

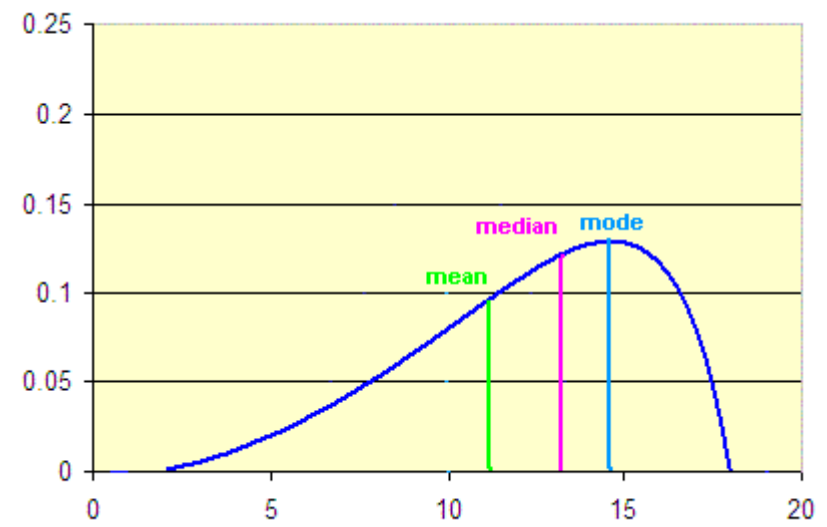
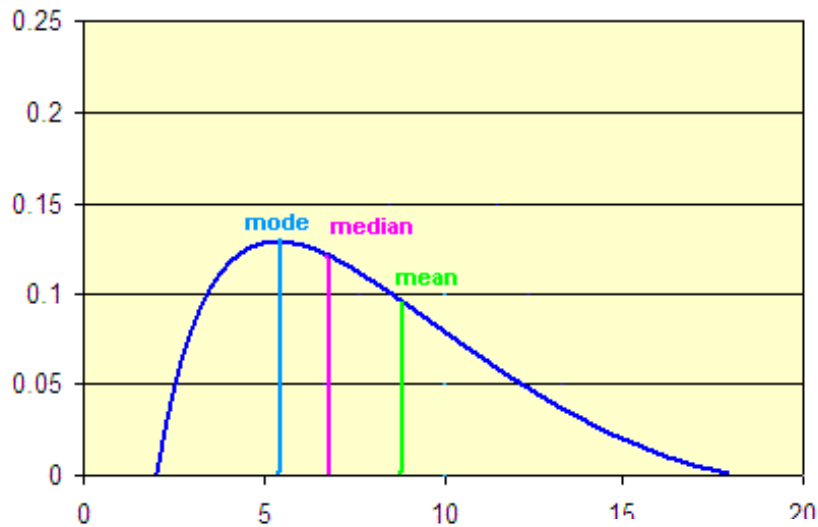
# The normal curve and standardisation

Percentiles, z-scores

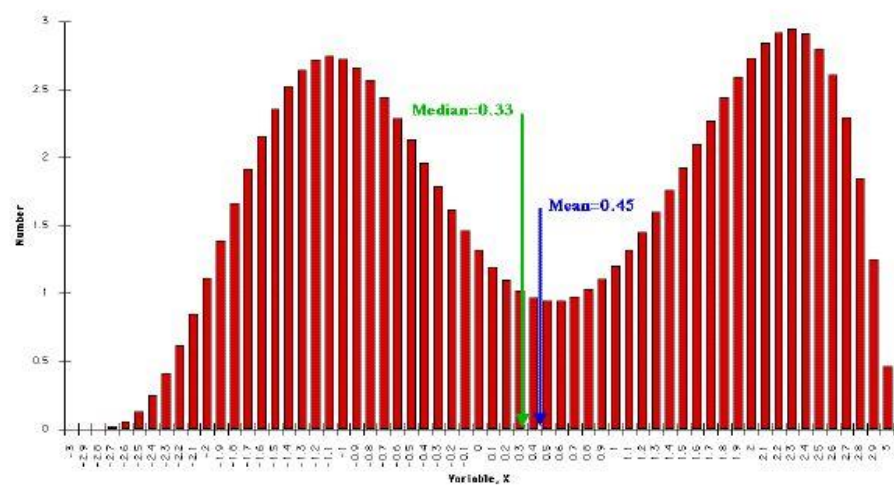
# The normal curve

- Frequencies (histogram)
- Characterised by:
  - Central tendency
    - Mean
    - Median
    - Mode
      - uni, bi, multi
      - Positively skewed, negatively skewed
  - Variability (spread, dispersion)
    - Range
    - Interquartile range
    - Standard deviation and Variance

Normal curves?  
Blue, pink and green marks?



Bimodal distribution



# Another key term

- Outlier

# The Normal Distribution

Described by:

- Shape: unimodal
- Central Tendency: mean = median = mode
- Variability??

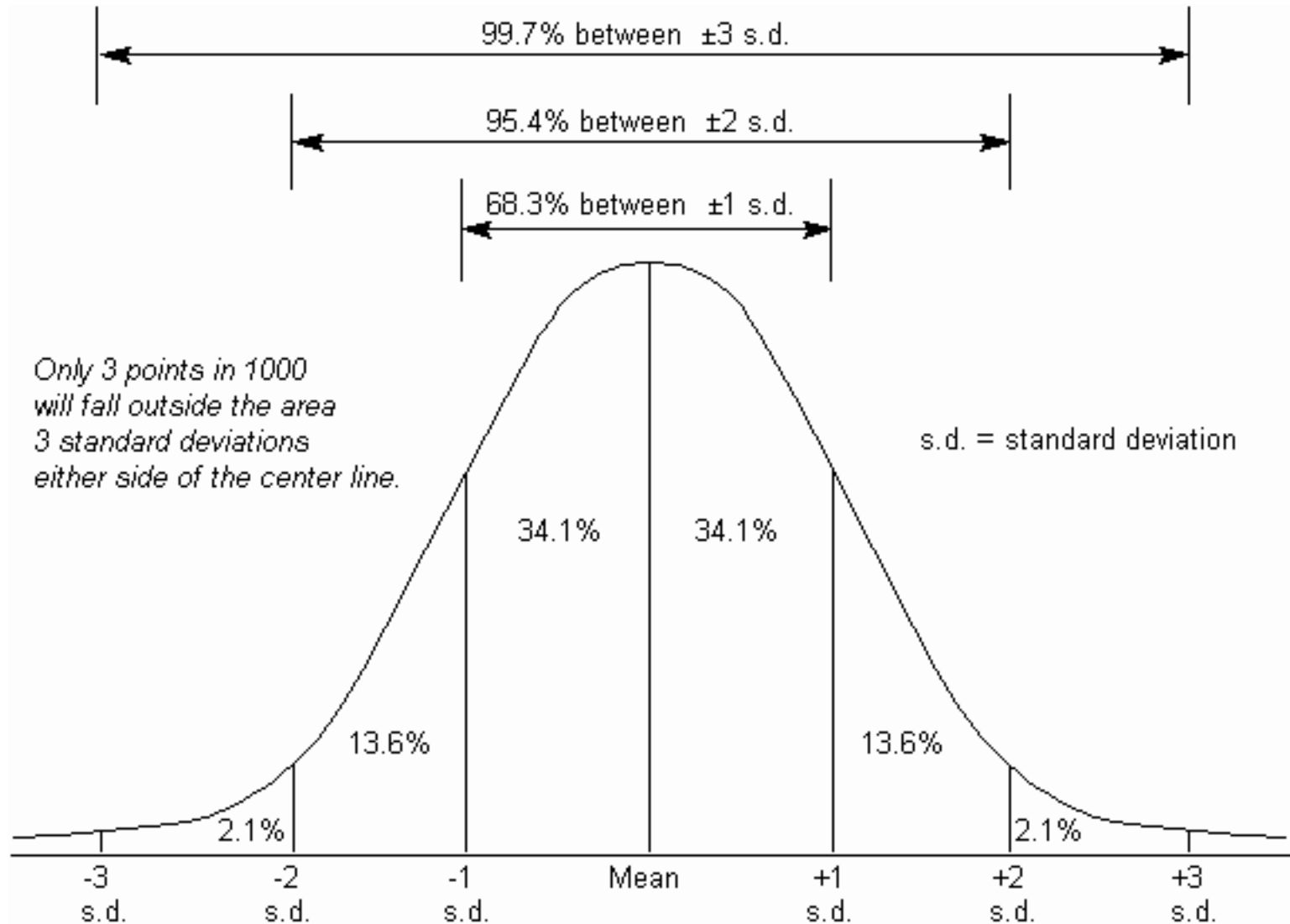
The normal curve and variability

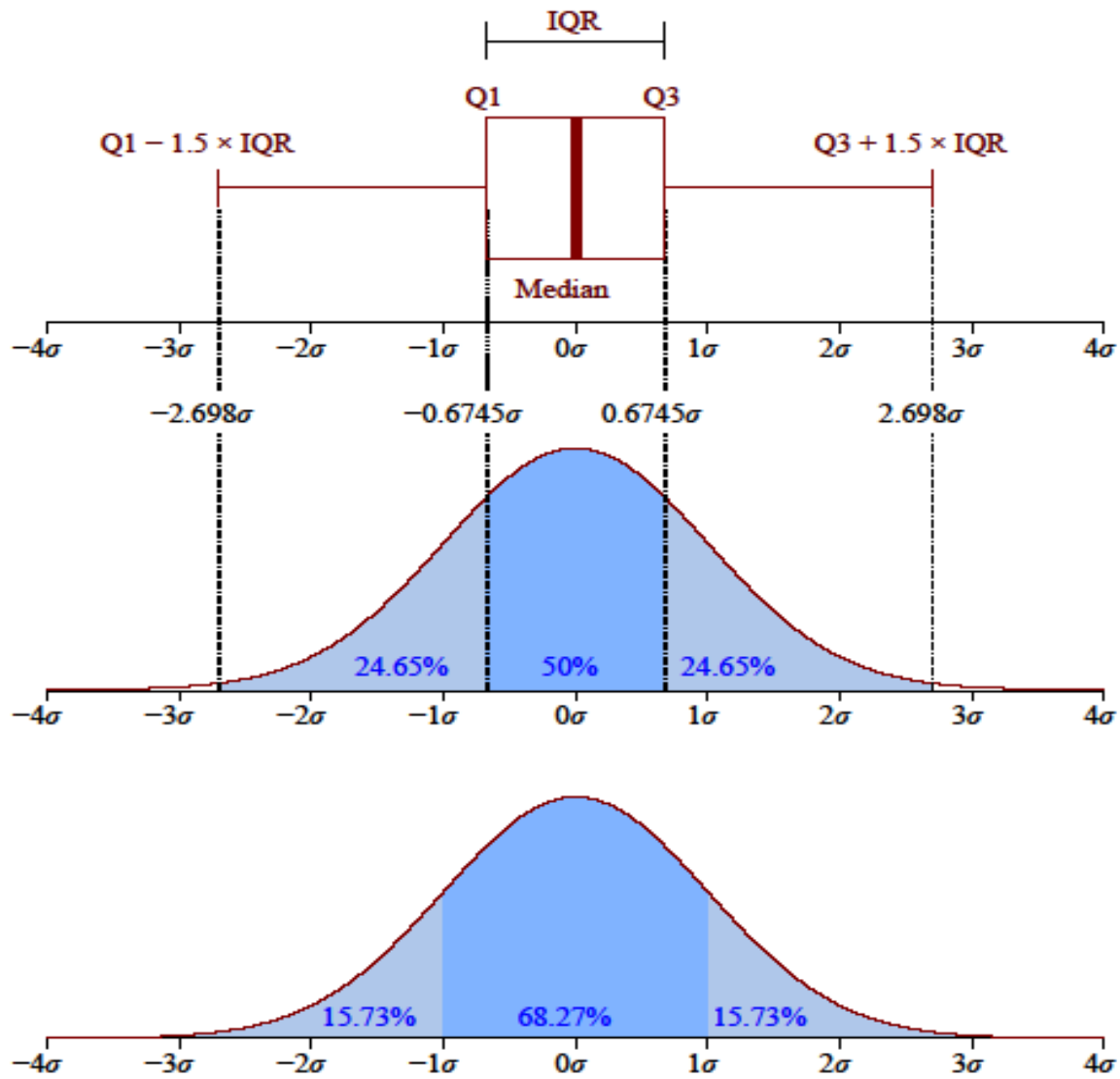
- Several natural distributions are described by the normal curve (if the sample is large enough)
- If the data are normally distributed, we can standardise the scores and compare different data sets.
- Lucy is 169 cm tall.
- Lucy has 673 Facebook friends.
- Is she as popular as she is tall?

- We need a measure for which we know the percentiles (what percentage of the data are below the score)
- The distance from the mean in Standard Deviations (for a normal curve)
- Standard score (z-score): the distance below or above the mean in Standard Deviations



# Probability Distribution





Boxplot and  
the normal  
distribution

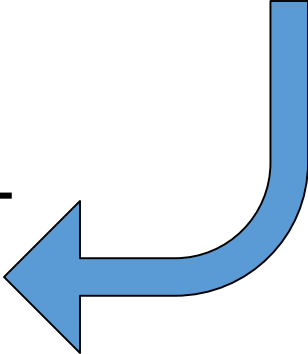
# Standard Scores ( $Z$ -scores)

- A way to put scores on a common scale
- Computed by subtracting the mean from the score and dividing by the standard deviation
- Interpreting the  $Z$ -score
  - Positive  $Z$ -scores are above the mean; negative  $Z$ -scores are below the mean
  - The larger the absolute value of the  $Z$ -score, the further the score is from the mean
  - A  $z$ -score of 0 means that the value = ?

# Z-Scores: The Standard Deviation “Meter”

- Use Z-scores to express values regardless of the original unit of measure
- Once you have the standard deviation, you can go from raw scores to z-scores, and from z-scores to raw scores.

$$SD = \sqrt{\frac{\Sigma(X - M)^2}{N - 1}}$$

$$z = \frac{(X - M)}{SD}$$


$$X = z(SD) + M$$

# How To Calculate Z-scores

$$z = \frac{(X - \mu)}{\sigma}$$

Wolf pups in den			
X	X-M	(X-M)^2	Z-score
5			
3			
6			
9			
...			

M =

SS =

s<sup>2</sup> =

SD =

# How To Calculate Z-scores

$$z = \frac{(X - M)}{SD}$$

Pups in den			
X	X-M	(X-M)^2	Z-score
5	-0.75	0.56	-0.3
3	-2.75	7.56	-1.1
6	0.25	0.06	0.1
9	3.25	10.56	1.3
...			

$$M = 5.75$$

$$SS = 18.75$$

$$s^2 = 6.25$$

$$SD = 2.5$$

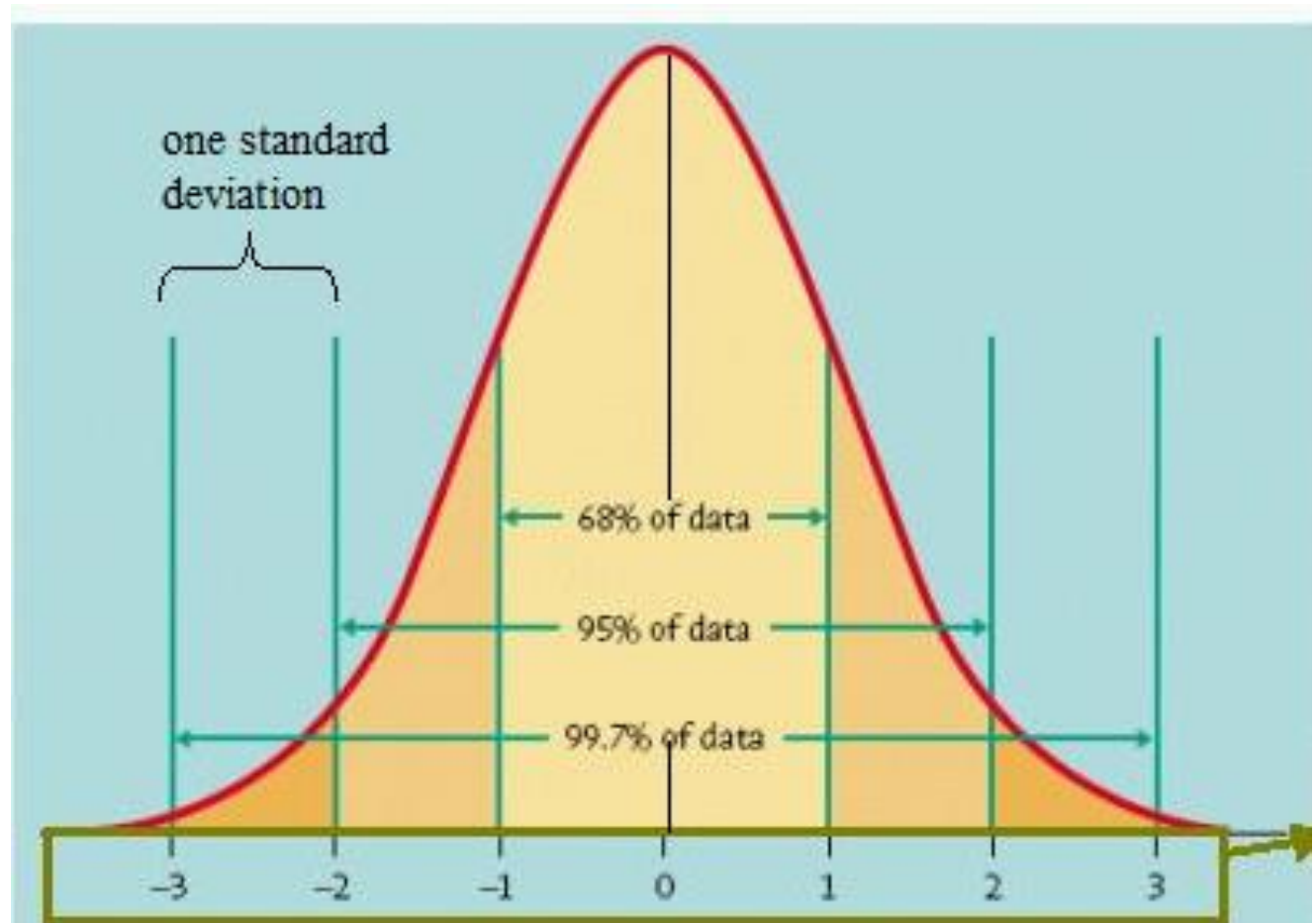
# exercise

- Take your data set (height, Facebook friends, etc.)
- Calculate the z-scores
  - SPSS: Analyze -> Descriptive Statistics -> Descriptives -> Save standardized values as variables
- Take an individual.
  - Are they as tall as they are popular?
  - Do they drink as much as they sleep?

# Z-Scores & Percentiles

<http://www.mathsisfun.com/data/standard-normal-distribution-table.html>

- Each z-score is associated with a *percentile*.
  - Z-scores tell us the percentile of a particular score
  - Can tell us % of population above or below a score, and the % of population between the score and the *mean* and the *tail*.





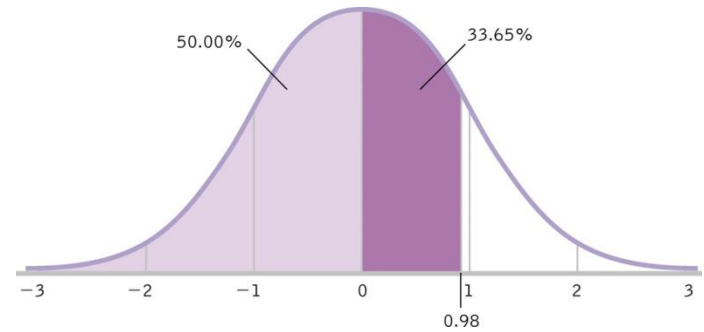
## Example: Fitness

- Jessica can run round Margit sziget in 45 minutes
- The mean for women is 38 minutes
- The standard deviation for women is 6.35

$$z = \frac{(X - M)}{SD} = \frac{45 - 38}{6.35} = 1.1$$

- According to z-score table, the percentile associated with  $z = 1.1$  is 36.4%

# Running Example:



- This tells us that
  - 36.4% of the population is between the mean and Jessica's score.
  - There is a 36.4% chance of Jess being slower than the average by this amount BY CHANCE ALONE
  - $50 + 36.4 = 86.4\%$  of women are faster than Jessica,
  - and there is a  $100 - 86.4\% = 13.6\%$  chance of someone being this slow by CHANCE ALONE. (The nearest tail is the positive tail (100<sup>th</sup> percentile), so you take the difference between 100 and 86.4)

# Running example: exercise 1

- Cecilia runs round the island in 33 minutes.
- What is the probability that someone is that slow by chance alone?  
(= What percentage of scores are between the Cecilia's score and the nearest tail of the distribution?)

(The mean for women is 38, The standard deviation for women is 6.35)

- **ESTIMATE THE ANSWER FIRST, THEN DO THE CALCULATIONS**

# Running example: exercise 2

- A person dressed in unisex winter clothes is known to run round the island in 22 minutes.
- What is the probability that this person is a member of the population of women?

(The mean for women is 38 minutes, The standard deviation for women is 6.35)

- **ESTIMATE THE ANSWER FIRST, THEN DO THE CALCULATIONS**

# Running example: Exercise 3

- What's the probability of someone being as far from the mean in **either direction** as
  - Jessica?
  - Cecilia?
  - The unidentified stranger?

hypothesis testing  
the z test

# Inferential statistics

- Your hypothesis predicts a difference between means
- If you look at the means, you can tell whether there is a difference or not but you cannot tell whether this difference is due to chance
  - because there is chance variation across people
  - every time you test the SAME people under the SAME circumstances, you'll have slightly different results by chance
- Inferential statistics will tell you how likely it is that the difference you observe is due to chance
  - In other words: how likely it is that one mean comes from the same population than the other mean
  - if the difference is probably not due to chance, you have **statistically significant** results

# What do we mean by “probably”

Statistical test tells you how likely it is that your observed means come from the same population

- Highest probability: 100% ( $p = 1$ )
  - meaning: there is a 100% chance that the difference between the means is due to chance
  - you definitely cannot reject the null hypothesis
- Low probability: 0.00000000001% ( $p = .0000000001$ )
  - meaning: there is an extremely low chance that the difference between the means is due to chance
  - you can definitely reject the null hypothesis



# What do we mean by “probably”

We have to decide on a cut-off point (critical value, alpha)

- Depends on how much risk of Type I error you are willing to take
- High stakes research: alpha = .001
  - you take your results to be statistically significant if  $p < .001$
  - this means that there is a 1% chance that the difference between the means is due to chance (Type I error)
- Low stakes research: alpha = .05
  - you take your results to be statistically significant if  $p < .05$
  - this means that there is \_\_\_\_\_ chance that the difference between the means is due to chance. (Type I error)

# One-tailed or two-tailed

- A one-tailed hypothesis: when the difference is predicted to be in a certain direction
- Two-tailed hypothesis: when the difference is predicted to be in either direction
- If your cut-off point is .05 for a one-tailed hypothesis, what is it for a two-tailed hypothesis?

# Statistical Significance

- A finding is statistically significant if the data differ from what we would expect from chance alone, if there were, in fact, no actual difference.
- They may not be significant in the sense of big, important differences, but they occurred with a probability below the critical cutoff value.
- Choice of statistical test to use depends on the distribution of the data and on the research design

The simplest test:  
z test

# The Central Limit Theorem

- The **central limit theorem** states that IF you take:
  - a. **several**
  - b. **random samples**
  - c. from ANY SHAPED population
- THEN the distribution of **sample means** will become approximately **normally distributed**
  - a) becoming more accurate the **larger the size of each sample**
  - b) with mean **M** and standard deviation **SD /  $\sqrt{N}$**



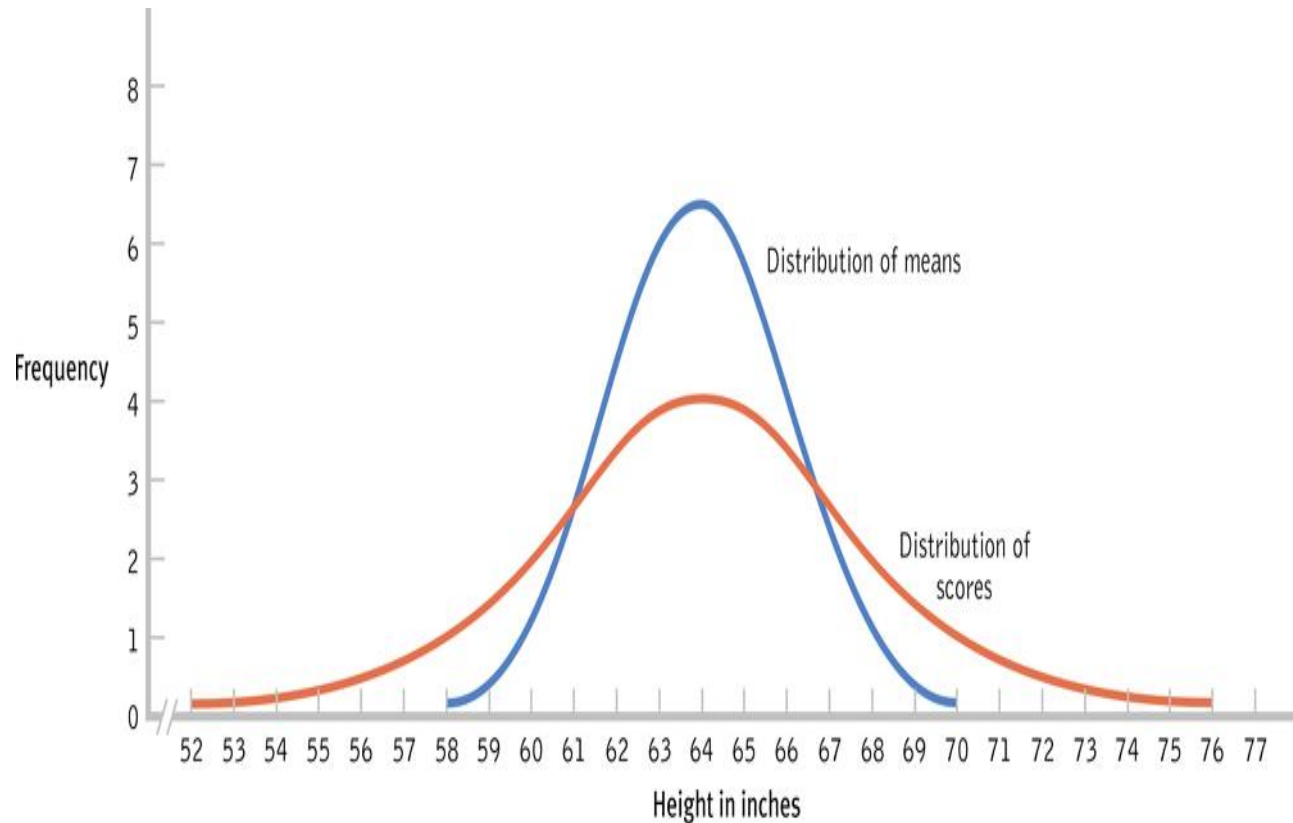
William Gosset  
(Student)  
and Guinness: how to  
select the best  
ingredients

# Why

- By using means rather than individual scores, we eliminate extreme scores

## Distribution of Means and Sample Size

As the sample size of each sample in the distribution of means increases, the normal curve becomes narrower and taller: smaller spread



For a distribution of means, the standard deviation from the mean is the Standard Error (of the Mean)

# Normal Distribution vs. Distribution of Means

- **Normal distribution** Standard Deviation & z-scores

$$SD = \sqrt{\frac{\Sigma(X - M)^2}{N - 1}} \quad z = \frac{(X - M)}{SD}$$

- **Distribution of means** Standard Error

$$SE = \frac{SD}{\sqrt{N}}$$



symbols

Distribution	Symbol for Mean	Symbol for Spread	Name of Spread
Scores	$\mu$	$\sigma$ (SD)	Standard Deviation
Means	$\mu_M$	$\sigma_M$ (SE)	Standard Error

# Compare the two distributions

- Facebook friends

1. Distribution of individual scores for whole class
  1. take everyone's data
  2. draw the histogram using the individual scores
2. Distribution of means for individual students'
  1. take only your group's data
  2. calculate the mean
  3. take everyone's means
  4. draw the histogram using the means

# Real life

- We rarely test a single individual
- We are more likely to compare means to test whether they could come from the same population
- But that's more complicated because the test population also has some variance
- To test whether a sample comes from a population, you can use the z-test

# The z-test

1. Take the mean and size of your sample
2. Take the mean and **standard deviation** of the comparison population
3. Calculate the Standard Error: divide the population SD by the square root of the sample N
4. Calculate the z-value:  
(the sample mean – the population mean)/Standard Error
5. Check the z-value against the Standard Normal Distribution to find out the probability of the sample coming from the comparison population

# Homework

Marie Claire data for the last time, I promise

1. Work with the entire data set (do not separate men and women)
  1. Get SPSS to calculate z-scores for both dependent variables
  2. Choose any 4 individuals and for each individual compare his/her z-score in distance with his/her z-score in time. Are the two z-scores similar for these 4 individuals?
  3. Why or why not?
2. We want to know whether this sample of men and women come from the population of anti-consumerists. We know that the population of anti-consumerists spend, on average, 9 minutes shopping with a standard deviation of 3 minutes.
  1. Calculate the z-score for our sample.
  2. Look up the associated percentages.
  3. Make a decision: are you confident that our sample come from the population of anti-consumerists?
  4. Why or why not?