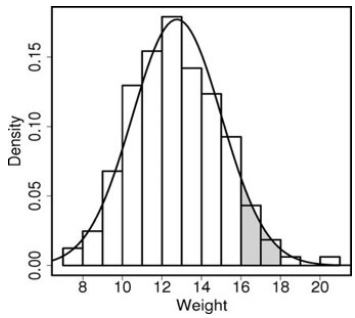
Example: Crab weights. The weights in grams of 162 crabs at a certain age were recorded as part of a larger experiment at the Royal Veterinary and Agricultural University in Denmark. The figure shows a relative frequency histogram of the observations, together with the graph of the normal density function.



Example: Crab weights, continued. The sample mean and standard deviation are given by

$$\bar{y} = 12.76\text{grams}, \quad s = 2.25\text{grams}$$

The function

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

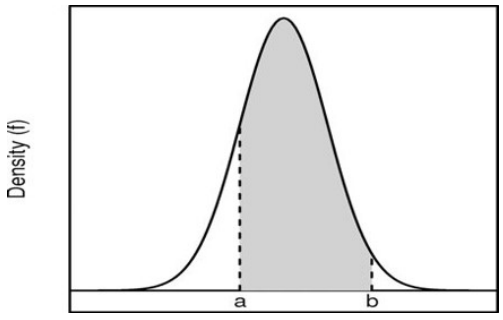
is called the density for the normal distribution with mean 12.76 and standard deviation 2.25. The point is that the curve approximates the histogram quite well. This means that the function $f(y)$ is a useful tool for description of the variation of crab weights.

Normal density function. If the sample is not too small, the histogram often looks quite smooth, and it is natural to approximate it with a smooth curve. The density for the normal distribution corresponds to a particular type of such a smooth curve; namely, the curve given by the function

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

Here, μ and $\sigma > 0$ are fixed numbers—the mean and the standard deviation, respectively. For the crab weight data we used $\mu = 12.76$ and $\sigma = 2.25$.

Normal density function, continued. The interpretation of the density is similar to that of the histogram. For an interval (a, b) the area under the curve from a to b is the probability that a random observation falls within the interval. This is illustrated in the area of the gray region, and it is interpreted as the probability that a random observation falls somewhere between a and b.



Normal density function, continued. Mathematically, the area is written as an integral, so the relationship between the probability and area can be written as

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

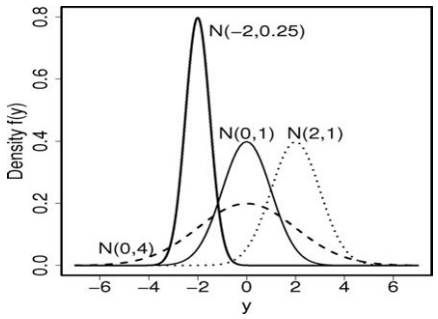
Here Y represents a random observation, and $P(a < Y < b)$ denotes the probability that such a random observation has a value between a and b. The right-hand side represents the area under the density curve over the interval from a to b. In particular we say that a variable Y is "normally distributed" (or "Gaussian") with mean μ and standard deviation σ. Then we write $Y \sim N(\mu, \sigma^2)$. Notice that we follow the tradition and use the variance σ^2 rather than the standard deviation σ in the $N(\mu, \sigma^2)$ notation.

Example: Crab weights. It seems reasonable to describe the variation of crab weights with the $N(12.76, (2.25)^2)$ distribution. Then the probability that a random crab weighs between 16 and 18 grams is

$$\int_{16}^{18} \frac{1}{\sqrt{2\pi(2.25)^2}} \exp\left(-\frac{1}{2(2.25)^2}(y - 12.76)^2\right) dy$$

This is turns out to be 0.065. Ten of the 162 crab weights are between 16 and 18 grams, corresponding to a relative frequency of $10/162 = 0.062$. The relative frequency and the probability computed in the normal distribution are close if the normal distribution describes well the variation in the sample, as in this example.

Normal density function. Recall that the density $f(y)$ is determined by the parameters μ and σ^2. The following figure shows the density for four different values of (μ, σ^2); namely, the densities for $N(0, 1)$, $N(0, 4)$, $N(2, 1)$, and $N(-2, 0.25)$. Note that all normal densities are "bell shaped."



Properties of normal density functions. The interpretations of μ and σ as the mean and standard deviation fit perfectly well with the "center" and "dispersion" interpretations in the properties below.

- **Symmetry.** $f(y)$ is symmetric around μ, so values below μ are just as likely as values above μ.
- **Center.** $f(y)$ has maximum value at $y = \mu$, so values close to μ are the most likely to occur.

- **Dispersion.** The density is "wider" for large values of σ compared to small values of σ (for fixed μ), so the larger σ the more likely are observations far from μ.

Standard normal distribution. The normal distribution with mean 0 and standard deviation 1, $N(0,1)$, is called the standard normal distribution and has the density function

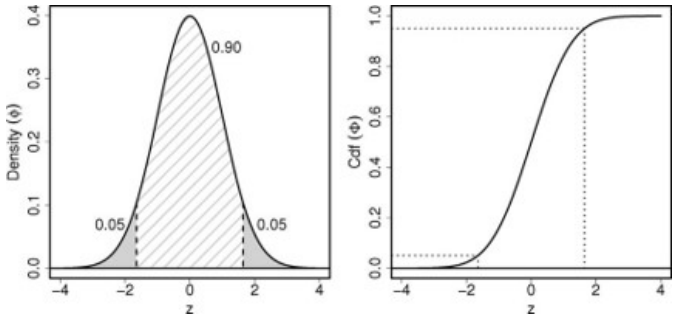
$$\phi(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)$$

Consider a $N(0,1)$ distributed variable Z. The probability that a random observation of Z falls within a certain interval is computed as the area. The area calculation cannot be solved explicitly, but certainly numerically. The result is usually denoted $\Phi(z)$; that is,

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) dy$$

$\Phi(z)$ is called the **cumulative distribution function** (cdf) of Z or the cumulative distribution function of $N(0,1)$.

Standard normal distribution, continued. The dashed lines in the right graph correspond to $z = -1.645$ and $z = 1.645$.



These values are selected because

$$\Phi(-1.645) = P(Z \leq -1.645) = 0.05$$
$$\Phi(1.645) = P(Z \leq 1.645) = 0.95$$

Probability calculations. Probabilities concerning an $N(0,1)$-distributed variable are easily expressed by

$$P(a < Z < b) = \int_a^b \phi(y)dy = \Phi(b) - \Phi(a)$$

Probabilities regarding an $N(\mu, \sigma^2)$-variable Y can be computed in the $N(0,1)$-distribution, and expressed by

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

Generally values $\Phi(z)$'s are looked up in the table of normal distribution, or computed by a computer program.

Example: Crab weights. Assume that crab weights are normally distributed with mean 12.76 and standard deviation 2.25. Then we get

$$\begin{aligned} P(16 < Y < 18) &= \Phi\left(\frac{18 - 12.76}{2.25}\right) - \Phi\left(\frac{16 - 12.76}{2.25}\right) \\ &= \Phi(2.33) - \Phi(1.44) \\ &= 0.990 - 0.925 = 0.065 \end{aligned}$$

where the Φ values are looked up in a table or computed with a computer. If we accept that the crabs in our sample are representative, we can conclude that there is a 6.5% chance of observing a weight between 16 and 18 grams for a randomly selected crab from the population.

Computing in R: Normal distribution. In the first command the first argument to pnorm() is the value x to compute the probability of outcome up to x, and the second and third arguments denote the mean and the standard deviation (sd), respectively, in the normal distribution. Note that the third argument is the standard deviation, not the variance.

```
pnorm(18,mean=12.76,sd=2.25)-pnorm(16,mean=12.76,sd=2.25)
```

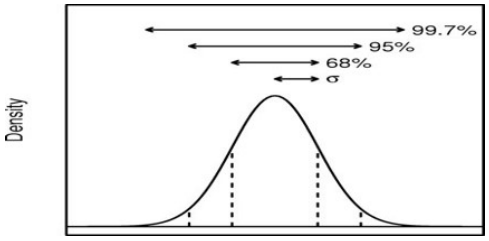
The default values of the mean and standard deviation are zero and one, respectively, for the standard normal distribution.

```
pnorm((18-12.76)/2.25)-pnorm((16-12.76)/2.25)
```

The results are the same, and computed in different ways.

Central part of distribution. We would expect the interval as $\mu \pm 1.96\sigma$ to contain approximately 95% of the observations. This corresponds to a commonly used rule-of-thumb that roughly 95% of the observations are within a distance of 2 times the standard deviation from the mean. Similar computations are made for other percentages.

Interval	$\mu \pm 1.645 \cdot \sigma$	$\mu \pm 1.960 \cdot \sigma$	$\mu \pm 2.576 \cdot \sigma$	$\mu \pm \sigma$	$\mu \pm 3 \cdot \sigma$
Probability	0.90	0.95	0.99	0.68	0.997



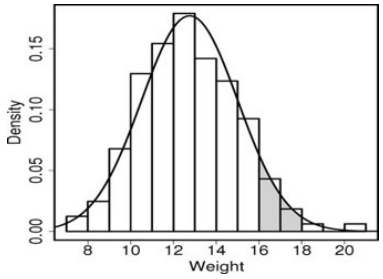
Computing in R: Normal distribution. Quantiles for the standard normal distribution are computed with the `qnorm()` function. For example, the 95% and the 97.5% quantiles are 1.645 and 1.960.

```
qnorm(0.95)
qnorm(0.975)
12.76 + qnorm(0.95)*2.25
12.76 + qnorm(0.975)*2.25
```

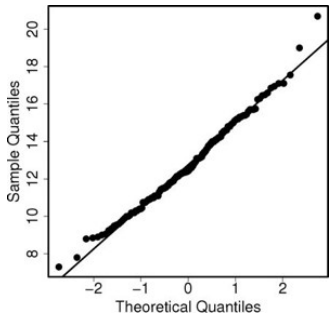
Like `pnorm()`, the function `qnorm()` can be used for normal distributions with non-zero mean and non-unit standard deviation (sd) by supplying the mean and sd as extra arguments. The above results are computed in different ways.

```
qnorm(0.95, mean=12.76, sd=2.25)
qnorm(0.975, mean=12.76, sd=2.25)
```

Are data normally distributed? For many applications it is important that the distribution is approximately a normal distribution, so we must carry out some kind of model validation. It would only rarely be correct to say that a certain variable is exactly distributed according to a normal distribution. If the sample is large enough that it makes sense to compare the histogram of the observations to the normal density with mean and standard deviation equal to the sample mean and sample standard deviation.



Quantile-quantile plot. Another relevant plot is the **QQ-plot**, or quantile-quantile plot, which compares the sample quantiles to those of the normal distribution. If data are $N(\mu, \sigma^2)$-distributed, the points in the QQ-plot should be scattered around the straight line with intercept μ and slope σ, so we can see whether there are serious deviations from the straight line relationship or not.



Construction of QQ-plot. First we have a sample $y_1, \ldots, y_n$ and that we want to check if the values could come from a normal distribution. Let $y_{(j)}$ denote the j-th smallest observation (order statistics) among $y_1, \ldots, y_n$ such that $y_{(1)} < y_{(2)} < \cdots < y_{(n)}$. Each interval between two $y_{(j)}$'s is ascribed probability $1/n$ and the intervals $(-\infty, y_{(1)})$ and $(y_{(n)}, +\infty)$ are ascribed probability $1/(2n)$ each. Let x_j be the $N(0,1)$-quantile corresponding to the accumulated probabilities up to $y_{(j)}$; i.e., let x_j be the $(j-0.5)/n$-th percentile of $N(0,1)$. Now, if the y_i's are normally distributed, then the ordered statistics $y_{(j)}$'s and the quantiles x_j's should be linear: $y_{(j)} \approx \mu + \sigma x_j$ for all $j = 1, \ldots, n$. Therefore, if we plot $y_{(j)}$'s the against x_j's, the points should be scattered around the straight line.

Example: Crab weights. QQ-plots are easily constructed with the `qqnorm()` function. Read a dataset into the data frame "databox" where the variable "wgt" contain the 162 measurements.

```
databox = read.csv(file.choose())
attach(databox)
qqnorm(wgt)
```

The command `qqline()` function adds the straight line corresponding to the normal distribution with the same 25% and 75% quantiles as the sample quantile values.

```
qqline(wgt)
detach(databox)
```

The central limit theorem. The **central limit theorem** (CLT) states that the mean of independent variables drawn from the same distribution is approximately normally distributed as long as the sample size is large no matter the distribution of the original variables. If $y_1, \ldots, y_n$ are independent and identically distributed variables (or, iid for short) with mean μ and standard deviation σ and n is large enough (usually larger than 30) then we can approximately observe

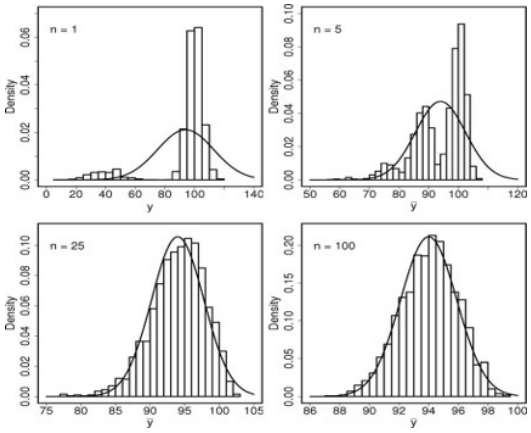
$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \sim N(\mu, \sigma^2/n)$$

The central limit theorem is quite astonishing: probabilities about an average can be approximately computed in the normal distribution, no matter how the original observations are distributed.

Example: Substance in cow milk. The concentration of a substance in cow milk is normally distributed with mean 100 and standard deviation 5 (in some unit) for healthy cows but normally distributed with mean 40 and standard deviation 10 for cows with a certain disease. A drug can be used for treatment, which brings back the concentration to the level for healthy cows. For 10% of diseased cows, however, the drug does not work. The distribution is clearly bimodal, and it has mean 94, variance 356.5, and standard deviation 18.9.

```
milk = data.frame(mean = c(100, 40), sd = c(5, 10))
cow = sample(c(1,2), 2000, replace=T, prob=c(0.9, 0.1))
data = rnorm(2000, mean=milk$mean[cow], sd=milk$sd[cow])
hist(data, breaks=20, col='yellow', freq=F)
x = seq(0, 140, length=100)
lines(x, dnorm(x, mean=94, sd=18.9), type='l', col='blue')
```

Example: Substance in cow milk. For the upper right part we simulated sample means of $y_1, \ldots, y_5$, $n = 5$. We repeated this 2000 times and plotted the histogram of the 2000 sample means. In the lower panels we did the same for an average of 25 and 100 values, respectively.



Practice problem 1.

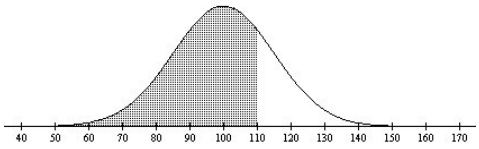
1. If z is a standard normal variable, find the probability that z is less than 1.13.
2. If z is a standard normal variable, find the probability that $z > 0.59$.
3. In the standard normal distribution, what value z separates the rest in the upper 5th percentile?
4. The grades on a chemistry midterm are normally distributed with a mean of 60 and a standard deviation of 12. Jim scored 42. Find the value z corresponding to $\Phi(z)$ in order to determine the proportion of students who scored lower than Jim.
5. Human body temperatures have a mean of 98.3°F and a standard deviation of 0.6°F. Express the probability that a patient got his body temperature of 99.5°F or higher in terms of $\Phi(z)$.

Answer key.

1. 0.871
2. 0.278
3. 1.645
4. -1.5
5. $1 - \Phi(2)$

Practice problem 2.

1. The graph depicts IQ scores of adults, and those scores are normally distributed with a mean of 100 and a standard deviation of 10. Find the area of the shaded region.



2. Assuming a normal distribution with mean μ and standard deviation σ, express the interval to contain 90% of observations in terms of μ and σ.

3. Assume that adults have IQ scores that are normally distributed with a mean of 100 and a standard deviation of 10. Find the probability that a randomly selected adult has an IQ between 90 and 120.

Answer key.

1. 0.841

2. $(\mu - 1.645\sigma, \mu + 1.645\sigma)$, or simply write $\mu \pm 1.645\sigma$

3. 0.818

Practice problem 3. Suppose that in 2004, the verbal portion of the Scholastic Aptitude Test (SAT) had a mean score of $\mu = 500$ and a standard deviation of $\sigma = 100$, while in the same year, the verbal exam from the American College Testing Program (ACT) had a mean of $\mu = 21.0$ and a standard deviation of $\sigma = 4.7$. Two friends, Mike and Tom, applying for college took the tests, Mike scoring 650 on the SAT and Tom scoring 30 on the ACT. Which of these students, Mike or Tom, scored higher among the population of students taking the relevant test? Justify your answer.

Answer key. Tom taking the ACT test performed better because the probability $1 - \Phi(1.91)$ of getting better than Tom is "smaller" than the probability $1 - \Phi(1.5)$ of the same case by Mike.

Practice problem 4.

1. The scores on the Graduate Management Admission Council's GMAT examination are normally distributed with a mean of 530 and a standard deviation of 100. What is the probability of an individual scoring above 500 on the GMAT?

2. Continue from the previous question, what is the interval on the GMAT in order to contain 95% of the scores?

3. An unknown distribution has a mean of 90 and a standard deviation of 20. Samples of size $n = 25$ are drawn randomly from the population. Find approximately the probability that the sample mean is between 85 and 92.

4. The annual precipitation amounts in a certain mountain range are normally distributed with a mean of 107 inches, and a standard deviation of 12 inches. (a) What is the standard deviation of mean annual precipitation during those 36 years? (b) What is the probability that the mean annual precipitation during 36 years will exceed 110 inches?

Answer key.

1. 0.618
2. (334, 726)
3. 0.586
4. (a) the standard deviation is 2 inches; (b) 0.067.