

Overview of Statistical Significance

Inferential Statistics: Determines how likely a given result occurred by chance alone. Since we can rarely study an entire population, we study a sample of the population and by inference apply that result to the entire population.

Null Hypothesis: The proposal that **no difference** exists between groups or that there is **no association** between risk indicator and outcome variables. If the null hypothesis is true then the findings from the study are the result of chance or random factors. The overall purpose of a typical study is to "reject the null hypothesis." Another example: there is less than a 1 in 20 chance that the differences between treatments seen in this trial could have occurred by chance; less than a 1 in 20 chance that the null hypothesis is true.

Chance: Random variation. Difference between the outcomes from a sample of the population and the true value obtained from looking at the outcomes from the entire population. Statistical methods are used to estimate the probability that chance alone accounts for the differences in outcomes.

Clinical vs. Statistical Significance: Statistical significance means the likelihood that the difference found between groups could have occurred by chance alone. In most clinical trials, a result is statistically significant if the difference between groups could have occurred by chance alone in less than 1 time in 20. This is expressed as a p value < 0.05. Remember that a trivial difference can have a very low p value if the number of subjects is large enough! Clinical significance has little to do with statistics and is a matter of judgment. It answers the question "Is the difference between groups large enough to be worth achieving?" Studies can be statistically significant yet clinically insignificant.

Level of Significance: The probability of incorrectly rejecting the null hypothesis, i.e. saying that there is a difference between two groups when actually there is none. Otherwise known as the probability of Type I error. By convention, the level of significance is often set to a p value of 0.01 or 0.05.

p Value: The measured probability of a finding occurring, i.e. rejecting the null hypothesis, by chance alone given that the null hypothesis is actually true. By convention, a p value < 0.05 is often considered significant. ("There is less than a 5% probability that the finding [null hypothesis rejected] was due to chance alone.")

Power: The probability of detecting an effect in the treatment vs. control group if a difference actually exists. Must also specify the size of the difference. For example, a paper describing a clinical trial with a new hypertension medication may contain the following statement - "The study had a power of 80% to detect a difference of 5 mm Hg in diastolic blood pressure between the treatment and control groups." Typical power probabilities are 80% or greater. Power = 1 - β (see Type II Error, below)

Type I Error: Mistakenly rejecting the null hypothesis when it is actually true. The maximum probability of making a Type I error that the researcher is willing to accept is called alpha (α). Alpha is determined before the study begins. False positive conclusion. Studies commonly set alpha to 1 in 20 ($=0.05$).

Type II Error: Mistakenly accepting (not rejecting) the null hypothesis when it is false. The probability of making a Type II error is called beta (β). Power = $1 - \beta$ (see above). False negative conclusion. For trials the probability of a β error is usually set at 0.20 or 20% probability. A 20% chance of missing a true difference.

Testing the Null Hypothesis to Assess Efficacy of Two Treatments (e.g. drug vs. placebo)

		Truth	
		Null hypothesis is true (no difference)	Null hypothesis is not true (difference)
Decision (based on statistical test)	Accept Null Hypothesis	Correct	Type II Error (beta)
	Reject Null Hypothesis	Type I Error (alpha)	Correct 1 - beta (Power)

Standard Error of the Mean (SEM): A measure of variability. The standard error of the mean quantifies how accurately the true population mean is known. A measure of the variability of the mean of the sample as an estimate of the true value of the population mean. The larger the sample size the smaller the standard error of the mean. Used in computing confidence intervals. In a clinical trial, the larger the sample size the tighter the 95% CI is around the point estimate of the study.

Standard Deviation: A measure of variability. The standard deviation quantifies how much the values vary from each other. A measure of the spread of individual observations around the mean value of the sample. A normal, unskewed curve will have 34% of the cases between the mean and 1 standard deviation above or below the mean; 68% of cases between 1 standard deviation above and 1 below the mean; 95.5% of cases will be within two standard deviations of the mean.

Confidence Interval: Often expressed as 95% confidence intervals. Studies are performed on a sample of the population, not the whole population. Confidence intervals give us some idea of how likely the sample mean represents the population mean. Expressed as the sample mean plus and minus a specified amount. A measure of the precision of the estimate. The 95% CI is the range of values within which we can be 95% sure that the true value lies for the whole population of patients from whom the study patients were selected. Most clinical trials study a sample of the population at risk. Because a sample is a subset of a population, the mean value obtained for the sample

studied may not be same as the mean value if the entire population was studied. Results from a sample population with a wider range of values will have broader confidence intervals than results from a study with a narrower range of values. Increasing the number of results (patients) within a sample population narrows the confidence intervals. The confidence interval (CI) quantifies uncertainty. Derived from the sample mean and the standard error.

Please see: [Use of confidence intervals to indicate uncertainty in research findings.](#)

Interobserver variability: Variability between observers. Do two or more radiologists give the same reading from the same radiograph?

Intraobserver variability: Variability by the same observer. Does a radiologist give the same reading of a radiograph when viewed on more than one occasion?

Survival Analysis: Statistical procedures for estimating survival (prognosis) in a population under study.

Cox Proportional-Hazard Model: A type of multivariate analysis that is used to identify a combination of factors that best predicts prognosis in the group of patients. Can also test the effect of individual factors independently. Analysis used when the outcome is the time to an event. The Cox proportional hazard model is used when practical considerations preclude observing survival time in all patients being studied.

Hazard (or Hazard Rate): Probability of an endpoint. A technical name for failure rate.

Hazard Ratio: Relative risk of an endpoint at any given time.

Multivariate Analysis: An analysis where the effects of many variables are considered. Can select a subset of variables that significantly contribute to the variation in outcome.

Kaplan-Meier Curve: Used for estimating probability of surviving a unit of time. Used to develop a survival curve when not all survival times are exactly known.