Samples and Data

Populations and Random Variables

A ***population*** is a large (possibly infinite) collection of individuals, objects or other instances which we want to describe probabilistically.

We assume that a population is described by a (scalar or multivariate) ***random variable*** in the following sense:

> For each member of the population, the random variable takes on a given, deterministic, value.
>
> The "randomness" of the random variable consists of the randomness of selecting an individual from the population.

Mathematically, we denote by $X = x$ the value of the random variable. Selecting a given member of the population and measuring X gives one instance of this random variable.

The probability density of X describes the likelihood of obtaining a value x within a given range.

Coin Flips

11.1. Example. Suppose we are flipping a coin with probability of heads p, $0 < p < 1$. The population might be

all flips of this coin conducted in the future

while the random variable X might be

$$X = \begin{cases} 1 & \text{coin turns heads up,} \\ 0 & \text{otherwise.} \end{cases}$$

Hence X follows a Bernoulli distribution with parameter p.

Each individual flip of the coin would represent an independent and identical copy of X.

We say that X describes the population, meaning that each member of the population gives an identical copy of X. The population size is indeterminate.

Student Height

11.2. Example. Suppose we are interested in the body height of the students of a certain university. Hence, our population might be described as

all students who were enrolled in the university in 2020.

The random variable X would be

the height (in cm) of the population.

It may well be that X is described by a normal distribution with certain mean and variance.

Each individual student would represent an independent and identical copy of X. The population size is possibly large, but is a well-defined number.

Again, the student height X describes the population, meaning that each student gives an identical copy of X.

Probability Theory vs. Statistics

Given a population and a random variable, probability theory and statistics are concerned with different questions:

Probability: The distribution of the random variable is fully known. What inferences can be drawn from the known information?

Statistics: The probability distribution is not known, but perhaps certain assumptions may be made. Data is gathered in order to make inferences on the distribution, e.g., its shape, expectation, variance etc.

In short: ***probability theory*** supposes one has complete knowledge of all parameters of a distribution, while ***statistics*** attempts to gain information on these parameters through experiments.

Probability Theory vs. Statistics

11.3. Examples.

(i) When considering coin flips, probability theory might answer the question: if $p = 1/2$, what is the likelihood of obtaining more than 60 heads when performing 100 coin flips?

A statistical question would be: If one obtains more than 60 heads in 100 coin flips, what can be said about p? Is there evidence that the coin is not fair? Can we give an interval $[p_0, p_1] \subset [0, 1]$ where we can be 90% sure that $p \in [p_0, p_1]$?

(ii) Given a student population whose height follows a normal distribution with mean μ and variance σ^2, probability theory would allow the calculation of the percentage of students whose height is above or below some threshold.

Statistics would attempt to gain information on μ and σ^2 by measuring the height of a certain number of students.

Random Sample (Mathematical Definition)

The remainder of this course is concerned with statistics, based on methods and techniques of probability theory.

The basis of all statistical approaches is a ***random sample***. The mathematical definition is straightforward:

11.4. Definition. A ***random sample of size n from the distribution of*** X is a collection of n independent random variables $X_1, \ldots, X_n$, each with the same distribution as X.

We say that $X_1, \ldots, X_n$ are independent, identically distributed (i.i.d.) random variables.

Each population member is an identical copy of X. A random sample comprises an independent selection of these copies.

Random Samples in Practice

Heuristically, a random sample is a subset of a population whose members have been selected in such a way, that the selection of one member does not influence the selection of any other member.

This means that each member of a random sample has been selected completely at random from the entire population and there is no ***bias*** in the selection.

For example, in obtaining a random sample of coin flips, one might just flip the coin. This is straightforward.

But to obtain a random sample of students enrolled at a university, it is not sufficient to just walk into a classroom for a course in, e.g., mathematics and measure the height of all students found there.

Question. Why is this?

Sample Size

We will generally discuss a random sample of size n from a population. How large should n be?

- The size n of a random sample should not be too small. However, a large population does not imply the need for a large random sample. In fact, given the need to make inferences to some specified degree of accuracy,

  the required minimum size of n is absolute,
  independent of the population size

- However, n should not be too large relative to the population size:

  If n is greater than 5% of the population,
  special care must be taken.

  We will suppose that our sample sizes are always smaller than 5% of the population.

Sample Sizes That Are Too Large

11.5. Example. Suppose that you are interested in a population of 100 students (e.g., all graduate students in a certain school). You wish to know what the proportion of students with body height greater than 180 cm is.

Suppose that (unknown to you), 20 out of 100 students have this height or greater. Suppose you take a sample of 50 students.

Then, using the hypergeometric distribution, we can calculate that there is 10% chance that the random sample includes 13 or more students with this height. Your guess for the proportion of students is then

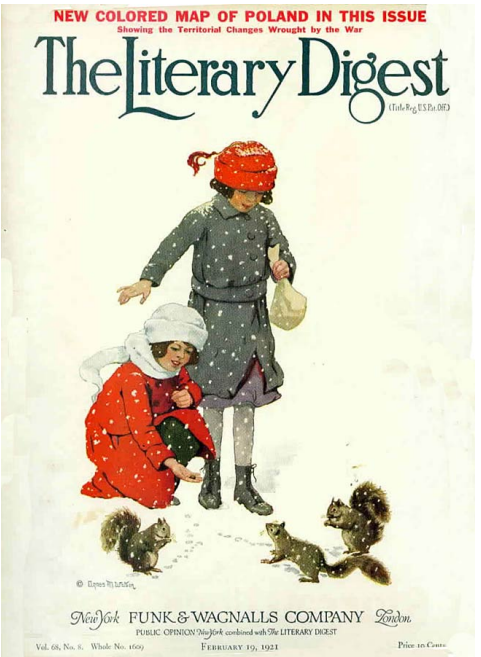
$$13/50 = 26\%.$$

However, in the remaining population, only $7/50 = 14\%$ actually have this height. The large sample has not only yielded a result that is different from the true proportion (that is to be expected in statistics), it has also ***perturbed the distribution of the remaining population***.

The 1936 US Presidential Election

In 1936, Alfred Landon, Republican and governor of Kansas, was seeking to unseat the incumbent president of the United States, Franklin D. Roosevelt.

The magazine ***The Literary Digest***, attempted to predict the outcome of the election a month before by conducting one of the biggest polls ever: Based on magazine subscription data, club membership lists and telephone directories, it queried more than 10,000,000 individuals, sending them a mock ballot.



Cover of the vol. 68, issue 8 (number 1609) of 19 February 1921 edition of the Literary Digest. File:LiteraryDigest-19210219.jpg. (2018, February 11). Wikimedia Commons, the free media repository.

It received 2.4 million responses and concluded that Landon would win, 57% – 43%.

However, it turned out that Roosevelt won, 62% – 38%. What had gone wrong?

Literature: https://www.math.upenn.edu/~deturck/m170/wk4/lecture/case1.html

Data

Suppose that we have obtained data from a random sample of size $n = 100$:

Data

```
{79, 141, 228, 3, 20, 14, 97, 194, 28, 56, 75, 37, 122, 27, 10,
 67, 23, 20, 103, 11, 92, 99, 64, 6, 118, 136, 682, 4, 70, 11,
 74, 40, 16, 114, 8, 149, 97, 7, 317, 346, 188, 149, 68, 150,
 88, 87, 155, 50, 26, 143, 126, 98, 153, 238, 30, 53, 132, 260,
 296, 25, 61, 87, 33, 51, 74, 111, 72, 178, 4, 67, 43, 229, 156,
 117, 104, 27, 23, 23, 186, 524, 107, 160, 41, 50, 352, 8, 153,
 142, 306, 320, 85, 44, 116, 39, 264, 360, 192, 142, 44, 29}
```

For most people this is just a "wall of numbers." The first step in statistical analysis is to ***understand and visualize the data***. FIn this section, we will use the above data for our examples.

Percentiles and Quartiles

We can characterize data by using ***percentiles***:

The xth percentile is defined as the value d_x of the data such that $x\%$ of the values of the data are less than or equal to d_x.

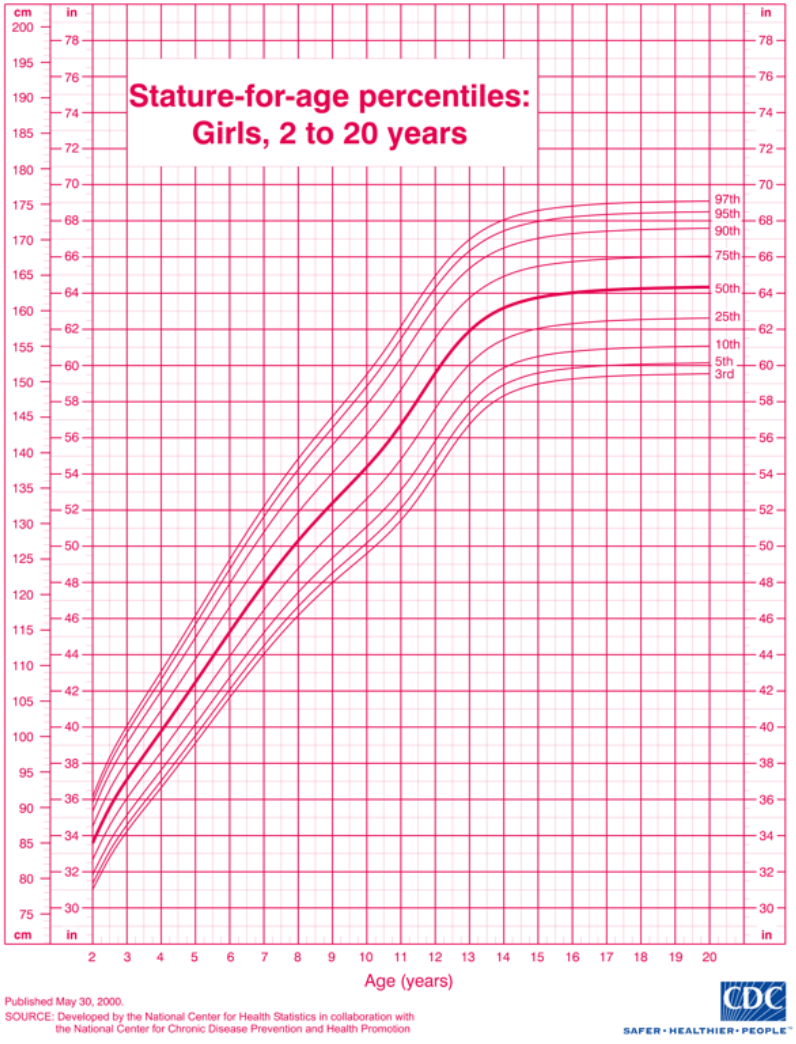
For instance, the 95th percentile is the datum such that 95% of the data is equal to or less than that value.

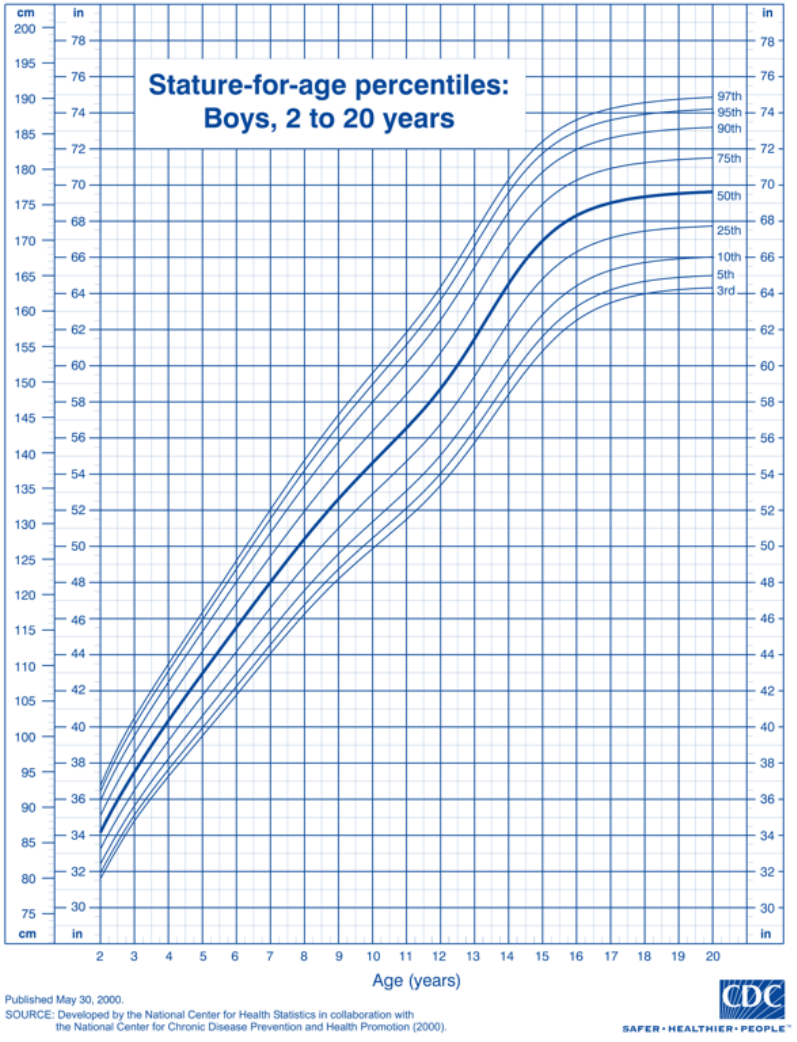
A special case are ***quartiles***:

- 25% of the data are no greater than the ***first quartile*** q_1,
- 50% are no greater than the ***second quartile*** q_2,
- 75% are no greater than the ***third quartile*** q_3.

The second quartile is also known as the ***median*** of the data.
(You may compare with the notion of the median of a continuous distribution, introduced previously).

Percentile Growth Curves for US American Children





Calculating Quartiles

Suppose that our list of n data has been ordered from smallest to largest, so that

$$x_1 \leq x_2 \leq x_3 \leq \cdots \leq x_n.$$

Then the median is given by

$$q_2 = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}) & \text{if } n \text{ is even} \end{cases}$$

Furthermore, the first quartile is defined as

- the median of the smallest $n/2$ elements if n is even.
- the average of the median of the smallest $(n-1)/2$ elements and the median of the smallest $(n+1)/2$ elements of the list if n is odd.

To calculate the third quartile, replace "smallest" with "largest" in the above definition.

Quartiles and Interquartile Range

Mathematica uses the definition of quartiles we have given here. Using the sample data shown before,

```
Quartiles[Data]
```

$$\left\{35,\ 87,\ \frac{299}{2}\right\}$$

The median (second quartile) is a measure of ***location*** of the data.

The difference between the third and first quartile is called the ***interquartile range***,

$$\mathrm{IQR} = q_3 - q_1,$$

and is a measure of ***dispersion*** of the data.

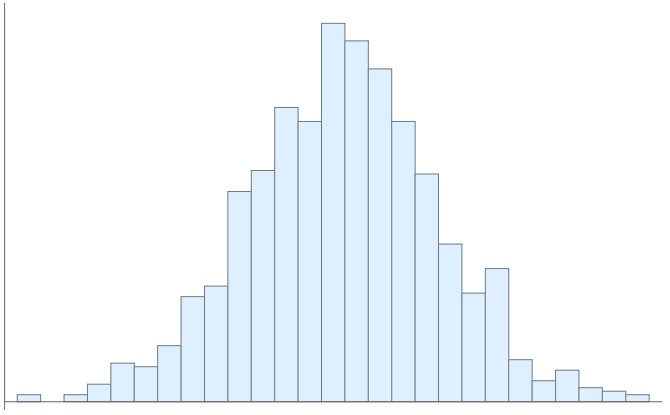
```
InterquartileRange[Data]
```

$$\frac{229}{2}$$

Histograms

A histogram is a (usually) vertical bar graph where each bar represents the (usually) the proportion or number of data in a given range.

The bars should show a rough silhouette of the underlying distribution's density function.



The histogram was first systematically introduced and analyzed by Karl Pearson.

Given data in a certain range, the first step is to select the number of categories, called ***bins***, and correspondingly the width of each bin.

- Too few bins: The shape of the distribution can not be clearly distinguished, important features will be "smoothed out."
- Too many bins: Individual bars are not supported by sufficiently many data points, spurious "features" may appear.

Number of Categories and Category Width

The traditional number of bins k is due to Sturges, which he proposed in 1926:

$$k = \lceil \log_2(n) \rceil + 1, \tag{11.1}$$

where the ***ceiling*** $\lceil x \rceil$ denotes the smallest integer greater than $x \in \mathbb{R}$.

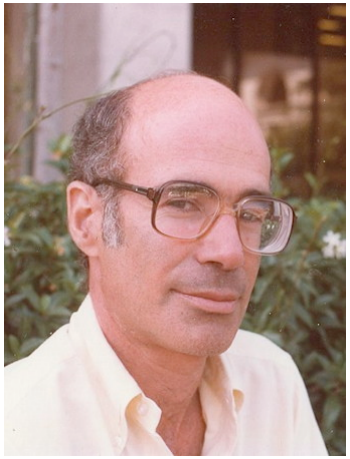
Sturges's rule is popular because it is simple and was based on one of the first serious analyses of this question. However, his derivation is flawed and the rule results in overly smoothed histograms for large n. Hence, various alternatives have been proposed.

We remark that the software Microsoft Excel uses the rule

$$k = \left\lceil \sqrt{n} \right\rceil.$$

Instead of the number k of categories, we can also fix the ***bin width h***.

The Freedman-Diaconis Rule



David A. Freedman (1938-2008). File:David A Freedman (statistician) 1984.jpg. (2018, October 23). Wikimedia Commons, the free media repository.

Persi W. Diaconis (1945-). File:Persi Diaconis 2010.jpg. (2014, April 18). Wikimedia Commons, the free media repository.

Among the various improvements that have been suggested, we will use the Freedman-Diaconis rule for the bin width h, which was presented in a publication in 1981.

The rule is designed to minimize the difference between the actual density of the distribution of the data and the height of the bars.

More precisely, if data of size n from the distribution of (X, f_X) is gathered on an interval I, then h should be chosen so that

$$\delta^2(h) = \mathrm{E}\left[\int_I |H(x) - f_X(x)|^2 \, dx\right]$$

is minimized, where H is the normalized height of the histogram bars.

The Freedman-Diaconis Rule

Analysis shows that this is realized if the bin width is

$$h \sim \frac{1}{\sqrt[3]{n}} \qquad \text{as } n \to \infty.$$

According to Freedman and Diaconis, numerical calculations show that

$$h = \frac{2 \cdot \mathrm{IQR}}{\sqrt[3]{n}}$$

yields good results for the estimation of the true density f_X from the histogram.

Literature: Freedman, D., Diaconis, P. On the histogram as a density estimator: L_2 theory. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* 57, 453–476 (1981). https://doi.org/10.1007/BF01025868

Determining the Bin Widths

The ***precision*** of the data $\{x_1, \ldots, x_n\}$ is the smallest decimal place of the values x_i.

The ***sample range*** is given by

$$\max_{1 \le i \le n} \{x_i\} - \min_{1 \le i \le n} \{x_i\}.$$

If the number of bins k has been determined (e.g., by Sturges's rule), then the bin width is calculated as

$$h = \frac{\max\{x_i\} - \min\{x_i\}}{k},$$

which should be rounded up to the precision of the data. If h is already at the precision of the data, one smallest decimal unit should be added to h.

If the bin width has been determined (e.g., by the Freedman-Diaconis rule), then nothing else needs to be done.

Binning the Data

Next, the actual bins need to be determined. Ideally, the bins should have the properties that

- The bins represent the data range well and do not go too far beyond it.
- Each datum should fall into exactly one bin.
- The bins should have the same width (in our approach here).

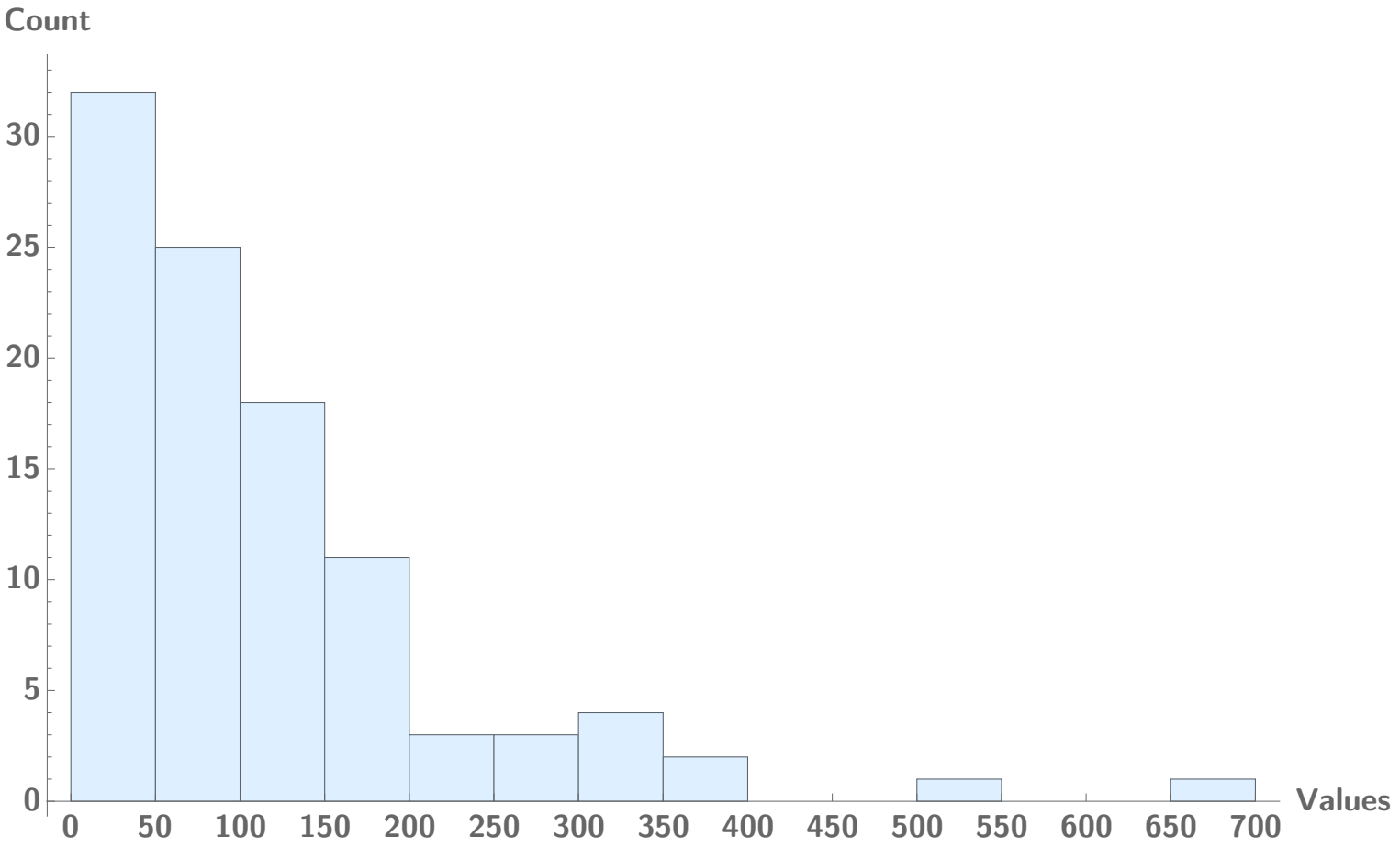
To achieve this, the ideal way is to take the smallest datum, subtract ***one-half of the smallest decimal of the data*** and then successively add the bin width to obtain the bins.

Since the bin boundaries are now at a higher precision than the data, no datum can lie on the boundary. The rounding up of the bin widths (if determined as above) will ensure that the data range is covered.

In practice, however, one often chooses "nice" values as bin boundaries.

Histogram (Freedman-Diaconis Bin Widths)

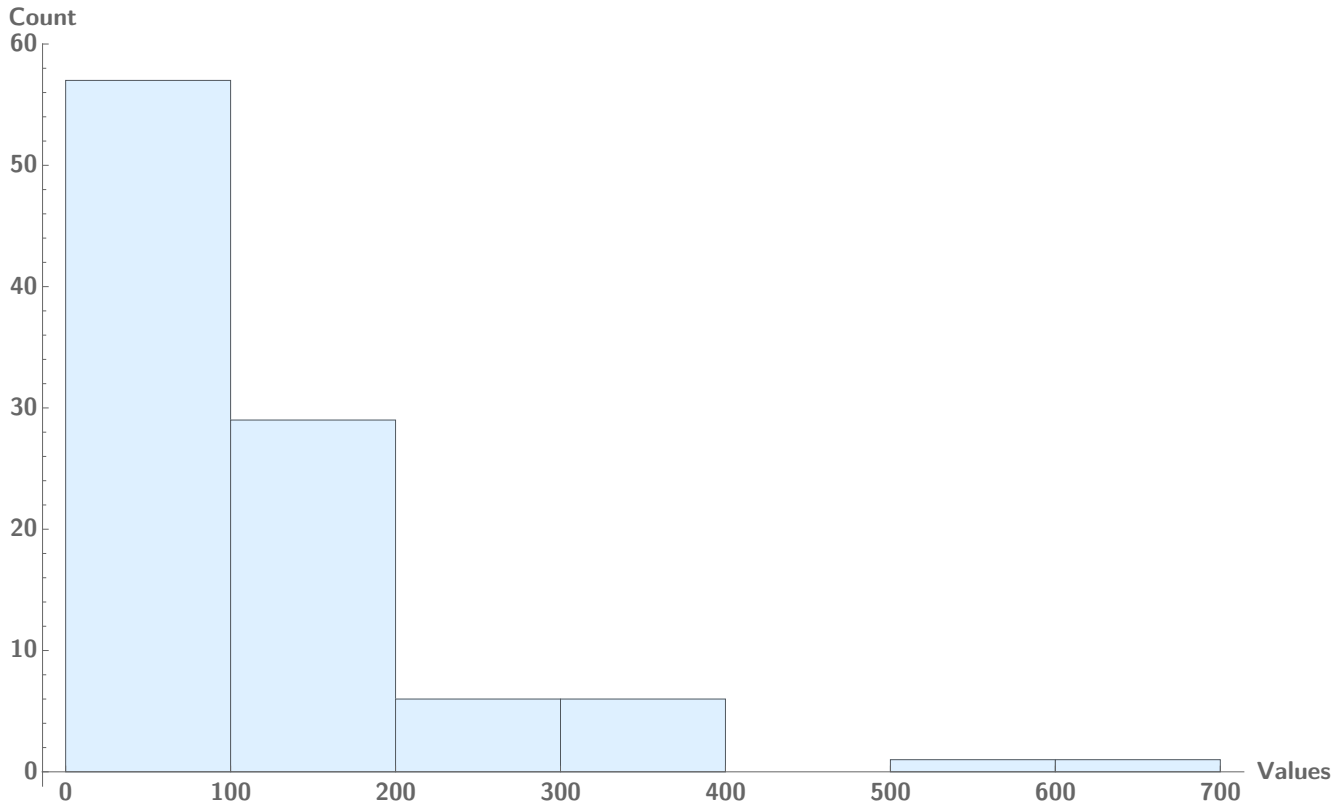
In our example, we have $\frac{2 \cdot \text{IQR}}{\sqrt[3]{n}} = 49.34$, which we round up to 50.



Mathematica: `Histogram[Data, "FreedmanDiaconis"]`

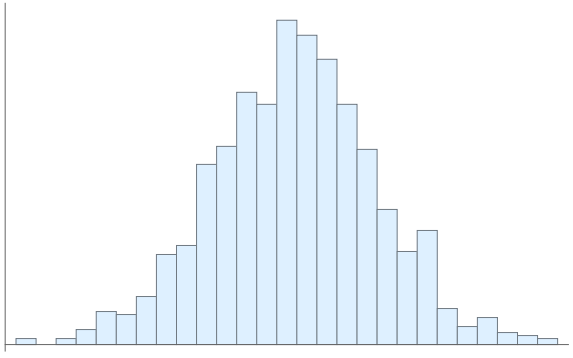
Histogram (Sturges's Rule Bin Number)

The data range is $682 - 3 = 679$ and Sturges's rule (based on 100 data) gives $k = 7$. We calculate $679/7 = 97$, which should be rounded up by one to $h = 98$.

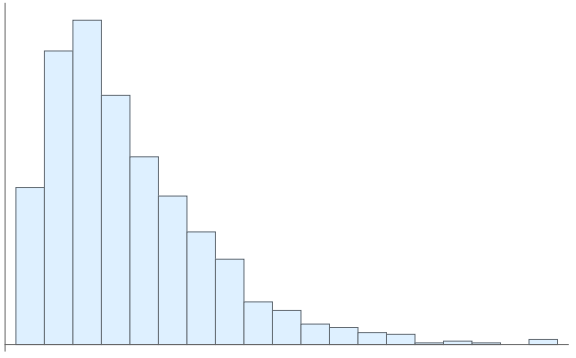


Mathematica: `Histogram[Data, "Sturges"]`

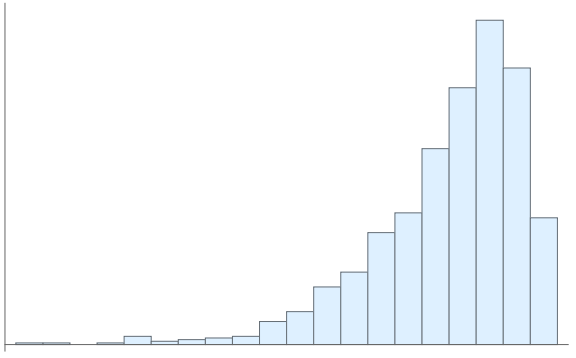
Describing a Histogram



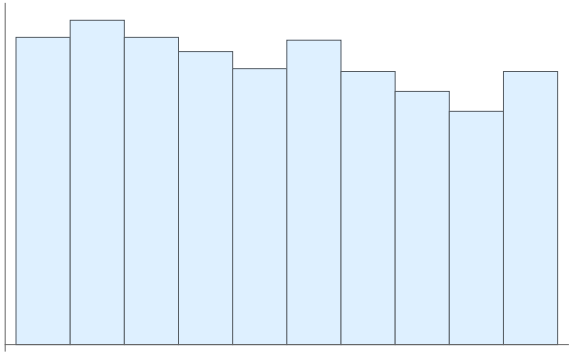
Symmetric, unimodal



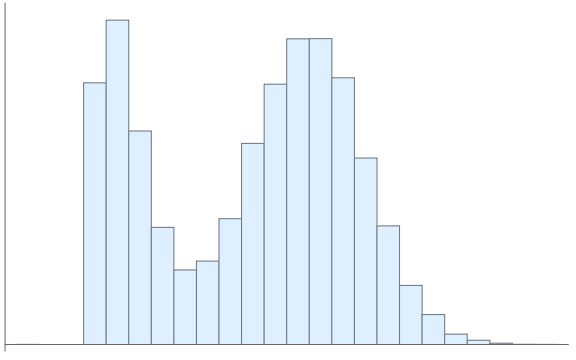
Positive skew, unimodal



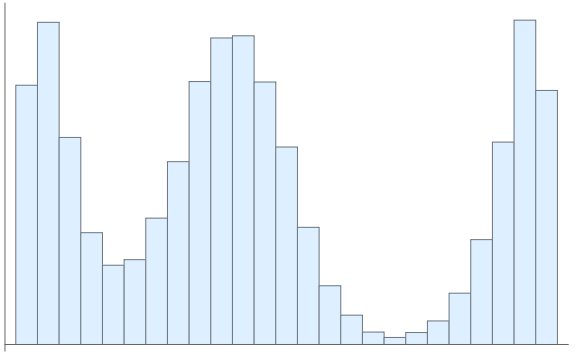
Negative skew, unimodal



Symmetric, no prominent mode



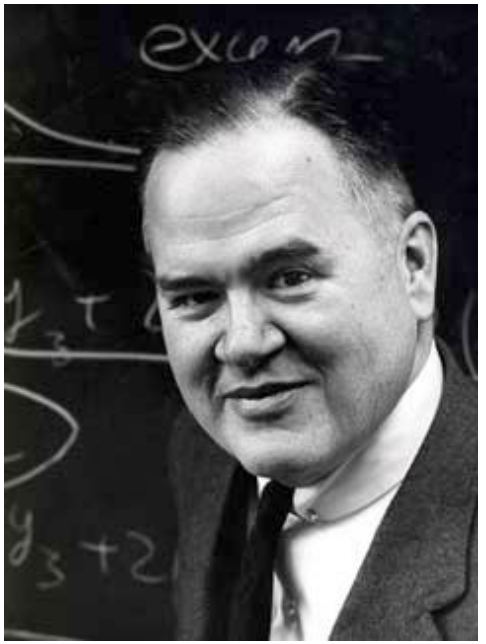
Bimodal



Multimodal

Stem-and-Leaf Diagrams

A ***stem-and-leaf diagram*** is a rough way to get an idea of the shape of the distribution of a random sample, while preserving some of its numeric information. It consists of labeled rows of numbers, where the label is called the stem and the other numbers are called leaves. This idea was introduced by Tukey in bis famous book ***Exploratory data Analysis*** in 1977.



John W. Tukey (1915-2000).
http://1stmuse.com/the_term_software/

To construct a stem-and-leaf diagram from a random sample, follow these steps:

(i) Choose a convenient number of leading decimal digits to serve as stems,
(ii) label the rows using the stems,
(iii) for each datum of the random sample, note down the digit following the stem in the corresponding row,
(iv) turn the graph on its side to get an impression of its distribution.

Stem-and-Leaf Diagrams

We will continue to use the data of Slide 12.

The package `StatisticalPlots` includes a command for stem-and-leaf plots:

```
Needs["StatisticalPlots`"]
```

```
StemLeafPlot[Floor[Data, 10], IncludeEmptyStems → True]
```

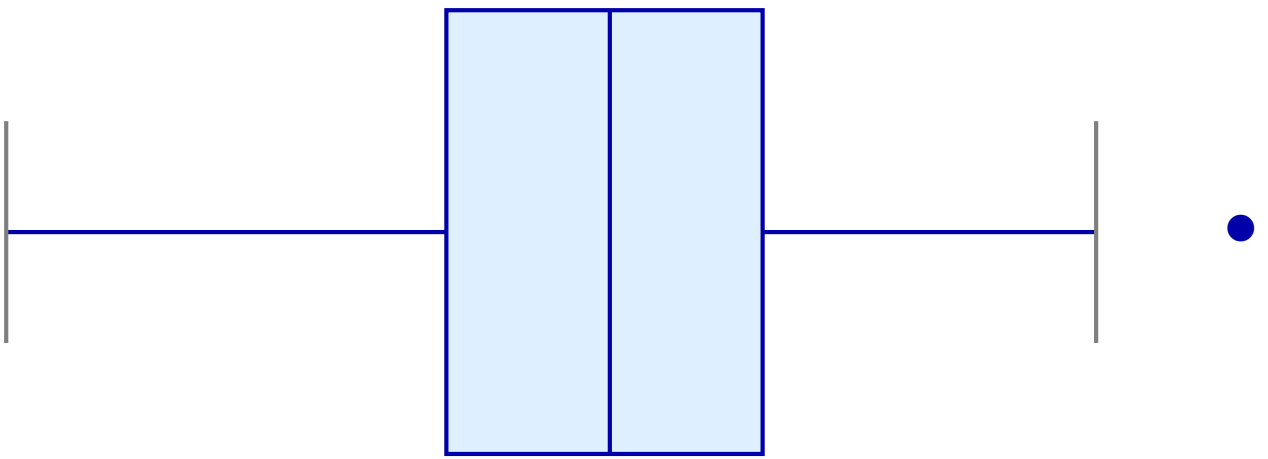
Stem	Leaves
0	000000011111222222222233334444455555666667777788899999
1	000111112233444444555556788 99
2	223669
3	012456
4	
5	2
6	8

Stem units: 100

Box-and-Whisker Plots (Boxplots)

A ***boxplot*** is a representation of data that is useful for checking for symmetry or skew and, in general, deviation of the data from that expected of a normal distribution. Boxplots were also introduced by Tukey in his 1977 book.

This is their general appearance:



Construction of Boxplots

A boxplot is drawn on an abscissa scale of values corresponding to the data. Often, the abscissa scale is not shown.

The central box has a center line, located at the median q_2, while the left and right sides of the box are located at the first and third quartiles q_1 and q_3, respectively.

We define the ***inner fences*** f_1 and f_2 using the interquartile range as follows:

$$f_1 = q_1 - \frac{3}{2}\,\mathrm{IQR}, \qquad f_3 = q_3 + \frac{3}{2}\,\mathrm{IQR}\,.$$

The "whiskers" (lines extending to the left and right of the box) end at the ***adjacent values***

$$a_1 = \min\{x_k : x_k \geq f_1\}, \qquad a_3 = \max\{x_k : x_k \leq f_3\}.$$

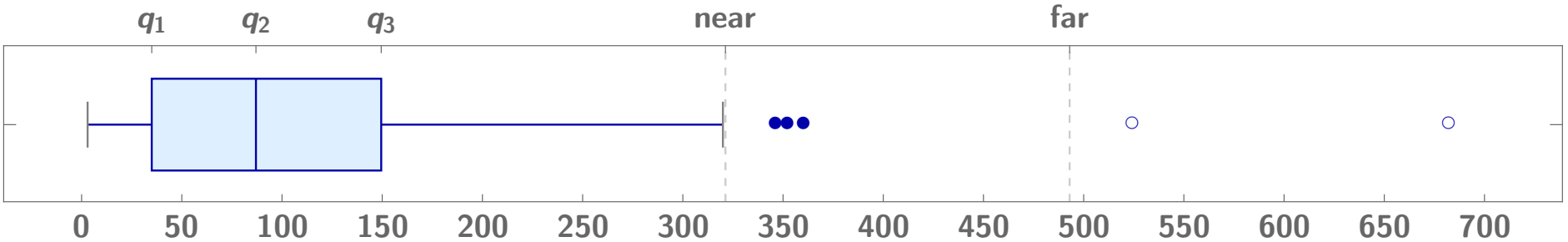
Construction of Boxplots

We define the ***outer fences***

$$F_1 = q_1 - 3\,\mathrm{IQR}, \qquad F_3 = q_3 + 3\,\mathrm{IQR}\,.$$

Measurements x_k that lie outside the inner fences but inside the outer fences are called ***near outliers***. Those outside the outer fences are known as ***far outliers***.

A boxplot generated from the example data of Slide 12 is shown below:



Interpreting Boxplots

If data is obtained from a normal distribution, one would expect to see

- a symmetric median line in the middle of the box;
- equally long whiskers;
- very few near outliers and no far outliers.

A rule of thumb states that:

> Of 1000 random samples of a normally distributed population, it can be expected that 7 will be outliers.

Data points lying between the inner and outer fences are called ***near outliers***, those lying outside the outer fences are called ***far outliers***. Far outliers are unusual if (and only if!) an approximately bell-shaped distribution of the random variable X of the population is expected. In this case, their origin should be investigated.

- If the outlier seems to be the result of an error in measurement or data collecting, it may be discarded from the data.
- If the outlier seems to be the result of a random measurement, it is recommended that statistics are reported twice: with the outlier included ***and*** without the outlier.

Interpreting Boxplots

A set of 10 data yields the following boxplot:

Which of the following sentences is the ***most appropriate*** conclusion?

1) There is no strong evidence that the data does not follow a normal distribution.
2) There is no strong evidence that the data follows a normal distribution.
3) There is strong evidence that the data does not follow a normal distribution.
4) There is strong evidence that the data follows a normal distribution.