



CLINICAL EPIDEMIOLOGY AND POPULATION HEALTH

Key Points – Interpreting Statistics in Clinical Research

1. Basic steps of statistical testing:

- a. State a hypothesis (as a “null” hypothesis) including the definition of outcome variable.
- b. Choose appropriate statistical test
- c. Calculate test statistic
- d. Determine (from a table or graph of its distribution – now estimated with statistical software) whether the value obtained is “extreme.” By extreme, we mean unlikely that the value of the statistical test has occurred by chance, if the null hypothesis were true. By convention, we consider a value extreme when it falls outside of 95% of the expected results under the null (in the “tails” of the distribution each containing 2.5%, for a two-tailed test).
- e. Create a confidence interval.

2. Interpretation:

- a. Statistically significant: null hypothesis is rejected.
- b. Not statistically significant: null hypothesis is not rejected. (Note that “is not rejected” is not the same as “is true.”)

Remember this table regarding two populations A and B. Population A has one exposure and population B has a different exposure (note: A and B can also have different levels of the same exposure, e.g., second hand smoke). The “populations” may be patients receiving either drug or placebo, children exposed to low vs. high levels of second hand smoke, or any other groups distinguished by an independent variable of interest. We observe a sample of each, and make a conclusion (our best guess) about the underlying populations (the TRUTH) they represent.

	TRUTH	
	A and B differ	A and B are the same (Null Hypothesis)
<u>Interpretation of study result</u>		
A and B different (reject null)	Correct (Statistical power = 1-beta)	Wrong (Type 1 error, “alpha” error)
No evidence that A and B different (do not reject null)	Wrong (Type 2 error, “beta” error)	Correct

Type 1 error (α = probability of making a type 1 error) – concluding that two populations are different (rejecting the null hypothesis), when they, in fact, are not different (that is, they are drawn from the same population). The **cutoff for α at 0.05 (the p-value)** is *generally* taken (arbitrarily) as the risk of alpha error below which we are comfortable rejecting the null.

Type 2 error (β = probability of making a type 2 error) – not concluding a difference exists (not rejecting the null), when the samples, in fact, are drawn from different populations.

3. Prototypical statistical tests:

a. One sample t-test

- i. Compares a sample mean (\bar{X}) to a known population mean (μ) or a pre-specified value (null = the sample mean and population mean or pre-specified value are the same)
- ii. For example, we may want to compare the average body mass index (BMI) of a sample of male smokers to the known average BMI of males in the US (which is 26.6).
- iii. t statistic is a comparison of the difference between the sample means and population mean compared to the variation in the population mean

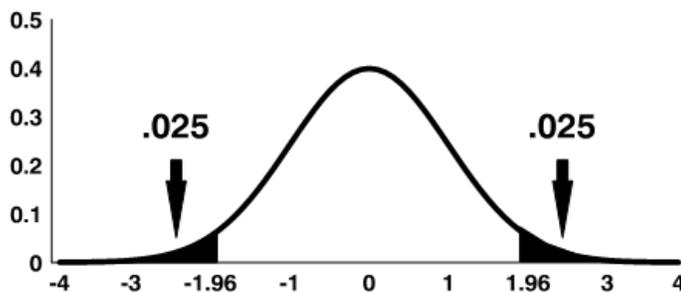
Calculated by (don't memorize, but understand. Note that the denominator is the SE = SD/\sqrt{n}):

$$t = \frac{\bar{X} - \mu}{SE}$$

Since we don't often know the standard deviation (SD) of a population, we use the sample SD as a proxy. This is why we use a t-statistic (from the t distribution, which is similar to a standard normal curve) instead of the Z-statistic (from the Z distribution, a standard normal curve). Don't worry about this level detail.

You may notice that this calculation looks a lot like the calculation of the Z-score! But rather than examining how many standard deviations an observation is away from a population mean (as we do in a Z-score), the one-sample t-test examines how many standard errors a sample mean is away from the population mean (or pre-specified value). Here, the distribution is of sample means (i.e., the sampling distribution).

Using the significance level of alpha 0.05, we reject the null hypothesis if $|t|$ is greater than the critical value, which is approximately 1.96 (i.e., 2 standard errors away). The critical region is shaded below.

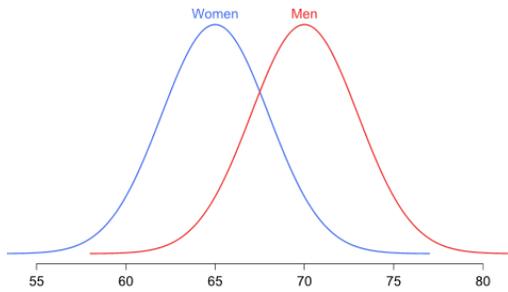


b. Two sample t-test of means

- i. Compares two means (null: the samples are from a single population).
- ii. For example, you may want to compare the average BMI for a sample of male smokers to the average BMI for a sample of female smokers.
- iii. t statistic (for 2 independent samples) is a comparison of the difference in the sample means compared to the expected variability in sample means.

Calculated by (don't memorize, but understand. Note that here the denominator is the SE of the mean difference, which you don't need to know how to calculate):

$$t = \frac{\bar{X}_A - \bar{X}_B}{SE}$$



The two-sample t-test tests whether the differences between the two samples (e.g. men vs. women) is zero. We use the same critical value rule as a one sample t-test. Using the significance level of alpha 0.05, we reject the null hypothesis if $|t|$ is greater than the critical value, which is approximately 1.96 (i.e., 2 standard errors away).

If you want to compare more than 2 samples, you can use a test called an **ANOVA** (or analysis of variance), a test that compares means of 2 or more samples.

[Note on non-parametric tests. One and two sample t-test assume a normal distribution and independence of observations. We will not go into these in detail, but there are hypothesis tests for variables that do not follow a normal distribution. These tests compare medians instead of means, often by comparing the “rank” of observations in two groups (whether higher-ranking observations are more common in one group). One common non-parametric test is the Wilcoxon rank sum test.]

c. Chi-squared test

- i. Seeks to determine whether an association exists between two categorical variables.
- ii. For example, we may want to test whether the probability of smoking (a yes/no binary variable) differs by sex (using a binary variable of male/female). Or, you may want to test whether the probability of smoking differs by race/ethnicity (a categorical variable with more than two levels).
- iii. Analyzes a 2x2 table (or a table with any number of rows and columns) by comparing observed vs. expected if the null hypothesis were true (i.e. if the proportion of outcomes were equal in the two groups).

The “expected” (E) value in any cell = (row total x column total) / total n of table.

The Chi squared statistic is calculated by (don't memorize, but understand what goes into the formula):

$$X^2 = \sum \frac{(O_i - E_i)^2}{E}$$

For a two by two chi-square table, the critical value is 3.84 at an alpha level of 0.05.

- For studies with small numbers in one or more cells (often taken as <5), we use Fisher's exact test, which is analogous but more difficult to calculate by hand (the computers don't groan).
- When comparing categorical outcomes between two groups ONLY, you can also use a two-sample Z test of proportions (you don't need to know how to calculate a two-sample Z-test for the course.)

4. Confidence Intervals:

Inferential statistics (above) are concerned with the certainty of rejecting a pre-specified null hypothesis. For example, is treatment the same as placebo? This doesn't help us answer questions such as:

- a. How much better (or worse)? What's the magnitude of benefit that we observed in our sample (the "point estimate")? Does the result exclude (with statistical certainty) a result that may be clinically meaningful.

With a confidence interval we report the best estimate of a difference based on our sample and the range of values statistically consistent with the data observed.

- b. Technical interpretation: if we repeated the trial many times and calculated the 95% CI for each, the true effect will be included 95% of the time.
- c. Confidence intervals show us the range of values compatible with the data. We can assess overlap and whether differences we deem clinically are included in the interval. That is, they help us determine not only whether a difference exists, but whether we are relatively certain that a clinically important difference exists.
- d. The 95% confidence interval around the mean (\bar{X}) of a normally distributed continuous variable is given by:

$$\text{Mean} \pm 1.96 \times \text{SE} \quad (\text{Where } \text{SE} = \frac{SD}{\sqrt{n}})$$

- e. The 95% confidence interval around a proportion (p) is given by:

$$p \pm 1.96 \sqrt{\frac{p(1-p)}{n}}$$