Basic Definitions in Statistics

Dr S.A.M. Heijke

UCT Dept of Anaesthesia & Perioperative Medicine

The fundamental idea of statistical analysis of medical studies is that we make observations on a sample of subjects and then draw inferences about the population from which the sample was drawn.

Measures of position

The mean, median and mode are measures of **location or** central tendency i.e. give an idea of *where* the middle of the information is.

MEAN:

The mean is the arithmetic average.
Add all the numbers and divide by the number of numbers.

$$\textbf{Average} = \frac{\textbf{Sum of observations}}{\textbf{Number of observations}}$$

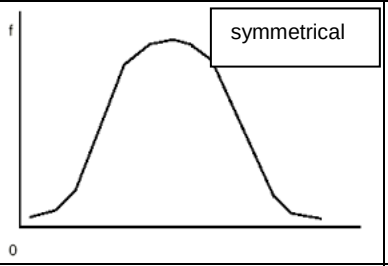
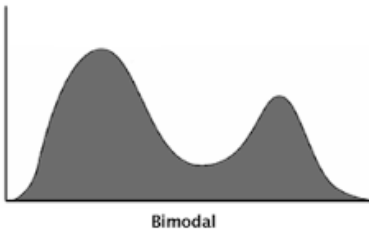
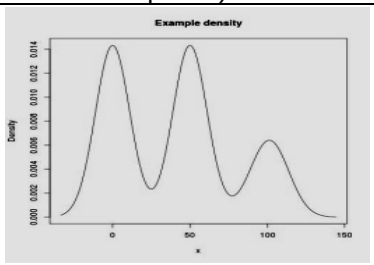
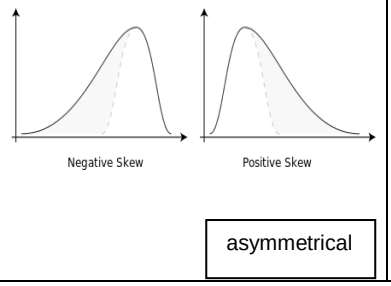
MEDIAN:

The value or quantity lying at the midpoint of a frequency distribution, i.e.

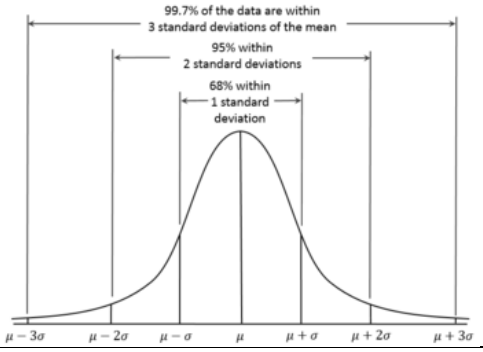
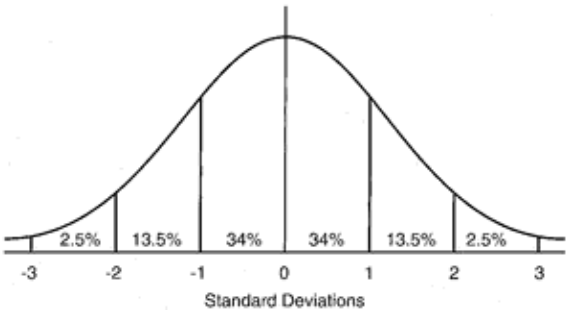
= the value of the middle observation if the number of observations is odd or
= the value of the average of the middle two observations if the number of observations is even

MODE:

The value or values that occur most frequently in the set of observations. i.e. results may be unimodal, bimodal or multimodal.

Unimodal (one peak)	Bimodal (two peaks)	Multimodal (two or more peaks)
		
		

A normal distribution: has the same mean, mode and median. The Standard Normal distribution has a normal distribution with a mean of 0.

Normal distribution	Standard Normal distribution
	

Measures of spread

The range, variance, standard deviation and the standard error of the mean give information about the degree of scatter or spread of the data.

RANGE:

The size of the interval that contains all the data in descriptive statistics or arithmetically the highest value minus the lowest value.

VARIANCE:

The variance measures how far the data points are from the mean. If you simply added them up the sum would be Zero because some lie on the positive side and some on the negative side so we can either take the absolute values of the distance from the mean, but this is difficult to work with so instead we can square the values of this difference and then add them together. To get the variance we then divide this sum by the number of observations n minus 1, i.e. n-1. This is because we are calculating the variance of a sample. If we were calculating the variance of the population, i.e. the entire set of data we would divide by N and we would use the Greek letter sigma or σ.

$$S^2 = \frac{\Sigma(X_i - \bar{X})^2}{n-1}$$

S - Variance

$\sum$ - Sum of

X_i - Data value

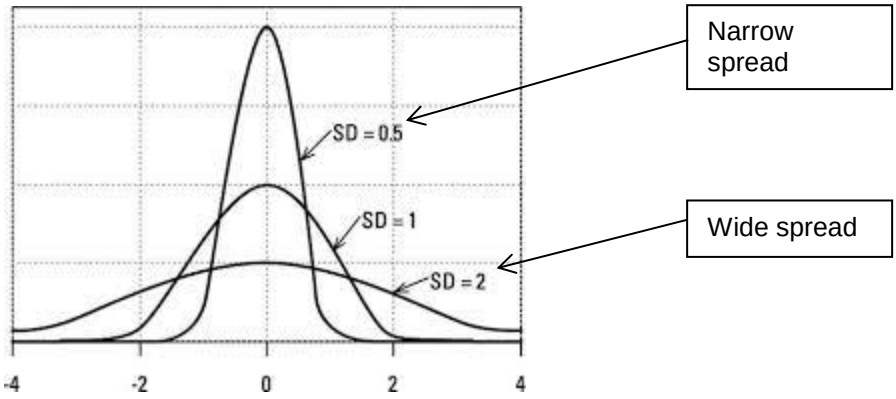
$\bar{X}$ - The mean

n - The number of observations

STANDARD DEVIATION (SD):

The standard deviation is the square root of the variance

$$S = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{n-1}}$$



STANDARD ERROR OF THE MEAN (SEM):

The standard error of the mean estimates how close the sample mean is to the population mean. It measures the variability of the sample means. The lower the value the more precisely the sample means reflect the true population mean. The lower the variance or the higher the number of observations the smaller this error will be.

$$SEM = \frac{S}{\sqrt{n}}$$

Biostatistics terms

PREVALENCE:

The percentage or proportion of the population that has a specific disease at a specific time, i.e. how common or widespread is the disease or condition (new and pre-existing cases).

$$Prevalence\ Rate = \frac{\begin{array}{c}Total\ number\ of\ new\ and\\ pre-existing\ cases\ of\ a\ disease\\ during\ a\ given\ time\ period\end{array}}{\begin{array}{c}Total\ population\ during\ the\\ same\ time\ period\end{array}} * 10^n$$

INCIDENCE:

The incidence indicates what the risks are of getting a disease (new cases).

$$Incidence\ Rate = \frac{\begin{array}{c}Total\ number\ of\ new\ cases\ of\ a\ disease\\ during\ a\ given\ time\ period\end{array}}{\begin{array}{c}Total\ population\ at\ risk\ during\ the\\ same\ time\ period\end{array}} * 100^n$$

E.g. A disease that is chronic may have a low incidence but high prevalence and a disease that has a short duration may have a high incidence but a low prevalence.

The Truth Table

		The "Truth"	
		Yes	No
Test Result	Yes	(A) True +	(B) False +
	No	(C) False –	(D) True –

Sensitivity: The ability of a test to correctly identify persons with the disease, i.e. detect the true positive rate.

$$\frac{tp}{tp + fn}$$

Specificity: The ability of a test to correctly identify persons without the disease, i.e. detect the true negative rate.

$$\frac{tn}{fp + tn}$$

True negative: Does not have the disease and tests negative for the disease.

False positive: Does not have the disease but tests positive for the disease.

		True class: Positive	True class: Negative	Measures
Predicted class	Positive	True positive *TP*	False positive *FP*	Positive predictive value (PPV) $\frac{TP}{TP+FP}$
	Negative	False negative *FN*	True negative *TN*	Negative predictive value (NPV) $\frac{TN}{FN+TN}$
Measures		Sensitivity $\frac{TP}{TP+FN}$	Specificity $\frac{TN}{FP+TN}$	Accuracy $\frac{TP+TN}{TP+FP+FN+TN}$

POSITIVE PREDICTIVE VALUES (PPV):

Positive predictive value is the probability that subjects with a positive screening test truly have the disease.

$$\frac{tp}{tp + fp}$$

NEGATIVE PREDICTIVE VALUES (NPV):

Negative predictive value is the probability that subjects with a negative screening test truly don't have the disease.

$$\frac{tn}{fn + tn}$$

RELATIVE RISK AND ODDS RATIO:

The relative risk (RR) is the ratio of the risk of an outcome in an exposed or treated group compared to a control group, e.g. how likely are you to get lung cancer if you smoke compared to if you do not. If there is no difference the risk will be one, if the risk is lower the ratio will be <1 and if it is higher the value will be >1.

The odds ratio (OR) is a measure of the association between exposure and an outcome. It represents the chance or **odds** that an outcome will occur given a particular exposure, compared to the **odds** of the outcome occurring in the absence of that exposure.

Odds Ratio (OR)

Contingency (or 2 x 2) Table

	Cases	Controls	Total
Exposed	a	b	a+b
Unexposed	c	d	c+d
Total	a+c	b+d	a+b+c+d

$$OR = (a/c) / (b/d) = (a*d) / (b*c)$$

Outcome

Predictor

	Yes	No
Yes	A	B
No	C	D

$$OR = \frac{A*D}{B*C}$$

If the prevalence of the disease is low then the odds ration approaches the relative risk

CONFIDENCE INTERVAL:

A confidence interval is a **range** of values within which we are fairly sure our true value lies. Usually we use 95%. This means a 95% "chance" that our value will be within this range and 5% that it will lie outside. If there are values above and below the range (two-tailed), then there will be 2.5% chance that the value will lie outside above the range and 2.5% chance that the value will be below the range. If the values are to one side only (one-tailed) then there will be a five percent chance that our result will fall there.

The mathematical calculation of this confidence interval depends on what we are measuring. For the mean we use a Z value (from tables)

Mean $\bar{x} \pm Z\frac{s}{\sqrt{(n)}}$ Margin of error

The size of the confidence interval depends on the variation within the population and the size of the sample

The size of the confidence interval depends on the variation within the population and the size of the sample

P-VALUE:

Probability is the likelihood of an event occurring by chance

"The **P value**, or calculated probability, is the probability of finding the observed, or more extreme, results when the null hypothesis (H_0) of a study question is true – the **definition** of 'extreme' depends on how the hypothesis is being tested."

In every experiment the researcher compares two or more groups. The null hypothesis (H_0) is the possibility that there is **no** difference between the groups being tested. The alternative hypothesis (H_1) is that any differences are real. Random sampling error may however show a difference between groups even if there really is not.

A small **p-value** (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis. A large **p-value** (> 0.05) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis.

- High P values: your data are likely with a true null.
- Low P values: your data are unlikely with a true null.

Setting up experiment steps:

1) State Hypothesis – Null (H_0) versus alternative (H_1)
2) Decide Significance level or α level – commonly 0.05
3) Collect sample
4) Calculate your p value (using tables or computer program)
5) Decide: if your value is less than p, reject the null hypothesis
   if your value is more than p, then accept the null hypothesis

STATISTICAL ERRORS:

	Reject H_0	Accept H_0
H_0 is true	Type 1 error α error	Correct
H_0 is false	Correct	Type 2 or β error

References

1. Numerous YouTube clips that are simplified for easier understanding though professional statisticians may be unhappy. e.g. P VALUE: https://www.youtube.com/watch?v=eyknGvncKLw
2. The mathsisfun site is easy to use and understand and has a great section on data mathematics https://www.mathsisfun.com/data/index.html