1. Is standard deviation is robust to outliers?

No.

The standard deviation is not robust to outliers.

A very high or a very low value would increase standard deviation as it would be very different from the mean. Hence outliers will effect standard deviation.

2. Suppose you have been given the following information about a variable V:

Mean(V) > Median(V)

And that the variable distribution is normal. Can you assume that that variable is right-skewed?

Yes, if the mean is greater than the median and distribution is normal, the variable is right-skewed

3. Suppose you have been given the following information about a variable V:

Mean(V) > Median(V). Now, can you assume that that variable is right-skewed?

No, since, its no where mentioned about the type distribution of the variable V, we cannot say whether it is left skewed or right skewed for sure.

Answer the following questions:

1. Please categorize the following variable by type based on the following statistical types

(Categorical Numerical, Categorical Nominal, Numerical Interval, Numerical Ratio) :

a. Temperature degrees [Celsius] - Numerical Interval

b. Temperature degrees [Kelvin] - Numerical Ratio

c. Exam grades [A-F] - Categorical Nominal

d. Exam grades [0 - 100] - Numerical Ratio

e. Height - Numerical Interval

f. Cities in Israel - Categorical Nominal

2. Assume that you have a population of 100,000 randomly generated numbers from a uniform distribution between the range 0 to 1, i.e ([0.72 , 0.98 … , 0.33]). Try to think about an experiment that generates a normal distribution out of the population.

The Central Limit Theorem is the sampling distribution of the sampling means approaches a normal distribution as the sample size gets larger, no matter what the shape of the data distribution. An essential component of the Central Limit Theorem is the average of sample means will be the population mean.

Mean of sample is same as mean of the population.

Standard deviation of the sample is equal to standard deviation of the population divided by square root of sample size.

The formula :

$$\mu_{\bar{x}} = \mu$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Where,
μ = Population mean
σ = Population standard deviation
$\mu_{\bar{x}}$ = Sample mean
$\sigma_{\bar{x}}$ = Sample standard deviation
n = Sample size

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

The "empirical rule" is that

• approximately 68% are in the interval [μ-σ,μ+σ]

• approximately 95% are in the interval [μ-2σ,μ+2σ]

• almost all are in the interval [μ-3σ,μ+3σ]

This says that if n is large enough, then a sample mean for the population is accurate with a high degree of confidence, since σ decreases with n. What constitutes "large enough" is largely a function of the underlying population distribution. The theorem assumes that the samples of size n which are used to produce sample means are drawn in a random fashion

3. You have a medical device that has a 0.9 probability to give the correct diagnosis of a certain disease (similar for both the negative and positive diagnosis). The probability of having the disease is 0.01 for the general public. You know that a person sample out of the population was diagnosed Positive for the disease by this medical device. Given this observation, What are the chances of that person having the disease?

If the device diagnoses correctly in 90% of cases.

90% of 1% of all diagnoses of “disease” are correctly diagnosed

i.e. 0.9% of all diagnoses will be diagnosed as having a disease

And out of 99% of healthy people, 9.9% will have the wrong diagnosis of "disease" i.e. :

total with a diagnosis of "disease" = 9.9% + 0.9% = 10.8%

The chances of that person having the disease = 0.9 / 10.8 = 0.083

i.e. 8.3% out of those who are with diagnosed disease

4. A test to check the effect of dieting vs. exercising has produced the following results:

Diet Only: sample mean = 5.9 kg

sample standard deviation = 4.1 kg

sample size = n = 42

Exercise Only:

sample mean = 4.1 kg

sample standard deviation = 3.7 kg

sample size = n = 47

Can you conclude in a statistically significant manner that one form of weight loss is better than the other

standard error 1 = $4.1/\sqrt{42} = 0.633$

standard error 2 = $3.7/\sqrt{47} = 0.540$

measure of variability = $\sqrt{[(0.633)2 + (0.540)2]} = 0.83$

H0: No difference in average fat lost in population for two methods. Population mean difference is zero.

H1: There is a difference in average fat lost in population for two methods. Population mean difference is not zero.

The sample mean difference = 5.9 – 4.1 = 1.8 kg

The standard error of the difference is 0.83

The test statistic z = (1.8 – 0)/0.83 = 2.17

p-value = 2 × [proportion of bell-shaped curve above 2.17]

Table 8.1 => proportion is about 2 × 0.015 = 0.03.

The p-value of 0.03 is less than or equal to 0.05, so:

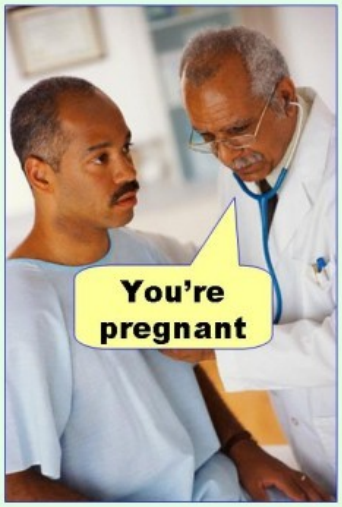
If really no difference between dieting and exercise as fat loss methods, would see such an extreme result only 3% of the time, or 3 times out of 100.

We can conclude in a statistically significant manner that one form of weight loss is better than the other

5. Given the experiment described what is the type I and type II error, what is the difference between them?

In statistical hypothesis testing, a type I error is the rejection of a true null hypothesis (also known as a "false positive" finding or conclusion), while a type II error is the non-rejection of a false null hypothesis (also known as a "false negative" finding or conclusion).

Type I error
(false positive)



Type II error
(false negative)