppose they are independent normal random variables with expected values μ_1, μ_2 and μ_3 and variances σ_1^2, σ_2^2 and σ_3^2, respectively.
What can be said about the distribution of $X_1 + X_2 + X_3$?

Answer: From the previous property, we can conclude that

$$X_1 + X_2 + X_3 \sim \mathcal{N}\left(\mu_1 + \mu_2 + \mu_3, \sqrt{\sigma_1^2 + \sigma_2^2 + \sigma_3^2}\right)$$

Some further properties of the Normal distribution

Example (ctd.)

If $\mu_1 = 40$ min, $\mu_2 = 50$ min and $\mu_3 = 60$ min, and $\sigma_1^2 = 10$ min^2, $\sigma_2^2 = 12$ min^2 and $\sigma_3^2 = 14$ min^2, what is the probability that the full task would take less than 160 min?

Answer: From the above, we have

$$X \doteq X_1 + X_2 + X_3 \sim \mathcal{N}\left(150, \sqrt{36} = 6\right)$$

Hence,

$$\mathbb{P}(X \leq 160) = \mathbb{P}\left(Z \leq \frac{160 - 150}{6}\right) = \mathbb{P}(Z \leq 1.67) \overset{\text{table}}{=} 0.9525$$

(using the standard normal table)

Checking if the data are normally distributed

Checking if the data are normally distributed

Fact

Many of the statistical techniques presented in subsequent chapters are based on an assumption that the distribution of the random variable of interest is **normal**

$\rightsquigarrow$ in many instances, we will need to check whether a data set appears to be generated by a normally distributed random variable

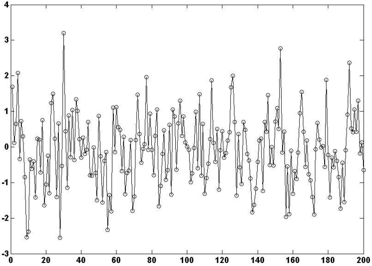
How do we do that ?

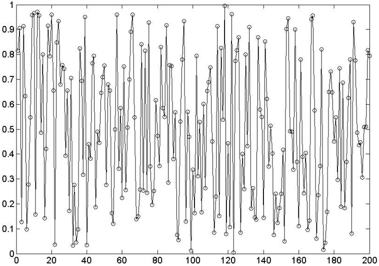
- Although they involve an element of subjective judgement, graphical procedures are the most helpful for detecting serious departures from normality

- Some of the visual displays we have used earlier, such as the density histogram, can provide a first insight about the form of the underlying distribution

Density histograms to check for normality

Realizations from two different distributions:





Density histograms to check for normality

- Think of a density histogram as a piecewise constant function $h_n(x)$, where n is the number of observations in the data set

$\leadsto$ then, if the r.v. X having generated the data has density f on a support S_X, it can be shown that, under some regularity assumptions, for any $x \in S_X$,

$$h_n(x) \to f(x)$$

as $n \to \infty$ (and the number of classes $\to \infty$)

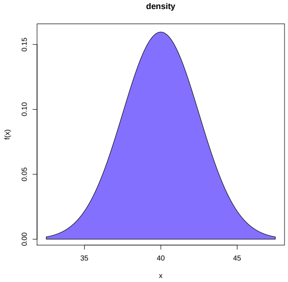
(the convergence "$h_n(x) \to f(x)$" has to be understood in a particular probabilistic sense, but details are beyond the scope of this course)

- Concretely, **the larger the number of observations, the more similar the density histogram and the 'true' (unknown) density f**

$\leadsto$ look at the histogram and decide whether it looks enough like the symmetric 'bell-shaped' normal curve or not

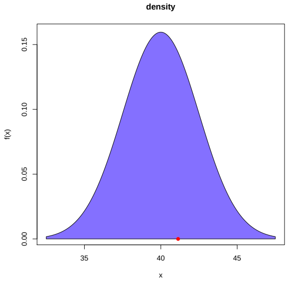
Density histograms to check for normality

Suppose we have a data set of size $n = 50$, drawn from a normal distribution (but this, we ignore)



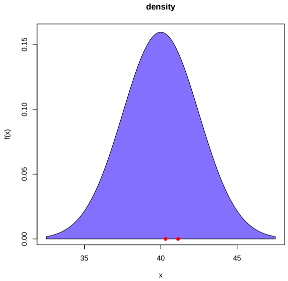
Density histograms to check for normality

Suppose we have a data set of size $n = 50$, drawn from a normal distribution (but this, we ignore)



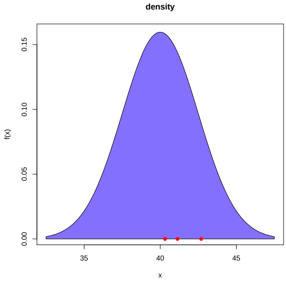
Density histograms to check for normality

Suppose we have a data set of size $n = 50$, drawn from a normal distribution (but this, we ignore)



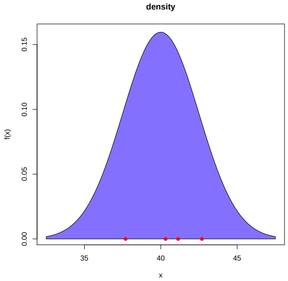
Density histograms to check for normality

Suppose we have a data set of size $n = 50$, drawn from a normal distribution (but this, we ignore)



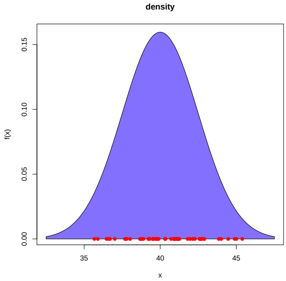
Density histograms to check for normality

Suppose we have a data set of size $n = 50$, drawn from a normal distribution (but this, we ignore)



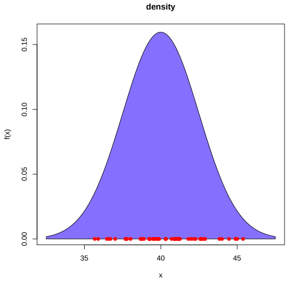
Density histograms to check for normality

Suppose we have a data set of size $n = 50$, drawn from a normal distribution (but this, we ignore)



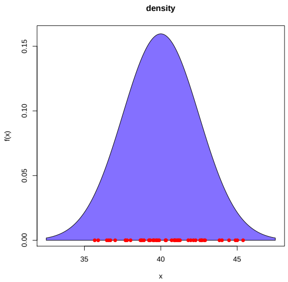
Density histograms to check for normality

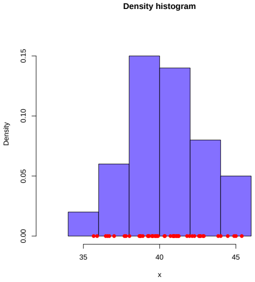
Suppose we have a data set of size $n = 50$, drawn from a normal distribution (but this, we ignore)



Density histograms to check for normality

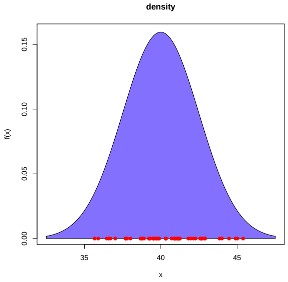
Suppose we have a data set of size $n = 50$, drawn from a normal distribution (but this, we ignore)

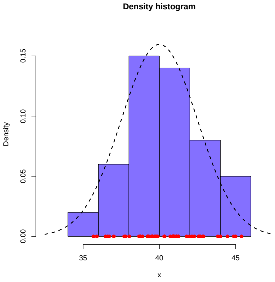




Density histograms to check for normality

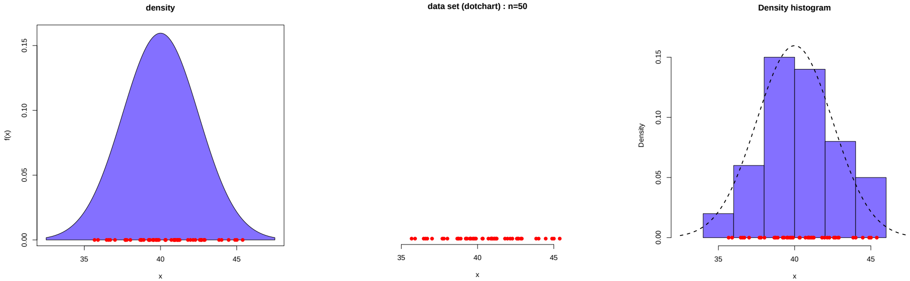
Suppose we have a data set of size $n = 50$, drawn from a normal distribution (but this, we ignore)





Density histograms to check for normality

Suppose we have a data set of size $n = 50$, drawn from a normal distribution (but this, we ignore)

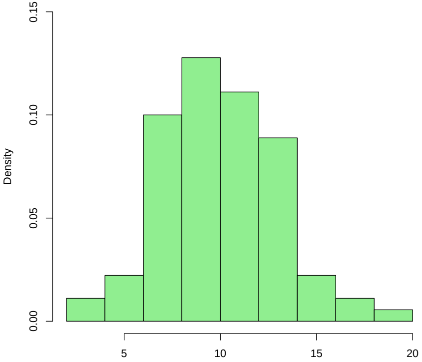


$\rightsquigarrow$ the density histogram looks like the bell-shaped curve

Besides, as both $f(x)$ and the density histogram are scaled such that the purple areas are 1, they are easily superimposed and compared

Density histograms to check for normality

Look again at the (density) histogram



$\leadsto$ the histogram is symmetric and bell-shaped, without outliers

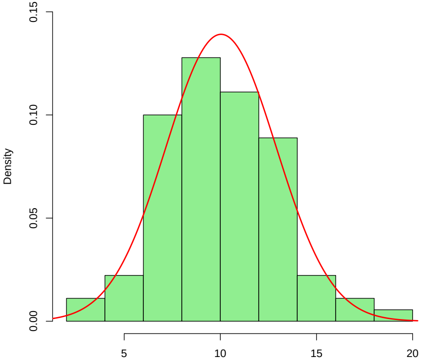
$\leadsto$ the normality assumption is reasonable

$\leadsto$ clear lack of symmetry (skewed to the right)

$\leadsto$ serious departure from normality

Density histograms to check for normality

Look again at the (density) histogram



$\rightsquigarrow$ clear lack of symmetry (skewed to the right)

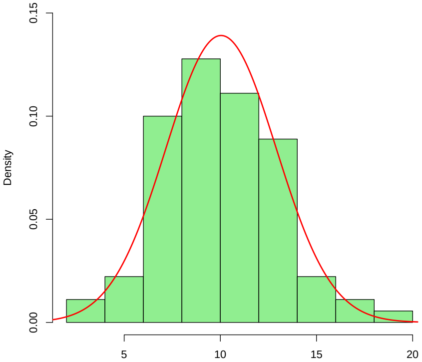
$\rightsquigarrow$ serious departure from normality

$\rightsquigarrow$ the histogram is symmetric and bell-shaped, without outliers

$\rightsquigarrow$ the normality assumption is reasonable

Density histograms to check for normality

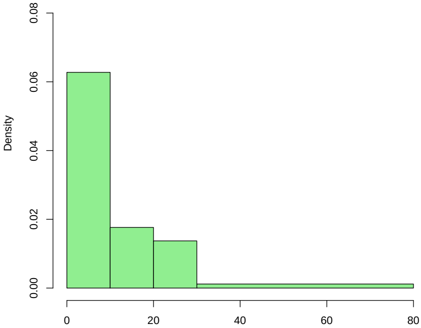
Look again at the (density) histogram



⤳ the histogram is symmetric and bell-shaped, without outliers

⤳ the normality assumption is reasonable

Look at the density histogram

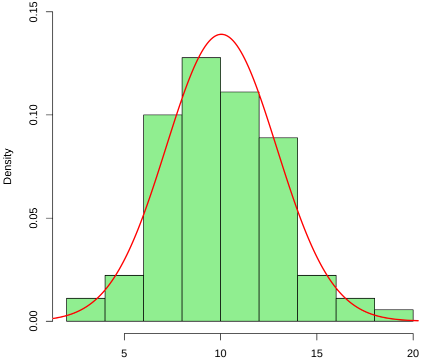


⤳ clear lack of symmetry (skewed to the right)

⤳ serious departure from normality

Density histograms to check for normality

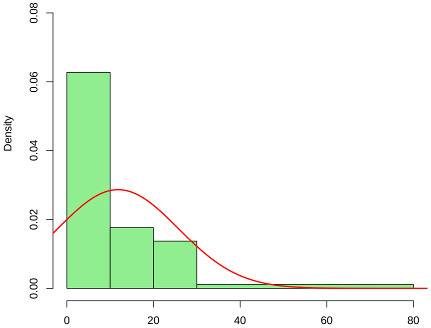
Look again at the (density) histogram



⇝ the histogram is symmetric and bell-shaped, without outliers

⇝ the normality assumption is reasonable

Look at the density histogram

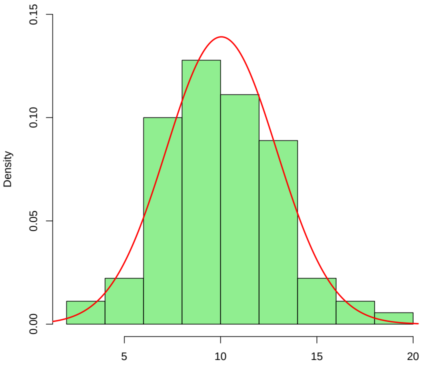


⇝ clear lack of symmetry (skewed to the right)

⇝ serious departure from normality

Density histograms to check for normality

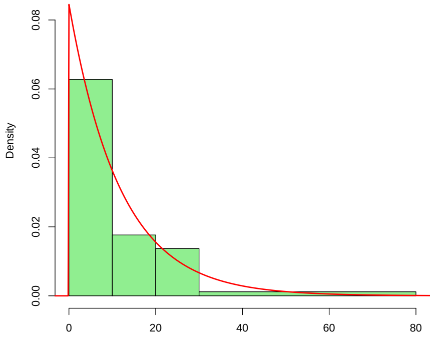
Look again at the (density) histogram



⤳ the histogram is symmetric and bell-shaped, without outliers

⤳ the normality assumption is reasonable

Look at the density histogram



⤳ clear lack of symmetry (skewed to the right)

⤳ serious departure from normality (Exponential?)

Quantile plots

Quantile plots

- Density histograms are easy to use, however they are usually not really reliable indicators of the distribution form unless the number of observations is very large

$\rightsquigarrow$ another special graph, called a **normal quantile plot**, is more effective in detecting departure from normality

- The plot essentially compares data ordered from smallest to largest with what to expect to get for the smallest to largest in a sample of that size if the theoretical distribut. from which the data comes is normal

$\rightsquigarrow$ if the data were effectively selected from the normal distribution, the two sets of values should be reasonably close to one another

Note : the quantile plot is also sometimes called **qq-plot**

($\rightsquigarrow$ instruction in Matlab : `qqplot`)

Quantile plots

Procedure for building a quantile plot :

- observations $\{x_1, x_2, \ldots, x_n\}$
- ordered observations : $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$
- cumulative probabilities $\alpha_i = \frac{i-0.5}{n}$, for all $i = 1, \ldots, n$
- standard normal quantiles of level α_i : for all $i = 1, \ldots, n$, z_{α_i} chosen such that $\mathbb{P}(Z \leq z_{\alpha_i}) = \alpha_i$, where $Z \sim \mathcal{N}(0, 1)$
- Quantile plot : plot the n pairs $(x_{(i)}, z_{\alpha_i})$

- If the sample comes from the Standard Normal distribution, $x_{(i)} \simeq z_{(i)}$ and the points would fall close to a 45°straight line passing by (0,0)

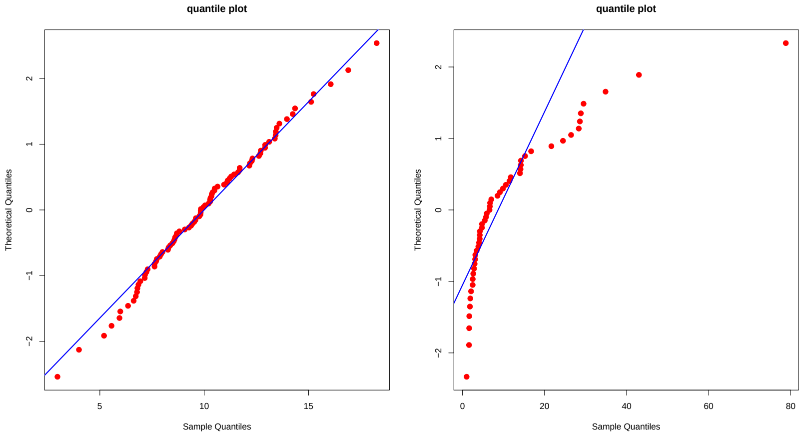
- If the sample comes from some other normal distribution, the points would still fall around a straight line, as there is a linear relationship between the quantiles of $\mathcal{N}(\mu, \sigma)$ and the standard normal quantiles

Fact

If the sample comes from some normal distribution, the points should follow (at least approximately) a straight line

Quantile plots : examples

The figure below displays quantile plots for the previous two examples



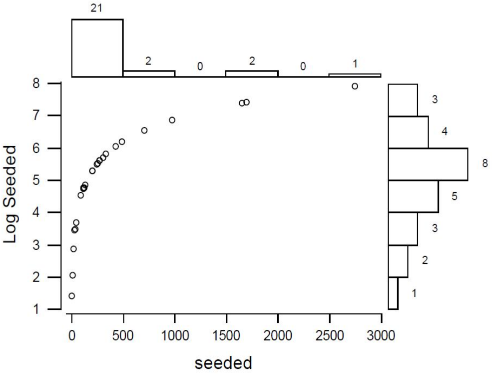
$\leadsto$ the normality assumption appears acceptable for the first data set, not at all for the second

Transforming observations

Transforming observations

- When the density histogram or the qq-plot indicate that the assumption of a normal distribution is invalid, transformations of the data can often improve the agreement with normality.
- Scientists regularly express their observations in natural logs

Let's look again at the seeded clouds rainfall data



Transforming observations

- Apart from $\log$, other transformations may be useful :

$$\frac{-1}{x}, \quad \sqrt{x}, \quad \sqrt[4]{x}, \quad x^2, \quad x^3$$

- If the observations are positive and the distribution has a long tail on the right, then concave transformations like $\psi(x) := \log(x)$ or $\sqrt{x}$ put the large values down farther than they pull the central or small values $\rightsquigarrow$ observations 'more symmetric'

- Convex transformations work the other way

- If the transformed observations are approximately normal (check with a quantile plot), it is usually advantageous to use the normality of this new scale to perform any statistical analysis

Note: transformations should be invertible (i.e. ψ^{-1} should exists)

Objectives

Now you should be able to :

- Calculate probabilities, determine mean, variance and standard deviation for normal distributions □
- Standardise normal random variables, and understand why this is useful □
- Use the table for the cdf of a standard normal distribution to determine probabilities of interest □
- Explain the general concepts of estimating the parameters of a population, in particular the difference between estimator and estimate, and the role played by the sampling distribution of an estimator □
- Illustrate those concepts with the particular case of the estimation of the mean in a normal population

Put yourself to the test ! $\rightsquigarrow$ Q20 p.131, Q30 p.195