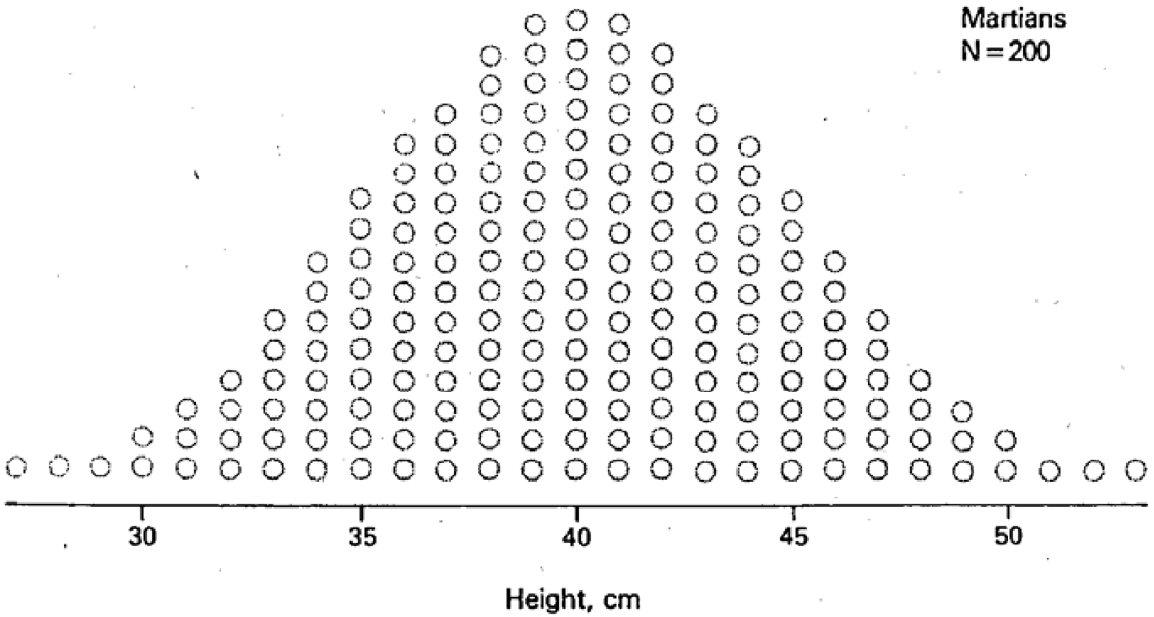
BME 3005 Biostatistics

Lecture 4 Part1: *Review, Introduction to ANOVA*

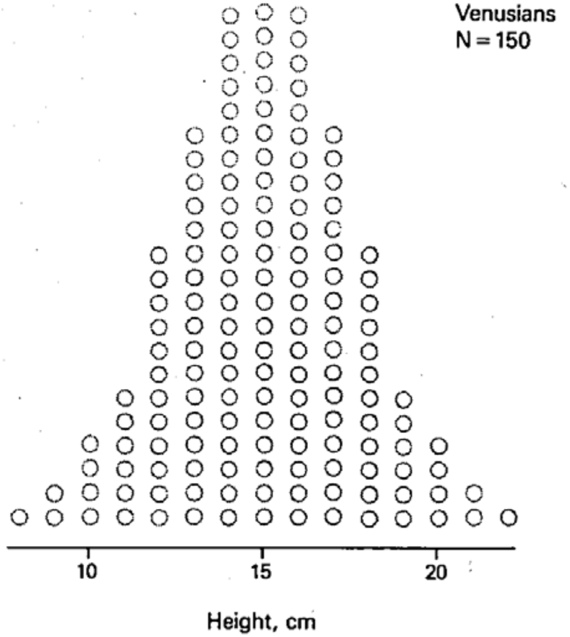
Asst. Prof. Burcu Tunç Çamlıbel

Martians



- Lets assume we observe **all the members the population**.
  - – The height of all the Martians
  - – 200 of them
  - – Figure 2.1 -- the height of all the Martians represented by a circle
- There is a *distribution* of heights.
- Most Martians are between 35 and 45 cm tall.
- Only a few are 30 cm or shorter or 50 cm and taller.

Venusians



- Now let’s check Venus.
- Figure 2.2 shows the heights of all the 150 Venusians.
- There is again a distribution of heights and all Venusians are around 15 cm tall, almost all of them are taller than 10 cm and shorter than 20 cm.

Venusians versus Martians

- Venusians are shorter and the variability of heights in Venusians is less than Martians.
- Notice that both of the distribution curves have a bell shape.
- Any member of the populations is more likely to be in the middle of the population than far out, and equally likely to be shorter or taller than the average.

Parameters of the distribution

- Since the shape of these two distributions are similar we only need to define how they differ.
- **Population** mean = average of all the heights of all the members of the population.

$$\mu = \frac{\Sigma X}{N}$$

- Mean height of Martians = 40 cm, Venusians = 15 cm
- Quantitative conclusion: The distribution of heights of Martians is higher than that of Venusians.

Variance

- Variability of a population about the mean
- Average squared (to take the absolute value of the difference) deviation from the mean
- Population variance=

$$\frac{sum(\text{value of a member - population mean})^2}{\text{number of population members}} = \sigma^2 = \frac{\Sigma(X-\mu)^2}{N}$$

- Standard deviation = square root of the average squared deviation form the mean

$$\text{standard deviation} = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$$

- σ of Martians = 5 cm, Venusians = 2.5 cm

Normal (Gaussian) distribution

- Bell shaped distribution
- Roughly 68 percent of the heights fall within 1 standard deviation from the mean, 95 percent falls within 2 standard deviations from the mean
- Its height at any given value of X is

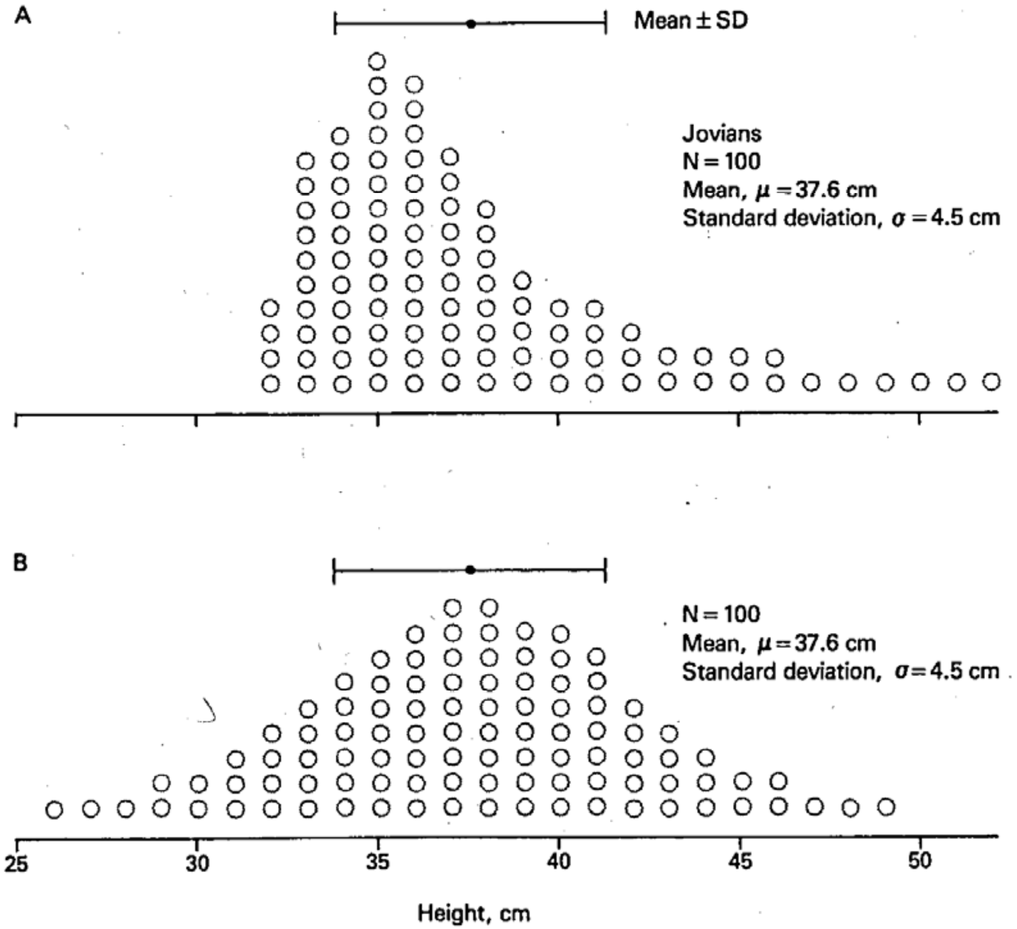
$$\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2\right]$$

- Note that the distribution is completely defined by the mean and the standard deviation.

Jupiter

- Now we look at Jupiter
- Mean height = 37.6 cm, standard deviation = 4.5 cm
- Similar to Martians according to the parameters
- **Raw data** tells a completely different story-- Figure 2.3

Fig 2.3


A
Mean ± SD
Jovians
N = 100
Mean, μ = 37.6 cm
Standard deviation, σ = 4.5 cm
B
N = 100
Mean, μ = 37.6 cm
Standard deviation, σ = 4.5 cm
25
30
35
40
45
50
Height, cm

Jovians

- The data distribution is not symmetric but it is skewed.

– A Jovian is not equally likely to have a height above average as below average.

– A few very tall individuals increase the mean and std in a way to make us think most of the Jovians are taller than they actually are.

- The mean and std are not enough to define this population.

Median

- The median is the value that half of the population falls below.
- Since %50 percent falls below the median value, it is also called the 50th percentile.
- List the *n* observations in order. The median is the (n+1)/2 observation. For odd number of observation, this would fall into a value, for even number of observations, the average of the surrounding observations.

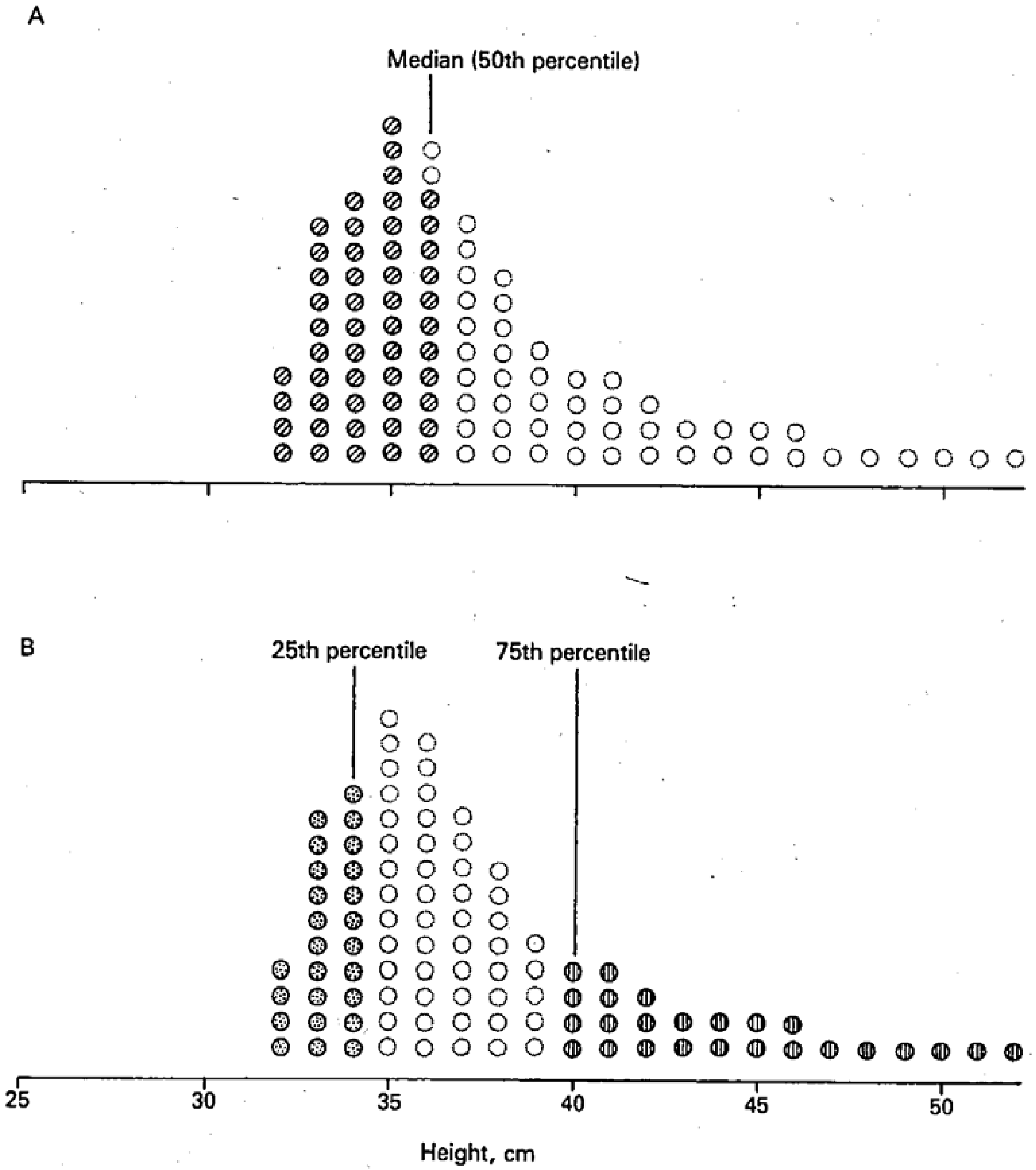
Percentiles

- Other percentiles are defined just like the median.
- 25th percentile - (n+1)/4observation
- For general, the pth percentile point is the *(n+1)/(100/ p)* observation.

- To give some indication of the dispersion of the heights of Jovians, one can report the 25th and 75th percentiles points. (34 cm and 40 cm)

- *The percentiles don't precisely define the distribution of heights, but they indicate what the range of heights is and that there are a few very tall Jovians, but not many short ones.*

Fig 2.4


A
Median (50th percentile)
B
25th percentile
75th percentile
25
30
35
40
45
50
Height, cm

Percentiles of normal distribution

- 2.5th percentile

  mean – 2 SD

- 16th percentile

  mean – 1 SD

- 25th percentile

  mean – 0.67 SD

- 50th percentile (median)

  mean

- 75th percentile

  mean + 0.67 SD

- 84th percentile

  mean +1 SD

- 97.5th percentile

  mean +2 SD

Fig 2.5

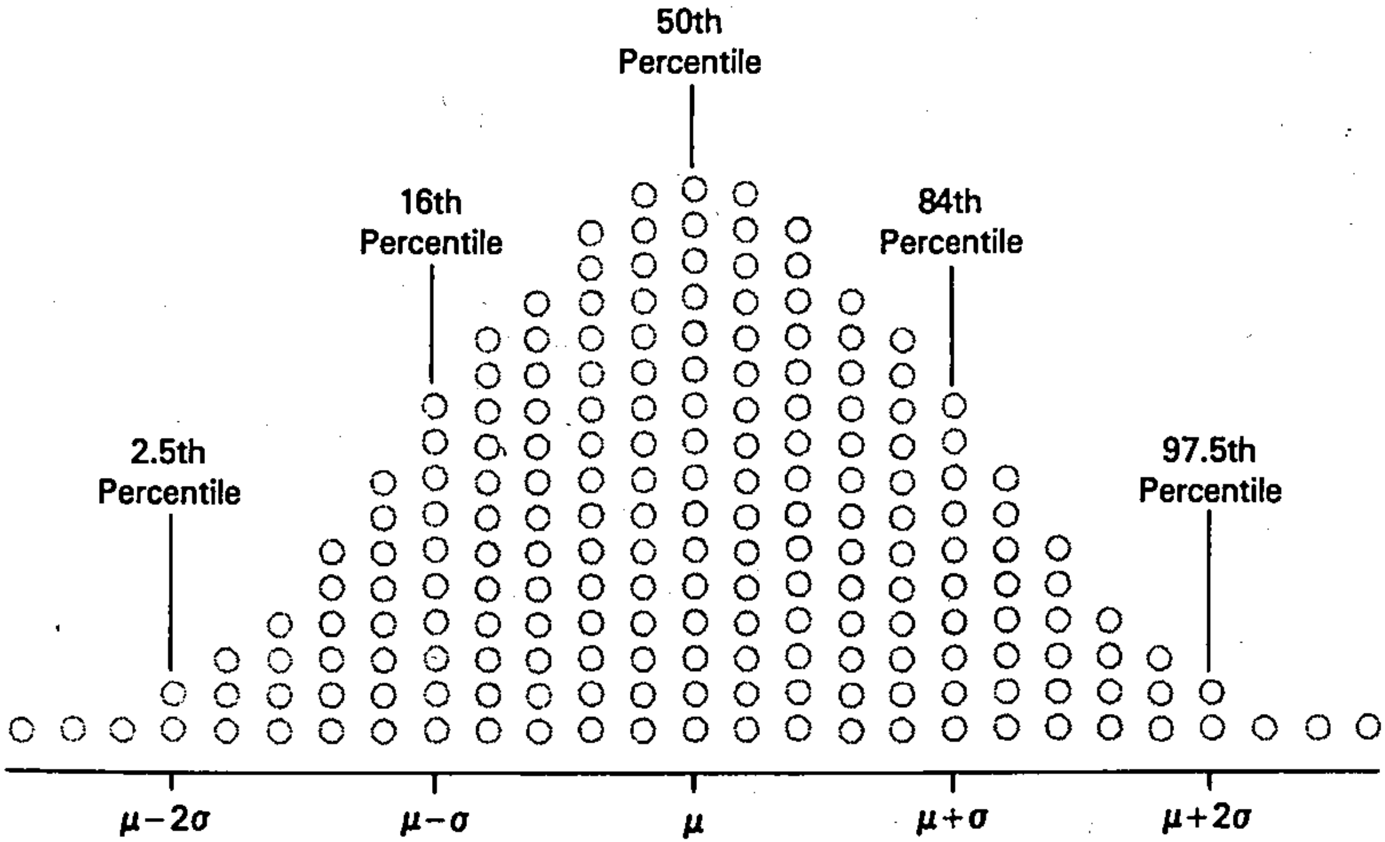


Figure 2-5 Percentile points of the normal distribution.

Why normal distribution is important?

- If the percentile values of a population are not very different than the expected values of a normal distribution based on mean and standard deviation, then the normal distribution is a good approximation to the true population.

- Some statistical tests can **only** be applied when the population is approximately normally distributed – like the t-test

- If not normally distributed, the data needs to be ranked and tests for **ordinal data** (chapters 10 and 11) should be used.

Sample

- In the real world, we can not observe a value for the whole population, but we get the value from a **sample** of the population in the hope that it represents the population well.

$$\text{Sample mean} \;=\; \overline{X} = \frac{\Sigma X}{n}$$

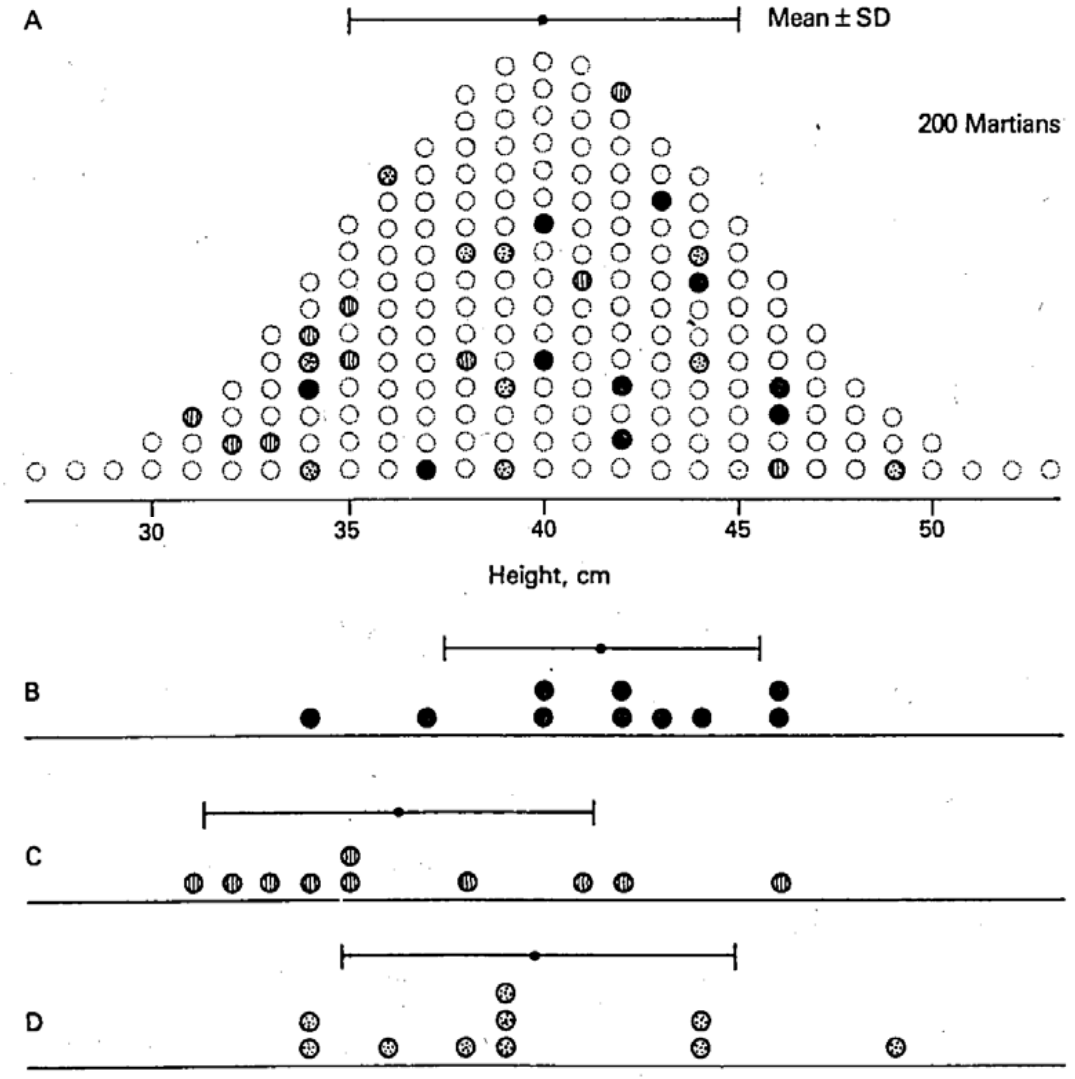
$$\text{Sample standard deviation} \;=\; s = \sqrt{\frac{\Sigma(X - \overline{X})^2}{n-1}}$$

- The sample mean is denoted by an X with a bar on top to show that it is an estimate of the real mean μ.
- The average squared deviation of a sample is computed by dividing by n-1 rather than n.
- The sample will never have as much a variation as the whole population, so it is compensated for by dividing by a smaller number.

How Good Are The Sample Estimates?

- Different samples will yield slightly different mean and SD for the whole population.
- We can compute standard errors of the mean and SD to see how representative the samples are of the main population.
- We can calculate the means of all possible random samples.

Fig 2.6

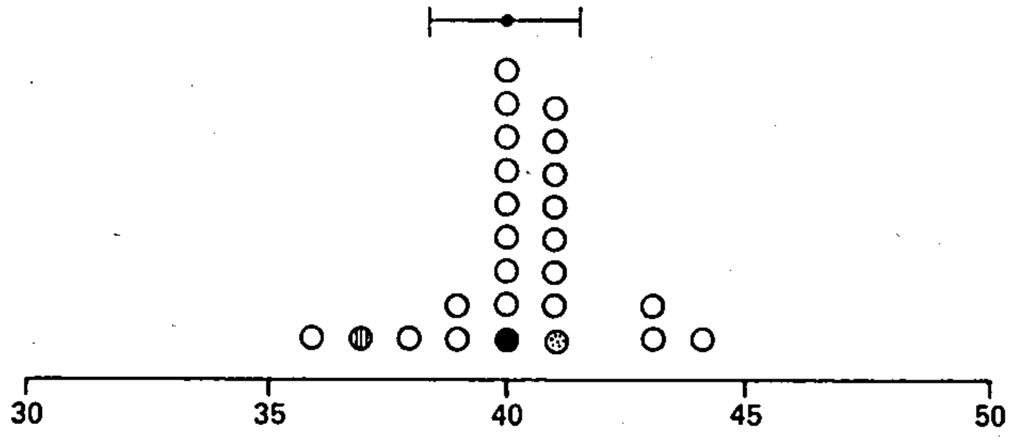


B $\overline{X} = 41.5cm, s = 3.8cm$

C $\overline{X} = 36cm, s = 5cm$

D $\overline{X} = 40cm, s = 5cm$

Fig 2.7



Means of 25 random samples of 10 Martians out of 200 Martians. The means of three specific samples of previous figure are shown with their corresponding patterns.

The new population of all possible sample means will be normally distributed regardless of the original population, its mean, **Xx (40 cm= µ)**, will be equal to the population mean and its standard deviation, **sx (1.6 cm)**, is called the ***standard error of the mean (SEM)***.

Some thoughts on samples

- The precision with which we can estimate the population mean increases as the sample size increases.

- So, the standard error of the mean decreases as the sample size increases.

- The more variability in the original population, the more variability will appear in the means of drawn samples, so SEM increases as the population standard deviation increases.

Central Limit Theorem

- The distribution of sample means will be approximately normal regardless of the distribution of the values in the original population from which the samples were drawn.
- The mean value of the collection of all possible sample means will be equal to the mean of the original population.
- The standard deviation of the collection of all possible means of samples of a given size, called the standard error of the mean, depends on both the standard deviation of the original population and the size of the sample.

SEM

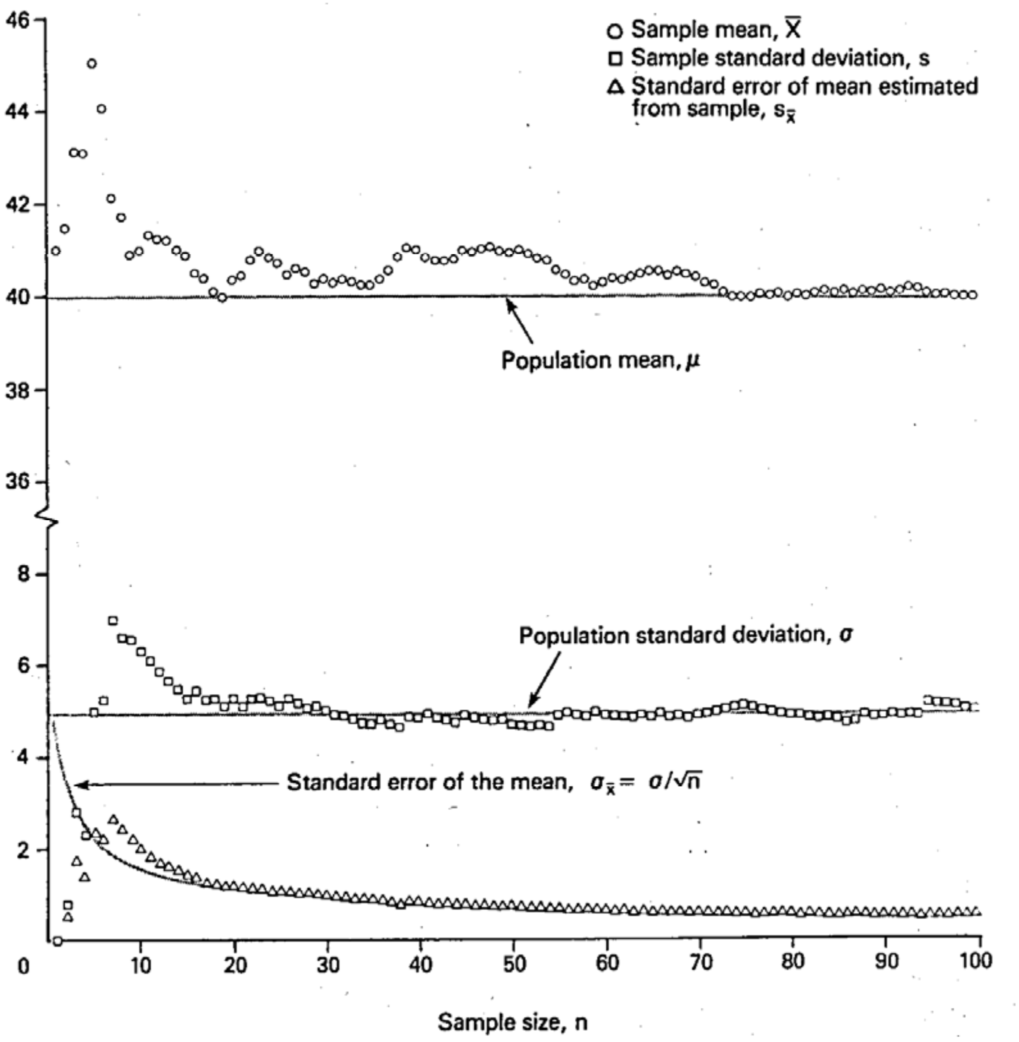
- The true SEM of samples of size n drawn from a population with a SD of σ is,

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

- The best estimate of SEM from a single sample is,

$$s_{\overline{X}} = \frac{s}{\sqrt{n}}$$

Fig 2.8



- The relationship between the sample mean, sample SD and SEM and how they vary with sample size.

- As the number of samples increase, the population mean and SD get estimated better and SEM decreases.

Standard deviation versus SEM

- Standard deviation and SEM tell two very different things.
- SD tells us about the variability in the population.
- SEM tells us about the uncertainty in the estimate of the mean using a given sample size.
- Readers are often interested in the data, so the data should never be summarized using SEM, but SD should be used.

Summary

- When a population follows a normal distribution, characterize it using mean and SD.

- If not normally distributed, use median and lower and upper percentiles.

- The standard error of the mean quantifies the precision of the mean population estimation using a sample of the whole population.

- SEM and standard deviation should not be confused or used interchangeably.