Data Analysis in Astronomy and Physics

Lecture 3: Statistics

M. Röllig

Statistics

- Statistics are designed to summarize, reduce or describe data.
- *A statistic is a function of the data alone!*
- Example statistics for a set of data X_1, X_2, … are: **average**, the **maximum value**, **average of the squares**,...
- Statistics are combinations of finite amounts of data.
- Example summarizing examples of statistics: **location** and **scatter**.

Location

Average (Arithmetic Mean)

Out[•]//TraditionalForm=

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N}$$

Example:

In[•]:=
```
Mean[{1, 2, 3, 4, 5, 6, 7}]
Mean[{1, 2, 3, 4, 5, 6, 7, 8}]
Mean[{1, 2, 3, 4, 5, 6, 7, 100}]
```

Out[•]= 4

Out[•]= $\frac{9}{2}$

Out[•]= 16

Location

Weighted Average (Weighted Arithmetic Mean)

Out[ ]//TraditionalForm=

$$\overline{X_w} = \frac{\sum_{i=1}^{N} w_i X_i}{\sum_{i=1}^{N} w_i}$$

In case of equal weights $w_i = w$: $\overline{X_w} = \frac{1}{\sum w_i} \sum w_i X_i = \frac{1}{\sum w} \sum \mathrm{w} X_i = \frac{1}{w N} w \sum X_i = \frac{1}{N} \sum X_i = \overline{X}$

Example: data: x_i, weights : $w_i = \frac{1}{x_i}$

In[ ]:=
```
x1 = {1., 2, 3, 4, 5, 6, 7}; w1 = 1 / x1; x1.w1 / Total[w1]
x2 = {1., 2, 3, 4, 5, 6, 7, 8}; w2 = 1 / x2; x2.w2 / Total[w2]
x3 = {1., 2, 3, 4, 5, 6, 7, 100}; w3 = 1 / x3; x3.w3 / Total[w3]
```

Out[ ]= 2.69972

Out[ ]= 2.9435

Out[ ]= 3.07355

Location

Median

Arrange X_i according to size and renumber. Then

Out[ ]//TraditionalForm=

$$X_{\text{med}} = \begin{cases} X_j & j = \frac{\text{N}}{2} + 0.5 \text{ where N odd} \\ 0.5\,(X_j + X_{j+1}) & j = \frac{\text{N}}{2} \text{ where N even} \end{cases}$$

Example:

In[ ]:=
```
Median[{1, 2, 3, 4, 5, 6, 7}]
Median[{1, 2, 3, 4, 5, 6, 7, 8}]
Median[{1, 2, 3, 4, 5, 6, 7, 100}]
```

Out[ ]= 4

Out[ ]= $\frac{9}{2}$

Out[ ]= $\frac{9}{2}$

Location

Weighted Median

For N distinct, ordered elements X_i with positive weights w_i such that $\sum w_i = 1$ the weighted median is the element k satisfying

Out[ ]//TraditionalForm=

$$\sum_{i=1}^{k-1} w_i \leq \frac{1}{2}, \text{ and } \sum_{i=k+1}^{N} w_i \leq \frac{1}{2}$$

If two elements satisfy the conditions then:

Lower weighted median

Out[ ]//TraditionalForm=

$$\sum_{i=1}^{k-1} w_i < \frac{1}{2}, \text{ and } \sum_{i=k+1}^{N} w_i = \frac{1}{2}$$

Upper weighted median

Out[ ]//TraditionalForm=

$$\sum_{i=1}^{k-1} w_i = \frac{1}{2}, \text{ and } \sum_{i=k+1}^{N} w_i < \frac{1}{2}$$

Location

Mode

X_{mode} is the value of X_i occurring most frequently; it is the location of the peak of the histogram of X_i

Example

In[•]:=
```
Commonest[{1, 2, 3, 4, 5, 6, 7}]
Commonest[{1, 2, 3, 4, 5, 6, 7, 8}]
Commonest[{1, 2, 3, 4, 5, 6, 7, 100}]
Commonest[{1, 2, 3, 4, 5, 6, 7, 3}]
```

Out[•]= {1, 2, 3, 4, 5, 6, 7}

Out[•]= {1, 2, 3, 4, 5, 6, 7, 8}

Out[•]= {1, 2, 3, 4, 5, 6, 7, 100}

Out[•]= {3}

Location

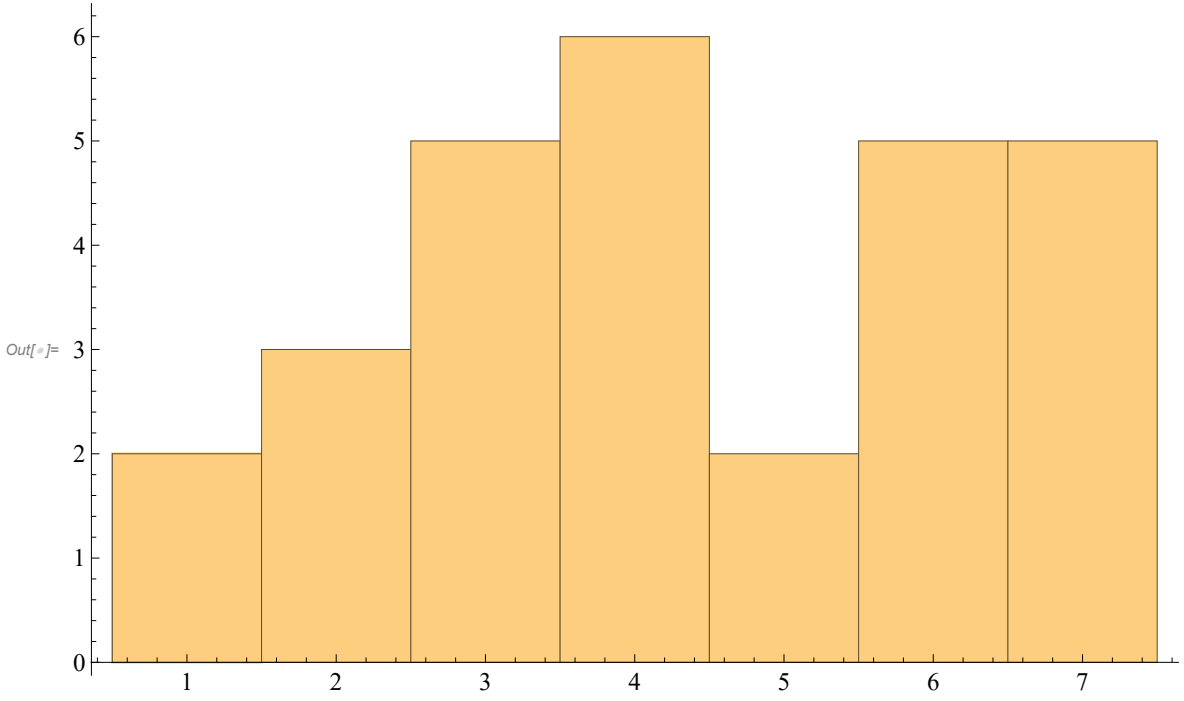
Mode

Example

In[ ]:=
```
Commonest[{1, 1, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 5, 5, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7}]
Histogram[{1, 1, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 5, 5, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7}, {1}]
```

Out[ ]= {4}

Out[ ]=



Location

Harmonic Mean

The harmonic mean is appropriate for situations when the **average of rates or ratios** is desired. Not defined if the data contains 0! The harmonic mean is the reciprocal arithmetic mean of the reciprocals

Out[•]//TraditionalForm=

$$\overline{X}_{\text{harmonic}} = \frac{1}{\frac{\sum_{i=1}^{N} \frac{1}{X_i}}{N}}$$

Example:

```
HarmonicMean[{a, b, c, d}]
```

$$\frac{4}{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

```
HarmonicMean[{1, 2, 3, 4, 5, 6, 7}] // N
HarmonicMean[{1, 2, 3, 4, 5, 6, 7, 8}] // N
HarmonicMean[{1, 2, 3, 4, 5, 6, 7, 100}] // N
```

```
2.69972
```

```
2.9435
```

```
3.07355
```

Location

Harmonic Mean - Example

Imagine a person who can rake the yard in three hours and another that can rake the yard in two hours. How long odes it take for both persons to do the job together?

Solution

One way to solve this is to look at the speed of each person. Person A needs 2 hrs, therefore he/she finishes 1/2 of the job per hour. Person B finishes 1/3 of the job per hour. Together they are finishing $\frac{1}{3} + \frac{1}{2} = \frac{5}{6}$ of the job / hour. The reciprocal gives the hours / job, i.e. together they need 6/5 hrs = 1 hr and 12 minutes. Looking at the formula:

The harmonic mean is the mean working time each person would need to finish the job in the same time (=6/5 hrs)

In[•]:=
```
HarmonicMean[{2, 3}]
```

Out[•]=
$$\frac{12}{5}$$

In[•]:=
```
(5 / 12 + 5 / 12)^-1
```

Out[•]=
$$\frac{6}{5}$$

Location

Harmonic Mean - Application

In a parallel circuit the total resistance is computed as:

Out[•]//TraditionalForm=

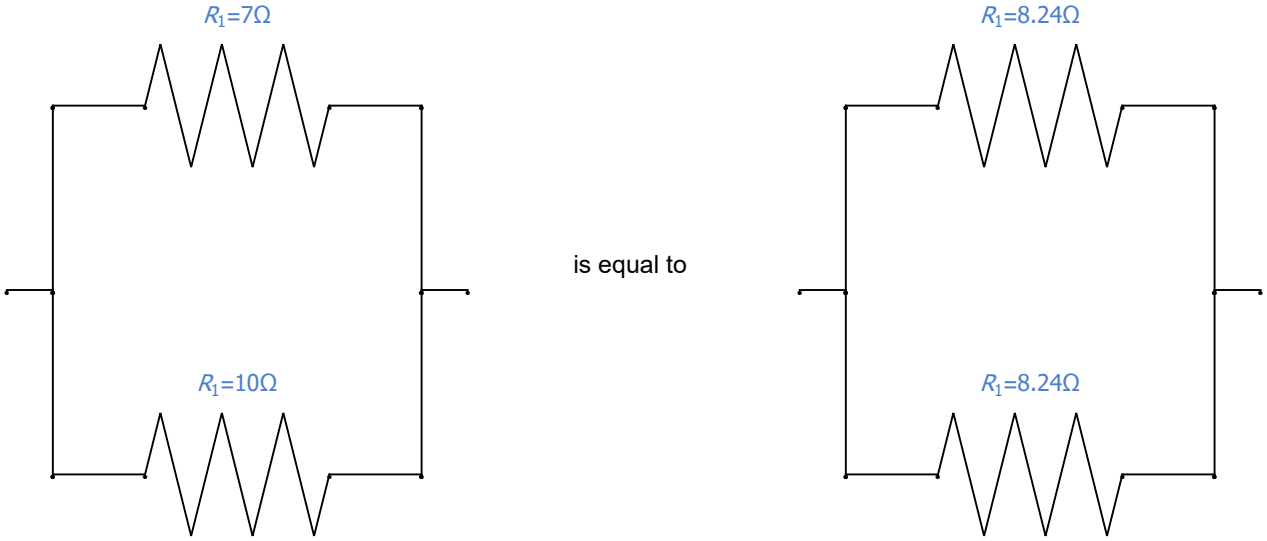
$$\frac{1}{R_{\text{tot}}} = \frac{1}{R_1} + \frac{1}{R_2}$$

Imagine R_1 = 7 ohms and R_2 = 10 ohms. Then R_{tot} = 4.12 ohms. The same total resistance would be achieved with two equal resistors of R = HarmonicMean[{7, 10.}] = 8.23529 ohms.

In[•]:=
```
(HarmonicMean[{7, 10.}]^-1 2)^-1
```

Out[•]=
```
4.11765
```

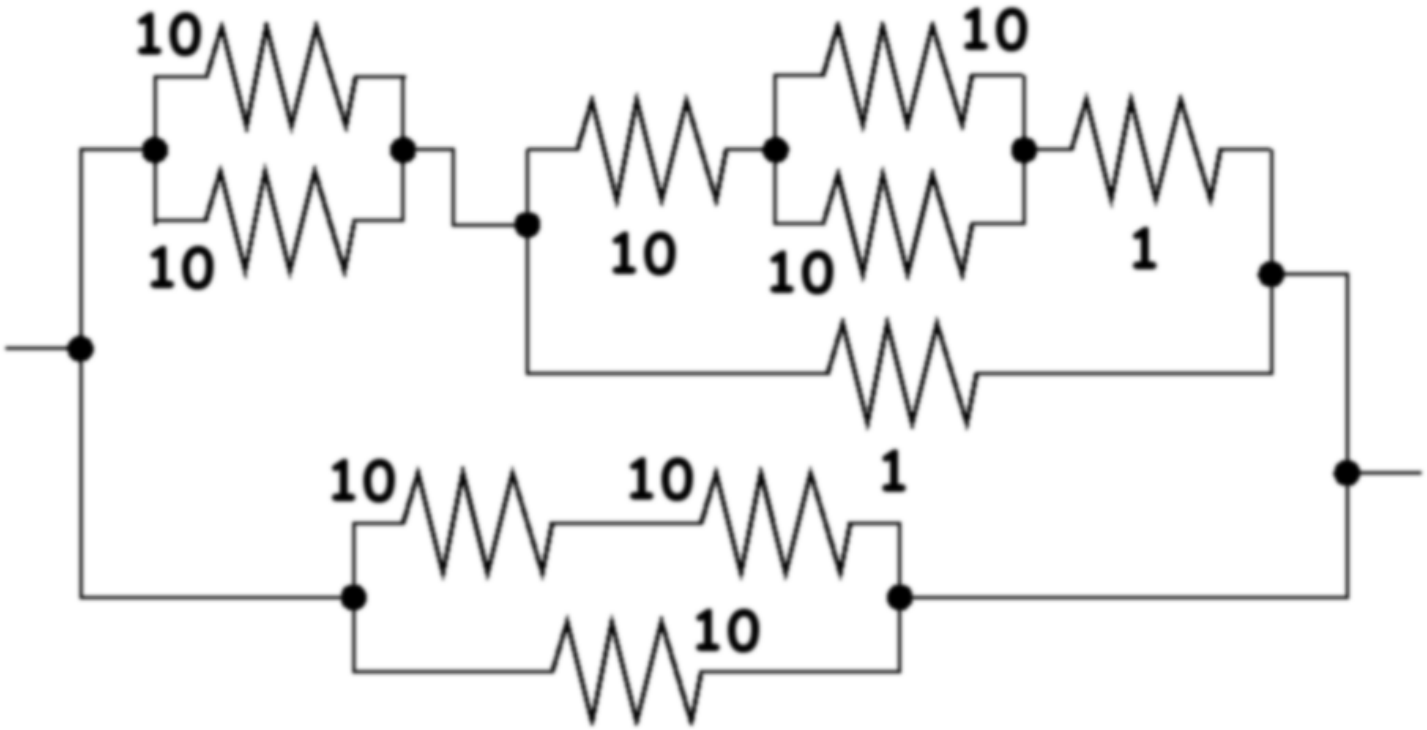
Out[•]=



Location

Harmonic Mean - Application

Task: create a circuit with a resistance of 3.14159 ohms using only 10 ohm and 1 ohm resistors:



Source: https://occupymath.wordpress.com/2019/07/11/working-together-the-harmonic-mean/

Quiz

1. Harmonic mean is particularly useful in averaging ---------- types of data.

a) Ratios
b) Rates
c) Both a & b
d) None

2. Harmonic mean gives more weight to --------- values

a) Smaller values
b) Larger values
c) Both the values
d) None of the above

3. Harmonic mean is particularly useful in computing ----------

a) Average speed
b) Average price
c) Average rate
d) All the above

Location

Geometric Mean

Out[ ]//TraditionalForm=

$$\overline{X}_{\text{geom}} = \left(\prod_{i=1}^{N} X_i\right)^{1/N}$$

The geometric mean of two numbers, a and b, is the length of one side of a square whose area is equal to the area of a rectangle with sides of lengths a and b.
Not defined if the data contains 0 or negatives!

Example:

```
GeometricMean[{a, b, c, d}]
```

$(a\ b\ c\ d)^{1/4}$

```
GeometricMean[{1, 2, 3, 4, 5, 6, 7}] // N
GeometricMean[{1, 2, 3, 4, 5, 6, 7, 8}] // N
GeometricMean[{1, 2, 3, 4, 5, 6, 7, 100}] // N
```

```
3.38002
```

```
3.76435
```

```
5.16183
```

Location

Geometric Mean

By using logarithmic identities to transform the formula, the multiplications can be expressed as a sum and the power as a multiplication.

Out[•]//TraditionalForm=

$$\overline{X}_{\text{geom}} = \left(\prod_{i=1}^{\text{N}} X_i\right)^{1/\text{N}} = \exp\left(\frac{\sum_{i=1}^{\text{N}} \log(x_i)}{\text{N}}\right)$$

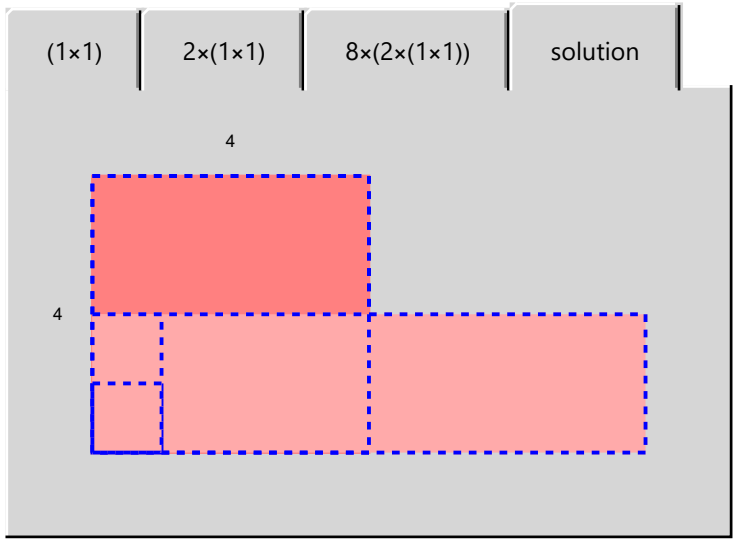
Quiz

A bacterial culture is growing. Its size doubles first and then grows again by a factor of 8. What is its mean rate of growth?

a) by a factor of 3
b) by a factor of 4
c) by a factor of 5
d) by a factor of 6

Solution

Out[ ]=



Location

Geometric Mean

An important property of the geometric mean (and only for the geometric mean) is that

Out[ ]//TraditionalForm=

$$\frac{\overline{X}}{\overline{Y}_{\text{geom}}} = \frac{\overline{X}_{\text{geom}}}{\overline{Y}_{\text{geom}}}$$

This makes the geometric mean the **only correct mean when averaging normalized results**, that is results that are presented as ratios to reference values. The next example demonstrates this:

Example from Wikipedia

Take the following comparison of execution time of computer programs:

Out[ ]=

	Computer A	Computer B	Computer C
Program 1	1	10	20
Program 2	1000	100	20
Arithmetic mean	500.5	55.	20.
Geometric mean	31.6228	31.6228	20.

The arithmetic and geometric means “agree” that computer C is the fastest. However, by presenting appropriately normalized values and using the arithmetic mean, we can show either of the other two computers to be the fastest. Normalizing by A’s result gives A as the fastest computer according to the arithmetic mean

Example from Wikipedia

Normalizing by A's result gives A as the fastest computer according to the arithmetic mean

Out[ ]=

	Computer A	Computer B	Computer C
Program 1	1.	10.	20.
Program 2	1.	0.1	0.02
Arithmetic mean	1.	5.05	10.01
Geometric mean	1.	1.	0.632456

while normalizing by B's result gives B as the fastest computer according to the arithmetic mean:

Out[ ]=

	Computer A	Computer B	Computer C
Program 1	0.1	1.	2.
Program 2	10.	1.	0.2
Arithmetic mean	5.05	1.	1.1
Geometric mean	1.	1.	0.632456

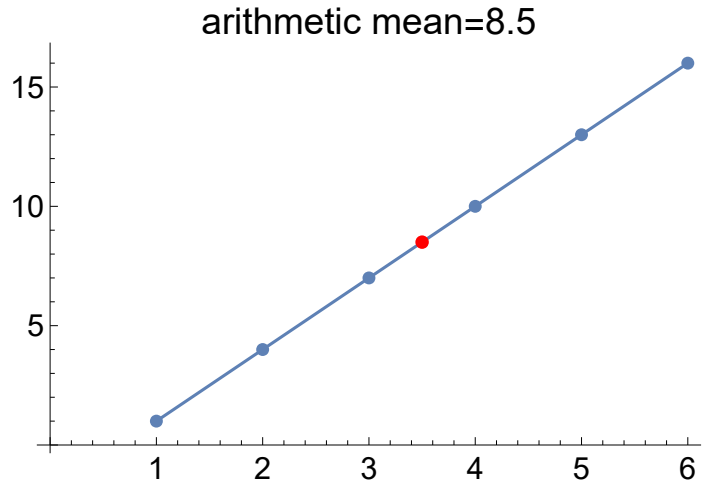
In all cases, the ranking given by the geometric mean stays the same as the one obtained with unnormalized values.

Which one is the correct mean?

Adding a constant (3) number:

In[ ]:=
```
additiveData = {1, 4, 7, 10, 13, 16.};
```

Out[ ]=

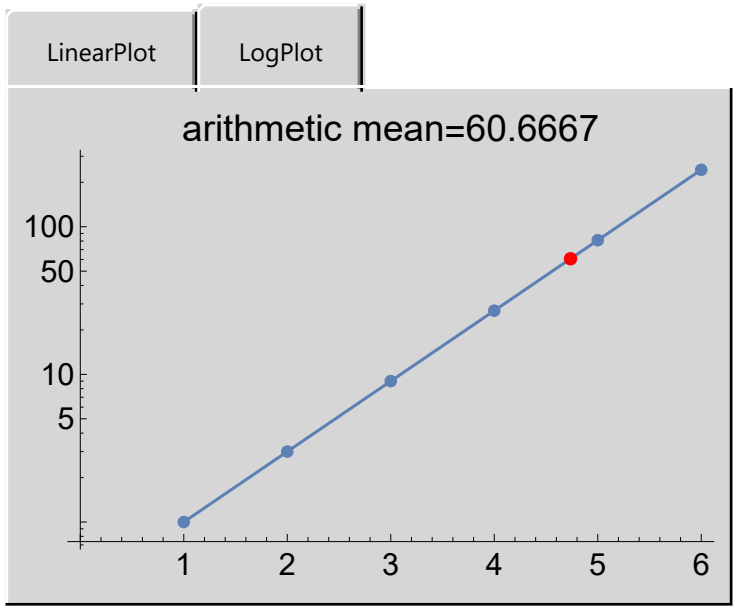


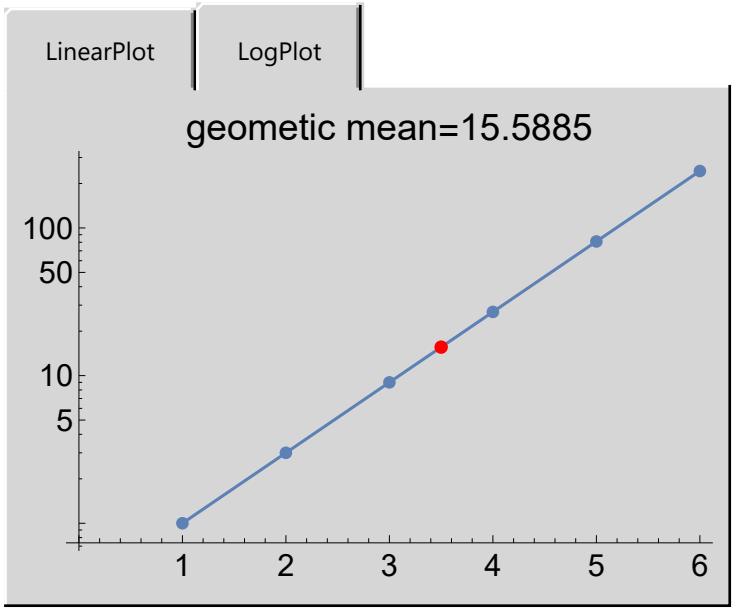
Which one is the correct mean?

Multiplying with a constant factor 3

In[ ]:=
```
multiplicativeData = {1, 3, 9, 27, 81, 243.};
```

Out[ ]=





Location

Percentile

The p-th percentile of an ordered data set is t**he value that has at most p% of the data below it** and at most (100-p)% above it.

Example:

Find the 67th percentile of the data set: {34,46,22,35,46,41,60,47,46,41,49,54,25,59,54}

Order the data {22,25,34,35,41,41,46,46,46,47,49,54,54,59,60}

Find 67% of 15 data points: 15×67/100=10.05, rounded up = 11

Find the 11th data point:

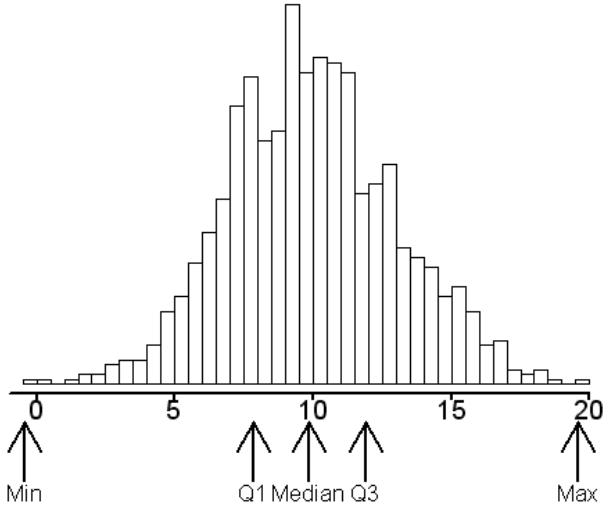
Out[ ]//TableForm=

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
22	25	34	35	41	41	46	46	46	47	49	54	54	59	60

So the 67th percentile is 49.

Location

Quartiles, Inter Quartile Range



Definition

1st quartile Q_1 = 25th percentile
3rd quartile Q_3 = 75th percentile

IQR = $Q_3 - Q_1$

Quiz

Medians and IQRs.

For each part, compare distributions (1) and (2) based on their medians and IQRs. You do not need to calculate these statistics; simply state how the medians and IQRs compare. Make sure to explain your reasoning.

(a) **(1)** 3, 5, 6, 7, 9
(2) 3, 5, 6, 7, 20

(b) **(1)** 3, 5, 6, 7, 9
(2) 3, 5, 8, 7, 9

(c) **(1)** 1, 2, 3, 4, 5
(2) 6, 7, 8, 9, 10

(d) **(1)** 0, 10, 50, 60, 100
(2) 0, 100, 500, 600, 1000

Spread or Scatter

Standard Deviation

The standard deviation (represented by the Greek letter sigma, σ) shows:

- how much variation or dispersion from the average exists.
- A low standard deviation indicates that the data points tend to be very close to the mean (also called expected value).
- A high standard deviation indicates that the data points are spread out over a large range of values.

Out[ ]//TraditionalForm=

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}} \text{ , where } \mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

Spread or Scatter

Standard Deviation

If, instead of having equal probabilities, the values have different probabilities, let x_1 have probability p_1, x_2 have probability p_2, ..., x_N have probability p_N. In this case, the standard deviation will be

Out[•]//TraditionalForm=

$$S = \sigma = \sqrt{\sum_{i=1}^{N} p_i \, (X_i - \mu)^2} \text{ , where } \mu = \sum_{i=1}^{N} p_i \, X_i$$

Spread or Scatter

Standard Deviation

Accordingly, the standard deviation of a continuous real-valued random variable X with probability density function $f(x)$ is

Out[•]//TraditionalForm=

$$S = \sigma = \sqrt{\int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx}\ , \text{where } \mu = \int_{-\infty}^{\infty} x f(x)\, dx$$

Another way to write this is in terms of the expectation value E

Out[•]//TraditionalForm=

$$S = \sigma = \sqrt{\mathrm{E}\left((X-\mu)^2\right)}\ , \text{where } \mathrm{E}(X) = \mu$$

Spread or Scatter

Variance

The variance is the squared standard deviation, or the expectation value of the squared deviation

Out[•]//TraditionalForm=

$$S^2 = \mathrm{var}(X) = \sigma^2 = \mathrm{E}\left((X-\mu)^2\right), \text{ where } \mathrm{E}(X) = \mu$$

or given the expectation value $\mu = \int x f(x)\, dx$

Out[•]=

$$S^2 = \mathtt{var}(X) = \sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$$

$$S^2 = \mathtt{var}(X) = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx$$

Example:

Out[•]=

data	mean	variance
{−10, 0, 10, 20, 30}	10	200
{8, 9, 10, 11, 12}	10	2

Spread or Scatter

Weighted Sample Variance

When a weighted mean μ_w is used, the variance of the weighted sample is different from the variance of the unweighted sample.

Out[ ]//TraditionalForm=

$$\sigma_w = \frac{1}{\sum_{i=1}^{N} w_i} \sqrt{\left(\sum_{i=1}^{N} w_i \, (X_i - \mu_w)^2\right)}, \text{ where } \mu_w = \frac{\sum_{i=1}^{N} w_i X_i}{\sum_{i=1}^{N} w_i}$$

Spread or Scatter

Properties of the variance

The variance is non-negative

Out[ ]//TraditionalForm=

$$\mathrm{Var}(X) \geq 0$$

The variance is invariant with respect to changes in the location parameter

Out[ ]//TraditionalForm=

$$\mathrm{Var}(X + a) = \mathrm{Var}(X)$$

If values are scaled by a constant, the variance is scaled by the square of that constant

Out[ ]//TraditionalForm=

$$\mathrm{Var}(a\,X) \geq a^2\,\mathrm{Var}(x)$$

If values X_i are statistically independent, i.e. uncorrelated:

Out[ ]//TraditionalForm=

$$\mathrm{Var}\left(\sum_i X_i\right) = \sum_i \mathrm{Var}(X_i)$$

Spread or Scatter

From the relations on the previous slide follows:

Out[ ]//TraditionalForm=

$$\mathrm{E}\left((X-a)^2\right) = \mathrm{Var}(X) + (\mu - a)^2$$

This simplifies in case of $a = 0$ to

Out[ ]//TraditionalForm=

$$\mathrm{Var}(X) = \mathrm{E}\left(X^2\right) - \mu^2$$

This is an extremely useful relation

Spread or Scatter

Variance of the mean

Out[ ]//TraditionalForm=

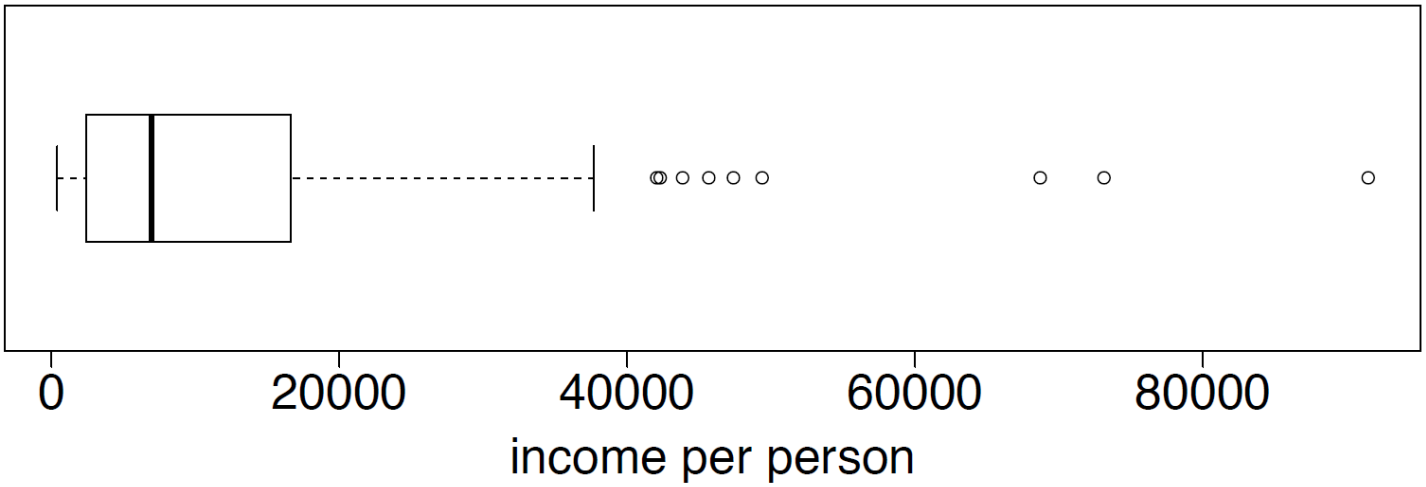
$$\mathrm{Var}(\overline{X}) = \mathrm{Var}\left(\frac{1}{N}\sum_i X_i\right) = \frac{1}{N^2}\,\mathrm{Var}\left(\sum_i X_i\right) = \frac{1}{N^2}\,\mathrm{N}\,\sigma^2 = \frac{\sigma^2}{N}$$

We see, that the variance of the mean decreases when N increases ⟶ standard error.

Quiz

Which of the following is **false** about the distribution of income per person in countries?

Min. = $403, Q1 = $2438, Median = $6975, Q3 = $16650, Max. = $91490



(a) 25% of the countries have incomes per person below $2438.
(b) 75% of the countries have incomes per person above $16650.
(c) IQR is 14212.
(d) The mean is expected to be greater than the median since the distribution is right skewed.

Spread or Scatter

Average Absolute Deviation

The absolute deviation of an element of a data set is the **absolute difference between that element and a given point**. Typically the deviation is reckoned from the central value, being construed as some type of average, most often the median or sometimes the mean of the data set.

Average Absolute Deviation from the Mean

Out[ ]//TraditionalForm=

$$\overline{\Delta X}_{\text{mean}} = \frac{\sum_{i=1}^{N} |X_i - \overline{X}|}{N}$$

Average Absolute Deviation from the Median

Out[ ]//TraditionalForm=

$$\overline{\Delta X}_{\text{med}} = \frac{\sum_{i=1}^{N} |X_i - \overline{X}_{\text{med}}|}{N}$$

Spread or Scatter

Median (absolute) Deviation

For the list $\{x_1, x_2, \ldots, x_n\}$, the median deviation is given by the median of

$$\left\{\left|x_1 - \tilde{x}\right|, \ldots, \left|x_n - \tilde{x}\right|\right\},$$

where $\tilde{x}$ is the median of the list.

Example:

Given the list {2, 2, 3, 4, 14}

3 is the median, so the absolute deviations from the median are

{|2-3|,|2-3|,|3-3|,|4-3|,|14-3|}

= {1, 1, 0, 1, 11}

reordered as {0, 1, 1, 1, 11}

with a **median of 1**, in this case unaffected by the value of the outlier 14, so the median absolute deviation (also called MAD) is 1.

Spread or Scatter

Root mean square (RMS)

The standard deviation is the **root mean square deviation**:

Out[ ]//TraditionalForm=

$$\text{rms} = \sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}} \text{ , where } \mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

When giving a rms, one needs to clarify what it is that you took the root mean square of. Beware any implicit definitions or conventions in your field.

For example, consider a spectral line with properly subtracted baseline, i.e. $\mu = 0$. Then the rms of the deviation turns into the **square root of the arithmetic mean of the squares of the original data**. It is especially useful when variates are positive and negative. (E.g effective value of alternative currents or voltage.

Out[ ]//TraditionalForm=

$$\overline{X}_{\text{rms}} = \sqrt{\frac{\sum_{i=1}^{N} X_i^2}{N}}$$

Please make sure that you understand what RMS is needed before you calculate it.

Convention

We will in the following use the convention that if applicable,

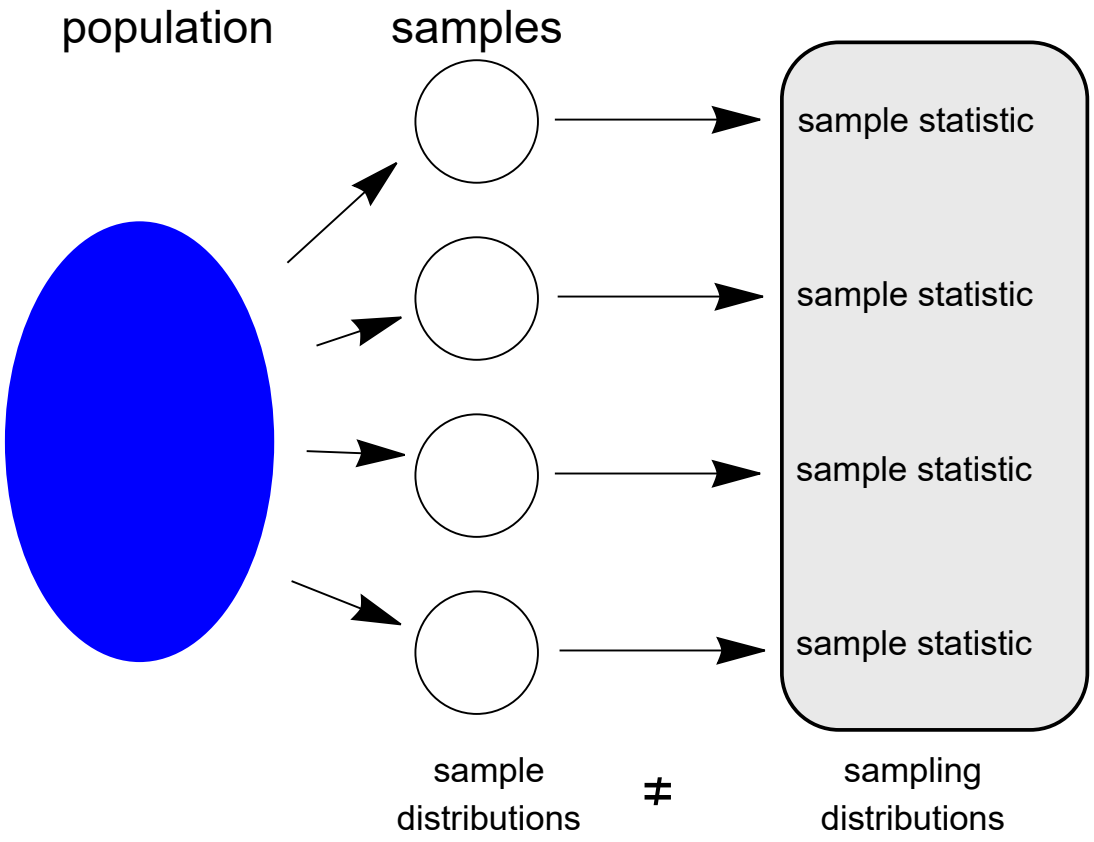
- Greek letters describe the whole population while
- Latin letters describe sample statistics,

e.g. a population might be normally distributed with **mean μ and variance σ^2**, while a sample from this sample will be distributed around the sample **mean $\bar{x}$ and the sample variance S^2**.

Sample vs. population

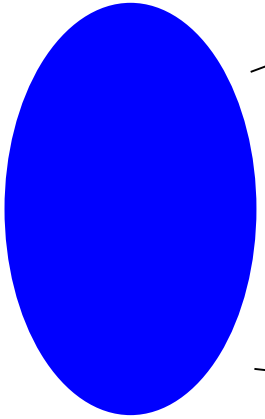
We have a few of the data X_i (**sample**) but we want to know about all of them (**population**).
We want their probability or frequency distribution cheaply (**efficiently**) and accurately (**robustly**, **unbiased**).

Sample vs. population



Sample vs. population

population



samples

sample 1: $\{x_{1,1}, x_{1,2}, \ldots, x_{1,n}\}$ $\overline{x_1}$

sample 2: $\{x_{2,1}, x_{2,2}, \ldots, x_{2,n}\}$ $\overline{x_2}$

sample 3: $\{x_{3,1}, x_{3,2}, \ldots, x_{3,n}\}$ $\overline{x_3}$

sample 4: $\{x_{4,1}, x_{4,2}, \ldots, x_{4,n}\}$ $\overline{x_4}$

$$\mu = \sqrt{x_1 + x_2 + \ldots + x_N}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

sampling distribution

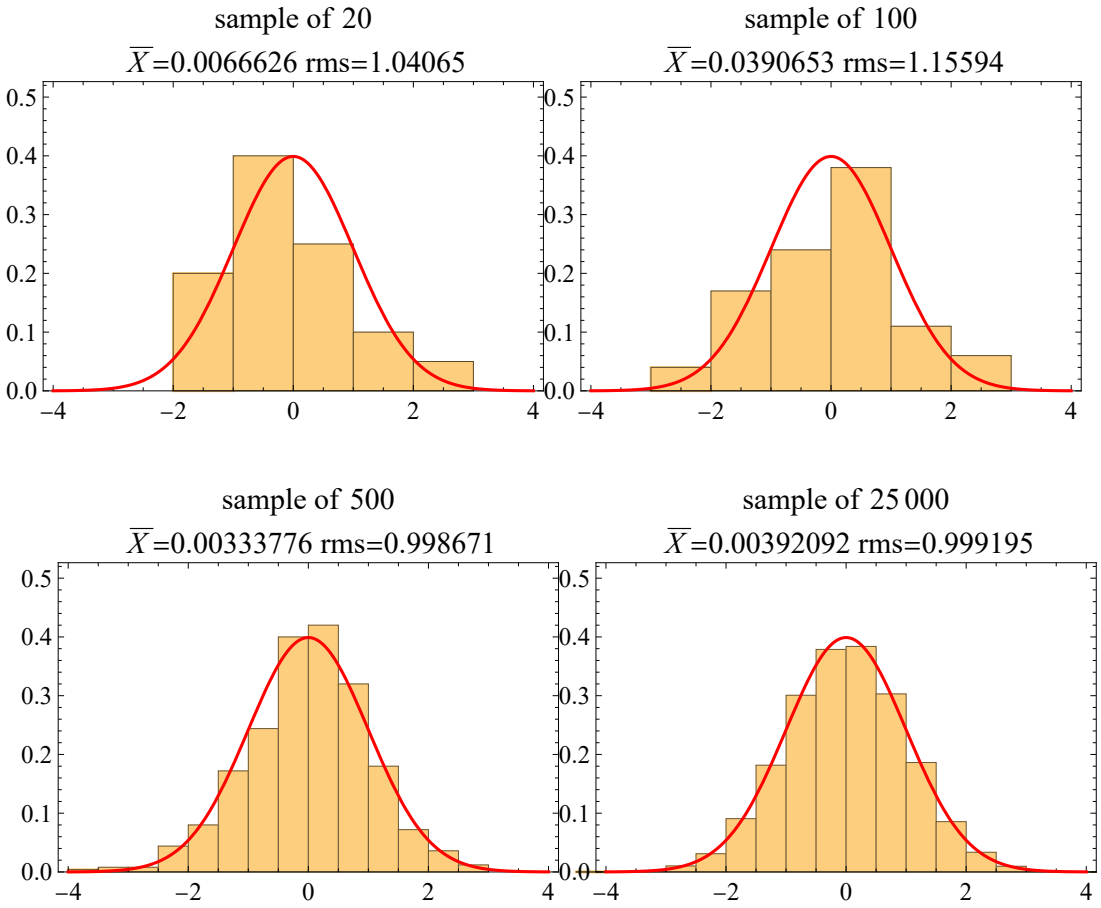
$$\text{mean}(\overline{x_i}) = \frac{\overline{x_1} + \overline{x_2} + \overline{x_3} + \overline{x_4}}{4} \approx \mu$$

standard error $SE = SD(\overline{x_i}) < \sigma$

Sample vs. population

Example: The larger a sample is, the closer its statistics resemble the population statistics.

Draw samples from a population obeying a Gaussian defined by $\mu = 0$, $\sigma = 1$. How does size of sample affect estimates?

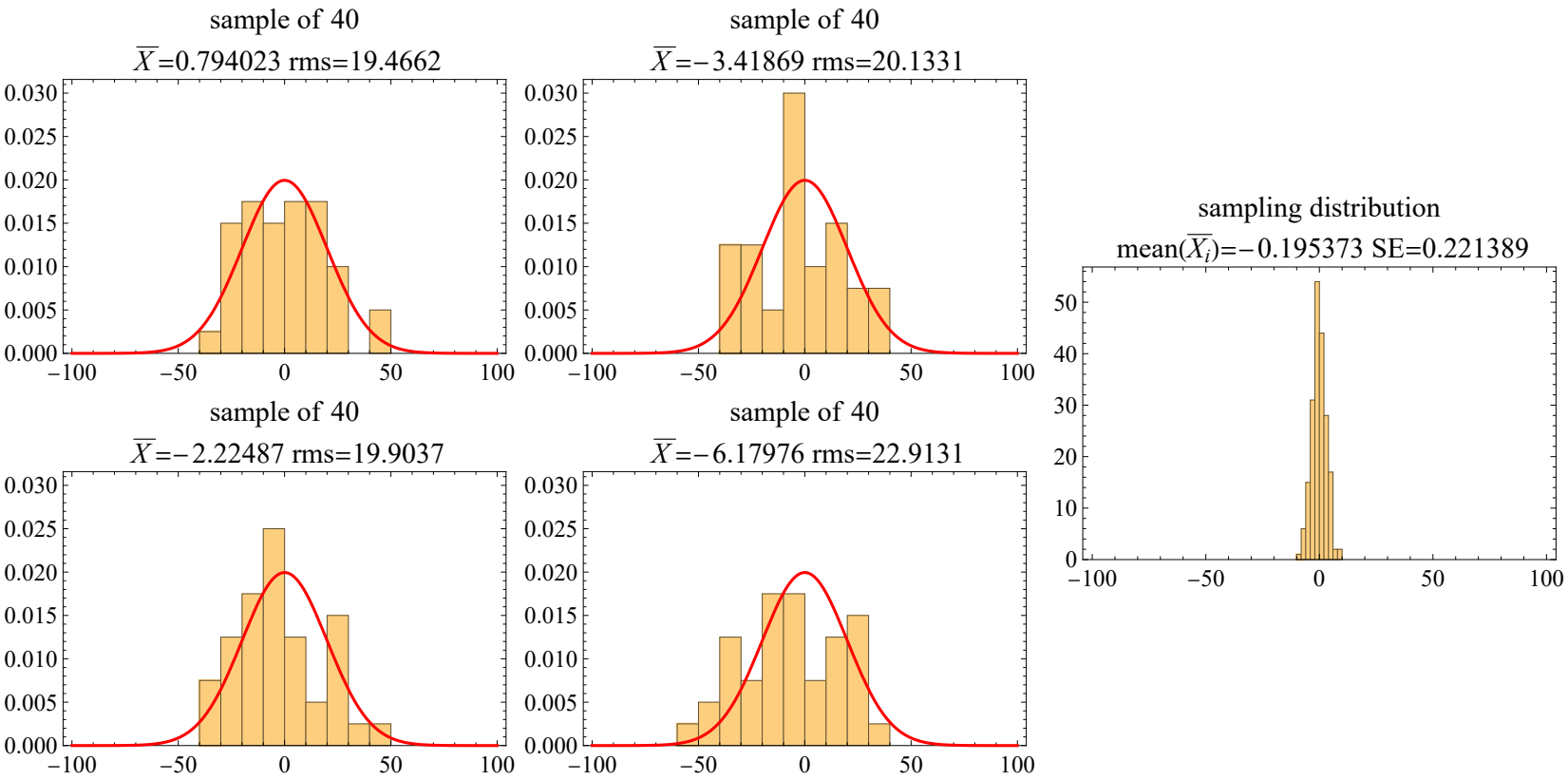


Sample vs. population

Example: Sample distributions versus sampling distribution.

Draw 200 samples of size 40 from a population obeying a Gaussian defined by $\mu = 0$, $\sigma = 20$. What is the sampling distribution?

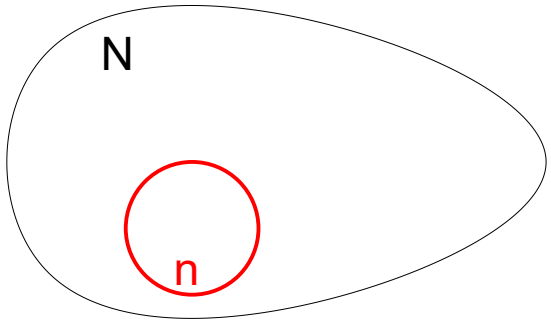
Out[ ]=



Requirements for statistics

- **unbiased**, meaning that the expectation value of the statistic turns out to be the true value.
- **consistent**, the case if the descriptor for arbitrarily large sample size gives the true answer
- should obey **closeness**, yielding smallest possible deviation from the truth.
- **robust**

Bias - Sample vs. population



Consider a data population (set of all existing data) of size N and a subset of the population, called sample, of size $n \leq N$. To describe the true properties of the population distribution (its parameters), we would need to collect all population data and calculate the mean value μ and the variance σ^2. In most cases this is impractical or just impossible.

Bias - Sample vs. population

Alternatively, we can try to estimate the parameter of the population by calculating statistics on the drawn sample.

Out[ ]//TableForm=

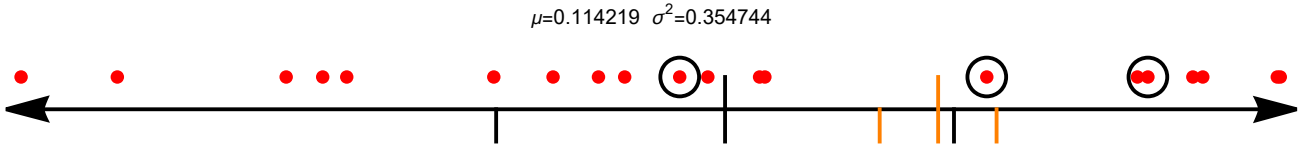
	population (parameter)	sample(statistic) biased	sample(statistic) unbiased
mean	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$		$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
variance	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$	$S_n^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}$	$S_{n-1}^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$

In the following numerical experiment, we draw a sample of 3 from a larger population and calculate the mean and the **biased** sample variance. They are shown as orange lines. The black lines show the **population mean** and **variance**.

Bias - Sample vs. population

Out[465]= Evaluate

Out[466]=



The mean value scatters around the true mean and will converge to if we repeat the sampling many times. The variance however, tends to a consistently smaller value compared to the true variance. In the Table below we repeat the sampling and take the average of the sample $\overline{x}$ and S_n^2. For a larger number of samples we see that $\overline{x} \to \mu$ but that S_n^2 does not tend to σ^2.

Sample size = 3

# of samples	$\overline{x}$	S_n^2
5	0.101286	0.123006
10	0.0611241	0.233868
100	0.155186	0.238613
1000	0.111925	0.236294
10 000	0.117746	0.23501

Sample size = 10

# of samples	$\overline{x}$	S_n^2
5	−0.0984042	0.205673
10	0.100987	0.363784
100	0.0995786	0.325872
1000	0.114148	0.318652
10 000	0.113945	0.31854

Bias - Sample vs. population

It turns out that for many samples, $S_n^2 \to \frac{n-1}{n}\sigma^2$, therefore, to get a better estimate for σ^2, the true variance of the underlying population, we need to take S_{n-1}^2.

Sample size = 3

samples	$\overline{x}$	S_n^2	S_{n-1}^2
5	0.178927	0.312733	0.4691
10	0.0162472	0.168092	0.252139
100	0.111182	0.227684	0.341526
1000	0.106398	0.230717	0.346075
10 000	0.115429	0.235869	0.353803

Sample size = 10

samples	$\overline{x}$	S_n^2	S_{n-1}^2
5	−0.0279401	0.21829	0.242544
10	0.103459	0.328238	0.364709
100	0.115999	0.310507	0.345008
1000	0.114237	0.318128	0.353475
10 000	0.117454	0.318679	0.354088

Consistency

A consistent estimator gives the true answer for an arbitrarily large sample.

For example: the RMS is a consistent measure of the standard deviation of a Gaussian distribution, because it gives the right answer for large *N*. But we showed before that it is a biased estimator because it underestimates the standard deviation for small samples.

Robustness

Consider a symmetric distribution with a few outliers (errors?). **As a measure of central location the median is far more robust than the average.** Also consider salaries – mean vs median.

Central Limit Theorem -CLT

"The distribution of sample statistics is nearly normal, centered at the population mean, and with a standard deviation equal to the population standard deviation divided by square root of the sample size."

Out[ ]//TraditionalForm=

$$\bar{x} \sim \mathcal{N}\left(\text{mean} = \mu,\ \text{SE} = \frac{s}{\sqrt{n}}\right)$$

Conditions for the CLT

1. **Independence**: Sampled observations must be independent.
   - random sample/assignment
   - if sampling without replacement, n < 10% of population
2. **Sample size/skew:** Either the population distribution is normal, or if the population distribution is skewed, the sample size is large (rule of thumb: n > 30).

Init

Computations