CognitiveSystems

Experimental Analysis

Prof. Christian Wallraven
wallraven@korea.ac.kr

Statistics?

- You need statistics in order to make concrete statements about data that has a random component to it
  - you asked random people
  - you repeated a question multiple times
  - your measurements contain noise
- There are two different kinds of statistics:
  - descriptive
  - inferential

Statistics?

- Descriptive statistics describe the data
  - obviously used in “Descriptive” experiments (surveys, etc.)
  - BUT: use it for every data you gather! plot it!
- Inferential statistics test hypothesis about the data
  - Are means different?
  - Do the two variables correlate?
  - Used to answer specific questions

What can it do for you?

- Provide objective criteria for testing hypotheses
  - by following statistical procedures, observer bias, for example, is greatly reduced, if not eliminated
- Provides a way to critically assess conclusions of other people
- When used beforehand, can save you a lot of time and effort
  - how many samples do you need to test?
  - what kinds of answers can you expect?

Where does it fail?

A few terms

- Population: the set of all possible items under question
- Sample: the subset of the population that is tested
  - obviously, the sample should be representative of the whole population
  - this is often NOT the case!
- Sample size: the number of data points
- Dependent variable: what you measure
- Independent variable: what you manipulate

What variable types are there?

- – Categories, such as “name”, “favorite food”
- – Sortable values, such as “movie ratings”, “class likability”
- – “Normal” types of numbers, such as “number of students”, “favorite number”

- Note: these are rough guidelines only

CognitiveSystems

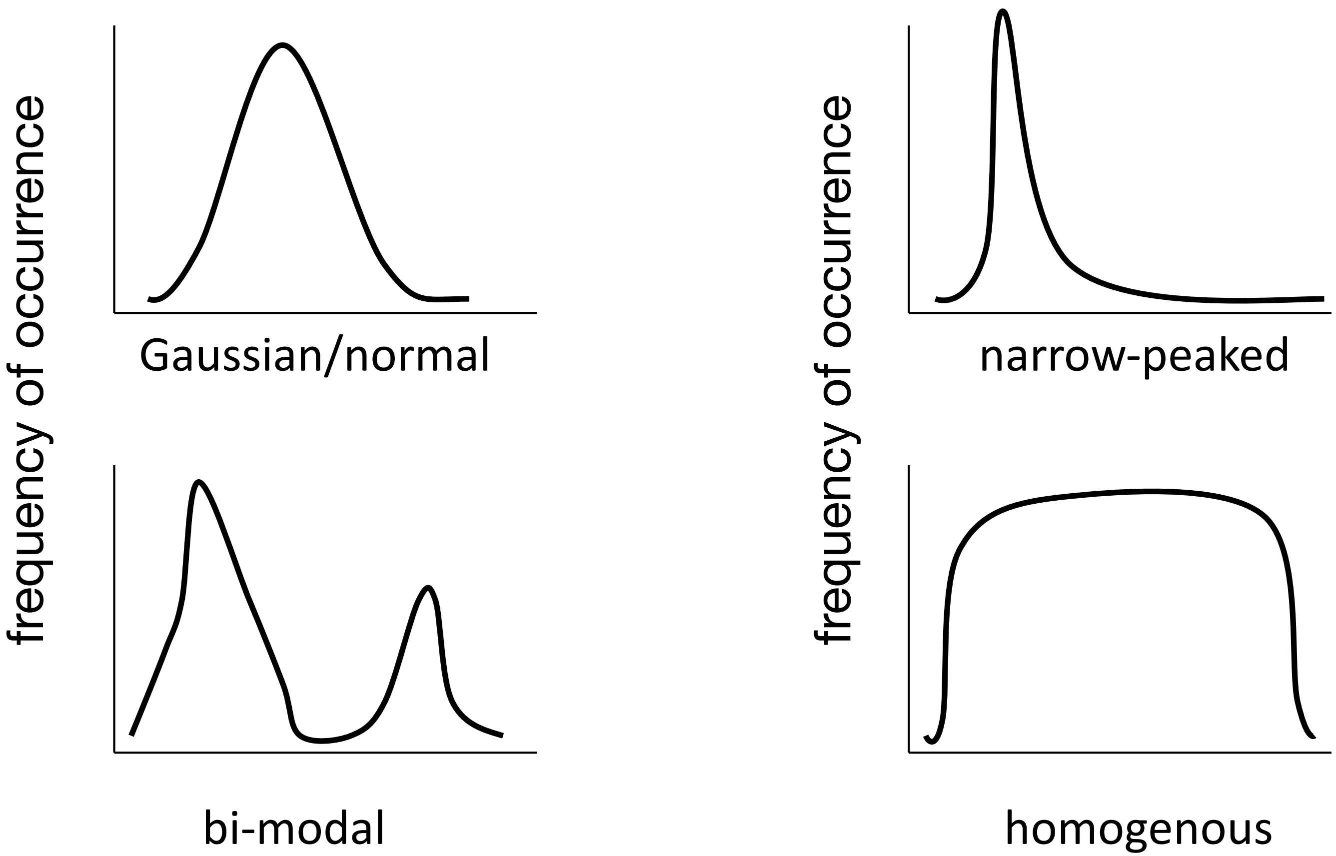
Probability distributions

Definition and Concepts

- Typically, due to noise, measurements in experiments are *random variables*. **Data is random!**
- This means that they do not always yield the same value, but vary according to a *probability distribution*
- The distribution describes how likely it is that the random variable takes on a certain value
  - the probability that I roll a 3 on a die
  - the probability that a neuron fires given a certain input
  - the probability that my reaction time in catching a ball is 0.4s
  - the probability that you answered 5000Won in the ultimatum game question

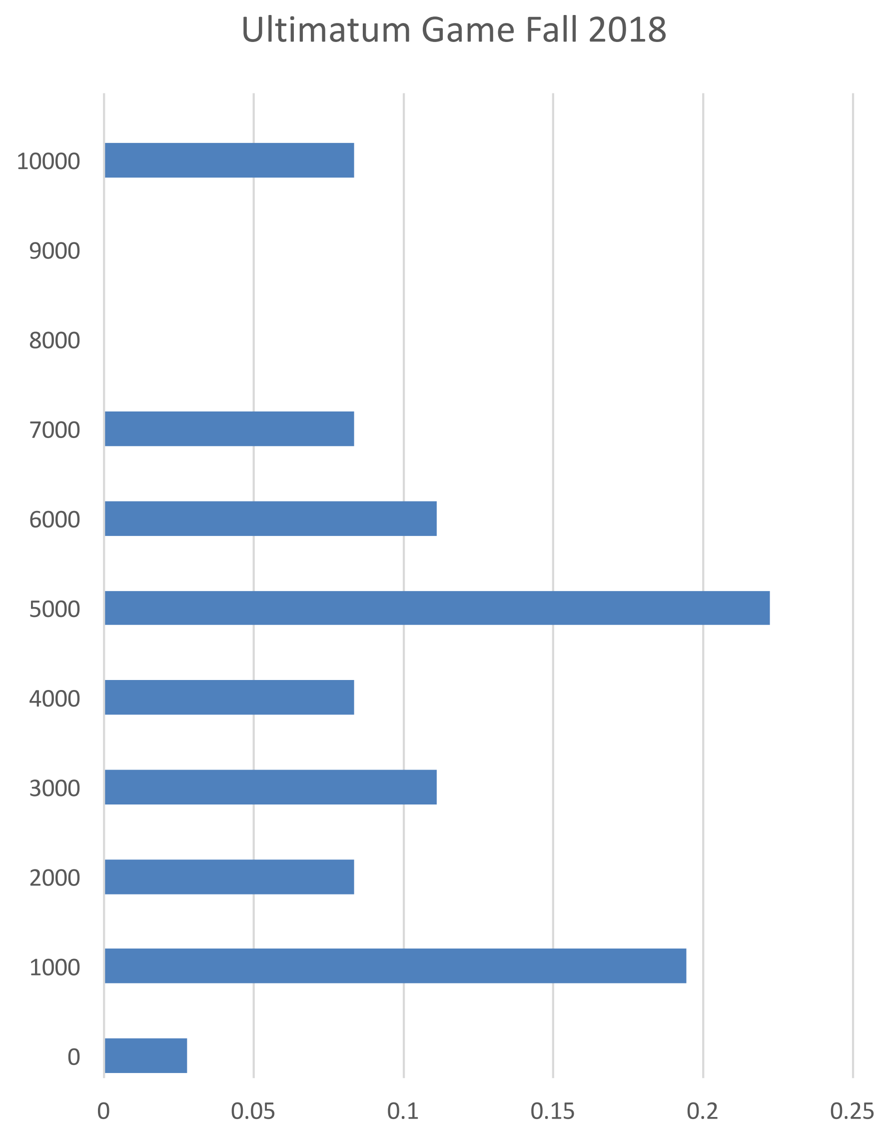
Some distributions

- How many times does a certain value occur?



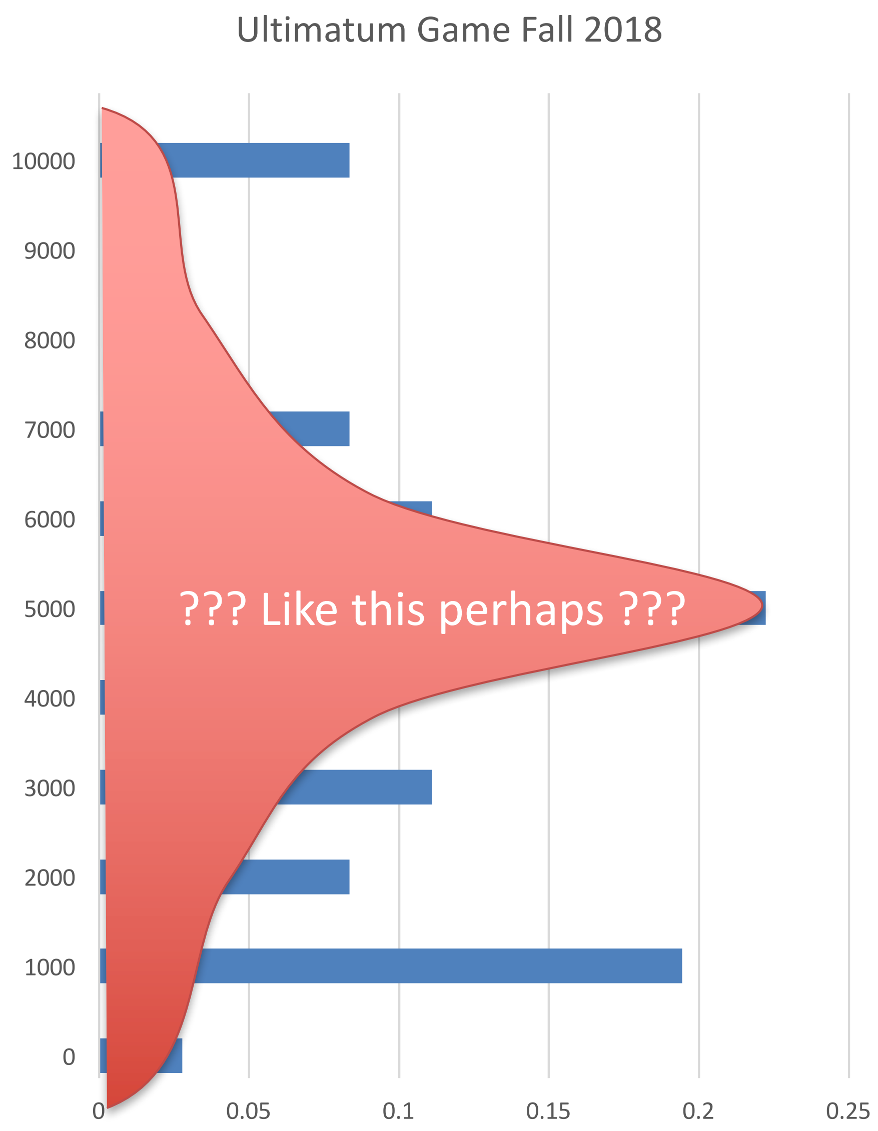
Remember?

- Your responses (36) for the ultimatum game make up a data distribution
- The *mode* of this distribution is 5000Won
  - this is the most often chosen response
- The *mean* of this distribution is 4222Won
  - multiply each money value by the probability and sum!

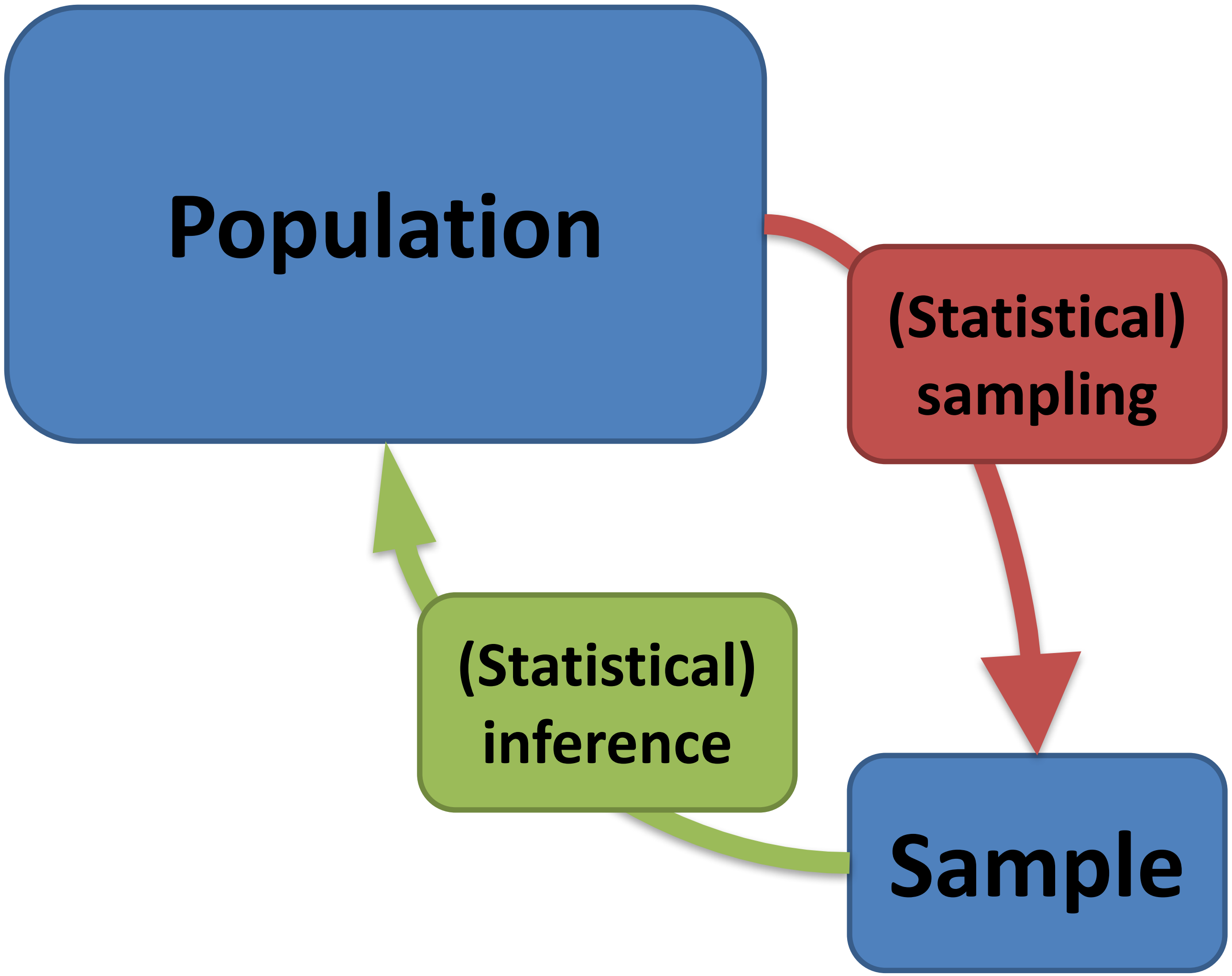


Difference between population and sample

- Important!
- This data distribution is one **sample** – I asked you!
- If I ask other people, this distribution may change



The Big Picture

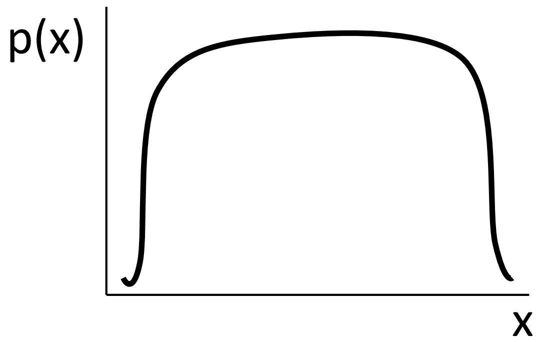


For now

- We will now talk about probability distributions in general
  - this includes population distributions (“all data”)
  - and sample distributions (“your measurements from all data”)

Definition and Concepts

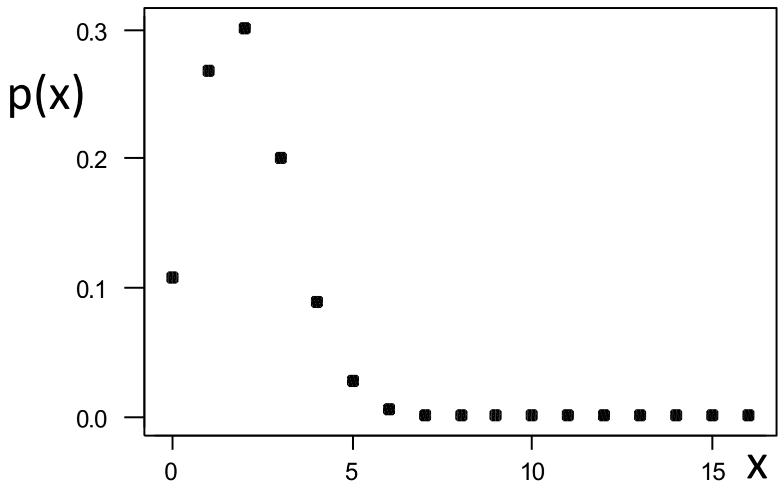
- A probability distribution describes how likely it is that the random variable takes on a certain value
- What is the probability that I obtain any value?
  - i.e., the probability of rolling a 1,2,3,4,5,6 on a die?
- If I sum up all probabilities from any distribution, the total probability must always be equal to 1, or 100%

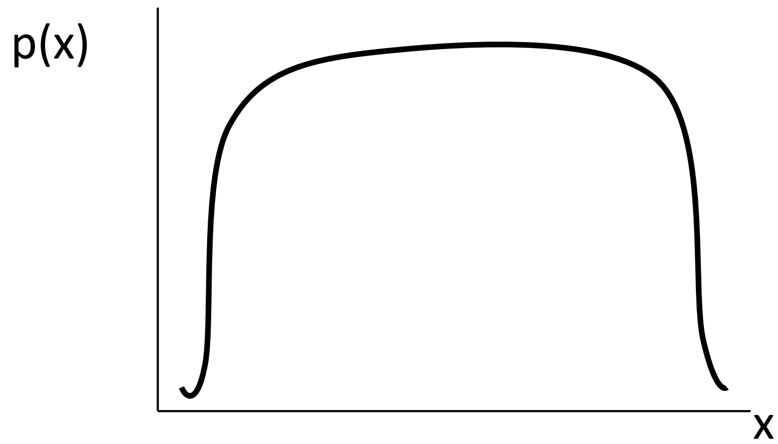


$$\sum_{\text{all } x} p(x) = 1$$

Definition and Concepts

- In general, there are two types of probability distributions
- *Discrete distributions*, for which the random variable can take on only certain values
  - rolling a die, tossing a coin, getting certain weather on a day
- *Continuous distributions*, for which the random variable can take on any value
  - stock prices, reaction time (>0s) and accuracy (any value between 0-100%) for humans





Discrete versus continuous

CognitiveSystems

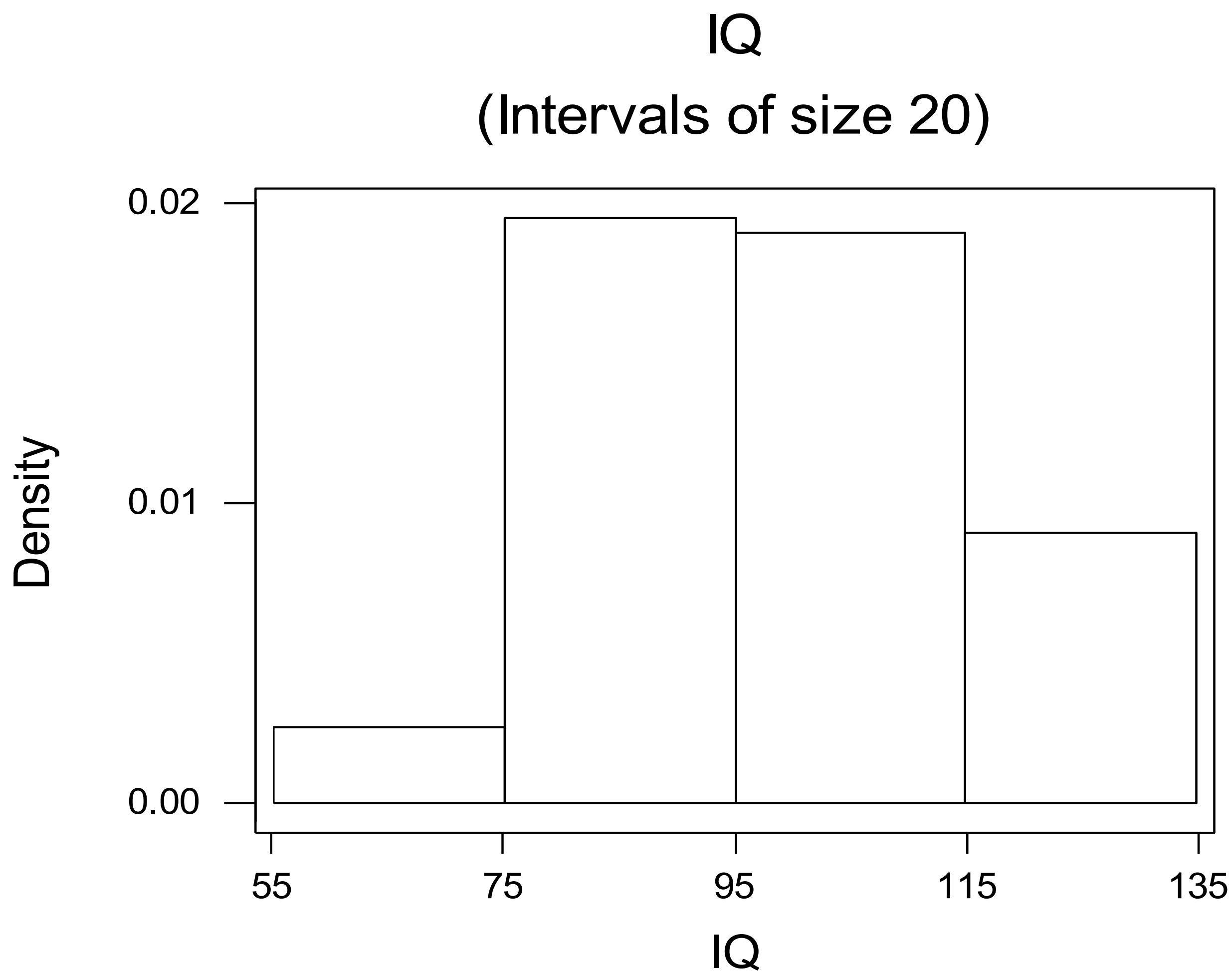
- When analyzing data, we will be interested often in making statements about continuous (ratio) measures
  - "What is the probability that the average grade in this class is more than 90%?"
  - "Are male students on average taller than female students?"
  - "Can people point faster to a red target than a blue one?"
- In the following, I will therefore skip the "discrete" distributions and focus instead on the most important continuous distribution!

Continuous probability distributions

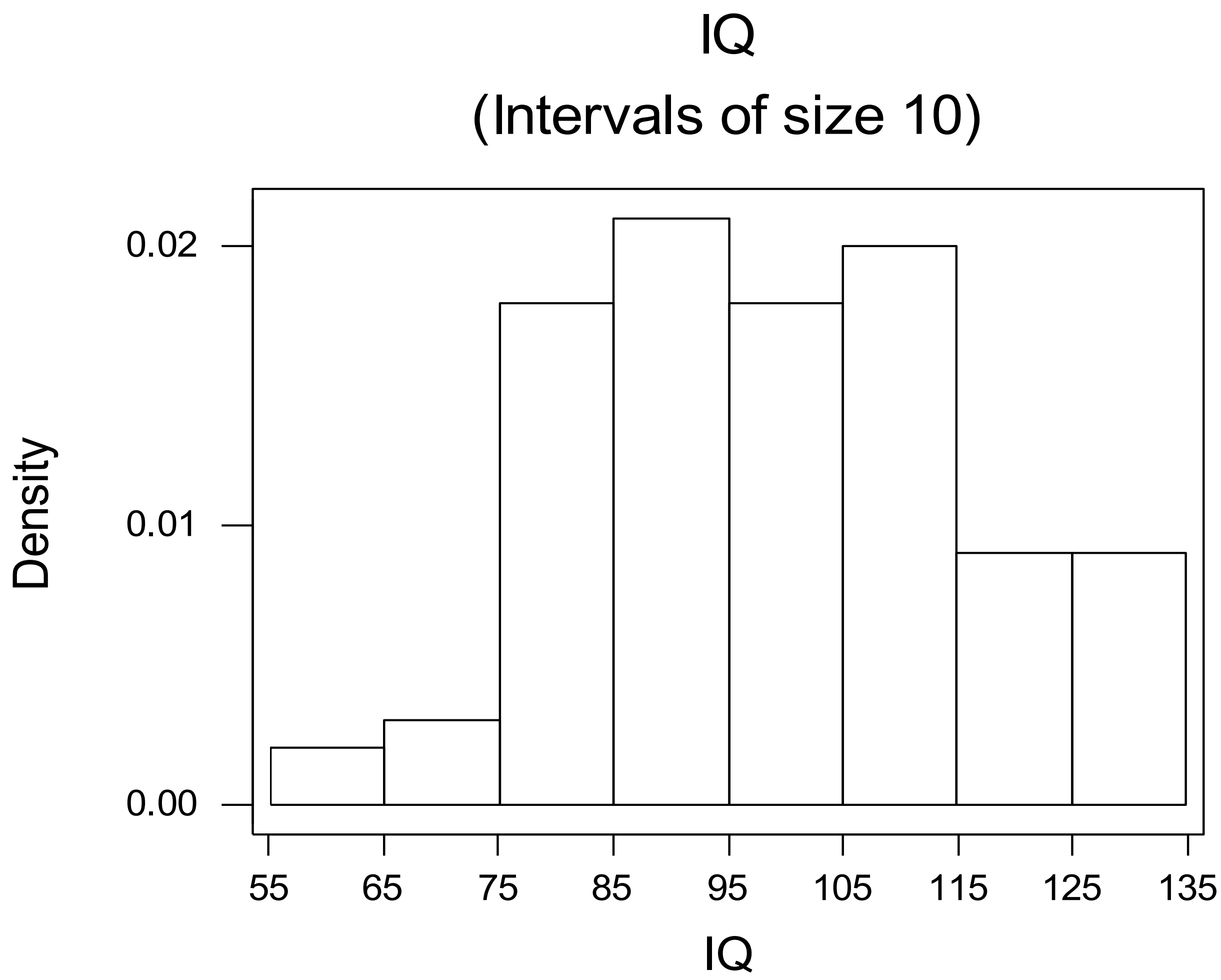
Continuous probability distributions

- A curve that describes the probability of getting any range of values of a random variable, say $P(X > 120)$, $P(X<100)$, $P(110 < X < 120)$
- The area under the curve is equal to the probability
- Since the curve is a probability distribution, it follows that the area under the whole curve most be equal to 1
- Similarly, the probability of getting a very specific number is 0, *e.g.* $P(X=120) = 0$

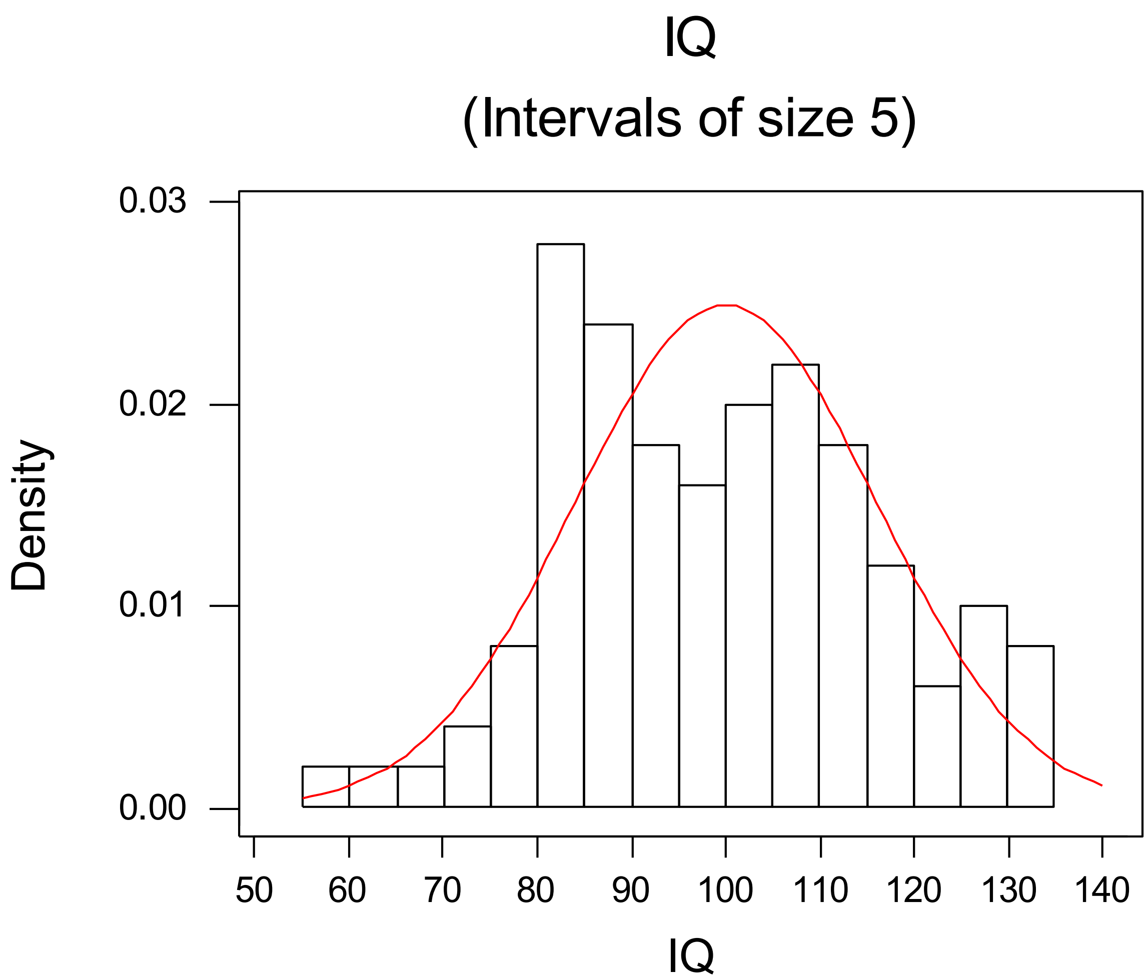
Histogram
(Area of rectangle = probability)



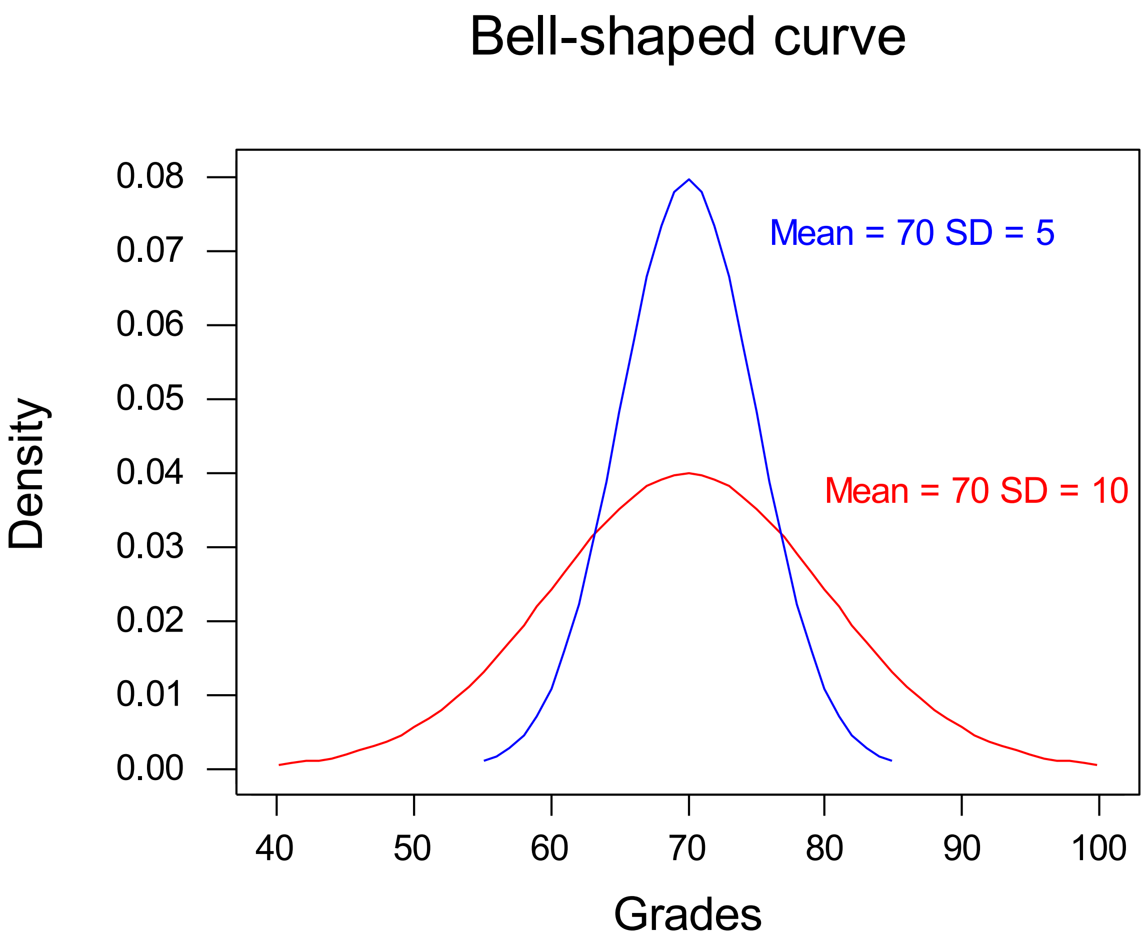
Decrease interval size...



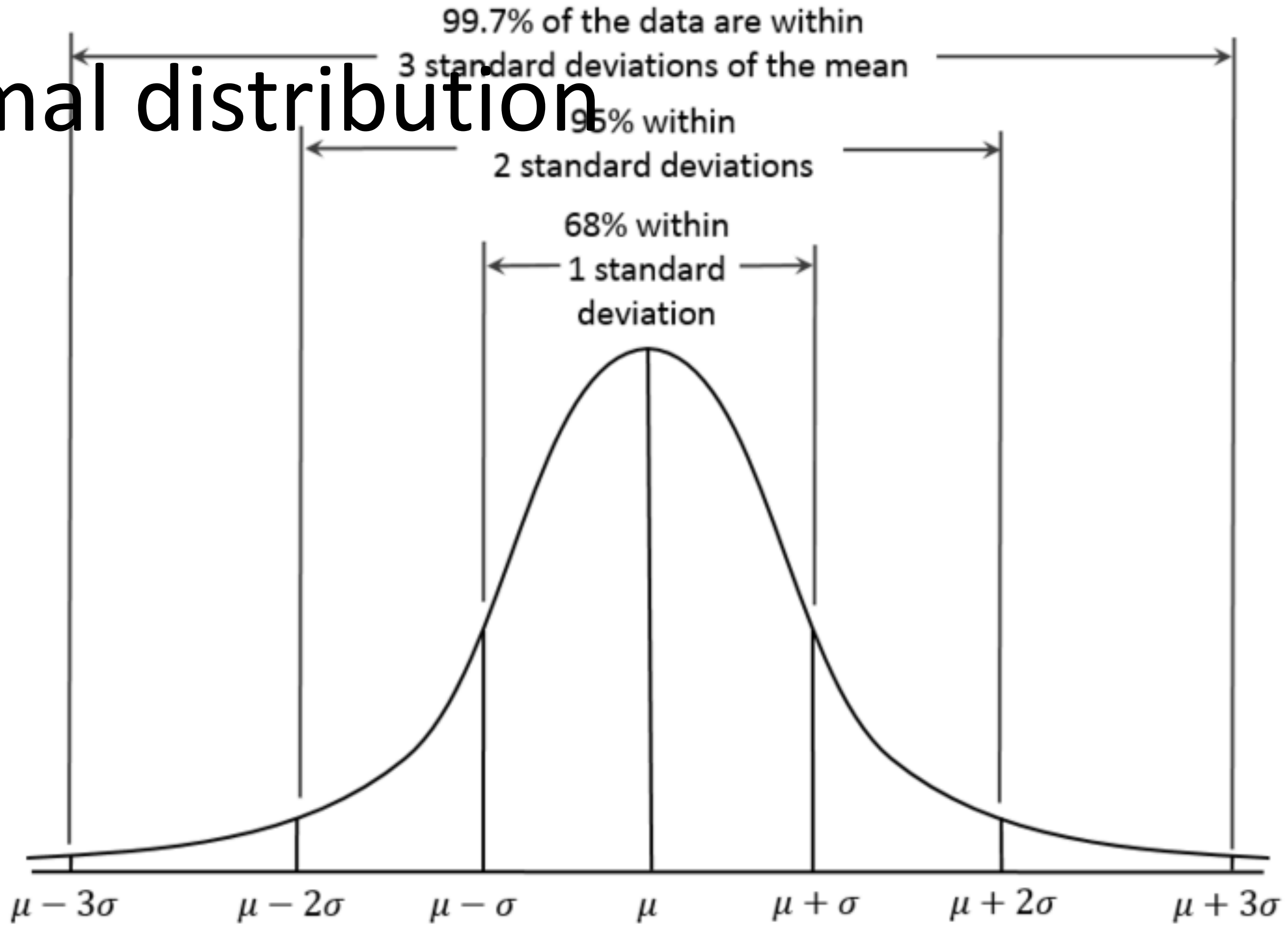
Decrease interval size more….



Normal distribution



Normal distribution

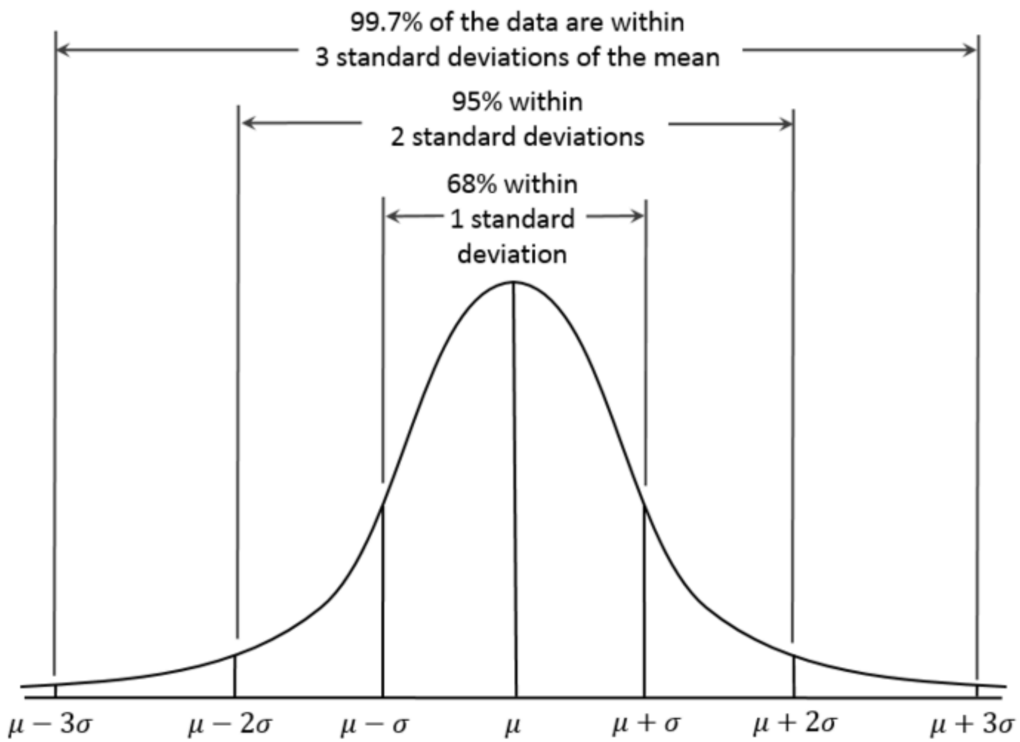


Characteristics of normal distribution

CognitiveSystems

- Symmetric, bell-shaped curve.
- Shape of curve depends on two parameters:
  - mean μ
  - standard deviation σ.
- The center of the normal distribution is at μ.
- The width or spread is determined by σ.

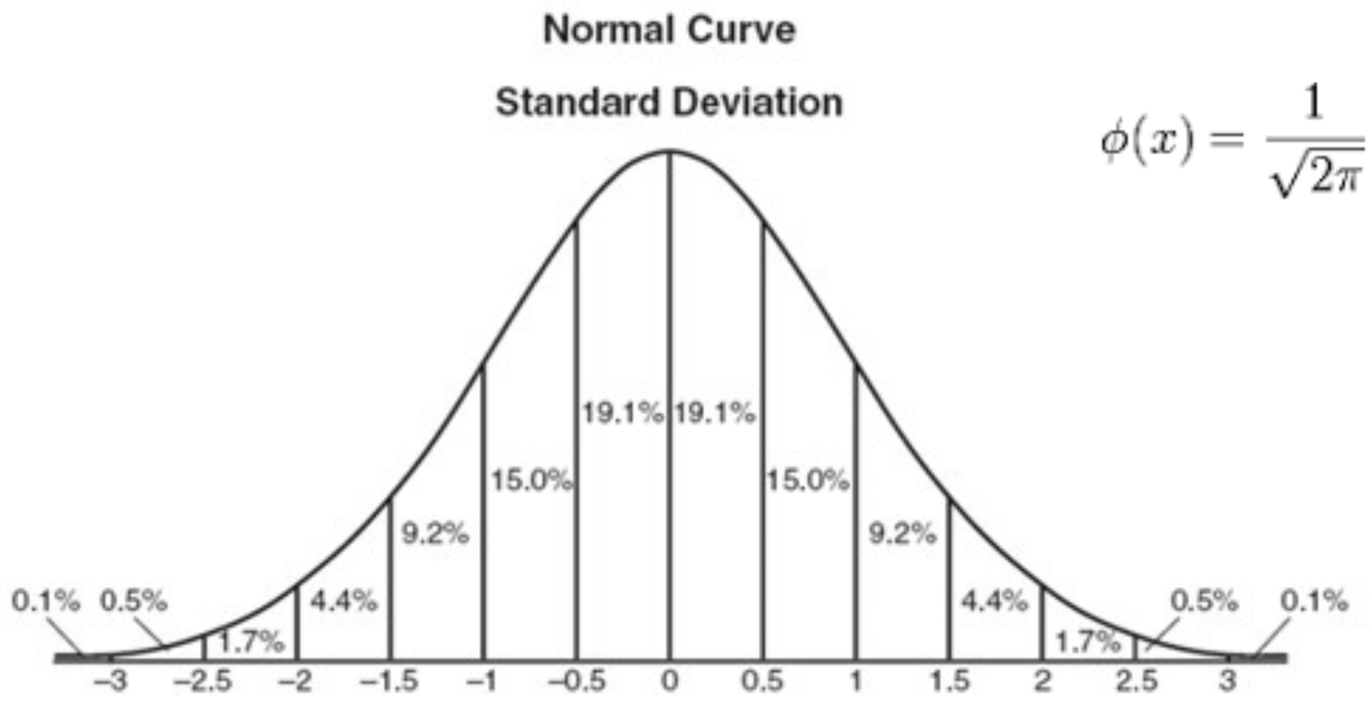
$$f(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Probability = Area under curve

- If you want to know the probability of a certain event, you will need to **integrate the area under the curve**
- Integrating the Gaussian analytically is not possible, so numerical integration methods must be used!
  - Matlab: erf
- Technically, you would need a table of probabilities for every possible normal distribution.
- But there are an infinite number of normal distributions (one for each μ and σ)!!
- Solution is to "**standardize**."

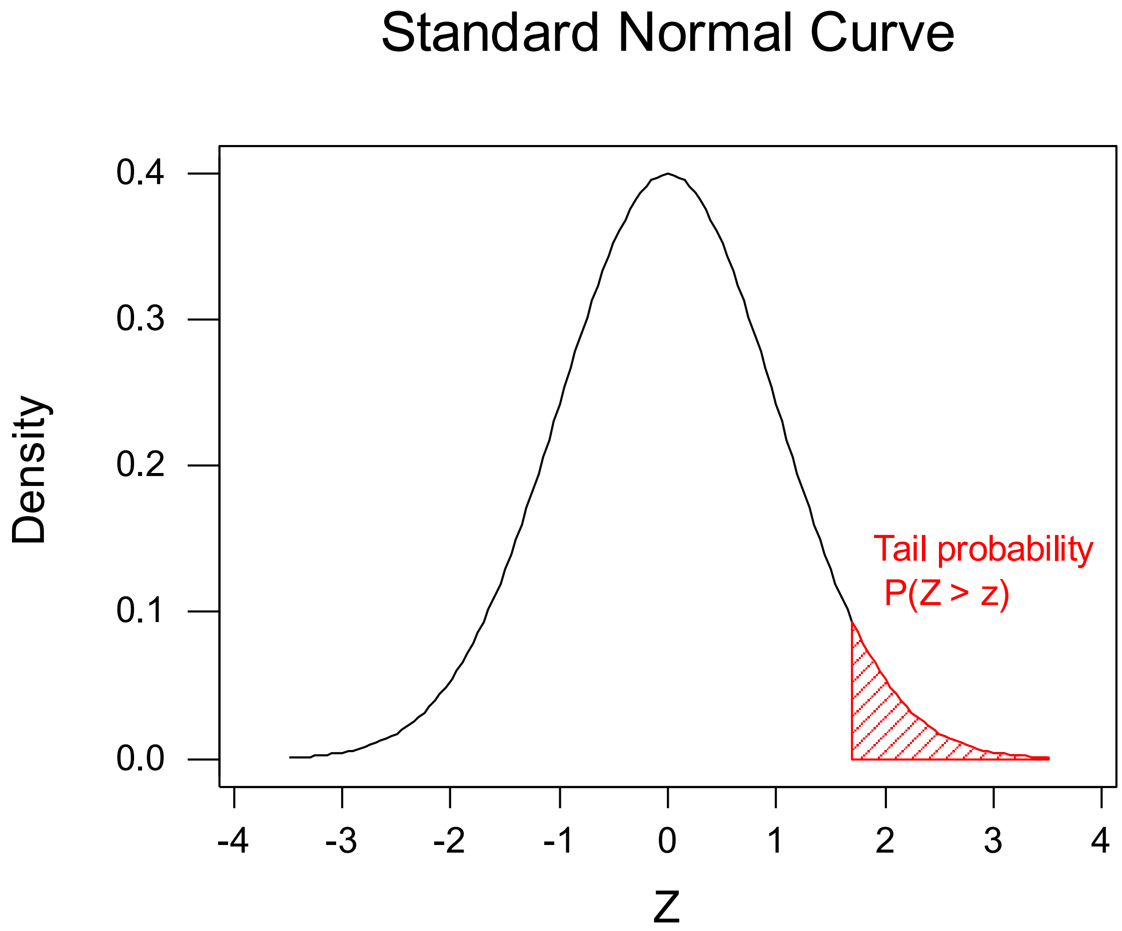
Standard Normal Distribution



$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

- Take value X and subtract its mean μ from it, and then divide by its standard deviation σ. Call the resulting value Z, $Z = (X- \mu)/\sigma$
- Z is called the **standard normal** (mean μ=0, and standard deviation σ=1).
- The areas under the curve need to be calculated only for this

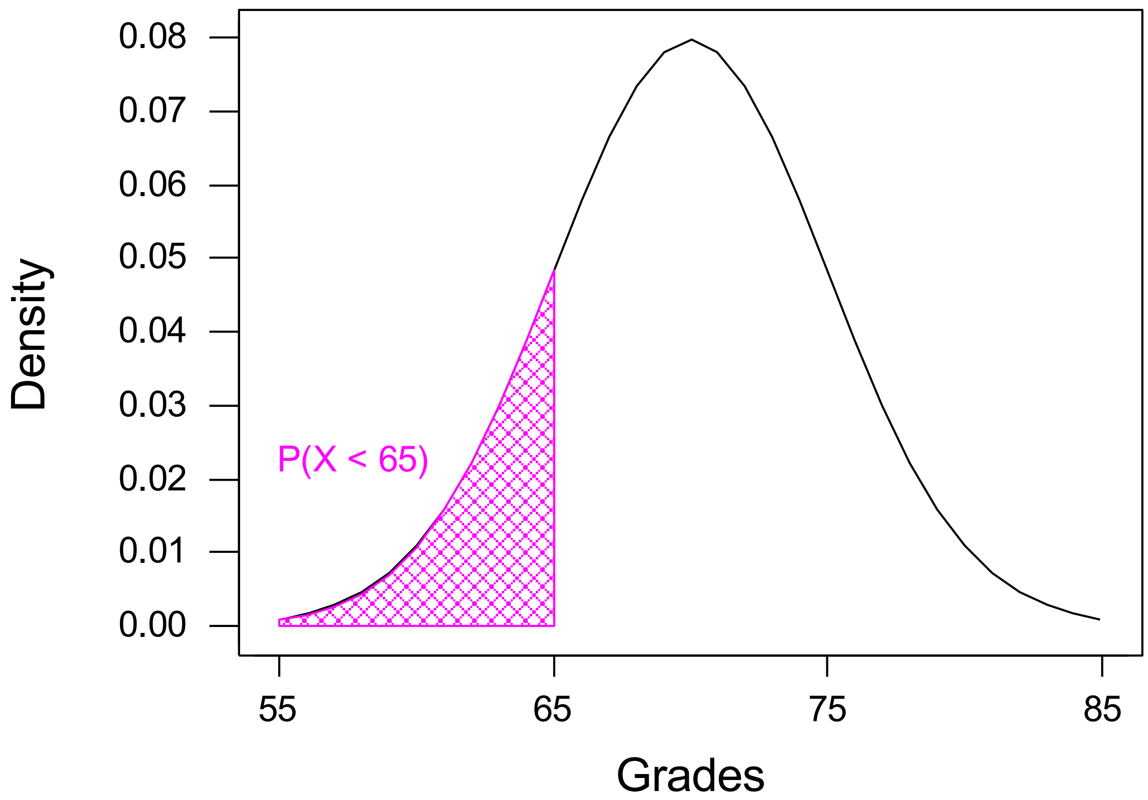
Standard Normal Distribution



Example of use

- Suppose we want to calculate $P[X \leq b]$
- We know that X is normally distributed $X \sim N(\mu, \sigma)$
- Then we calculate $z = \dfrac{b - \mu}{\sigma}$
- Using $P[X \leq b] = P[Z \leq z]$, we can look up the probability $P[Z \leq z]$ from a table of z-values

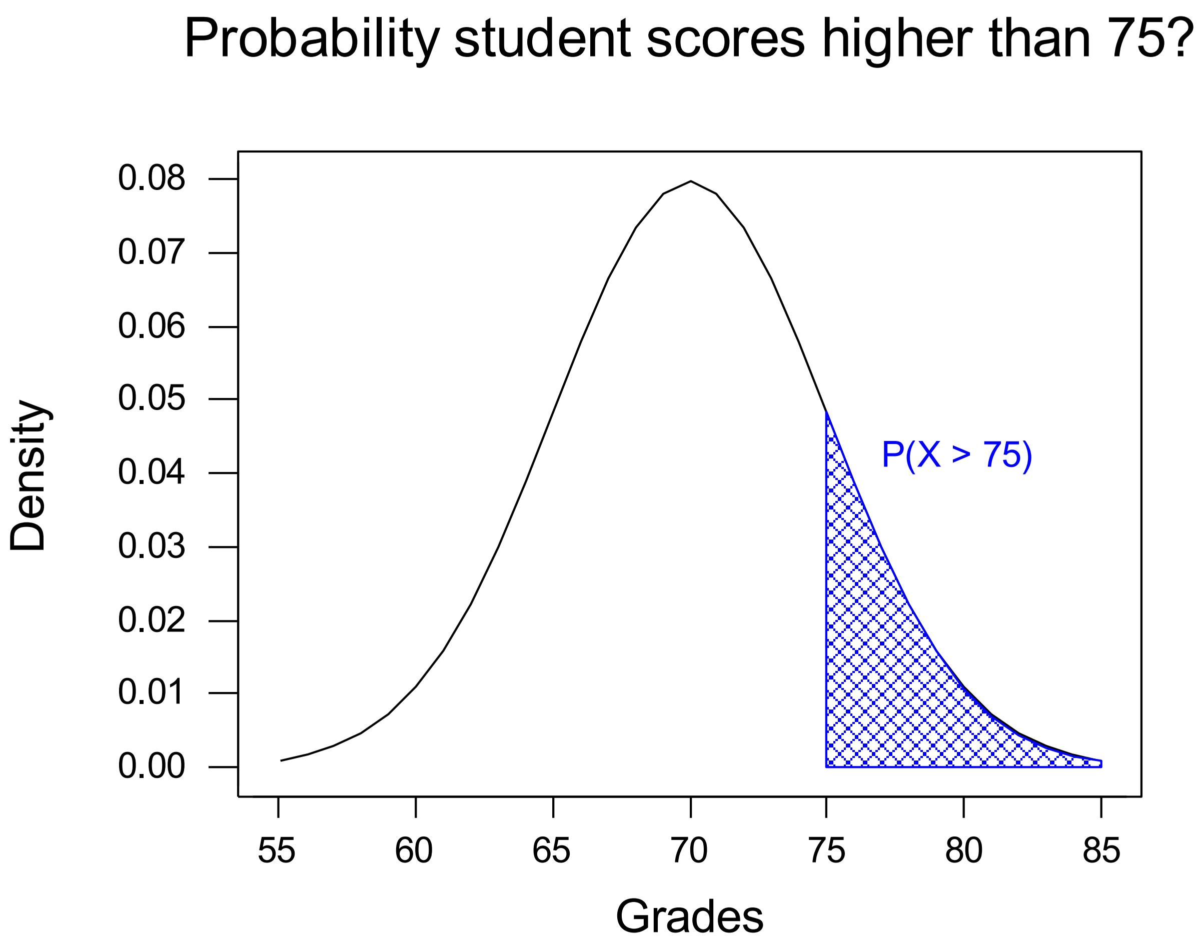
Probability of grade score <65?



Example of use

- Suppose we want to calculate $P[Z > z]$
- Using the fact that probability distributions have to be normalized, we know that this must be $1\text{-}P[Z \leq z]$
- And this of course corresponds to the area to the right of *z*

Probability of score >75?



Example of use

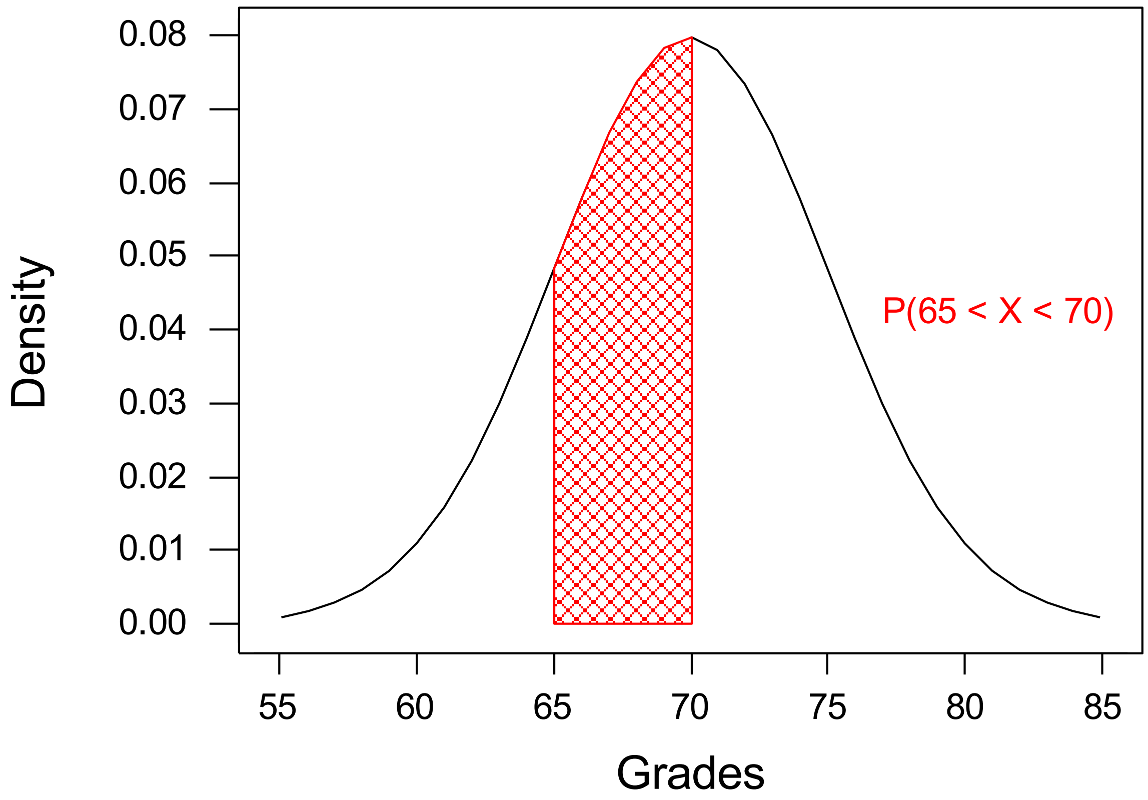
- If we want to know the probability of a range of values

$$P[a \leq Z \leq b]$$

- This is simply the area between a and b and we do:

$$P[a \leq Z \leq b] = P[Z \leq b] - P[Z \leq a]$$

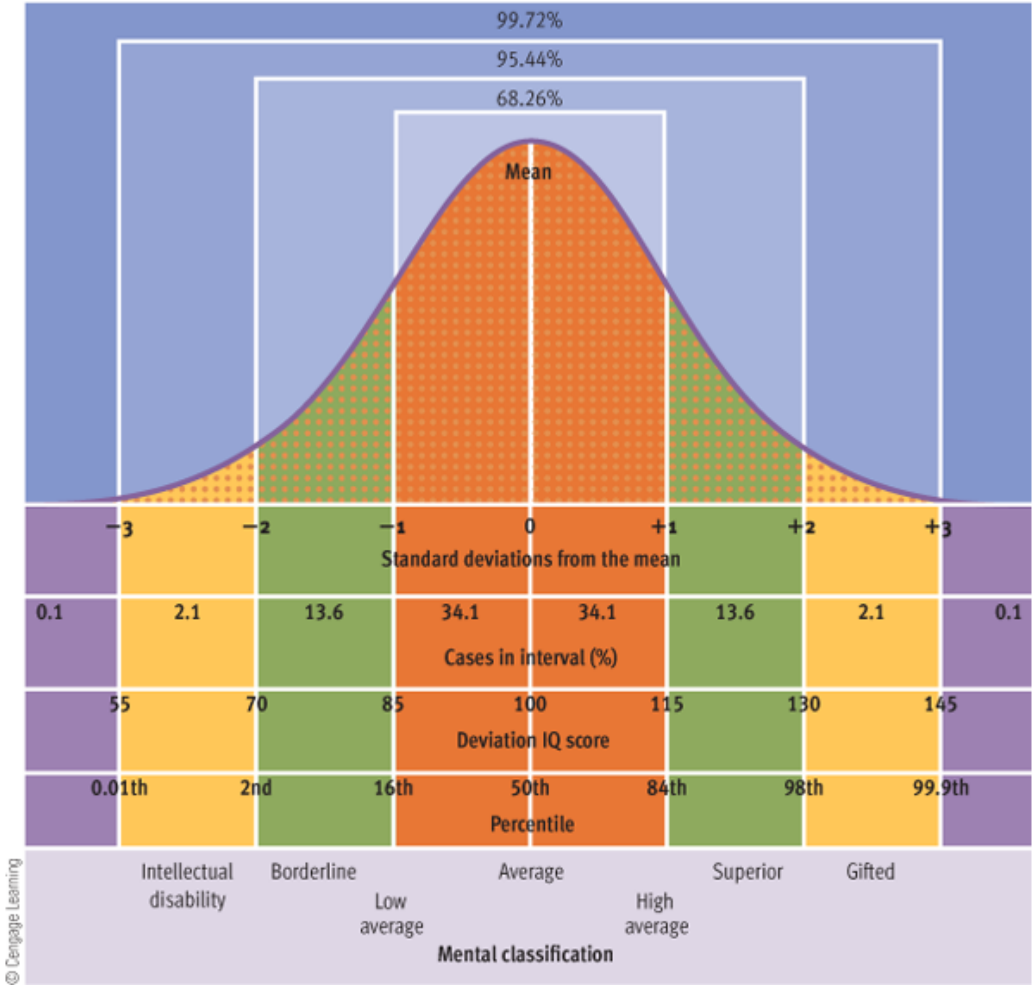
Probability of 65<score<70?



Transportation Example

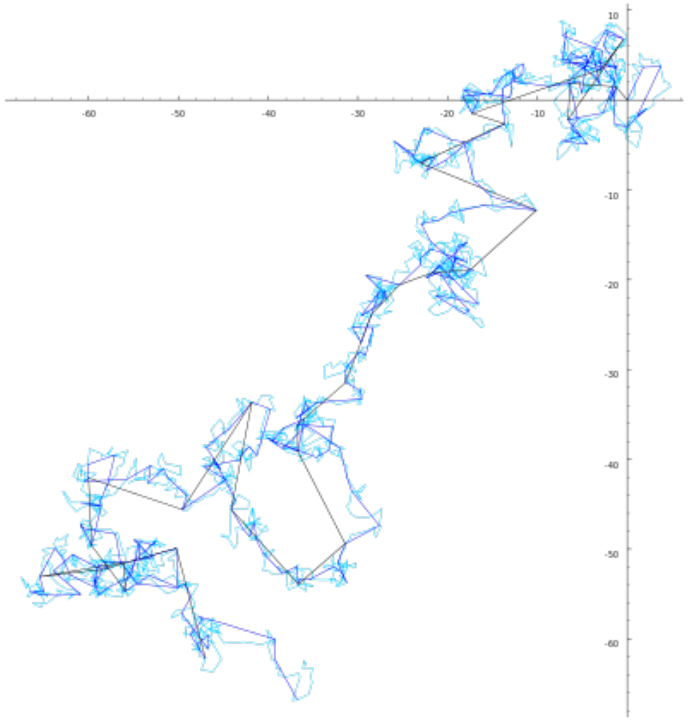
Why is the normal distribution important?

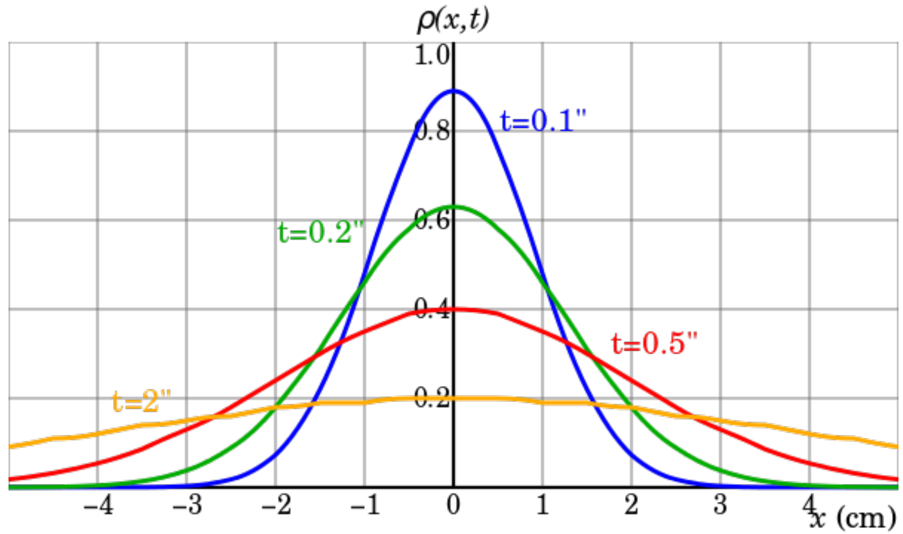
- Use in intelligence testing:
  - Wechsler in 1939 proposed his famous Wechsler Adult Intelligence Scale (WAIS), which is a series of tests that make up one score
  - Wechsler showed that this score was roughly normally distributed
  - In fact, today's IQ tests are defined to be with mean 100 and std of 15!
  - With your score, you can now compare in which range of the population you are



Why is the normal distribution important?

- Use in physics:
  - put a particle in some solution, let it float around, after a certain time the position of the particle will be normally distributed
  - this is called **diffusion** and the associated movement is called Brownian Motion
  - it was first successfully modeled by Einstein



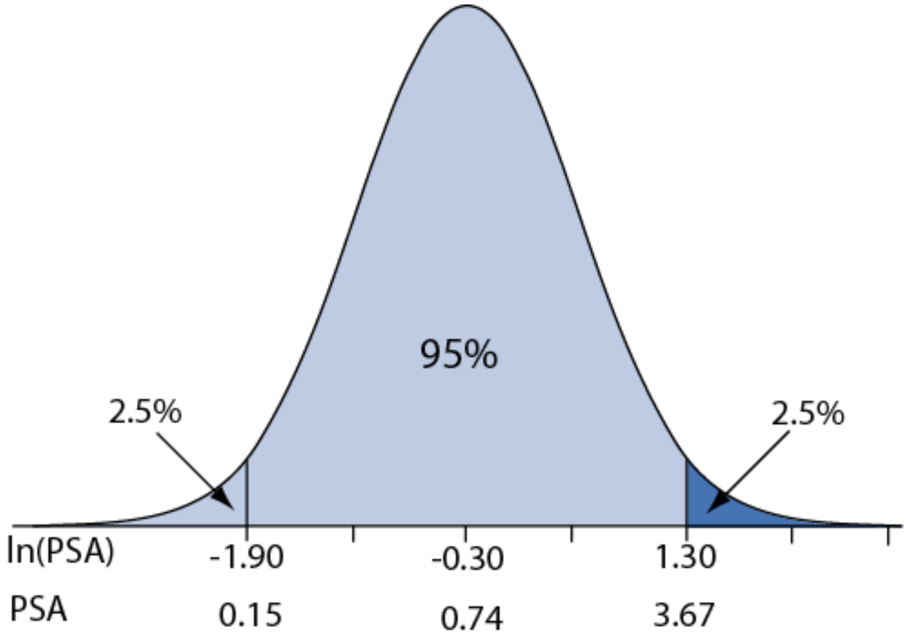


wikipedia

Why is the normal distribution important?

- Use in biology:
  - growth-processes and spread of epidemics are **approximately log-normally distributed**
- Prostate Specific Antigen (PSA) is used to screen for prostate cancer
- In non-diseased populations, it is not normally-distributed, but its **logarithm** is:
  - $\ln(\text{PSA}) \sim N(-0.3, 0.8)$
- We therefore know that 95% of ln(PSA) are within
  $= \mu \pm 2\sigma$
  $= -0.3 \pm (2)(0.8)$
  $= -1.9$ to 1.3

Take exponents of "95% range"
$\Rightarrow e^{-1.9,1.3} =$ 0.15 and 3.67
$\Rightarrow$ Thus, 2.5% of non-diseased population have values greater than 3.67 $\Rightarrow$ use 3.67 as screening cutoff

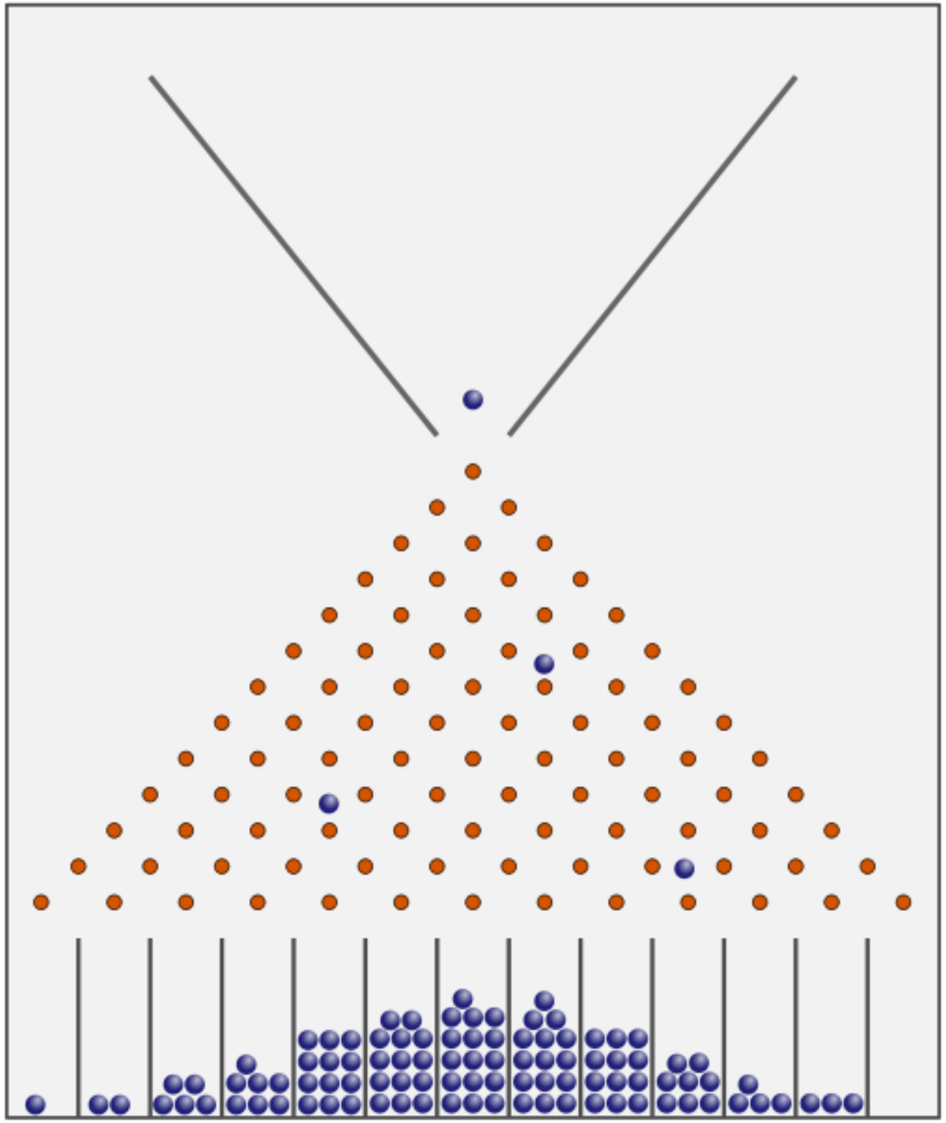


Why is the normal distribution important?

- The bean machine by Francis Galton
- With n rows, and k bins, and p being the probability of left vs right, you get:

$$\binom{n}{k} p^k (1-p)^{n-k}$$

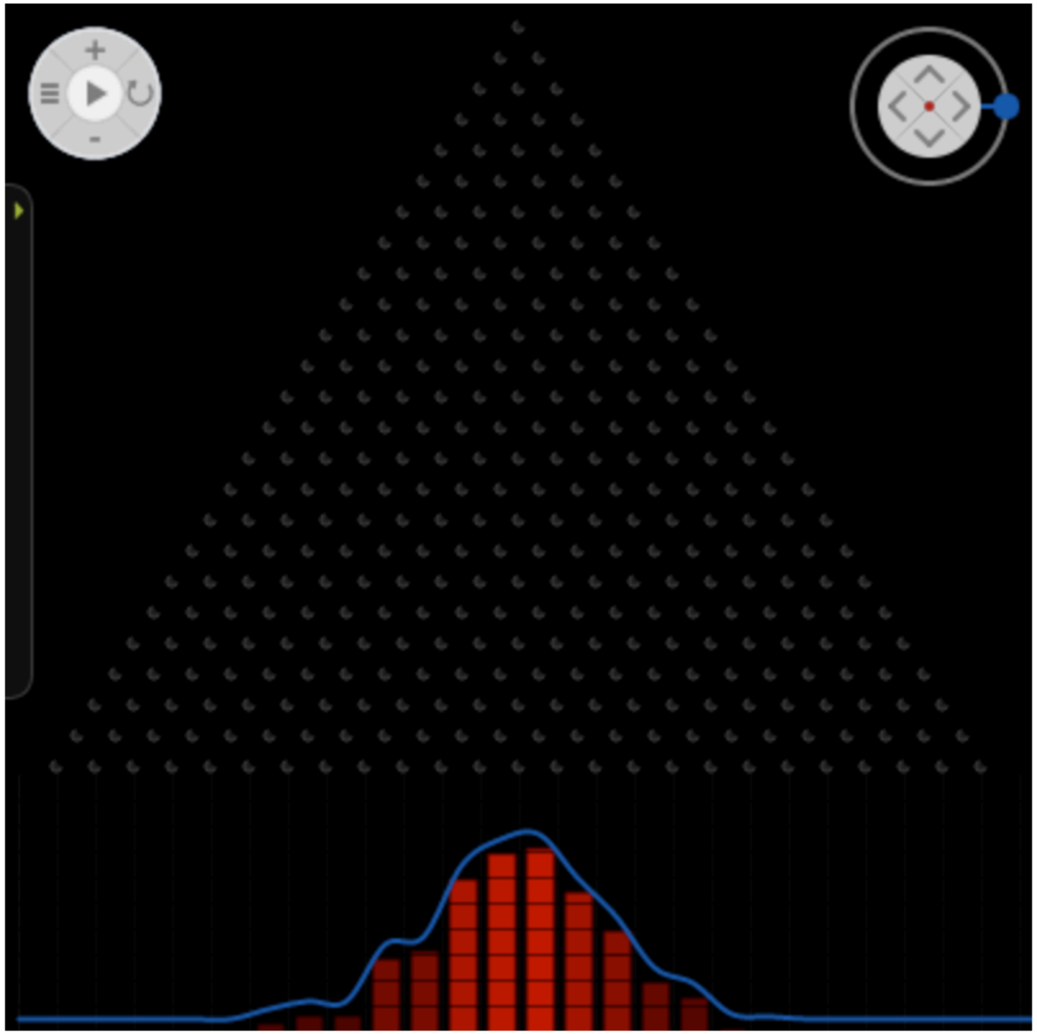
- This is called a Bernoulli distribution
- But it turns out that for **large n**, we get this final distribution – looks **normal**



http://upload.wikimedia.org/wikipedia/commons/7/78/Galton_Box.svg

Why is the normal distribution important?

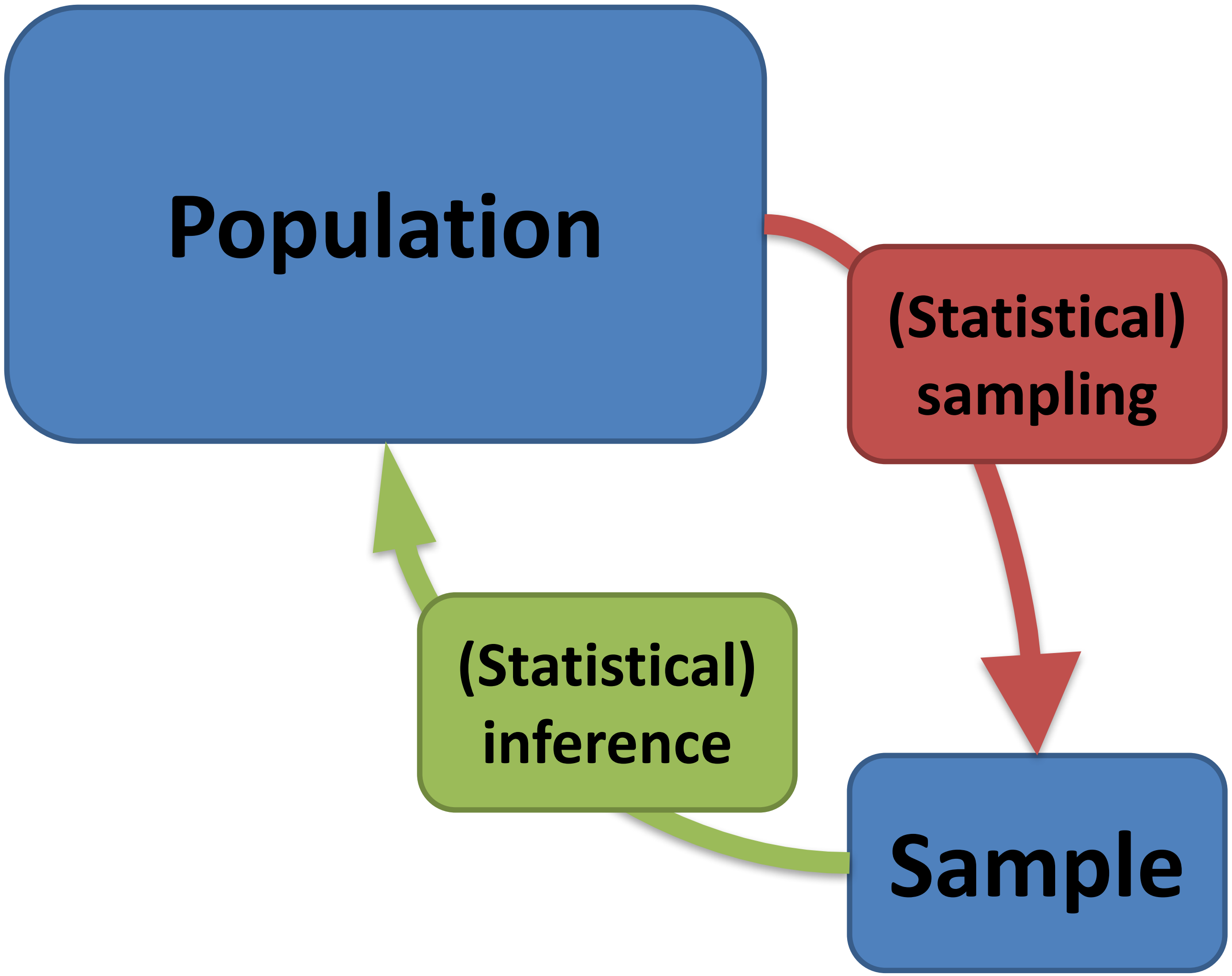
- The live simulation shows the same thing
- It therefore seems like the normal distribution can **approximate** another – totally – different distribution
- This is an important example of the “central-limit theorem” to which we will come back



https://www.khanacademy.org/math/recreational-math/math-warmup/distribution-warmup/a/bean-machines

Sampling Distributions

The Big Picture



Definitions

- Let's say I'm interested in knowing these PARAMETERS of the POPULATION:
  - The average height of a male person
  - The average shoe size of a Korea University undergraduate
- Since testing the POPULATION is not possible, I have to resort to taking a SAMPLE
- My SAMPLE is this class and so I ask you for your heights and your shoe sizes, giving me 80+ SAMPLE POINTS
- I then average the heights and the shoe sizes, giving me two numbers that I call the STATISTIC

Definitions

- Parameter: A number describing a population
  - The average height of a male person
  - The average shoe size of a Korea University undergraduate
- Statistic: A number describing a sample
  - The average height of 86 people in this class
  - The average shoe size of 86 people in this class
- Random Sample: every unit in the population has an equal probability of being included in the sample
  - When looking for the average height of a male person, this class is NOT a good random sample
  - When looking for the average shoe size of a Korea University undergraduate, perhaps this class is a better random sample
- **Sampling Distribution**: the probability distribution of a statistic
  - Every time I have a new class, the actual heights and shoe sizes I ask for change, and so will their average value

Assumptions

- A random sample should represent the population well, so sample statistics from a random sample should provide reasonable estimates of population parameters
  - So the average height may not be a good estimate, but the average shoe size may be!
- All sample statistics have some error in estimating population parameters
  - Both average height and shoe size will NOT be exactly the same as their "real" population value
- If repeated samples are taken from a population and the same statistic (e.g. mean) is calculated from each sample, the statistics will vary, that is, they will have a distribution

Sampling from a normal distribution

CognitiveSystems

- Let's take a population with a normal distribution (IQ, heights,...) and take samples from it
- If we are interested in the mean $\overline{X}$ as the sample statistic, then:

- $\overline{X}$ has a normal distribution with
- mean = $\mu_{\overline{x}} = \mu$
- and
- standard deviation = $\sigma_{\overline{x}} = \dfrac{\sigma}{\sqrt{n}}$

This factor is very important

Sampling from a normal distribution

- In other words, if the we know that the population distribution is normal, take samples, calculate their mean, and do this many times, then:
- The mean of the distribution of those means will approximate the population mean
- The standard deviation of the distribution of those means will depend on the original population standard deviation, but will be **reduced by the square root of the sample size**

Transportation Example

- Speed is normally distributed with mean 45 km/h and standard deviation 6 km/h.
- Take random samples of **n = 4**.
- Then, sample means are normally distributed with mean **45 km/h** and standard error **3 km/h** [from 6/sqrt(4) = 6/2].

Using empirical rule...

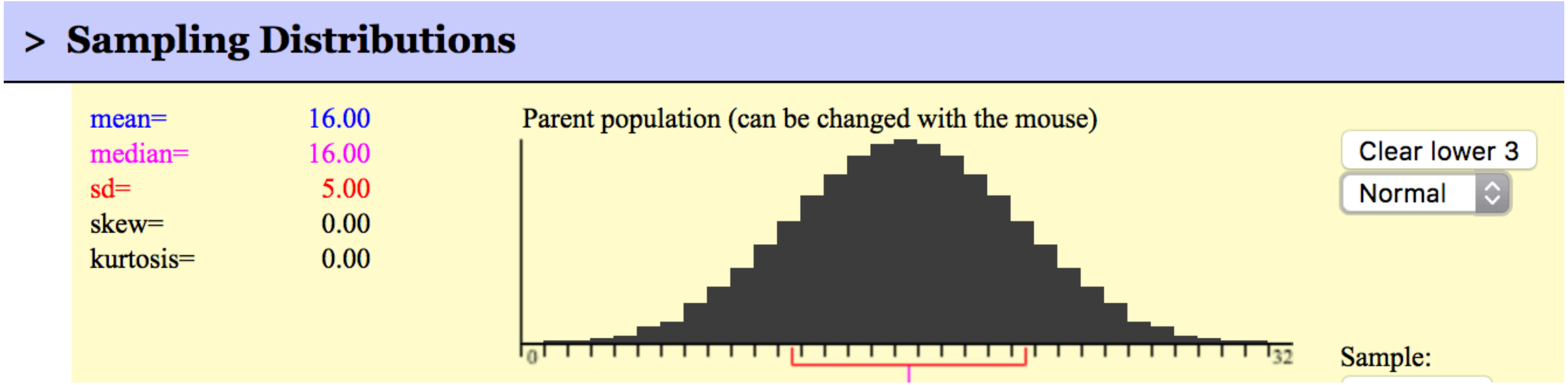
- 68% of samples of **n=4** will have an average speed between 42 and 48 km/h.
- 95% of samples of **n=4** will have an average speed between 39 and 51 km/h.
- 99% of samples of **n=4** will have an average speed between 36 and 54 km/h

What happens if we take larger samples?

- Speed is normally distributed with mean 45 km/h and standard deviation 6 km/h.
- Take random samples of **n = 36** .
- Then, sample means are normally distributed with mean **45 km/h** and standard error **1 km/h** [from 6/sqrt(36) = 6/6].

Let’s test this

- Here’s a javascript simulation that you can use to sample from a normal distribution



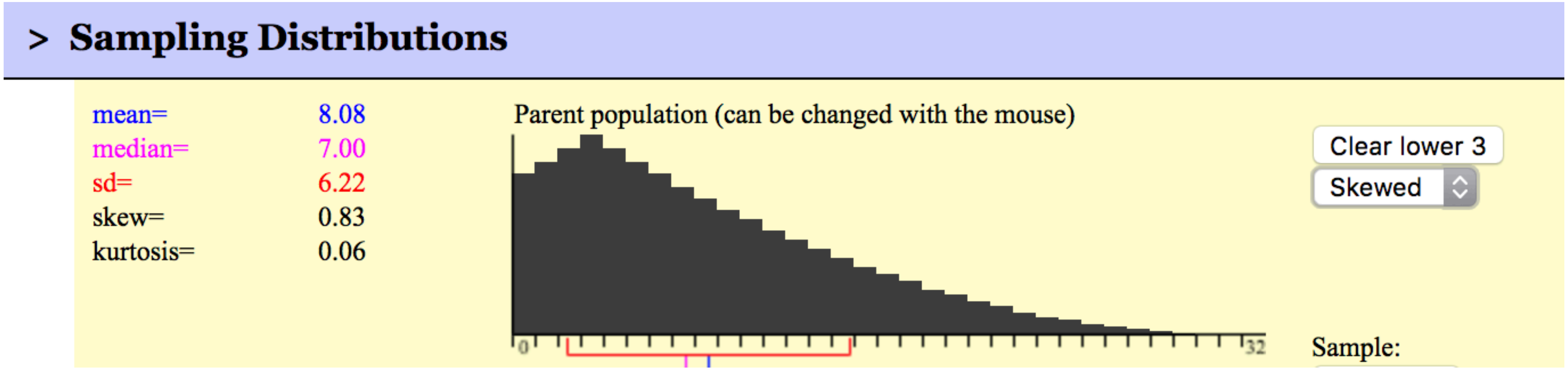
http://onlinestatbook.com/stat_sim/sampling_dist/

Central Limit Theorem

- If the sample size (n) is large enough, $\bar{X}$ has a normal distribution with
- mean = $\mu_{\bar{x}} = \mu$
- and
- standard deviation = $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

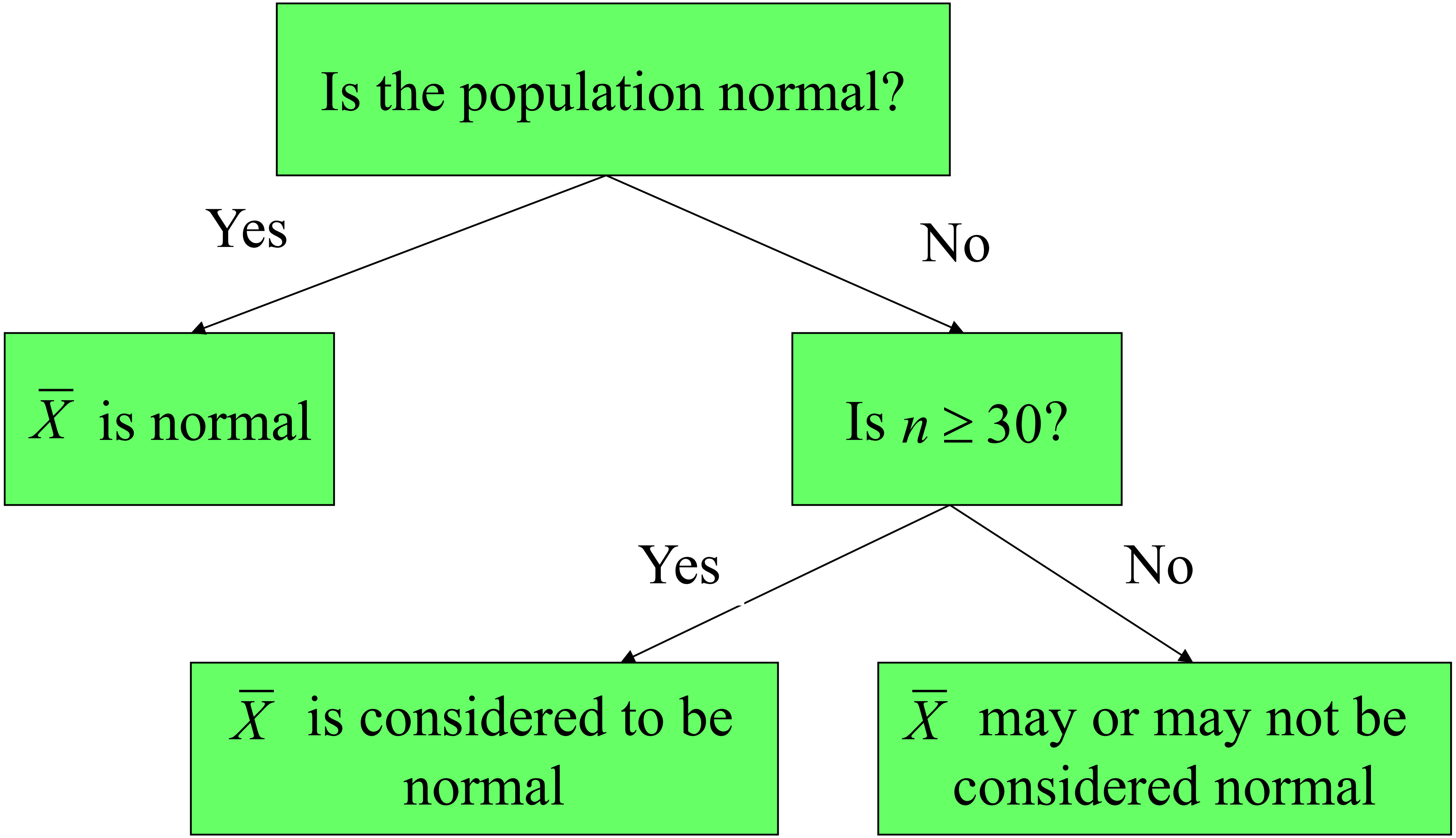
Let's test this

- Here's a javascript simulation that you can use to sample from OTHER distributions



http://onlinestatbook.com/stat_sim/sampling_dist/

What does it mean for n to be large? How large?



Proportion “heads” in 50 tosses

- Bell curve for possible proportions
- Curve centered at true proportion (0.50)
- SD of curve = Square root of [p(1-p)/n]
- SD = sqrt [0.5(1-0.5)/50] = 0.07
- By empirical rule, 68% chance that a proportion will be between 0.43 and 0.57

- **Note that the ”true” underlying distribution of this question was NOT normal – we just used the normal approximation since our number of trials was n=50!**

Can people read minds?

CognitiveSystems

- Let's do an experiment:
- Five cards are randomly shuffled. A random card is picked by the researcher and the participant has to guess which of the cards it was
- Since one guess will not tell us much, we repeat this process n = 80 times

Can people read minds?

CognitiveSystems

- We are talking about people, so we need to test hundreds of people, and each person does n = 80 trials and we calculate the proportion correct
- To answer our question we should be able to tell
  - What sample proportions go beyond luck?
  - Or equivalently: What proportions are within the normal guessing range?

Can people read minds?

- We have 5 cards, so if I also randomly choose a card, then my probability of getting the card correct is $p = 0.2$
- So therefore, typical guessers should get $p=0.2$
- And we also know from the coin toss example, that such a process has a $SD = \mathrm{Sqrt}\,[0.2(1-0.2)/80] = 0.035$

Can people read minds?

- This therefore describes a normal distribution centered around 0.2 with a SD of 0.035
- From that, we immediately know that 99% of all people will be found within proportions correct of 0.095 and 0.305 (+/- 3SD)
- When doing hundreds of tests, we may find several people whose values lie above these boundaries
  - does this mean they have ESP?
- And we could increase our confidence by having people do more trials!

Why is the normal distribution important?

- Use in statistics and for data analysis:
  - since all data has **measurement errors** (noise), and we may assume that these errors are obtained by **many different** kinds of processes, the resulting error can be **approximated** by the normal distribution (central limit theorem!)
  - we saw that if you take a lot of measurements from any distribution and average them, the resulting distribution is approximately normal (central limit theorem!)
    - means or standard deviations from samples are normally distributed

Key concepts

YOUR DATA

- Please everybody fill out the ”Getting data” assignment. It is a short ANONYMOUS survey that asks you some questions – yes, I ask your height and shoe size ☺
- I need this data from you by Sunday, so that our next session can be filled with YOUR input!
- THANKS!