Getting your data: Sources and samples

Sources of psychological data and Data collection methods

Data sources

- Behavior
- Physiological data
- Self-reports
- Peer-reports
- Activity reports (objective/projective)
- Biographical or archival data

Data collection methods

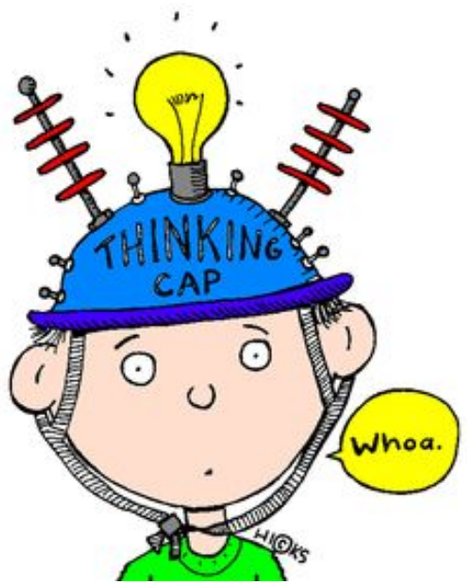
- Observation
- Measurement
- Focus-groups
- Survey
- «Archival data»: databases, papers

Why experiment is not a method of data collection?

Because it is a method of study organization

Data collection exercise - 15 mins -

- In groups of 4 think of a Research Question/ Hypothesis
- What type of data is the most suitable for your RQ or H?
- What data collection method is the most suitable?
- WHY?

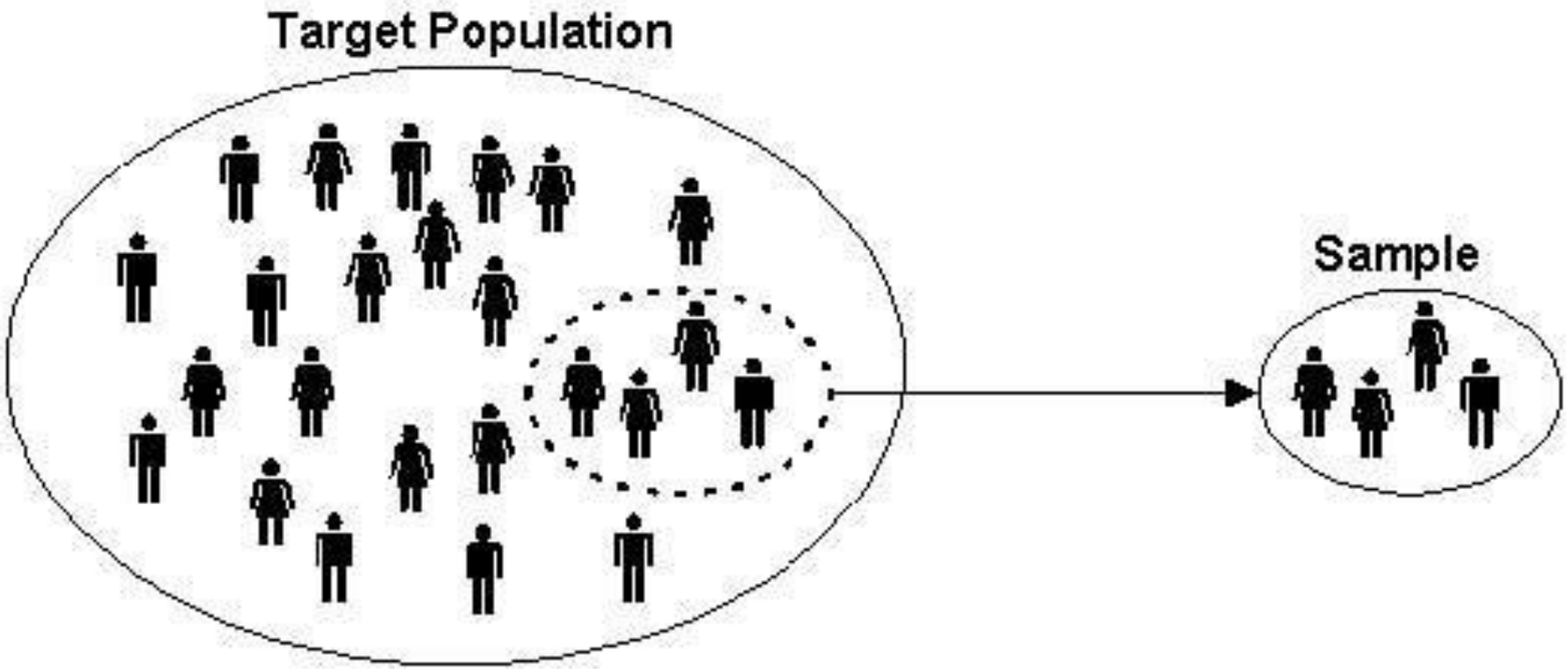


Sample

What does Sample mean?

Sample is a limited set of research objects (units) which we use to make general conclusions about the whole population.

Why do we need samples?



Sample and distribution

What is distribution?

- a relationship between the values of a random variable and the frequency (or the probability) with which each of these values can be found in a sample (or a population).

Values of variable

Distribution of values

Descriptive statistics…

MEAN

Commonly used in sport to find out a score in sports like Football, Basketball and Cricket

Is also known as the "average"

1. Add up all the values to get the total
2. Then divide the total by the number of values you added together

3 + 4 + 8 + 7 + 5 + 3 = 30

30 ÷ 6 = 5

The average for these values is 5

MEDIAN

Used when comparing house prices.

The "middle" number in a set of values

1. First put all the values in order
2. Find the middle number in the set of data
3. If there are two values in the middle, find the mean of these two.

1, 2, 4, 5, 6, 8, 9

The median is 5.

Mode

Eg. What is the mode of goals kicked by a footballer after each round?

The number which occurs the most

1. Count how many of each value appears
2. The mode is the value which appears the most
3. There can be more than 1 mode

1, 2, 2, 5, 6, 6, 9

2 and 6 are the mode for these values

range

Measures difference between all the values. Used in weather.

The range is the difference between the highest and lowest value

1. Find the highest and lowest values
2. Subtract the lowest value from the highest value.

1, 2, 2, 5, 6, 6, 9

9 − 1 = 8 The range is 8

Exercise

A survey of 20 students was conducted to find out how many books they had read during the past three months (including books for school). The results from those 20 students are shown below. Find the mean, median, and mode for this data.

2, 4, 5, 1, 3, 2, 5, 6, 1, 2, 4, 3, 6, 10, 12, 10, 2, 8, 6, 7

Answers:

Mean = 4.95.

Median = 4.5

Mode = 2.

Normal distribution

Properties of any theoretical normal distribution:

1) The curve never approaches horizontal axis.

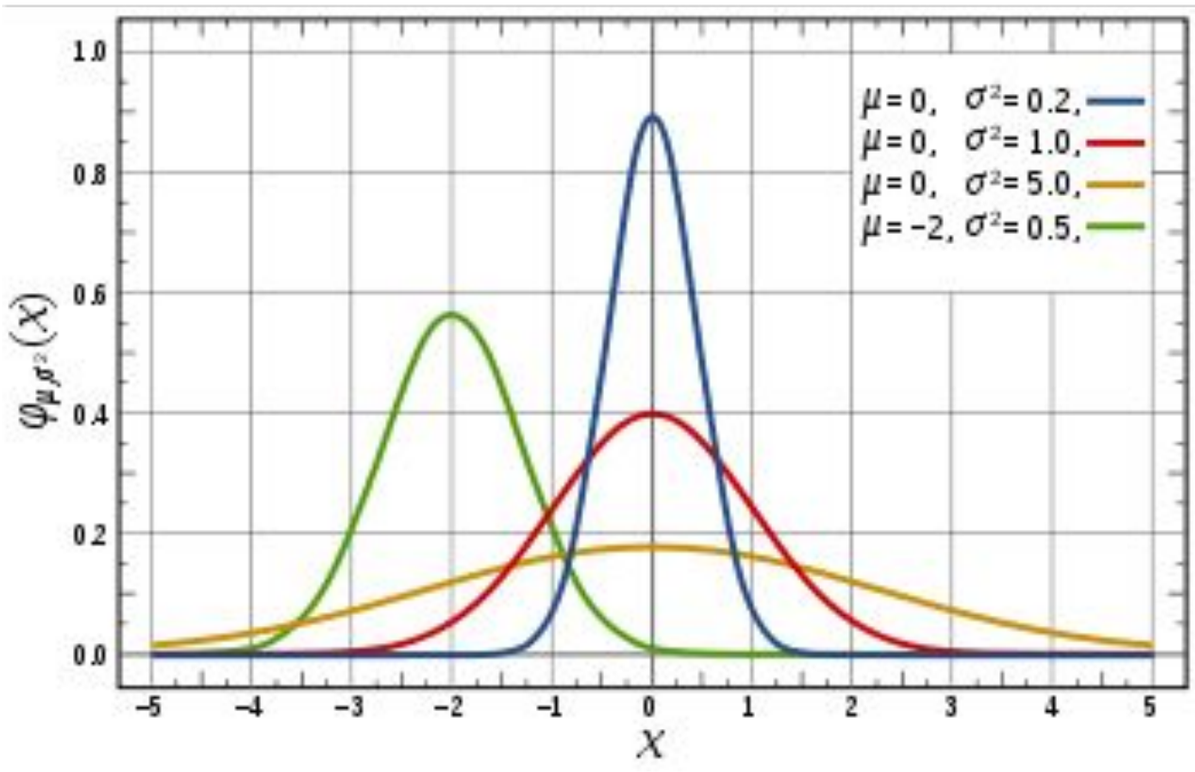
2) Symmetrical around the mean.

3) Skewness = 0 and kurtosis = 0.

Standard normal distribution is a special case of theoretical n.d. with 2 properties:

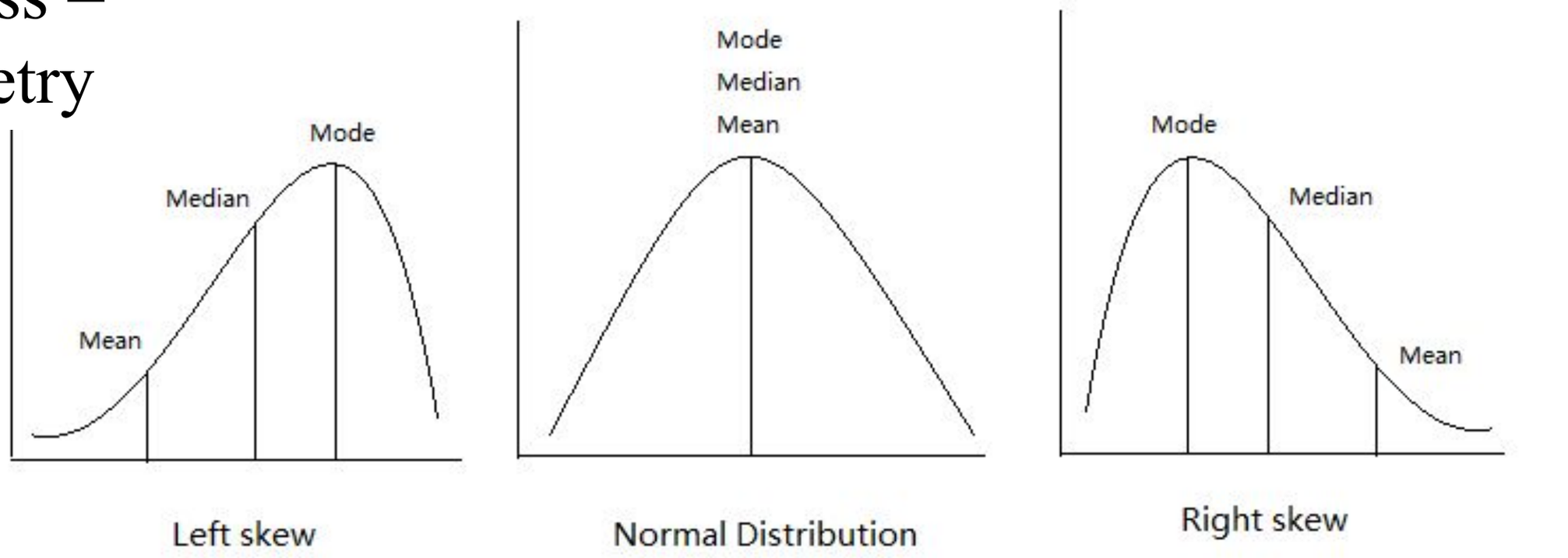
1) μ = 0, σ = 1;

2) area under the curve = 1, and integral of (-∞; z] can be interpreted as probability of finding values equal to or below Z.

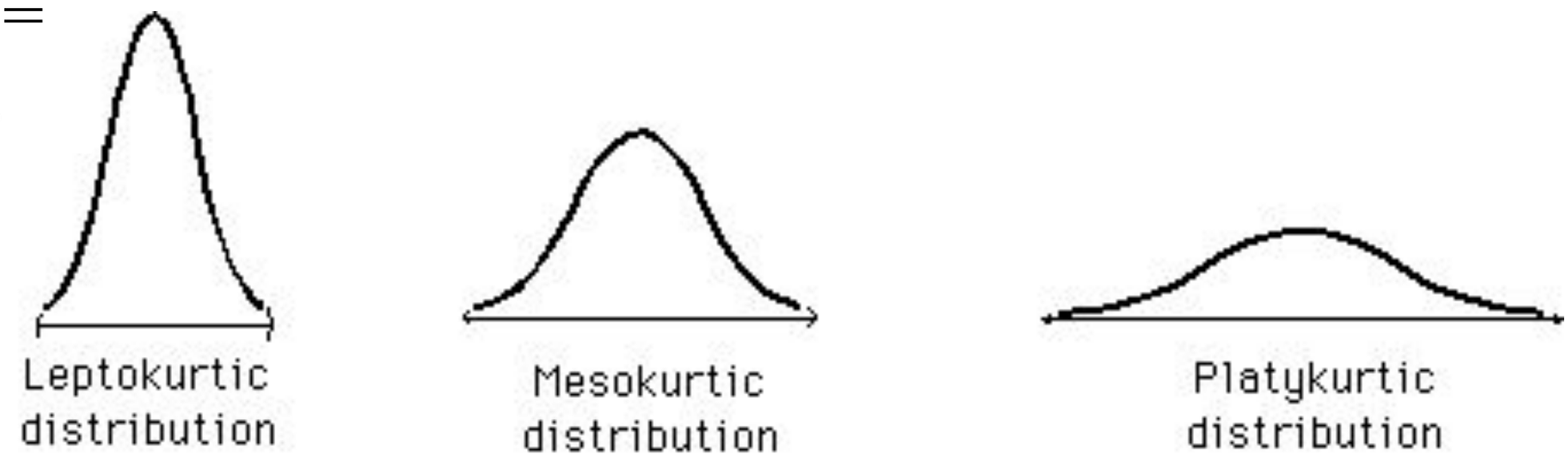


Normal distribution

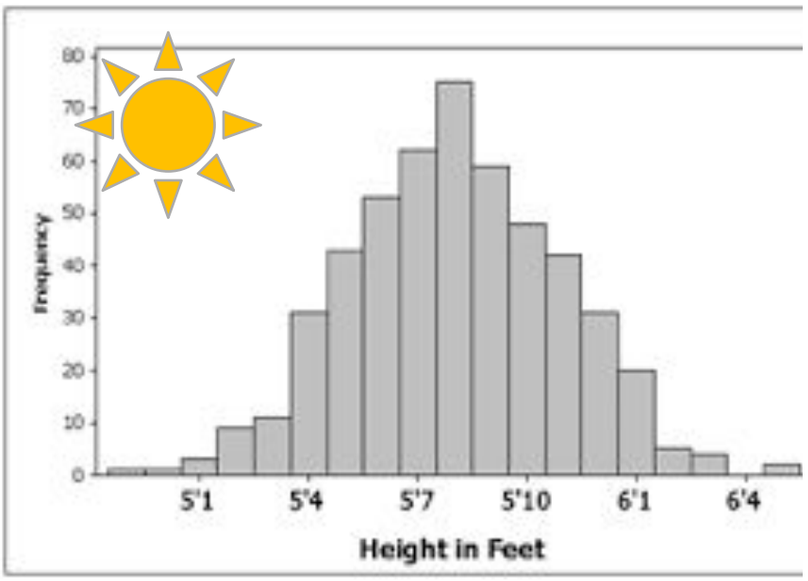
Skewness = asymmetry

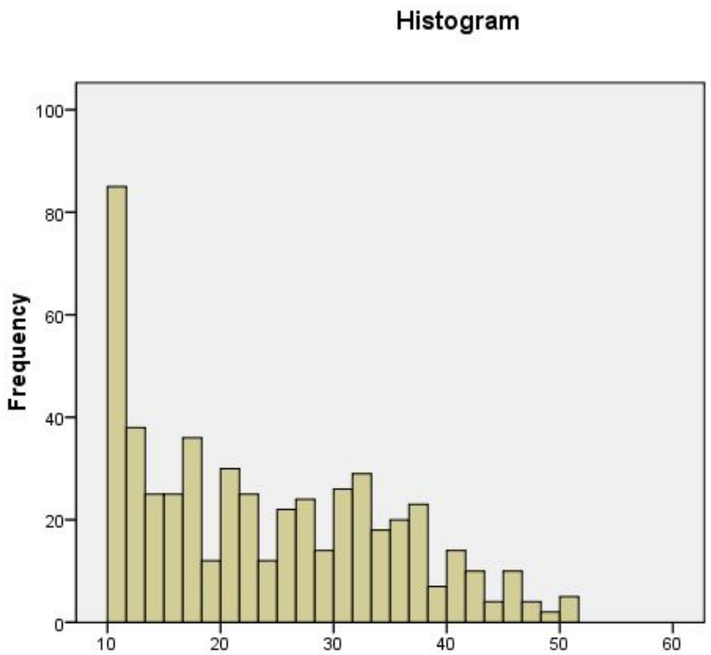


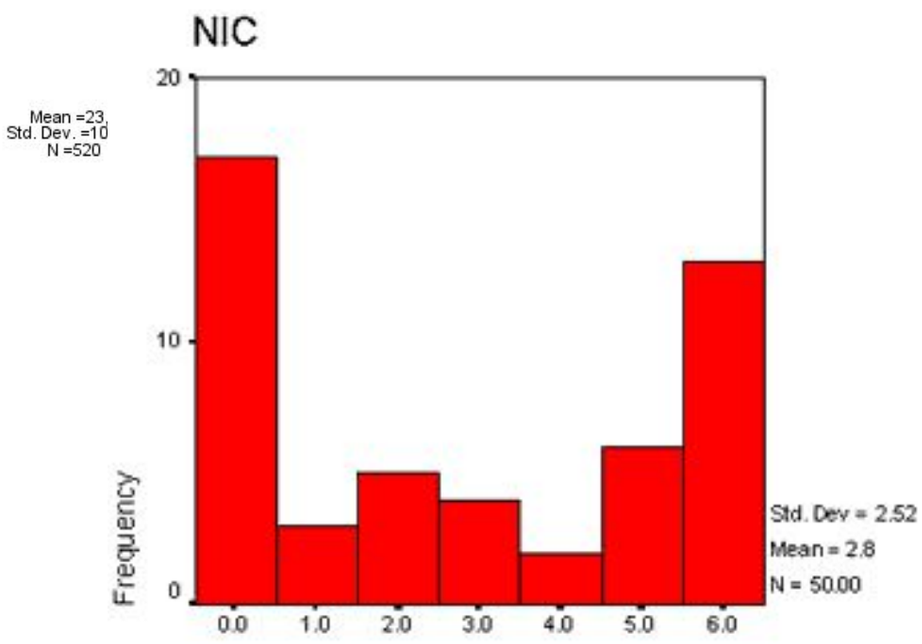
Kurtosis = flatness

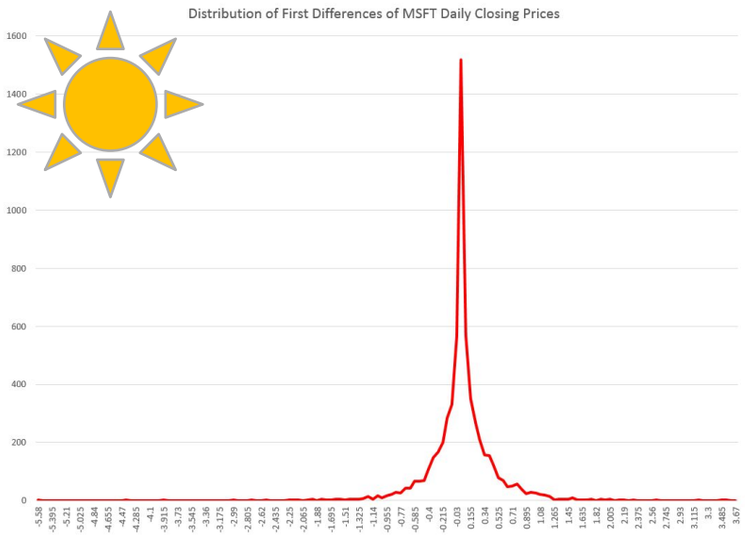


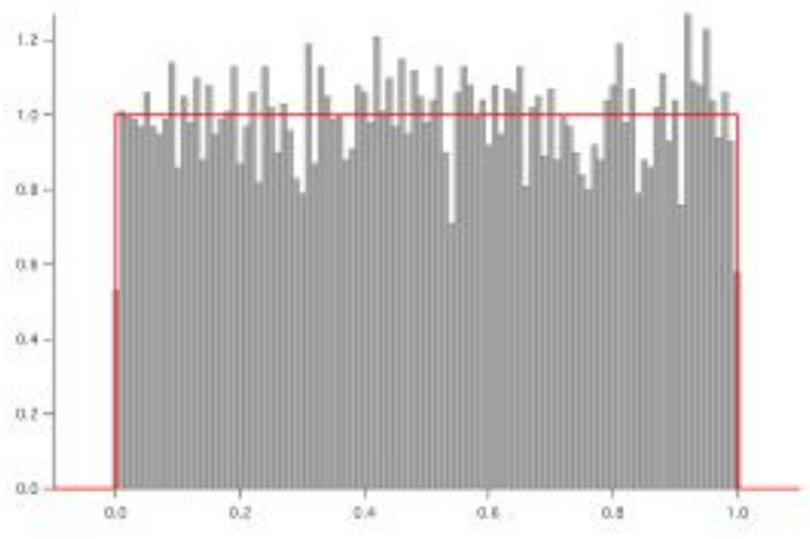
Where is NORMAL distribution?









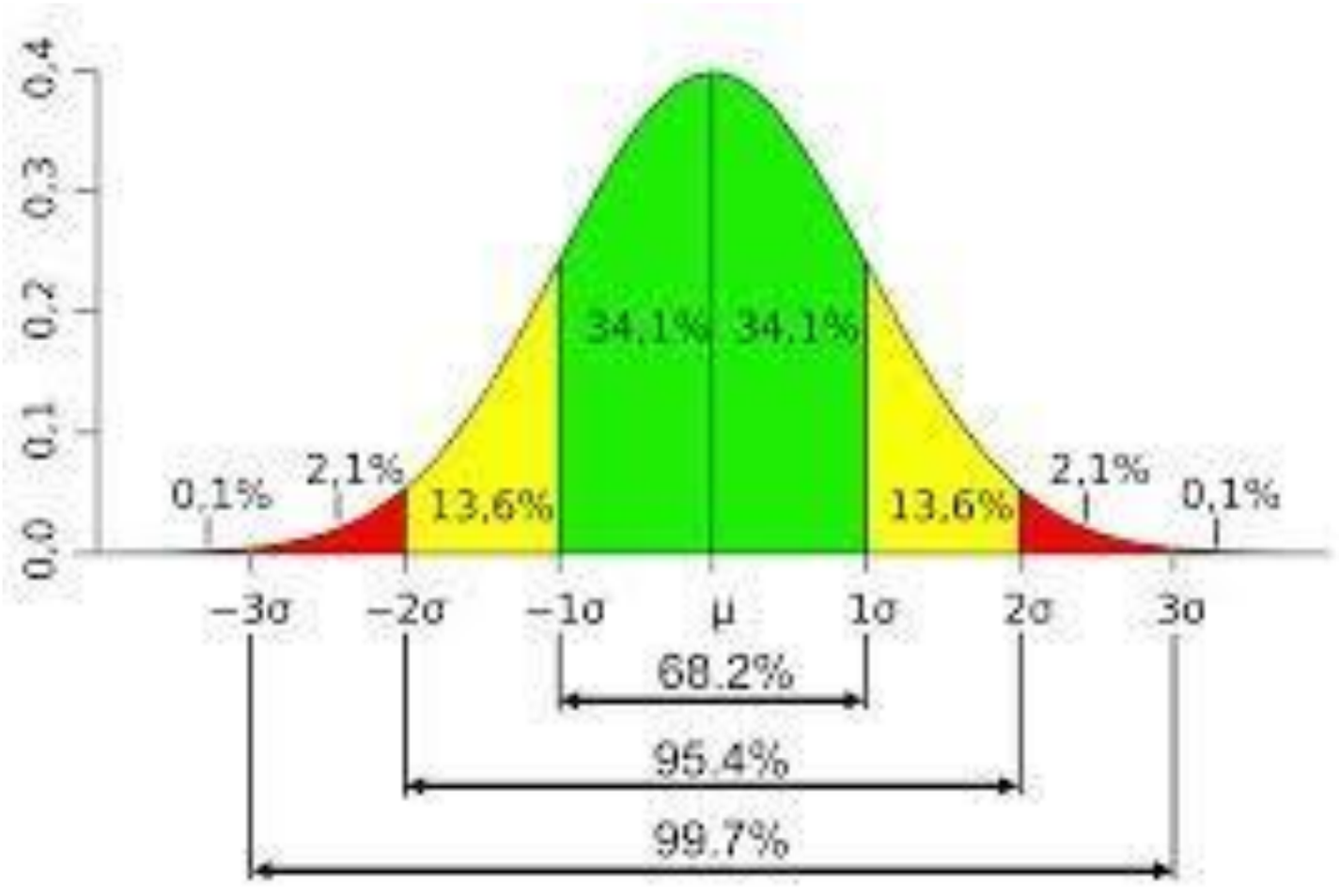


NIC

What do we know about STANDARD normal distribution?

1) The curve never approaches horizontal axis
2) Symmetrical around the mean
3) Skewness = 0 and kurtosis = 0
4) Mean = 0, SD = 1
5) Mean= mode= median =0

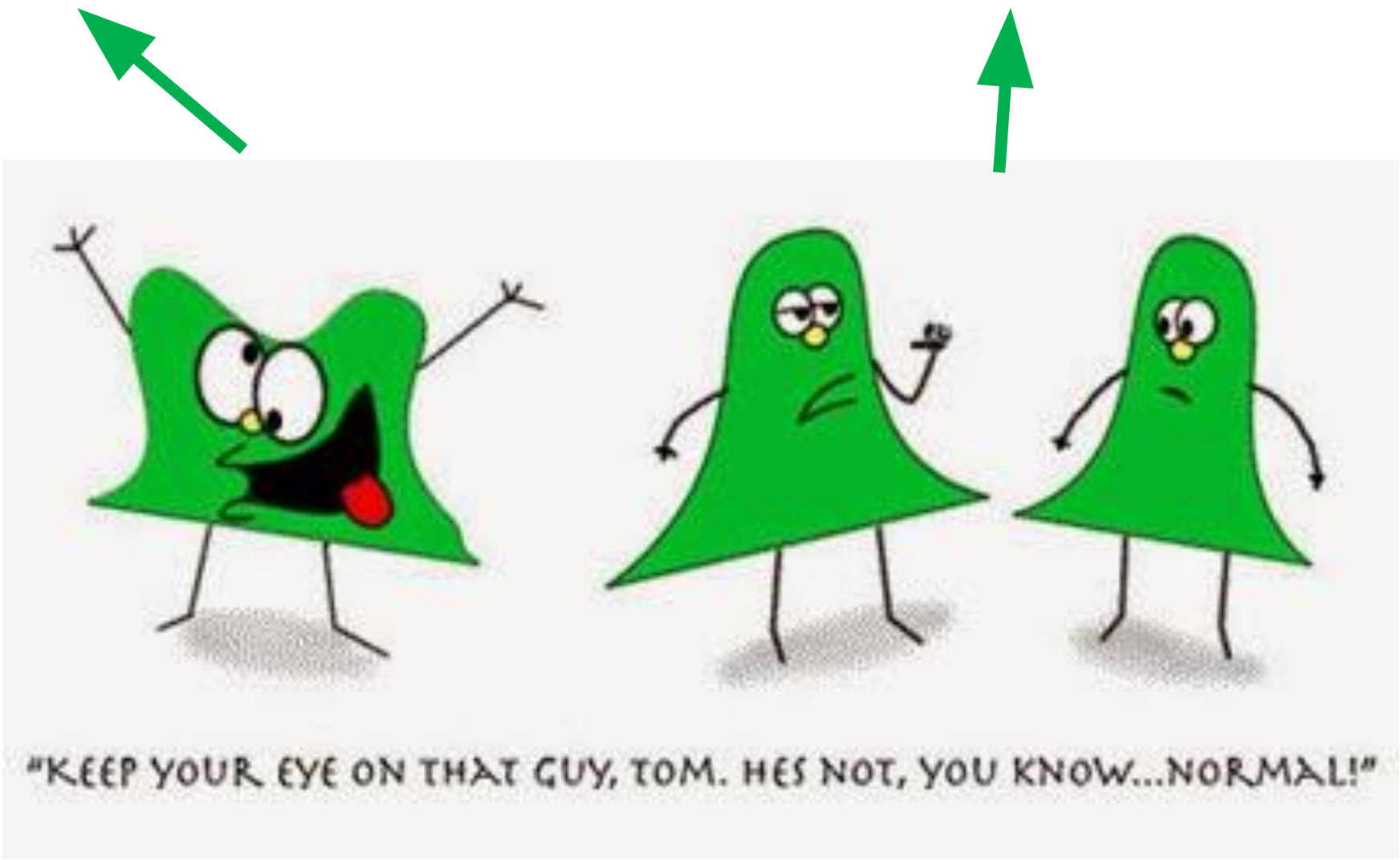
Example 1. If you get a score of 90 in Math and 95 in English, you might think that you are better in English than in Math. However, in Math, your score is 2 standard deviations above the mean. In English, it's only one standard deviation above the mean. It tells you that in Math, your score is far higher than most of the students (your score falls into the tail)



Why is it important to know what kind of distribution your variables have?

Non-parametric tests

Parametric tests



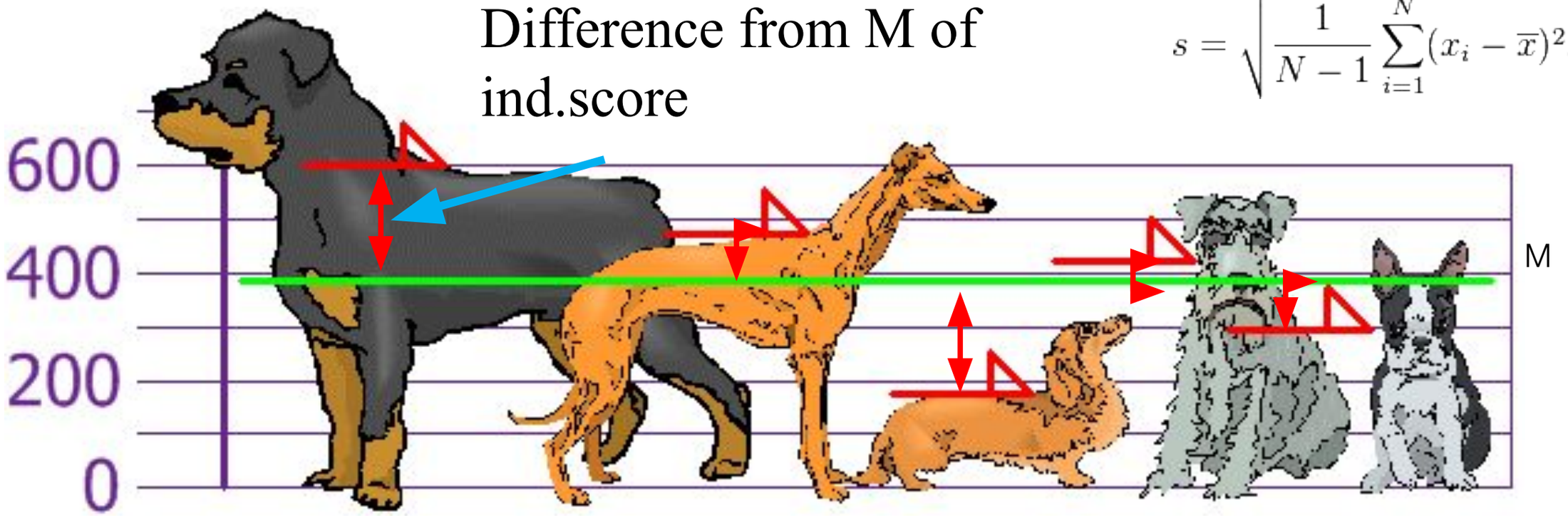
Descriptive statistics…

Variance - is the sum of the squared differences from the M of each score, divided by the total number of scores minus 1
Provides info HOW FAR scores are spread out

Standard deviation(SD) - square root of variance
It is a quantification of scores variation, and **it's expressed in the same units as the data**

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}.$$

Difference from M of ind.score



Are you tall?

When you know so much about distributions, you can compute a height distribution in your group

Sample Variance

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

Sample Standard Deviation

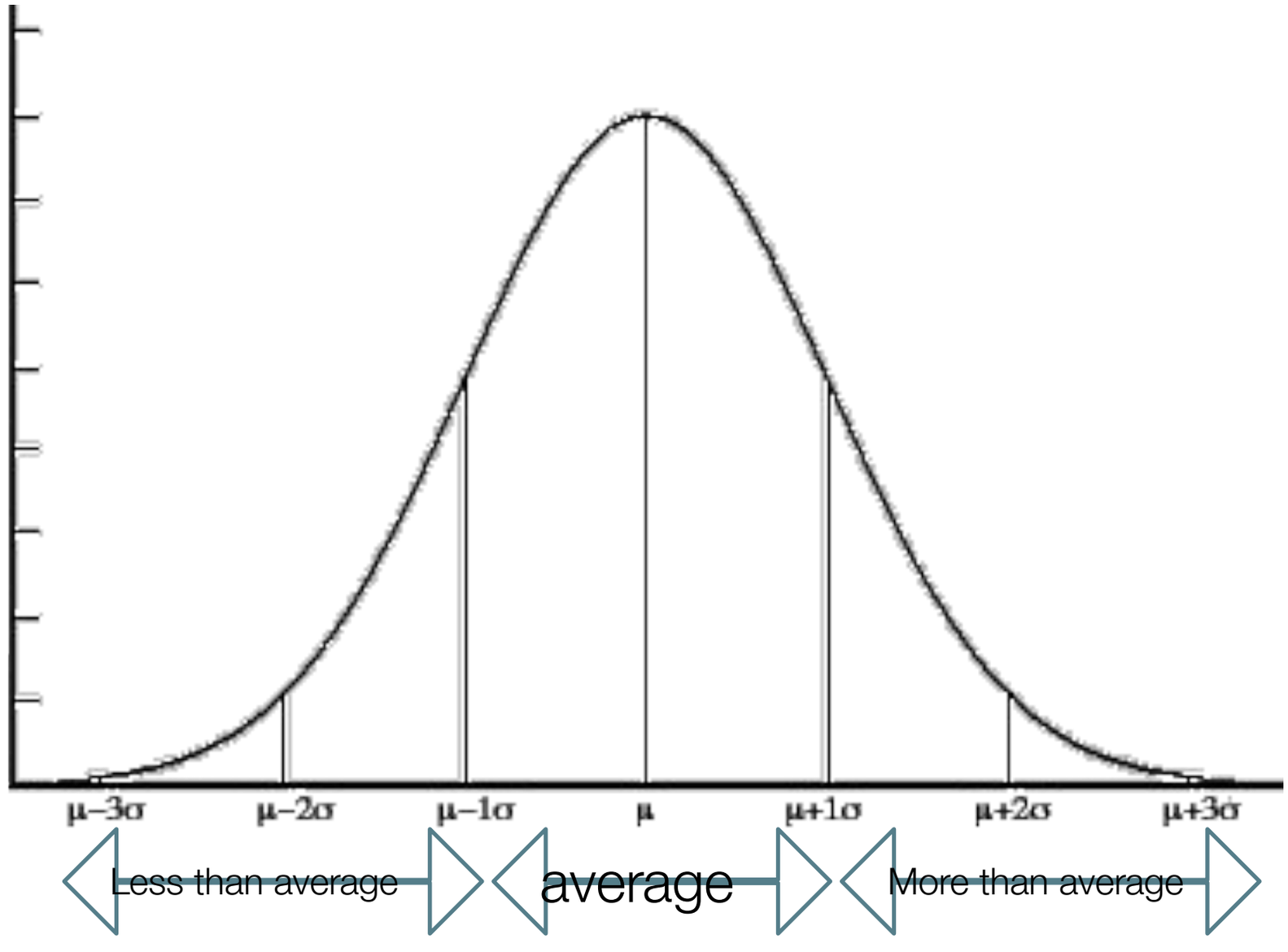
$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

your personal height

mean height

sample size

When you know Mean and SD, you can estimate whether you are tall or not



But…

Is this result applicable in other groups?

Are you tall in other groups?

In HSE?

In Russia?

To answer this question we should use standard scores

Standard scores (Z-scores)

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

your individual height

mean height in a given sample

standard deviation in a given sample

A very good explanation of Z-scores: https://statistics.laerd.com/statistical-guides/standard-score.php

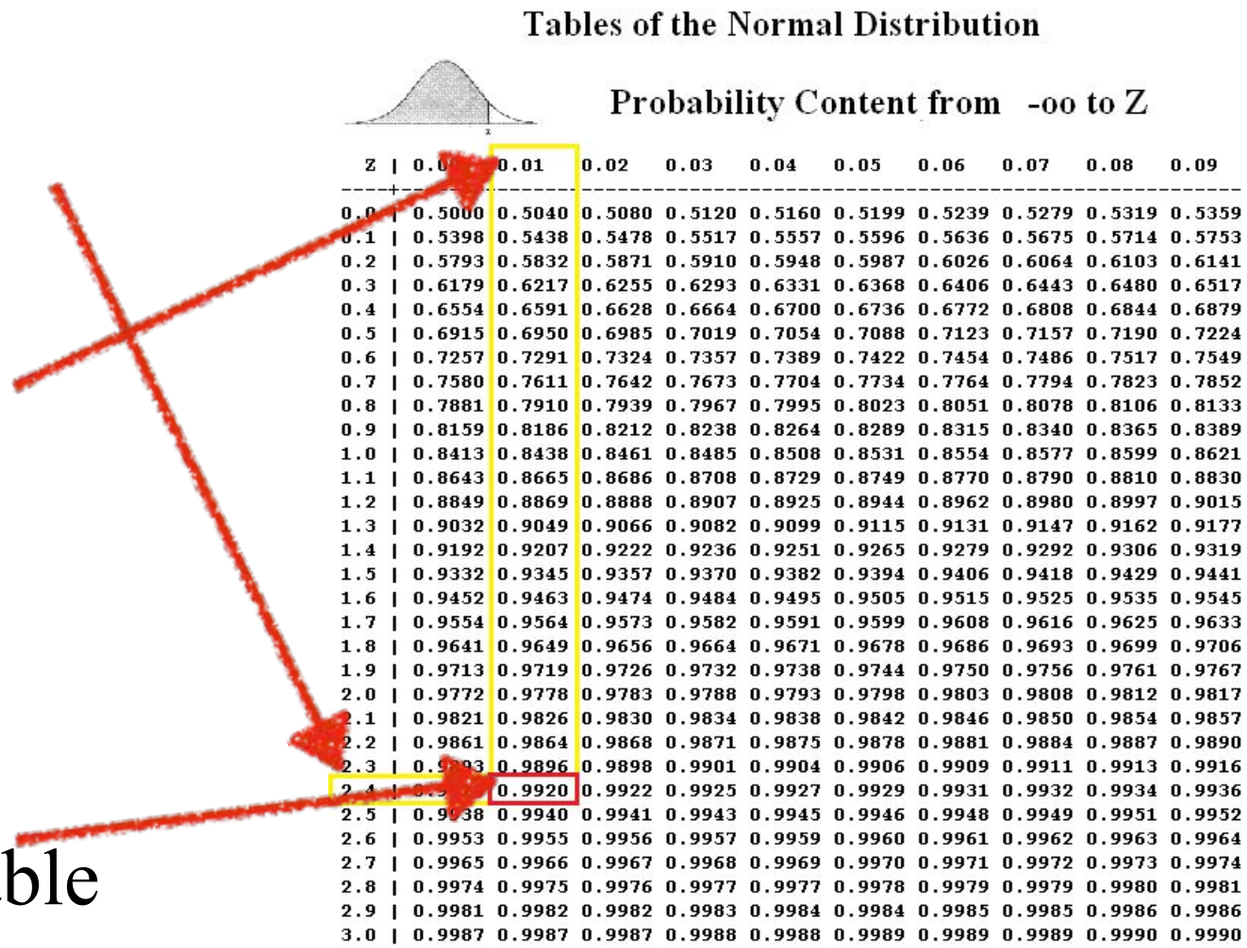
Standard normal table

Shows you a PROBABILITY that all observed values in your sample are lower than Z

The label for rows contains the integer part and the first decimal place of Z.

The label for columns contains the second decimal place of Z.

The values within the table are the probabilities corresponding to the table type.



Tables of the Normal Distribution

Probability Content from -oo to Z

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9[illegible]	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

What is the probability to find people taller than you in…

…Guatemala?

Mean = 147.3 cm

SD = 6.3

your Z = (your cm - 147.3)/ 6.3

Then look in Z-table

…Hong Kong?

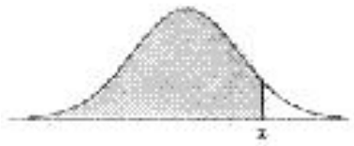
Mean = 160.1 cm

SD = 5.7

your Z = (your cm - 160.1)/ 5.7

Then look in Z-table

Tables of the Normal Distribution



Probability Content from -oo to Z

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Sample size and standard error

We know M and SD in your group

And we know M and SD in Guatemala

Which stats provide more trustworthy description of height in a country?

Why?

Standard error

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

SD

Sample size

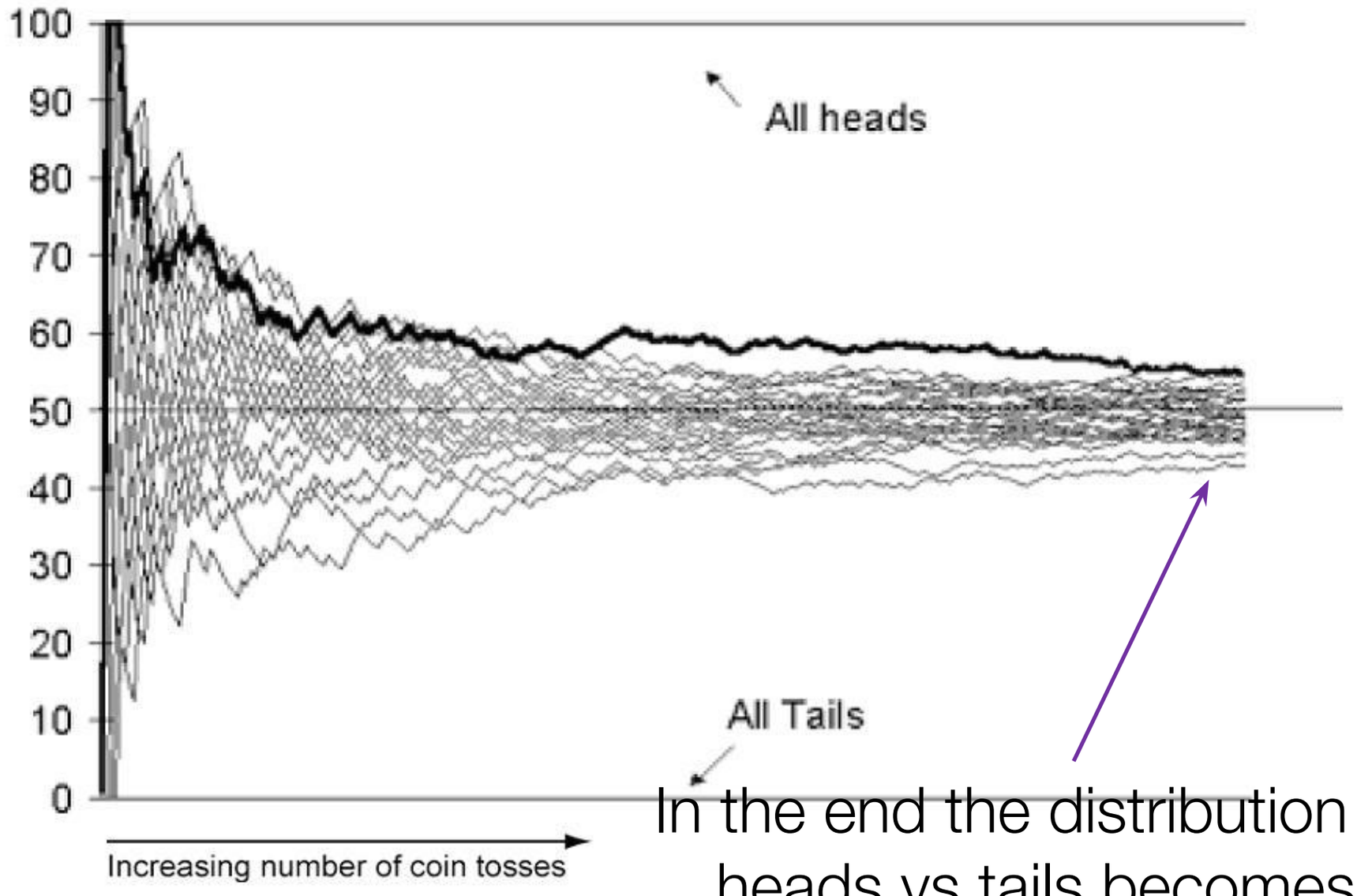
Guatemala:
SE = 6.3/ sqrt(15000) = .05

Our group:
SE =?

1. *SE depends on a sample size*
2. *The bigger the sample the smaller the SE*
3. *The smaller SE the more trustworthy estimations you have*

Why do bigger samples provide better estimation?

Law of Large Numbers



In the end the distribution of heads vs tails becomes NORMAL (50/50)

Sampling strategies

Probability strategy

True random sampling
using a random number table (a computer) to select people from a list, a phone book, etc. (a variety is called 'systematic random sampling' = select every nth person);

Stratified sampling / quota sampling
we define the target groups (strata) within our sample (genders, age groups, etc.) and collect respondents from each stratum to get the % you need

Cluster sampling
select the most representative group from a set (a class from a school, a neighborhood from a city

Multi-stage strategies
different strategies used at different sampling stages: e.g., 1) select a school from a city, and 2) select a number of students from that school

Non-probability strategy

Snowball approach:
start with some respondents (e.g., friends), asking each to recruit more people to the study.

Convenience sample:
people at work, students, etc.

Self-selecting sample:
those who agrees to take part in the study; «volunteer bias».

Exercise: Match the statement with the appropriate term

A. The process of random sampling	A	1. Get a list of everyone in the population 2. Select every Nth (e.g. 10th) person in the list until you have enough participants.
B. The process of stratified sampling	B	1. Get a list of everyone in the population 2. Identify relevant sub-groups, and divide up the population into these groups. 3. Select randomly from these groups in the correct proportions until you have enough participants.
C. The process of systematic sampling	C	1. Ask known individuals to take part. 2. Ask these participants to identify others that should participate in the study.
D. The process of snowball sampling	D	1. Get a list of everyone in the population 2. Put all the names into a spreadsheet 3. Use software to select randomly from the spreadsheet until you have enough participants.

I want to study cultural differences…/ I want to study how culture influence…

This is possible only with representative samples collected in few countries!!!!

A non-representative or a sample from 1 country only cannot help you with this kind of RQ

Open access data:
European Social Survey http://www.europeansocialsurvey.org/

World Values Survey http://www.worldvaluessurvey.org/wvs.jsp

European Values Survey http://www.europeanvaluesstudy.eu/

Recommended reading:

Howitt & Cramer, 2011, p. 232-246 (Samples).

Supplementary reading:

Bakeman, 2000 (Chapter 7 in Reis & Judd, 2000) (Observation)

Cramer, 2007 (in Robins, Fraley, Krueger, 2007) (Archival method)

Diamond & Otter-Henderson, 2007 (in Robins, Fraley, Krueger, 2007) (Physiological measures)

Fraley, 2007 (in Robins, Fraley, Krueger, 2007) (Internet studies)

Wilkinson, Joffe, & Yardley, 2004 (Interviews and focus groups)

Why Standardize ... ?

Example 2. Here are the students results (out of 60 points):
20, 15, 26, 32, 18, 28, 35, 14, 26, 22, 17
Most students didn't even get 30 out of 60, and most will fail.

The test must have been really hard, so the Prof decides to Standardize all the scores and only fail people 1 standard deviation below the mean.
How many students will fail?

Answer:
The Mean is 23, and the Standard Deviation is 6,6, and these are the Standard Scores:
-0,45, -1,21, 0,45, 1,36, -0,76, 0,76, 1,82, -1,36, 0,45, -0,15, -0,91
Only 2 students will fail (the ones who scored 15 and 14 on the test)

Next time

Psychological measurement: Psychometrics and psychophysics