

Hypothesis Testing

Hypothesis: conjecture, proposition or statement based on published literature, data or a theory which may or may not be true

Statistical Hypothesis: conjecture about a population parameter

- usually stated in mathematical terms
- two types, **null** and **alternate**

Null Hypothesis (H_0): states that there is NO difference between a parameter and a specific value or among several different parameters

Alternate Hypothesis (H_1): states that there is a “significant” difference between a parameter and a specific value or among several different parameters

Examples:

- $H_0: \mu = 82 \text{ kg}$ $H_1: \mu \neq 82 \text{ kg}^*$
- $H_0: \mu \leq 150 \text{ cm}$ $H_1: \mu > 150 \text{ cm}$
- $H_0: \mu \geq 65.0 \text{ s}$ $H_1: \mu < 65.0 \text{ s}$
- $H_0: \mu_0 = \mu_1$ $H_1: \mu_0 \neq \mu_1^*$
- $H_0: \mu_0 \geq \mu_1$ $H_1: \mu_0 < \mu_1$

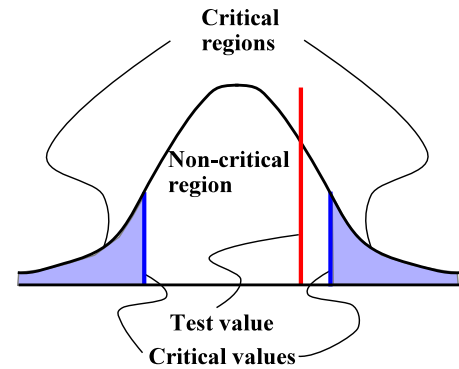
Notice that the equality symbols are always with the null hypotheses.

* These are called two-tailed tests; others are all “directional” or one-tailed tests.

Two-tailed vs One-tailed Tests

Two-tailed: -also called a non-directional test

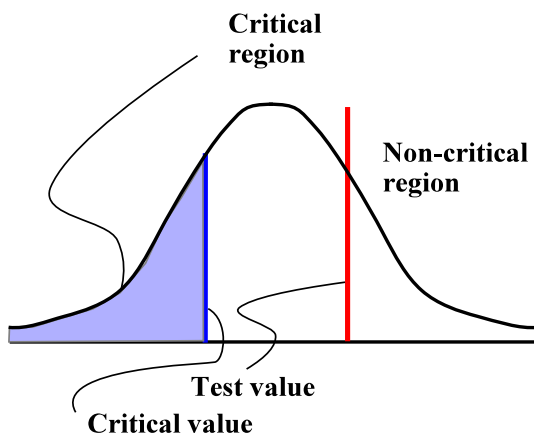
- null hypothesis rejected if sample mean falls in either tail
- most appropriate test especially with no previous experimentation
- less powerful than one-tailed



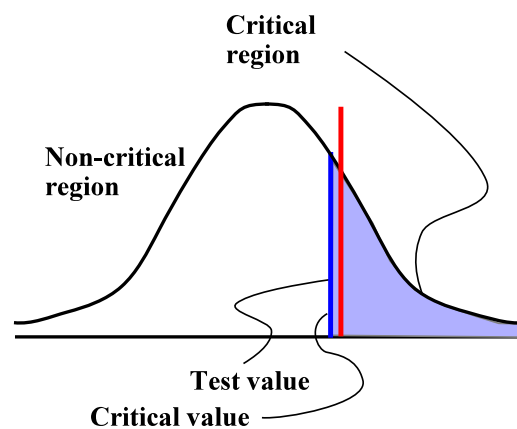
One-tailed: -also called a directional test

- researcher must have reason that permits selecting in which tail the test will be done, i.e., will the experimental protocol increase or decrease the sample statistic
- more powerful than two-tailed since it is easier to achieve a significant difference
- fails to handle the situation when the sample means falls in the “wrong” tail

One-tailed, left



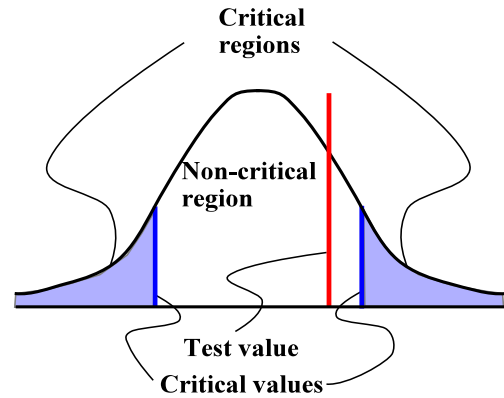
One-tailed, right



Statistical Testing

To determine the veracity (truth) of an hypothesis a **statistical test** must be undertaken that yields a **test value**. This value is then evaluated to determine if it falls in the **critical region** of a appropriate probability distribution for a given **significance** or **alpha () level**.

The critical region is the region of the probability distribution that rejects the null hypothesis. Its limit(s), called the **critical value(s)**, are defined by the specified confidence level (CL). The CL **must be selected in advance** of computing the test value. To do otherwise is statistical dishonesty. It is usual to perform a two-tailed test.



Instead of reporting significance levels (= 0.05) or equivalent probabilities (P<0.05) many researchers report the test values as probabilities or **P-values** (e.g., P = 0.0455, P = 0.253, P < 0.001, Not P=0.000). Advanced statistical programs report P-values, if not, use P<0.05 or P<0.01.

Truth table:	H ₀ is true and H ₁ is false	H ₀ is false and H ₁ is true
Test rejects H ₀ (accepts H ₁)	Error () ☹ Type I error	Correct (1 -) ☺ (experiment succeeded)
Test does not reject H ₀ (accepts H ₀)	Correct (1 -) ☹ (experimental treatment failed)	Error () ☹ Type II error

z-Test and t-Test

Test for a Single Mean:

- used to test a single sample mean (\bar{X}) when the population mean (μ) is known
- Is the sample representative of the population or is it different (greater, lesser or either)?

z-Test:

- when population s.d. (σ) is known

Test value:
$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

- if z is in critical region defined by critical value(s) then sample mean is “significantly different” from the population mean,
- if σ is unknown then use sample, s , as long as sample size is greater than 30

Test value:
$$z = \frac{\bar{X} - \mu}{s / \sqrt{n}}$$

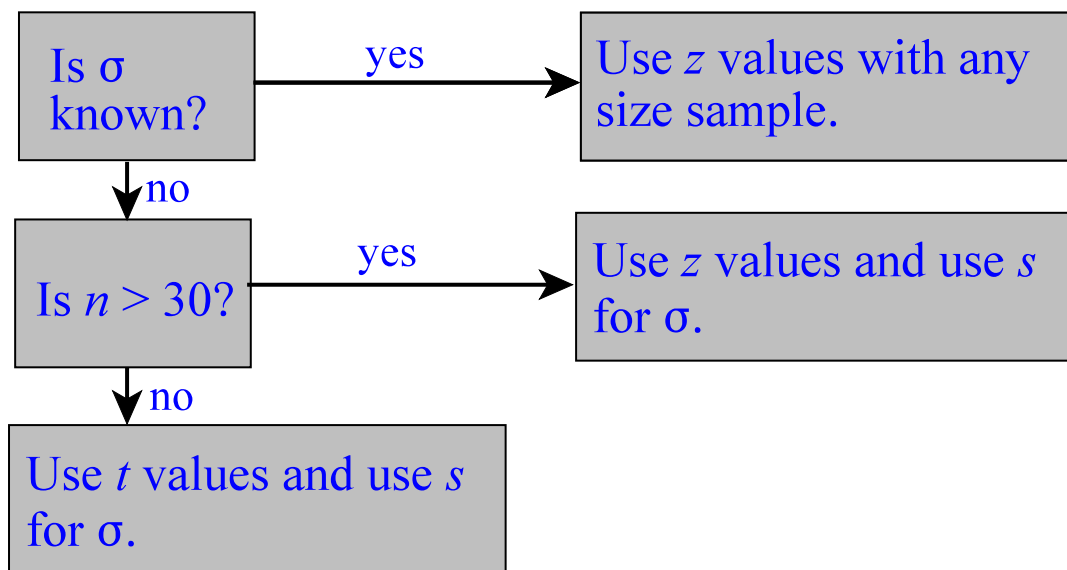
t-Test:

- if σ is unknown and $n < 30$ then use t -test and t -distribution with d.f. = $n-1$

Test value:
$$t = \frac{\bar{X} - \mu}{s / \sqrt{n}}$$

Flow Diagram for Choosing the Correct Statistical Test

Same as flow diagram used for confidence intervals. Generally the sample's mean and standard deviation are used with the t -distribution. The t -distribution becomes indistinguishable from the z -distribution (normal distribution) when $n > 30$.



Power of a Statistical Test

- Power:** -ability of a statistical test to detect a real difference
- probability of rejecting the null hypothesis when it is false (i.e., there is a real difference)
 - equal to $1 - \beta$ ($1 - \text{probability of Type II error}$)

Ways of increasing power

- **Increasing α** (e.g., $\alpha = 0.10$ vs 0.05) will increase power but it also increases chance of a Type I error.
- **Increase sample size (n)**. Increases cost.
- **Use ratio or interval data** versus nominal or ordinal data.
- **Parametric tests** are more powerful than equivalent non-parametric tests (e.g., t -test vs Wilcoxon).
- Using “**repeated-measures**” tests, such as, the dependent-groups t -test, repeated-measures ANOVA or Wilcoxon signed-ranks test. By using the same subjects repeatedly, variability is reduced.
- If variances are equal use **pooled estimates** of variance (e.g., independent groups t -test).
- Using samples that represent **extremes** (e.g., 18-24 year olds vs 65-70 year olds). Reduces generalizability of experiment results.
- **Standardizing testing procedures** and trained testers reduces variability.
- Using **one-tailed** vs. two-tailed tests. Problem occurs if results are in wrong tail. Not recommended.