Looking at Data-Distributions

Introduction to Statistics
Carl von Ossietzky Universität Oldenburg
Fakultät III - Sprach- und Kulturwissenschaften

Statistics

- **Statistics**: is the study of the collection, organization, analysis, interpretation and presentation of data.
- It deals with all aspects of data, including the planning of data collection in terms of the design of surveys and experiments.
- **Data**: population sizes, increase of income, migration rates, number of unemployed people, rise of annual turnovers of companies, average temperature for today's date, etc.
- **Why statistics in general?** Wherever studies are **empirical** (involving data collection), and where that data is **variable**.
- **Goal of statistics**: to gain understanding from data. Graphs and calculations help to answer the question: "What do the data tell me?"
- Most areas of applied science require statistical analysis.

Statistics

- **Descriptive Statistics**: describe data without trying to make further conclusions.
  - Example: describe average, high and low scores from a set of test scores.
  - Purpose: characterizing data more briefly, insightfully.
- **Inferential Statistics**: describe data and its likely relation to a larger set.
  - Example: scores from a **sample** of 100 students justify conclusions about all.
  - Purpose: learn about a large **population** from study of smaller, selected **sample**, esp. where the larger population is inaccessible or impractical to study.
  - Note 'sample' vs. 'population.'

Statistics

The most common error in arguments involving statistics is not mathematical or even technical. Most common error: getting off track:

- "L is a better cold medicine. It kills 10% more germs."
- "Retail food is a rough business. Profit margins are as low as 2%!"
- "X is completely normal. 31.7% of the population reports that they have engaged in X."

Of course, this is **not** limited to statistical argumentation!

Data

Any set of data contains information about some group of *individuals*. The information is organized in *variables*.

- **Individuals** are the objects described by a set of data. Individuals may be people, but they may also be animals or things.
- A **variable** is any characteristic of an individual. A variable can take different values for different individuals.

Data

Examples of variables:

Variable	**Typical Values**
height	170 cm, 171 cm, 183 cm, 197 cm, ...
sex	male, female
reaction time	305 ms, 376.2 ms, 497 ms, 503.9 ms, ...
language	Dutch, English, Urdu, Khosa, ...
corpus frequency	0.00205, 0.00017, 0.00018, ...
age	19, 20, 25, ...

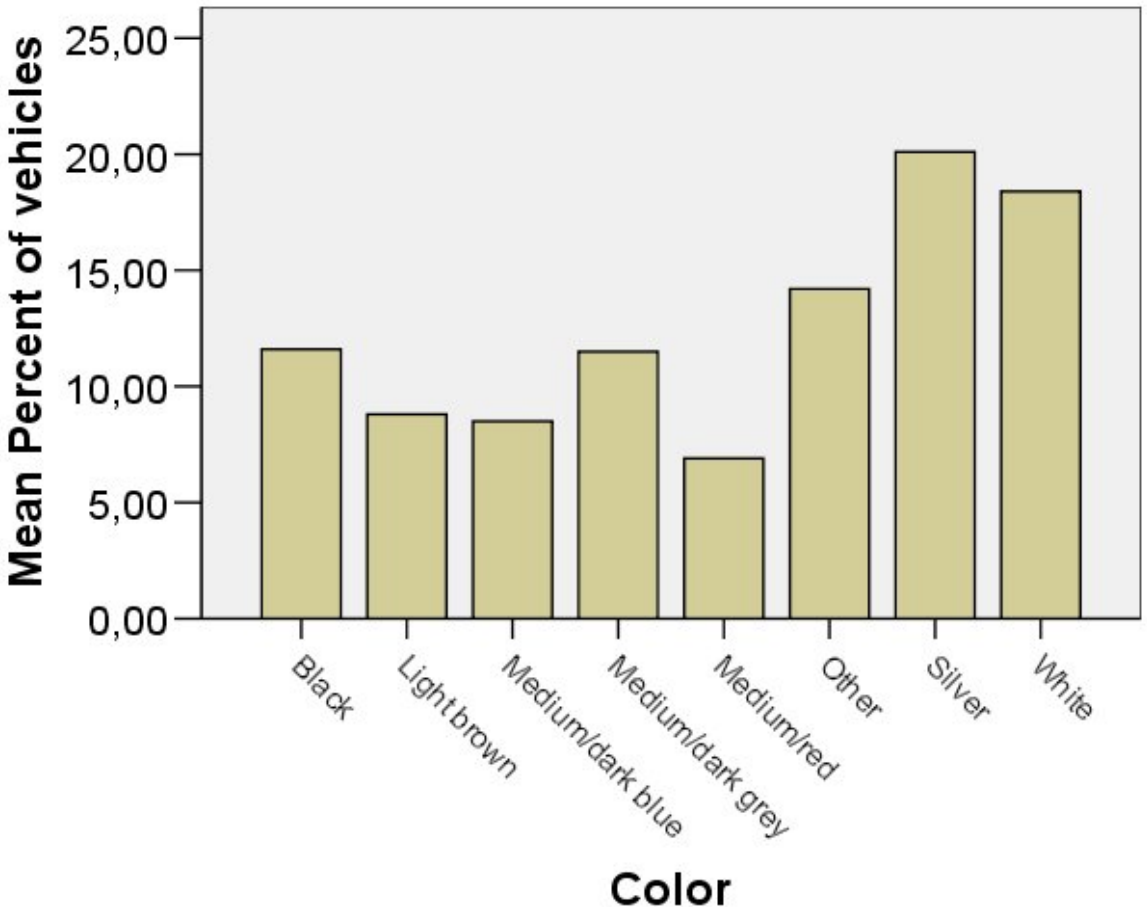
Variables tell us the properties of **individuals** or **cases**.

Data

- A **categorical variable** places an individual into one of several groups or categories.
- A **quantitative variable** takes numerical values for which arithmetic operations such as adding and averaging make sense.
- The **distribution** of a variable tells us what values it takes and how often it takes these values.

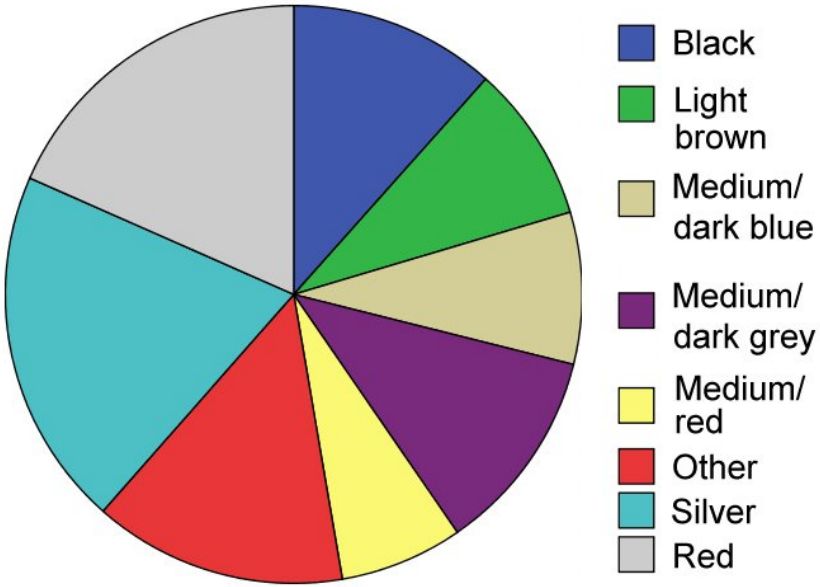
Graphs

- **Exploratory data analysis**: examination of data in order to describe their main features.
- Graphs are useful for displaying distributions: bar graph, pie chart, stem-and-leaf plot, histograms, time plot.
- Look for the **overall pattern** and for striking deviations from that pattern, e.g. outliers: individual value that falls outside the overall pattern.
- Describe the overall pattern of a distribution by its **shape** (major peaks (modes), symmetric, skewed), *center* and *spread*.



Bar graph. Percentages of most popular colors for vehicles made in North America during the 2003 model year. (Source: DuPont Automotive Color Survey).

Pie graph



Pie graph. Percentages of most popular colors for vehicles made in North America during the 2003 model year. Source: DuPont Automotive Color Survey.

Stemplot

A **stem-and-leaf plot** shows a quick picture of the shape of a distribution while including the actual numerical values in the graph. Stem-and-leaf plots work best for $small$ numbers of observations that are all greater than 0. To make a stem-and-leaf plot:

- split each data value into a **leaf** (usually the last digit) and a **stem** (the other digits). Stems may have as many digits as needed, but each leaf contains only a single digit;
- write the stems in a vertical column with the smallest at the top, and draw a vertical line at the righ of this column;
- write each leaf in the row to the right of its stem, in increasing order out from the stem.

Stem-and-leaf plot

- Example: growth season: average number of days between last frost day in the spring and the first frost day in autumn. For 57 years the growth seasons are given for cities in the USA (source: Old Farmers Almanac / National Climatic Center):

```
279 244 318 262 335 321 165 180 201 252 145 192 217 179 182 210 271 302
169 192 156 181 156 125 166 248 198 220 134 189 141 142 211 196 169 237
136 203 184 224 178 279 201 173 252 149 229 300 217 203 148 220 175 188
160 176 128
```

- Sort the values:

```
125 128 134 136 141 142 145 148 149 156 156 160 165 166 169 169 173 175
176 178 179 180 181 182 184 188 189 192 192 196 198 201 201 203 203 210
211 217 217 220 220 224 229 237 244 248 252 252 262 271 279 279 300 302
318 321 335
```

Stem-and-leaf plot

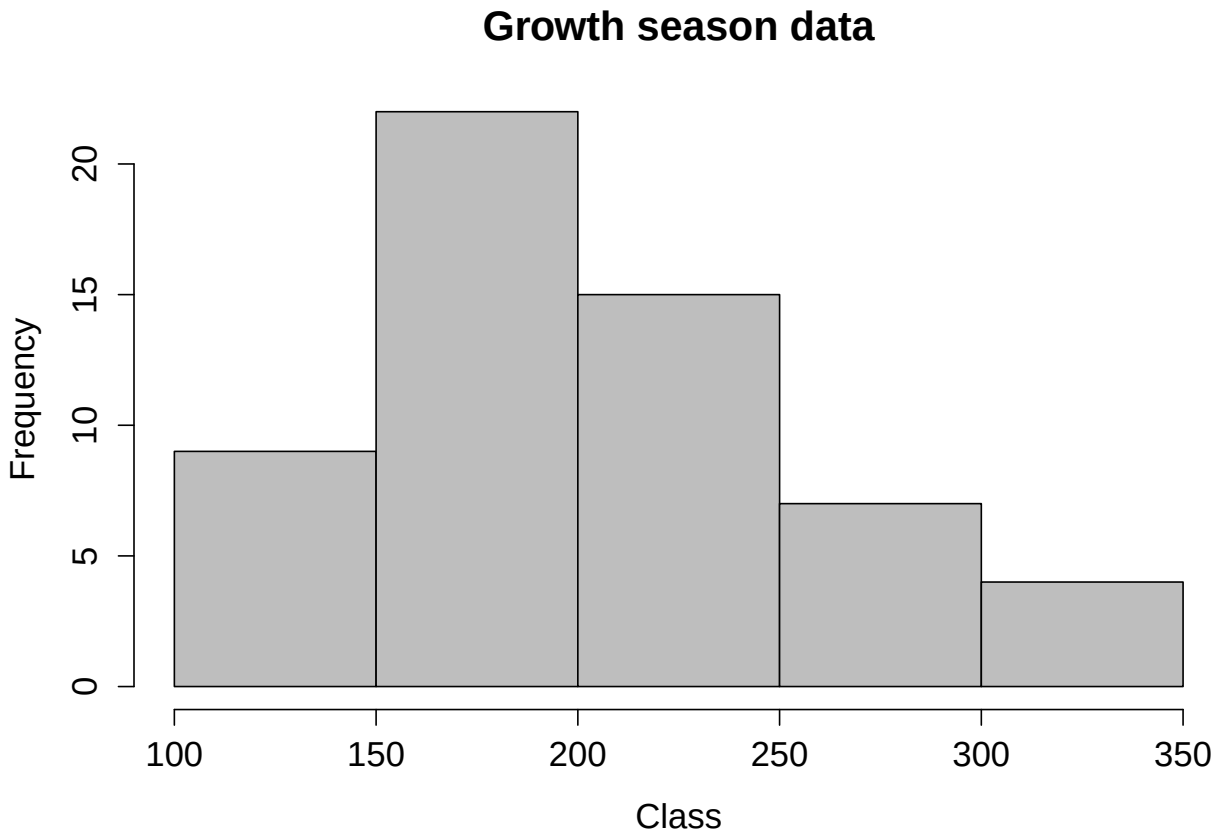
- Stem-and-leaf plot:

```
12 | 5846
14 | 1258966
16 | 0569935689
18 | 0124892268
20 | 11330177
22 | 00497
24 | 4822
26 | 2199
28 |
30 | 028
32 | 15
```

Histogram

- A **histogram** breaks the range of values of a variable into classes and displays only the count or percent of the observations that fall into each class.
- Choose convenient number of classes, always choose classes of equal width.
- Histograms do not display the actual values observed. Use stem-and-leaf plots for small data sets.
- In our example: values range from 125 to 335, so we choose as our classes:

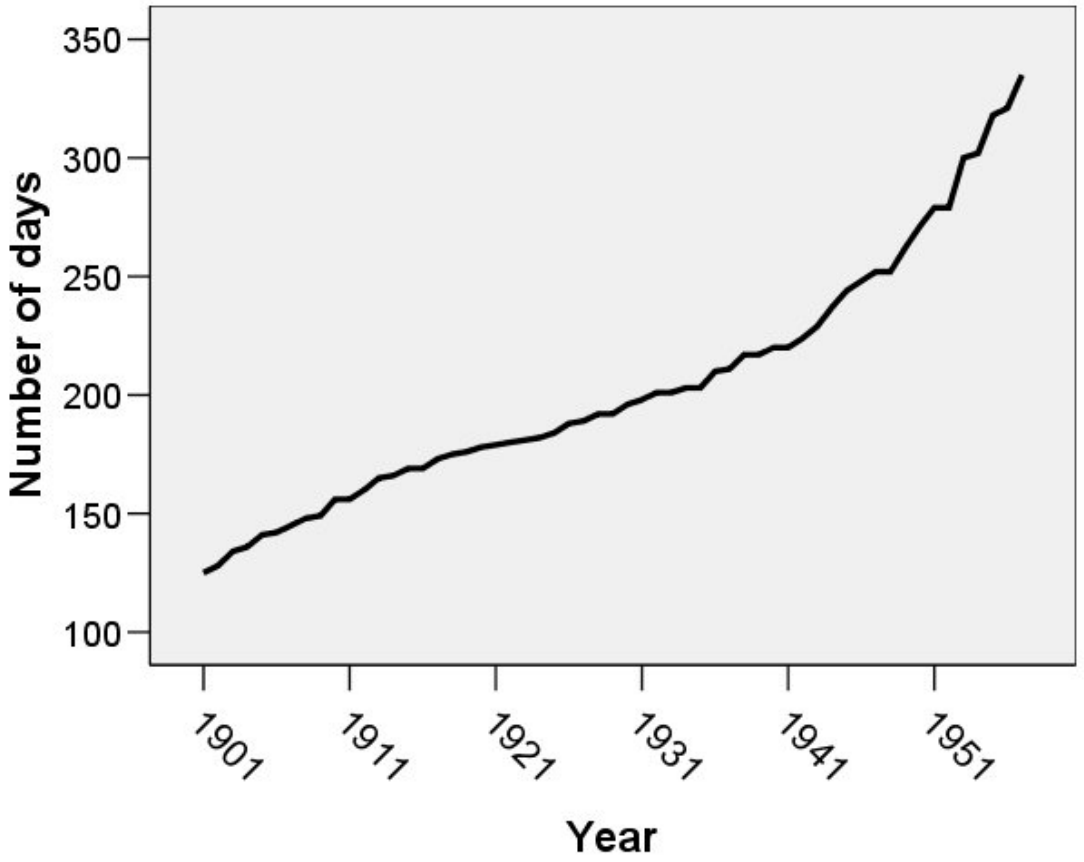
class	frequency
100 - 149	9
150 - 199	22
200 - 249	15
250 - 299	6
300 - 349	5



The graph is not perfectly symmetric but a litte bit left skewed. There are no outliers. The graph has one peak (or mode), so the pattern is unimodal.

Time plot

- Displays of the distribution of a variable that ignore time order (e.g. stem-and-leaf plots and histograms), can be misleading when there is systematic change over time.
- A **time plot** of a variable plots each observation against the time at which it was measured
- Always put time on the horizontal scale and the variable you are measuring on the vertical scale.
- Connecting the data points by lines helps emphasize any change over time.
- In our example: assume that the 57 growth season measurements are measured for the years 1901 to 1957.



The time plot shows the trend toward an increasing number of days per growth season, a trend that cannot be seen in the stem-and-leaf plot or histogram.

Measures of center and spread

- A brief description of a distribution should include its **shape** (graphs) and its **center** and **spread** (numbers).
- Measures of center: mean, median
- Measures of spread: quartiles, standard deviation

Mean

- To find the **mean** $\overline{x}$ of a set of observations, add their values and divide by the number of observations. If n observations are $x_1, x_2, \ldots, x_n$, their mean is:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n} \sum x_i$$

- In our example of growth seasons:

$$\overline{x} = \frac{279 + 244 + \ldots + 128}{57} = \frac{11608}{57} = 203.65$$

  which means that a grow season consists of 203.65 days on average in the period from 1901 to 1957.

- Mean is sensitive to the influence of a few extreme observations, it is not a **resistent measure**.

Median

The **median M** is the midpoint of a distribution, the number such that half of the observations are smaller and the other half are larger. To find the median of a distribution:

- Arrange all observations in order of size, from smallest to largest.
- If the number of observations n is odd, the median M is the center observation in the ordered list. Find the location of the median by counting $(n+1)/2$ observations up from the bottom of the list.
- If the number of observations n is even, the median M is the mean of the two center observations in the ordered list. The location of the median is again $(n+1)/2$ from the bottom of the list.

Median

- In our example: sort the observations:

```
125 128 134 136 141 142 145 148 149 156 156 160 165 166 169 169 173 175
176 178 179 180 181 182 184 188 189 192 192 196 198 201 201 203 203 210
211 217 217 220 220 224 229 237 244 248 252 252 262 271 279 279 300 302
318 321 335
```

- Location of median: $(n+1)/2 = (57+1)/2 = 58/2 = 29$. The 29th value is: 192.
- The median is more resistent than the mean, i.e. less sensitive to extreme observations (outliers).
- If the distribution is exactly symmetric, the mean and median are the same. In a skewed distribution, the mean is farther out in the long tail than is the median.

Quartiles

The p**th percentile** of a distribution is the value such that p percent of the observations fall at or below it. Median: 50th percentile, first quartile: 25th percentile, third quartile: 75th percentile. To calculate the quartiles:

- Arrange the observations in increasing order and locate the median M in the ordered list of observations.
- The **first quartile** Q_1 is the median of the observations whose position in the ordered list is to the left of the location of the overall median.
- The **third quartile** Q_3 is the median of the observations whose position in the ordered list is to the right of the location of the overall median.

Quartiles

- Observations left of the overall median (192):

  125 128 134 136 141 142 145 148 149 156 156 160 165 166 169 169 173 175
  176 178 179 180 181 182 184 188 189 192

- Location of Q_1: $(n+1)/2 = (28+1)/2 = 29/2 = 14.5$.
  Average of values at the 14th and 15th position: (166+169)/2 = 167.5.

- Observations right of the overall median (192):

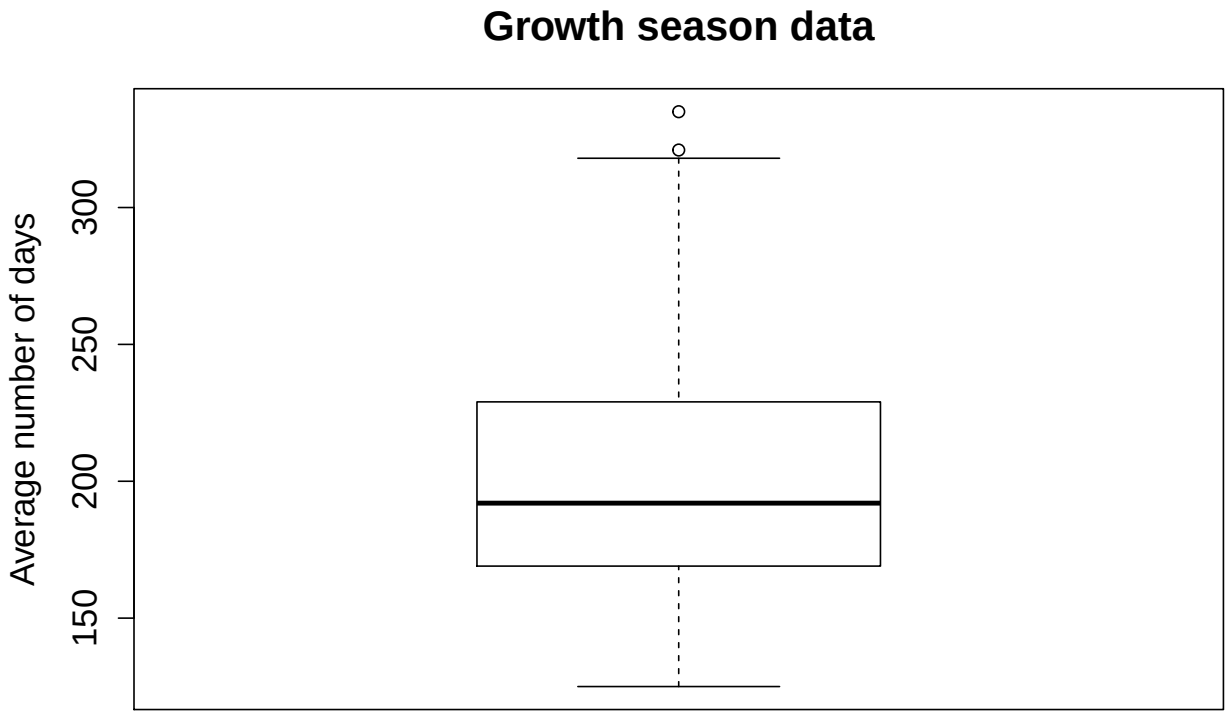
  196 198 201 201 203 203 210 211 217 217 220 220 224 229 237 244 248 252
  252 262 271 279 279 300 302 318 321 335

- Location of Q_3: $(n+1)/2 = (28+1)/2 = 29/2 = 14.5$.
  Average of values at the 14th and 15th position: (229+237)/2 = 233.

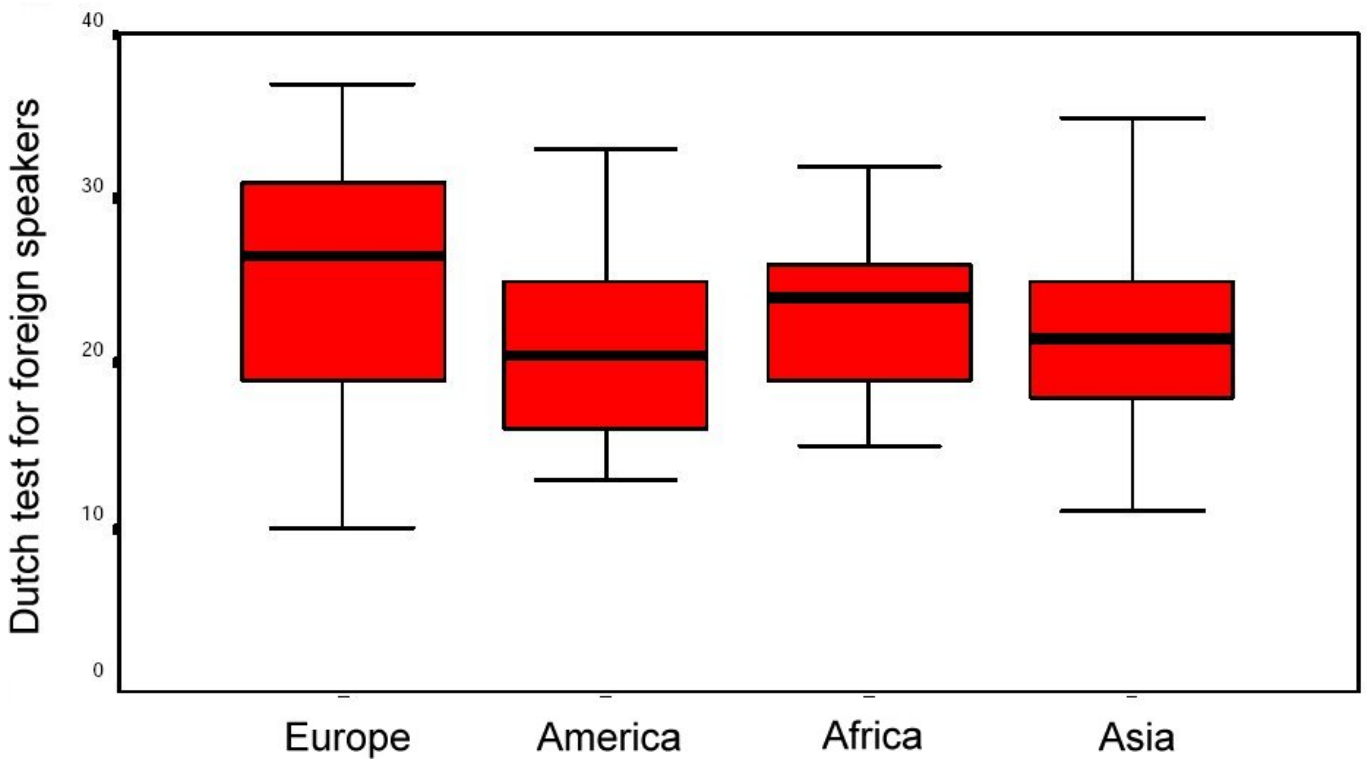
- Some statistical software programs may give slightly different values.

Five-number summary

- **Five-number summary**: Minimum, Q_1, M, Q_3, Maximum.
- A **boxplot** is a graph of the five-number summary:
  - A central box spans the quartiles Q_1 and Q_3.
  - A line in the box marks the median M.
  - Lines extend from the box out to the Minimum and Maximum.
- **Interquartile range** IQR: $Q_3 - Q1$.
- In our example: 233 - 165.5 = 65.5.
- Call an observation a suspected outlier if it falls more than $1.5 \times IQR$ above the third quartile or below the first quartile.
- In our example: above $233 + (1.5 \times 65.5) = 331.28$, and below $165.5 - (1.5 \times 65.5) = 67.25$. One outlier: 335.



Boxplot for the growth season data. We found 335 (57th observation) to be an outlier. SPSS and R also consider 321 (56th observation) an outlier.



A Dutch test was made by a European, American, African and Asian group. Boxplots are useful for side-by-side comparison.

Standard deviation

- **The variance** s^2 of a set of observations is the average of the squares of the deviations of the observations from their mean. In symbols, the variance of n observations $x_1, x_2, \ldots, x_n$ is:

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n-1} = \frac{1}{n-1}\sum(x_i - \overline{x})^2$$

- The **standard deviation s** is the square root of the variance s^2.
- In our growth season example:

$$s = \sqrt{\frac{150673.6}{57-1}} = \sqrt{2690.6} = 51.9$$

Standard deviation

- **Basic idea**: the deviations $x_i - \overline{x}$ display the spread of the values x_i around their mean $\overline{x}$.
- **Squared deviations**: since the sum of unsquared deviations would be 0!
- Squared deviations point to the mean in a way unsquared deviations do not.
- In this way the standard deviation turns out to be the natural measure of spread for the *normal distributions*.
- **Standard deviation rather than variance**: the standard deviation s measures the spread around the mean in the original scale, the variance s^2 does not.
- **Degrees of freedom**: $n - 1$. Because the sum of the deviations is always 0, the last deviation can be found once we know the other $n - 1$. Only $n - 1$ of the deviations can vary freely.

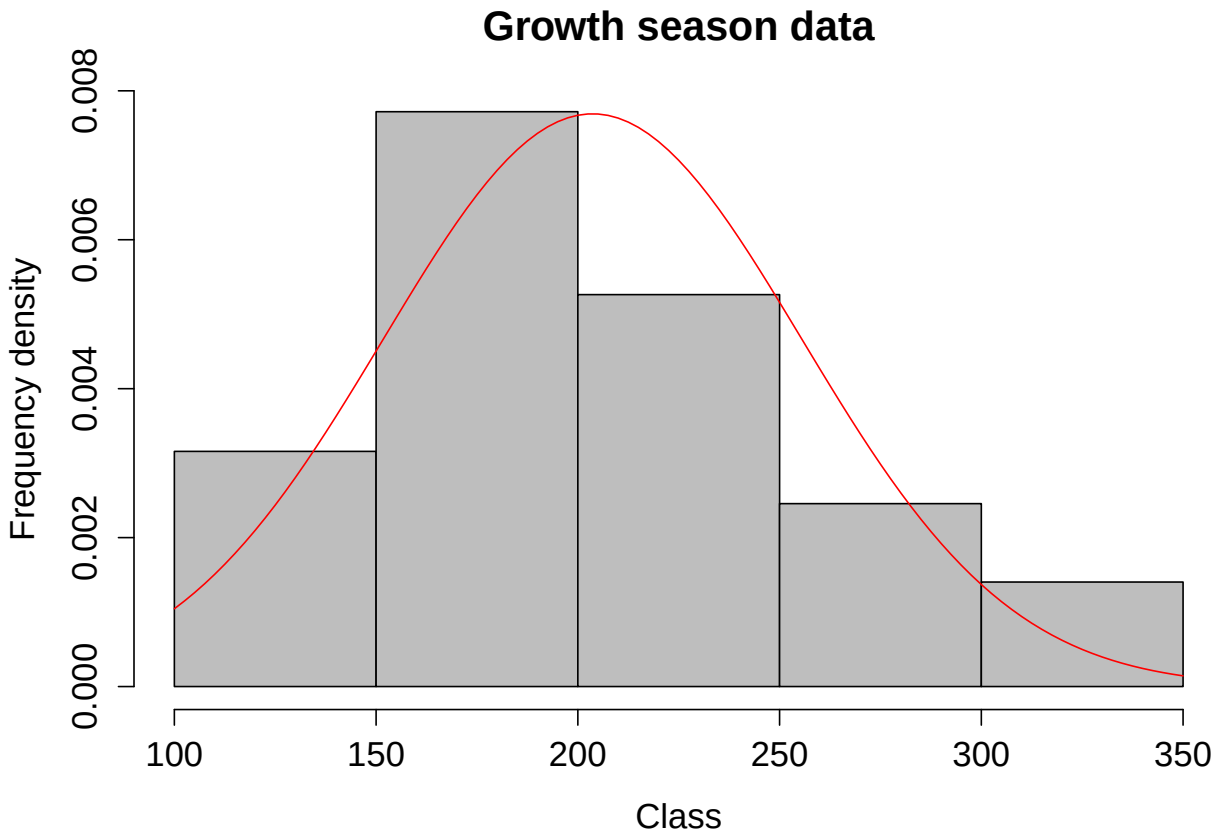
Standard deviation

Properties of the standard deviation:

- s measures spread around the mean and should be used only when the mean is chosen as the measure of center.
- $s = 0$ only when there is $no\ spread$. This happens only when all observations have the same value. Otherwise $s > 0$. As the observations become more spread out around their mean, s becomes larger.
- s, like the mean $\overline{x}$, is not resistant. A few outliers can make s very large.

Density curve

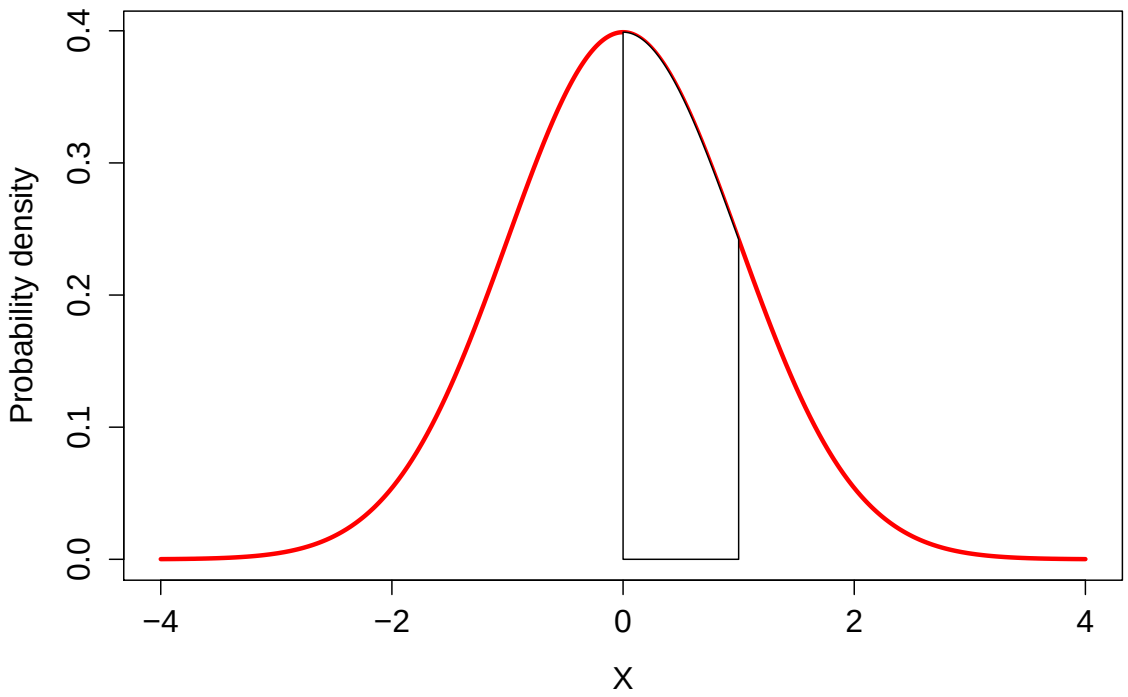
- A smooth **density curve** is an idealization that pictures the overall pattern of the data but ignores irregularities as well as outliers.
- A density curve is a curve that:
  - is always on or above the horizontal axis and
  - has area exactly 1 underneath it.
- A density curve describes the overall pattern of a distribution. The area under the curve and above any range of values is the proportion of all observations that fall in that range.
- **Probability density**: the relative likelihood for a random variable to take on a given value.



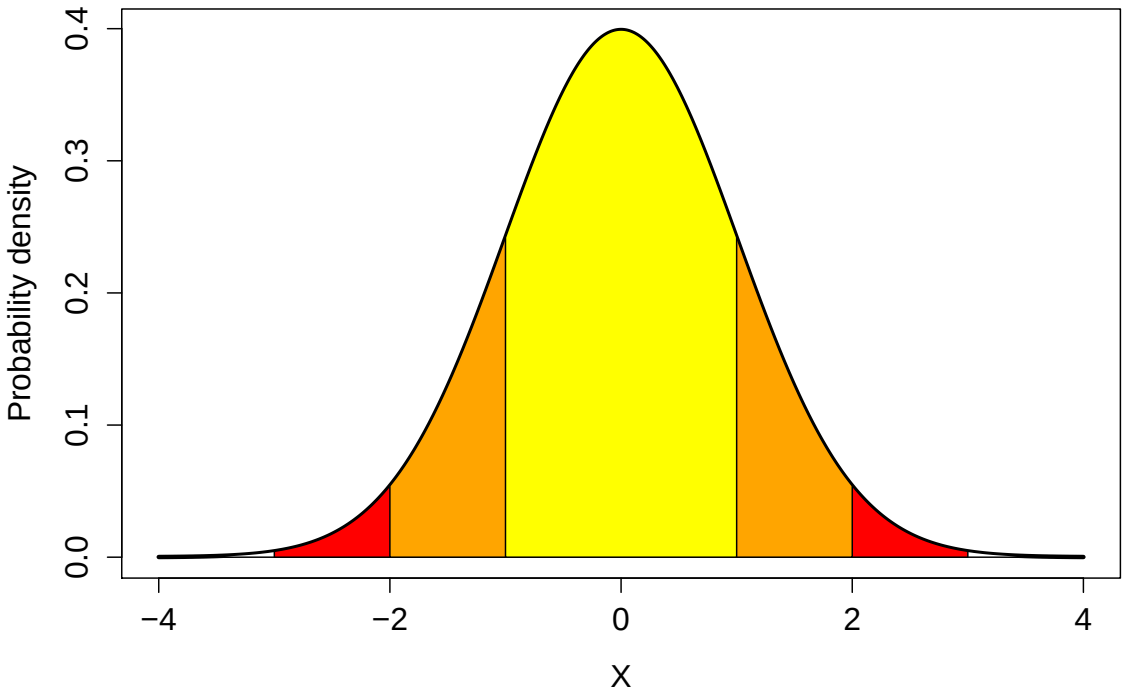
The distribution of the number of days of 57 growth seasons divided in 5 classes. The distribution is a little bit left-skewed in comparison with the bell-shaped curve.

Density curve

- The **median** of a density curve is the equal-areas point, the point that divides the area under the curve in half.
- The **mean** of a density curve is the balance point, at which the curve would balance if made of solid material.
- The median and mean are the same for a symmetric density curve. They both lie at the center of the curve.
- The mean of a skewed curve is pulled away from the median in the direction of the long tail.
- A density curve is an idealized description of a distribution of data. Distinguish between mean and standard deviation of the density curve (μ and σ) and the mean and standard deviation of the actual observations ($\overline{x}$ and s).
- **Normal curve**: symmetric, unimodal, bell-shaped.



Each **normal curve** is specified by its μ and σ: **N(μ,σ)**. Here μ=0, and σ=1. Changing μ moves curve along horizontal axis. Changing σ changes the spread of the curve. Location of σ: where the curve changes from falling ever more steeply to falling ever less steeply.



The **68-95-99.7 rule**: 68% of the observations fall within 1 standard deviation to the right and left of the mean; 95% of the observations fall within 2 standard deviations to the right and left of the mean; 99.7% of the observations fall within 3 standard deviations to the right and left of the mean.

Standardizing observations

- If x is an observation from a distribution that has mean μ and standard deviation σ, the **standardized value** of x is:

$$z = \frac{x - \mu}{\sigma}$$

- A standardized value if often called a z-**score**.
- A z-score tells us how many standard deviations the original observation falls away from the mean, and in which direction.
- Observations larger than the mean are positive when standardized, and observations smaller than the mean are negative.

Standardizing observations

- **Example**: heights of young women are approximately normal with mu=164 cm and σ=6 cm. The standardized height is:

$$z = \frac{height - 164}{6}$$

- A woman 173 cm tall has standardized height:

$$z = \frac{173 - 164}{6} = 1.4$$

which is 1.4 standard deviations above the mean.

Standardizing observations

- A woman 152 cm tall has standardized height:

$$z = \frac{152 - 164}{6} = -1.8$$

  which is 1.8 standard deviations les than the mean.
- Variables such as height are represented by capital letters near the end of the alphabet, e.g. X, observations such as the height of a particular woman are represented by lowercase letter, e.g. x.
- Standardizing a variable that has any normal distribution produces a new variable that has the **standard normal distribution**.

Standardizing observations

- The **standard normal distribution** is the normal distribution $N(0, 1)$ with mean 0 and standard deviation 1.
- If a variable X has any normal distribution $N(\mu, \sigma)$ with mean μ and standard deviation σ, then the standardized variable:

$$Z = \frac{X - \mu}{\sigma}$$

has the standard normal distribution.

Cumulative proportions

- **Cumulative proportions**: the proportion of observations in a distribution that lie at or below a given value.
- When the distribution is given by a density curve, the cumulative proportion is the area under the curve to the left of a given value.
- Our example of growth seasons: assume μ=204 days and σ=52.28 days. What is the chance that a growth season will be shorter than 290 days? Or: $P(X < 290)$?
- Find the proportion under the $N(204, 52.28)$ curve left of x=290.

Cumulative proportions

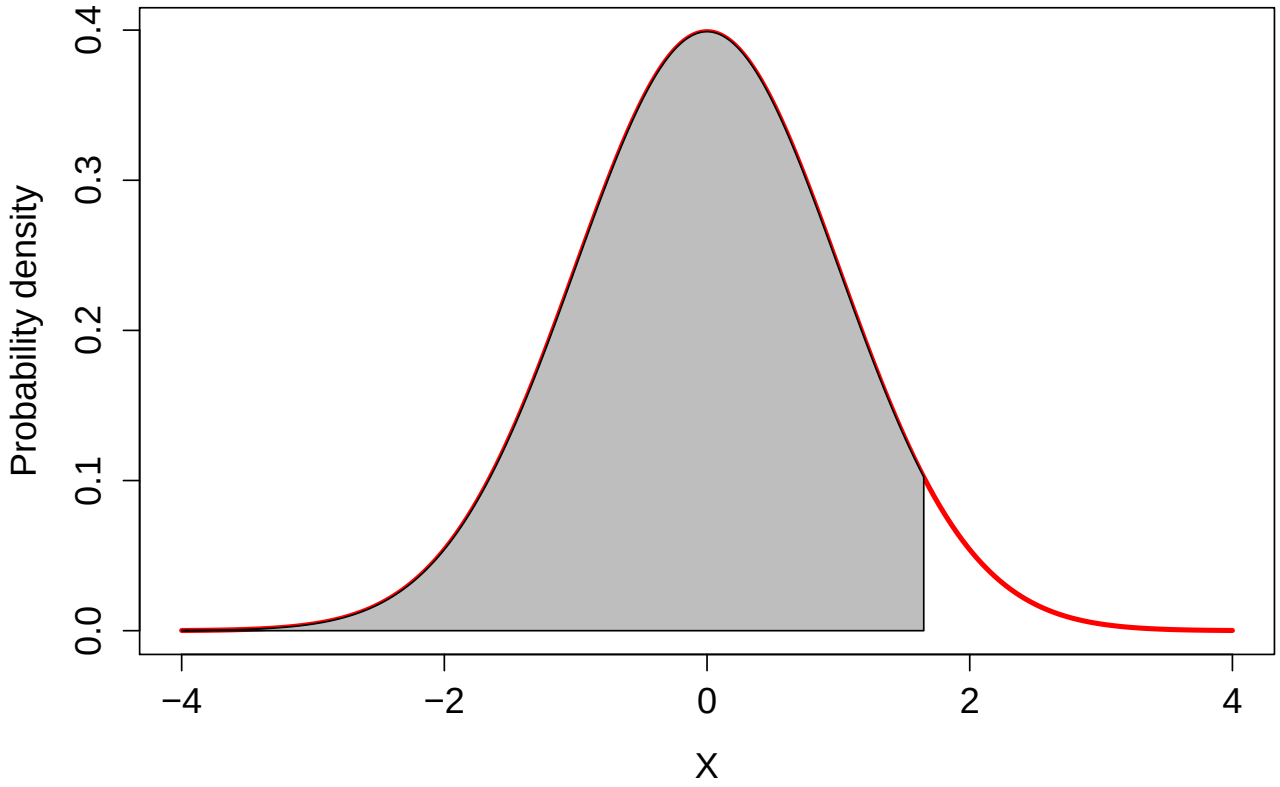
- Convert x=290 to a z-score:

$$z = \frac{x - \mu}{\sigma} = \frac{290 - 204}{52.28} = 1.645$$

- Find 1.65 in a table with standard normal probabilities, for example at:

`http://clas.sa.ucsb.edu/staff/binh/stdNormalTable.pdf`

- We need z to round to two places when we use this table: 1.65.
- According to this table the area left of this value has a surface of 0.9505. So p=0.9505 or 95%.



In the standard normal distribution the area left of z=1.640 has a surface of 0.9495. So p=0.9505 or 95%.

Cumulative proportions

- What is the chance that the growth season will last between 190 and 290 days?
- More formally:

$$P(190 < X < 290)$$

- Expressed in z-values:

$$P(-0.27 < Z < 1.65)$$

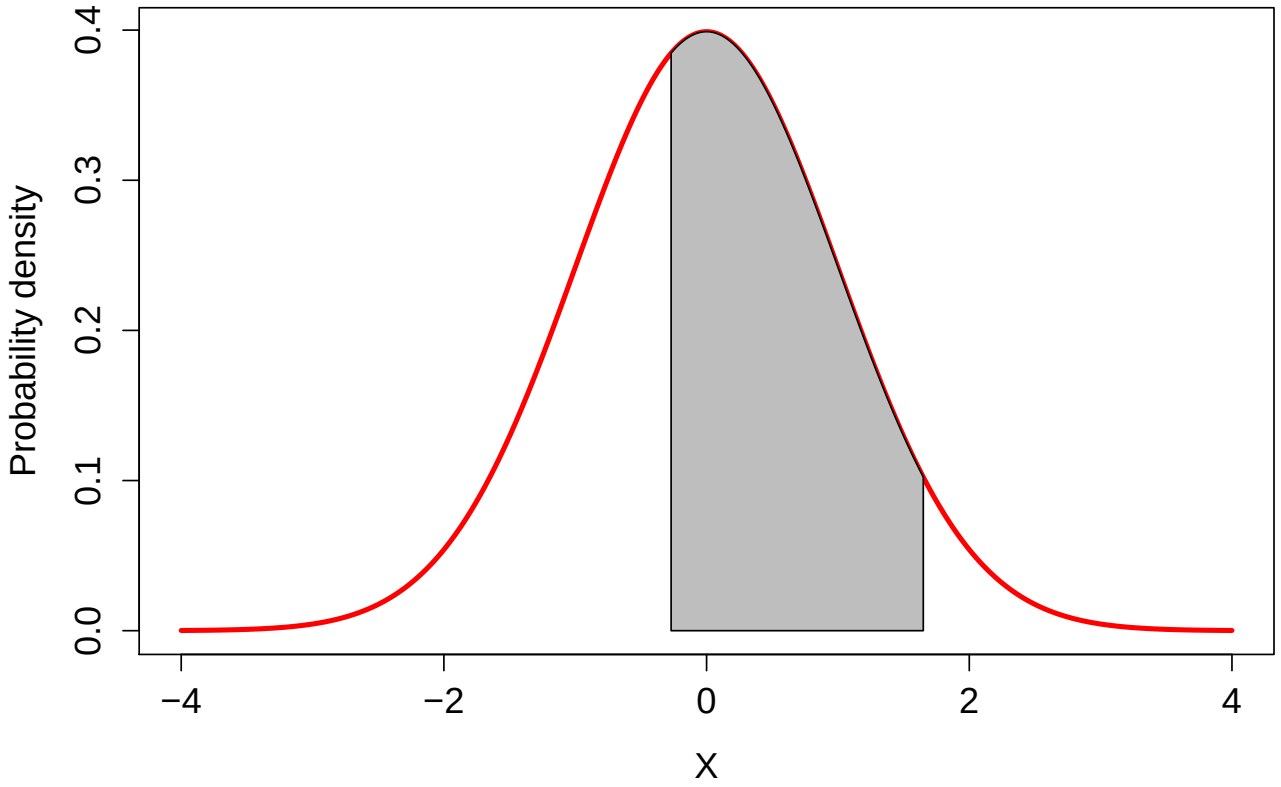
- This is equal to:

$$P(Z < 1.65) - P(Z < -0.27)$$

- From the table we find:

$$0.9505 - 0.3936 = 0.5569$$

- So p=0.5569 or 56%.



In the standard normal distribution the area between z=-0.27 and z=1.65 has a surface of 0.5569. So p=0.5569 or 56%.

Cumulative proportions

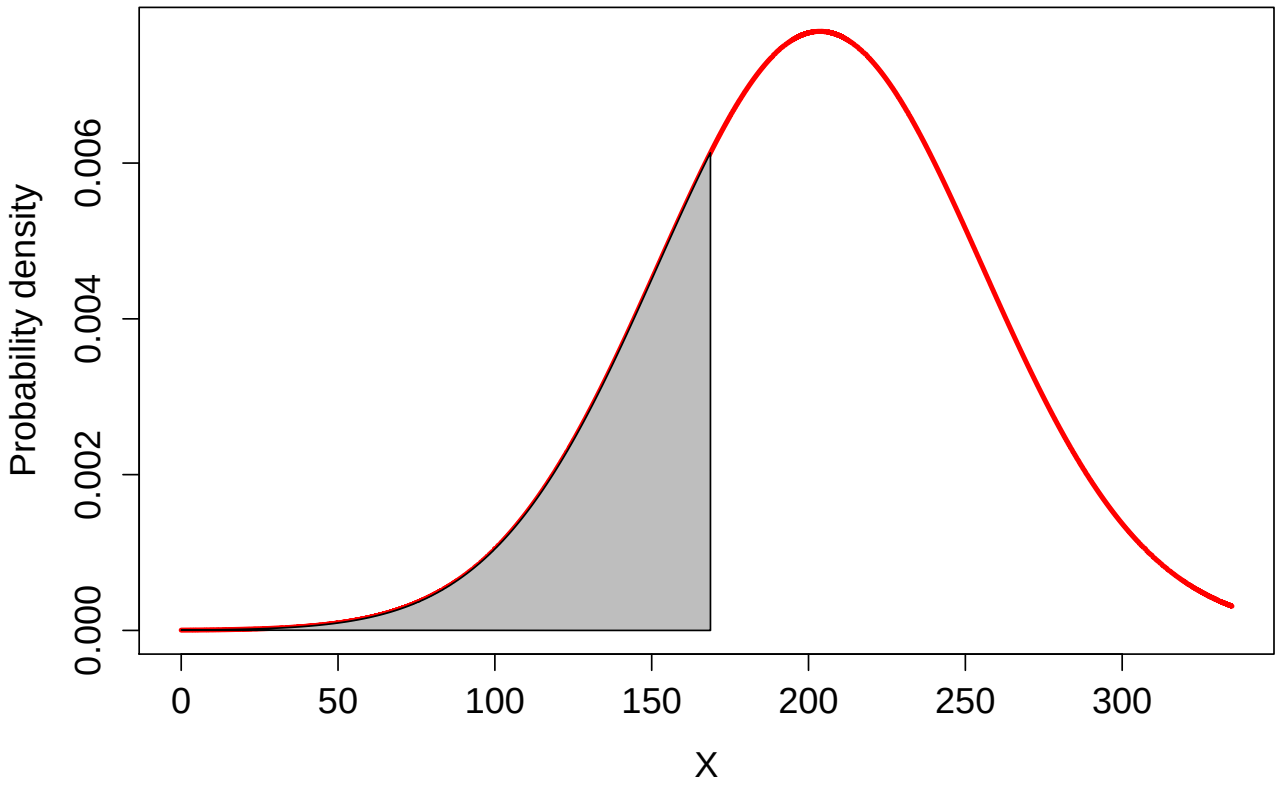
- Which growth seasons form the 25% of shortest growth seasons?
- 25% is 0.25. In the table the entry closest to 0.25 is 0.2514. This entry is corresponding to z=-0.67. So z is the standardized value with area 0.25 to its left.
- We have to transform this z-value to the original x-value. We know:

$$z = \frac{x - 204}{52.28} = -0.67$$

- Solving this equation for x gives:

$$x = 204 + (-0.67)(52.28) = 168.97$$

- So all growth seasons of less than 169 days form the 25% of shortest growth seasons.



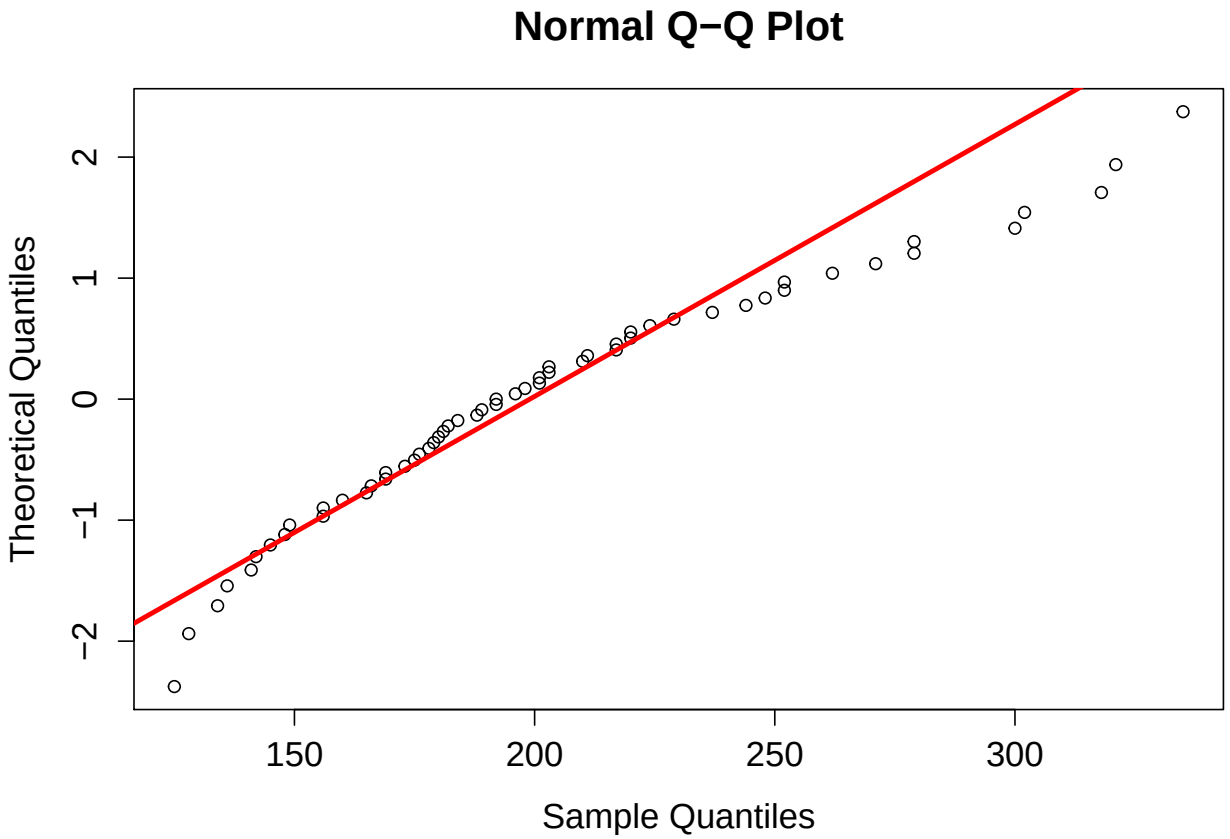
In the normal distribution with μ=204 and σ=52.28 we select the first 25% of the observations. The area ends at x=168.7377, which is rounded to 169.

Normal quantile plot

- A histogram or stem-and-leaf plot can reveal distinctly nonnormal features of a distribution.
- But we need a more sensitive way to judge the adequacy of a normal model. Most useful tool: **normal quantile plot**. Basic idea:
  - Arrange the observed data values from smallest to largest.
  - Record what percentile of the data each value occupies. For example, the smallest observation in a set of 20 is at the 5% point, the second smallest is at the 10% point, and so on.
  - Do normal distribution calculations to find the z-scores at these same percentiles. For example, z=-1.645 is the 5% point of the standard normal distribution, z=-1.282 is the 10% point, and so on.
  - Plot each data point x against the corresponding z.

Normal quantile plot

- If the data distribution is close to standard normal, the plotted points will lie close to the 45-degree line $x = z$.
- If the data distribution is close to any normal distribution, the plotted points will lie close to some straight line.
- Systematic deviations from a straight line indicate a nonnormal distribution.
- Outliers appear as points that are far away from the overall pattern of the plot.
- Right-skewed distribution: the largest observations fall distinctly above a line drawn through the main body of points. Similarly, left skewness is evident when the smallest observations fall below the line.



Normal quantile plot for the number of days of growth seasons. The distribution is a little bit left-skewed.