Faculty of Life Sciences

The Normal distribution

Ib Skovgaard & Claus Ekstrøm
E-mail: ims@life.ku.dk

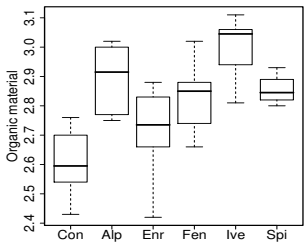
Program

- The normal distribution (continued)
- Density function
- Comparing data with the normal distribution
- Scaling and standardization
- The standard normal distribution
- The normal distribution in R
- Sums of normals are again normal
- Normal probability calculations
- Probabilities of centered intervals
- Summary of main points

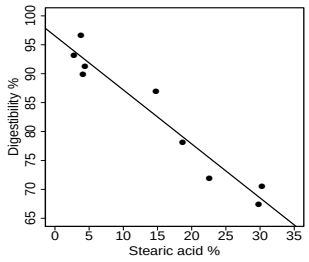
The normal (or Gaussian) distribution

The normal distribution is used to model residual variation

One-way ANOVA



Linear regression



One sample:

Blood pressure								
96	119	119	108	126	128	110	105	94

Why the normal distribution?

The normal distribution

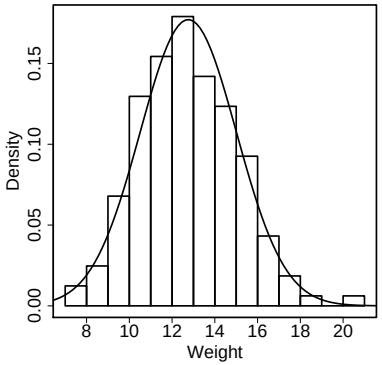
- often fit (biological) data
- has nice mathematical properties
- arises as a sum of "random errors" from several sources of variation. This is the result called the Central Limit Theorem (CLT): sum or mean of many terms resemble a normal distribution.

Also called the Gaussian distribution after

Carl Friedrich Gauss — German mathematician and physicist, 1777–1855.

Weights of crabs

- Weights of 162 crabs of a certain age: $y_1, \ldots, y_{162}$.
- R: $\bar{y} = 12.76$, $s = 2.25$
- Histogram normalized to have total area 1
- Curve for f, where f is the density of the normal distribution



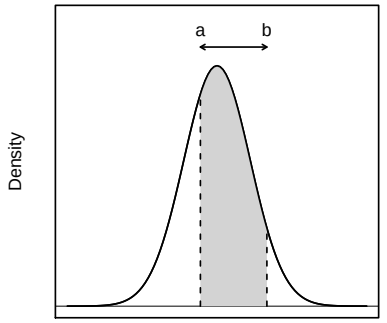
$$f(y) = \frac{1}{\sqrt{2\pi \cdot 2.25^2}} \exp\left(-\frac{1}{2 \cdot 2.25^2}(y - 12.76)^2\right)$$

The curve fits nicely to the histogram.

Density and probabilities

Density for the normal distribution with mean μ and standard deviation σ:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y - \mu)^2\right)$$



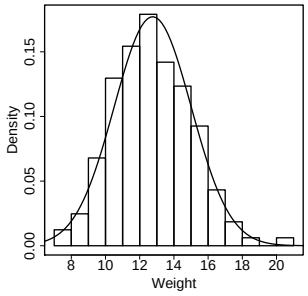
The probability (= population frequency) of the interval between a and b is equal to the area under the density curve within that interval:

$$P(a < Y < b) = \int_a^b f(y)\,dy$$

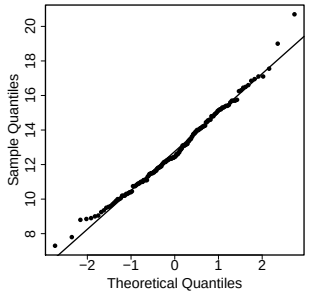
Are the data normally distributed?

Weight of 162 crabs: $y_1, \ldots, y_{162}$. Sample mean is $\bar{y} = 12.76$ and standard deviation is $s = 2.25$.

Histogram and density



QQ-plot



- Histogram compared with the density for $N(\bar{y}, s^2)$
- QQ-plot: quantile-quantile plot. Compare with a straight line with intercept $\bar{y}$ and slope σ.
  R: `qqnorm(wgt); qqline(wgt) or abline(..)`

Scaling and standardization

- Scaling preserves normality: If Y is normal then $a + b \cdot Y$ is normal with the "natural" mean $a + b\mu$ and standard deviation $|b|\sigma$.
- Standardization: In particular, the standardized version, $Z = \frac{Y-\mu}{\sigma}$ has mean zero and standard deviation one.

The standard normal distribution

Probabilities from $N(\mu,\sigma^2)$ may be converted to probabilities from $N(0,1)$. For example,

$$P(Y \le a) = P\left(\frac{Y-\mu}{\sigma} \le \frac{a-\mu}{\sigma}\right) = P\left(Z \le \frac{a-\mu}{\sigma}\right) = \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Density of $N(0,1)$:

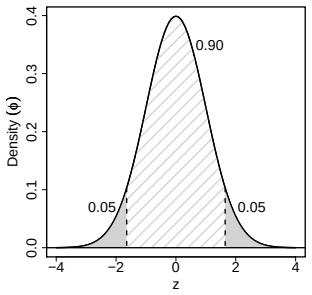
$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right)$$

Distribution function — "area to the left of z":

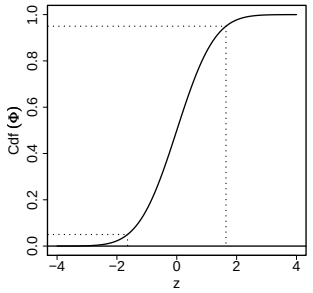
$$\Phi(z) = P(Z \le z) = \int_{-\infty}^{z} \phi(x)\,dx$$

Standard normal distribution

Density



Distribution function



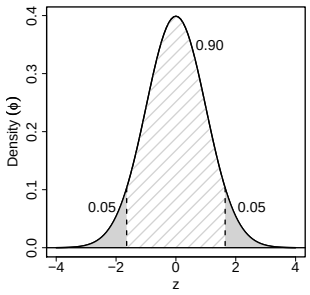
The 95%-quantile is 1.6449:

$\Phi(1.6449) = P(Z \le 1.6449) = 0.95$

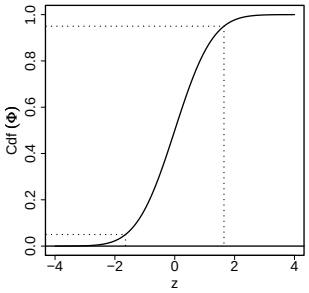
$P(-1.6449 \le Z \le 1.6449) =$

Standard normal distribution

Density



Distribution function



The 97.5%-quantile is 1.960:

$\Phi(1.960) = P(Z \le 1.960) = 0.975$

$P(-1.96 \le Z \le 1.96) =$

Normal distribution function in R

The distribution function for $N(\mu,\sigma)$ is

$$\Phi_{\mu,\sigma}(x) = P(Y \le x),$$

which in R is the function `pnorm(..)`:

```
> pnorm(x, mean= mu, sd=sigma)
```

Examples (crab weights):

```
> pnorm(14, mean= 12.76, sd=2.25)
[1] 0.7092212 # Prob. of a value less or equal to 14

> pnorm(14, mean= 12.76, sd=2.25, lower.tail=FALSE)
[1] 0.2917888 # Prob. of a value greater than 14
```

Normal distribution quantiles in R

Quantiles in $N(\mu,\sigma)$ are values of the inverse distribution function.
In R it is written

```
> qnorm(q, mean= mu, sd=sigma)
```

Example (crab weights):

```
 # 0.75 = Prob. of a value less or equal to ?
 > qnorm(0.75, mean= 12.76, sd=2.25)
[1] 14.27760
 # Answer: 75% of the population has values below 14.28
```

R: The standard normal distribution

In `pnorm(..)` and `qnorm(..)` in R, `mu=0` and `sigma=1` are default values, corresponding to the standard normal distribution.

- Distribution function: `pnorm`, fx.

```
> pnorm(1.6449)
[1] 0.9500048
```

- Quantiles: `qnorm`, fx.

```
> qnorm(0.975)
[1] 1.959964
```

Compare with the table in Appendix C2, p. 400 in the book.

Sum of normally distributed variables

Sum of normals is again normal:

If Y_1 and Y_2 are independent and both normal then their sum is again normal with

- mean of sum = sum of means,
- variance of sum = sum of variances,
- BUT NOT standard deviation of sum = sum of standard deviations,

This holds also for the sum of more than two variables.

N-probabilities

Crab data: assume that crab weight $\sim N(12.76, 2.25^2)$

What is the probability that

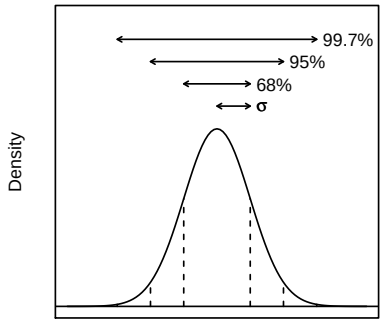
- a randomly selected crab weighs at most 14 gram?
- en a randomly selected crab weighs at least 10 gram?
- a randomly selected crab weighs between 10 and 14 gram?
- the sum of the weights of two randomly selected crabs is at least 26 gram?

```
> pnorm(1.227)
[1] 0.8900887
> pnorm(-1.227)
[1] 0.1099113
> pnorm(0.151)
[1] 0.5600121
```

```
> pnorm(0.5511)
[1] 0.7092174
> pnorm(-0.5511)
[1] 0.2907826
```

Probabilities of centered intervals

For any normal distribution:



$$P(\mu-\sigma<Y<\mu+\sigma)$$

$$=P(-1<Z<1)$$

$$=0.68$$

- 68% of the population is within $\mu\pm\sigma$
- 95% of the population is within $\mu\pm2\cdot\sigma$
- 99.7% of the population is within $\mu\pm3\cdot\sigma$

Chapter 4 summary: main points

- Comparing data with a normal distribution
  - Compare histogram with the density of $N(\bar{y},s^2)$.
  - Compare the QQ-plot with a straight line
- Linear transformation of a normal is again normal
- Standardization and the standard normal distribution, $N(0,1)$
  - Standardization: $Z=\frac{Y-\mu}{\sigma}\sim N(0,1)$
  -
  - N-probabilities calculated from $N(0,1)$
  - R: Normal probability functions
- Sums of normal variables are again normal.
- Intervals $\mu\pm\sigma$, $\mu\pm2\cdot\sigma$ or $\mu\pm3\cdot\sigma$