Statistics

2023 Lectures
Part 10 - Testing Statistical Hypotheses

Institute of Economic Studies
Faculty of Social Sciences
Charles University

Testing

- Assume having a random sample $X_1, \ldots, X_n$ from distribution with $f(x, \theta)$, θ unknown.
- Now the key question instead of estimating θ is whether the true value θ belongs to some specific subset Θ_0 of Θ.
- Test statement H_0 (null hypothesis) that $\theta \in \Theta_0$ against another statement (alternative hypothesis) $H_1 : \theta \notin \Theta_0$.
- The observed sample $X = x$ provides evidence, usually in favor of one of the two hypotheses
- The process is trivial if the sets of possible values of X are disjoint for $\theta \in \Theta_0$ and $\theta \notin \Theta_0$...often not the case, some values of X can occur for both $\theta \in \Theta_0$ and $\theta \notin \Theta_0$

Example 94: Assume we are interested in question whether a certain coin is fair. Then we can construct null hypothesis as $H_0 : \theta \in \Theta_0 = \{\frac{1}{2}\}$ and $H_1 : \theta \notin \{\frac{1}{2}\}$, where θ is the unknown probability of heads (or tails).

Error types and critical region

- There may be observations of X which would lead us to conclusion that $\theta \notin \Theta_0$ while that is true, the so called type I error, or even when $\theta \notin \Theta_0$ there may be observations of X suggesting that $\theta \in \Theta_0$, type II error.

Example 95: Consider the case of a prisoner. If you find evidence about his/her innocence, the prisoner will be set free. H_0 : prisoner innocent, H_1 : prisoner is guilty.
Type I error occurs when innocent prisoner is found guilty and remains in prison, while type II error occurs when guilty prisoner is set free. Which error is more serious to be avoided?

- there is an asymmetry in hypotheses: we set H_0 considering type I error as more painful.
- to make a decision about "$\theta \in \Theta_0$" or "$\theta \in \Theta_1 := \Theta \setminus \Theta_0$" we specify a set C of points in the space of values of X such that whenever $X \in C$, we decide that $\theta \in \Theta_1$. Such set C is called a critical region for the hypothesis $\theta \in \Theta_0$.

Power function

- Consider $\theta \in \Theta_0$ and we decide that $\theta \in \Theta_1$.
  Such event has the probability

$$P_\theta(X \in C) = \alpha(\theta) \quad \text{for } \theta \in \Theta_0,$$

  while probability of type II error is

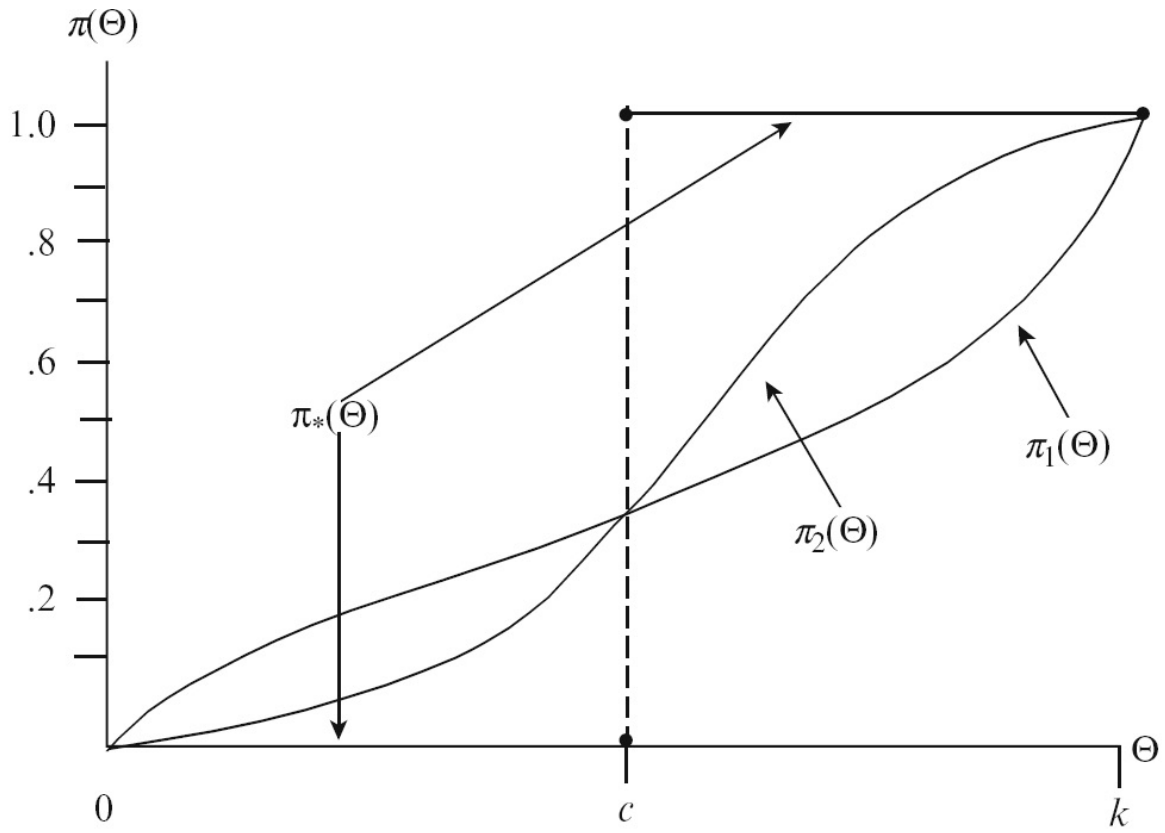
$$P_\theta(X \notin C) = 1 - P_\theta(X \in C) = \beta(\theta) \quad \text{for } \theta \in \Theta_1.$$

- both α and β are functions of the parameter θ, depending on the power function

$$\pi_C(\theta) := P_\theta(X \in C).$$

- ideally, we choose the "best" critical region such that we minimize both types of error, i.e. minimize $\pi(\theta)$ for $\theta \in \Theta_0$ and maximize $\pi(\theta)$ for $\theta \in \Theta_1$
- we cannot satisfy both criteria at once in most of the cases (Pareto optima)

Power function



Test and a simple/composite hypothesis

Definition 47: A decision rule that tells us what to do in case of observing sample $X = x$ is called a test if the only actions allowed are "reject H_0" and "not reject H_0".

- Please, read the book on the discussion about "not rejecting H_0" and "accepting H_0"!
- rejection or acceptance of a hypothesis are not statements about truth or falsehood. "The terms "accepting" and "rejecting" a statistical hypothesis are very convenient and are well established...to accept hypothesis H means only to take an action A rather then B. This does not mean that we necessarily believe that the hypothesis H is true" (Neymann, 1950)

Definition 48: A hypothesis $H_i : \theta \in \Theta_i, i = 0, 1$, is called simple if it completely specifies the distribution of the sample. Otherwise it is called composite.

Example

Example 96: A supermarket buys oranges such that there is a guarantee that less than 3% are rotten. Testing procedure consists of randomly choosing 30 oranges and shipment will be accepted only if there is no more then 1 unacceptable fruit, otherwise rejected. Find the values of the power function, if the test null hypothesis is true and if the alternative is true.

Example: testing the mean in $N(\theta, 1)$

Example 97: Test hypothesis about the mean in $N(\theta, 1)$.

$$H_0 : \mu = \mu_0 \quad \text{... simple hypothesis}$$
$$H_1 : \mu \neq \mu_0 \quad \text{... composite hypothesis}$$

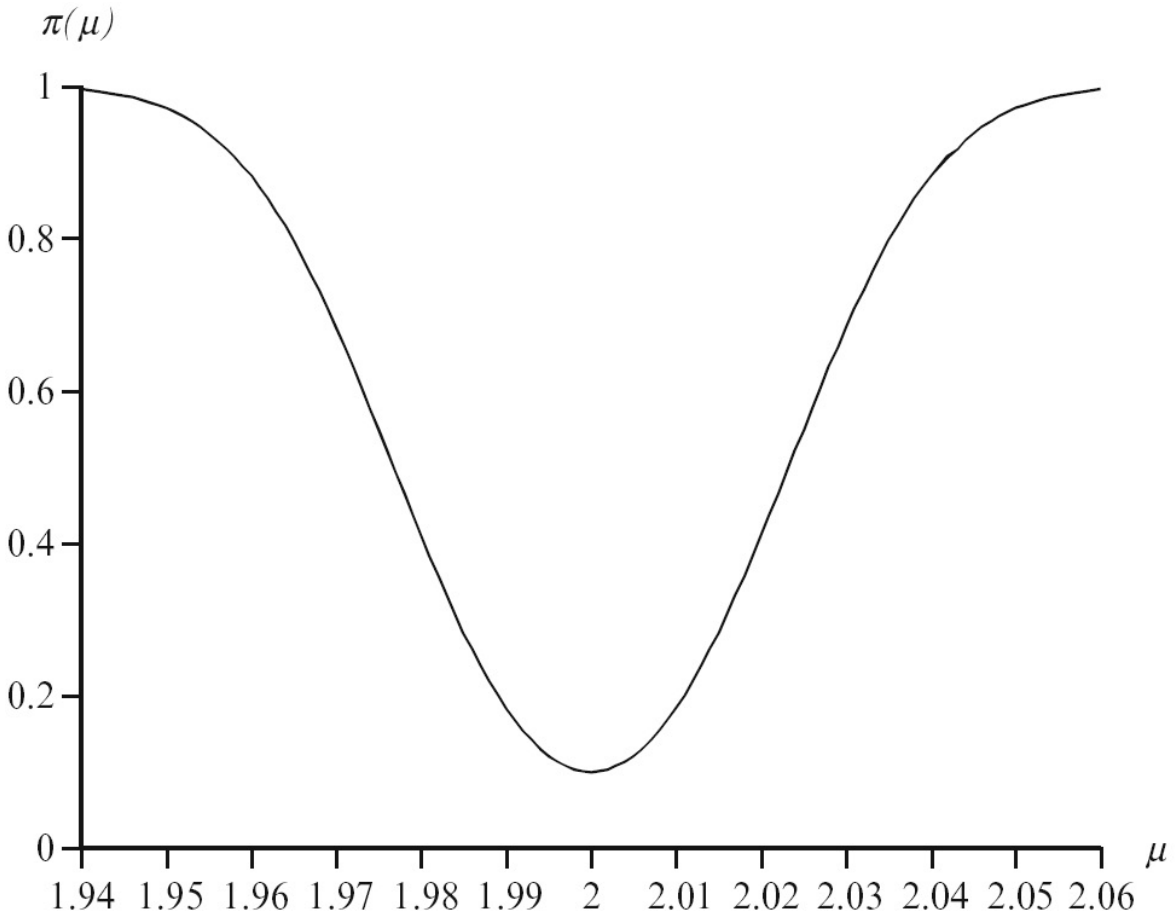
The idea: if the sample mean deviates "too much" from μ_0, we should reject H_0, hence for given threshold K we set

$$C = \{X : |\bar{X} - \mu_0| > K\}$$

$$\pi_C(\mu) = P_\mu(|\bar{X} - \mu_0| > K) = 1 - P_\mu(|\bar{X} - \mu_0| \leq K) =$$
$$= 1 - P\left(\frac{-K + \mu_0 - \mu}{1/\sqrt{n}} \leq \frac{\bar{X} - \mu}{1/\sqrt{n}} \leq \frac{K + \mu_0 - \mu}{1/\sqrt{n}}\right)$$

If H_0 is true then $\pi_C(\mu_0) = 1 - P(|\frac{\bar{X}-\mu_0}{1/\sqrt{n}}| \leq K\sqrt{n})$. Thus, for any fixed n we can choose C such that α is equal to a preassigned level and β approaches 0 as μ moves away from μ_0.

Modified example: testing the mean in $N(\theta, 0.04)$, $\mu_0 = 2$ and $\alpha = 0.1$



Both hypotheses simple

- Suppose both H_0 and H_1 are simple, i.e., $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$.
- In the discrete case, this means that for some x

$$P(X = x|H_0) \neq P(X = x|H_1).$$

Analogously, in the continuous case, for some A

$$\int_A f(x, \theta_0)\mathrm{d}x \neq \int_A f(x, \theta_1)\mathrm{d}x.$$

- If the test rejects H_0 if $X \in C$ then

$$\alpha = P(X \in C|H_0)$$
$$\beta = P(X \notin C|H_1)$$

- To find the most powerful test (with minimal α and minimal β, in the Pareto sense), we should improve C. For that, it is convenient to partition the set of values of $X = x$ into four disjoint sets.

Partition of values of x

For continuous case:
Set $A_1 = \{x : f(x,\theta_0) = 0, f(x,\theta_1) > 0\}$ and let $C^* = C \cup A_1$. Then

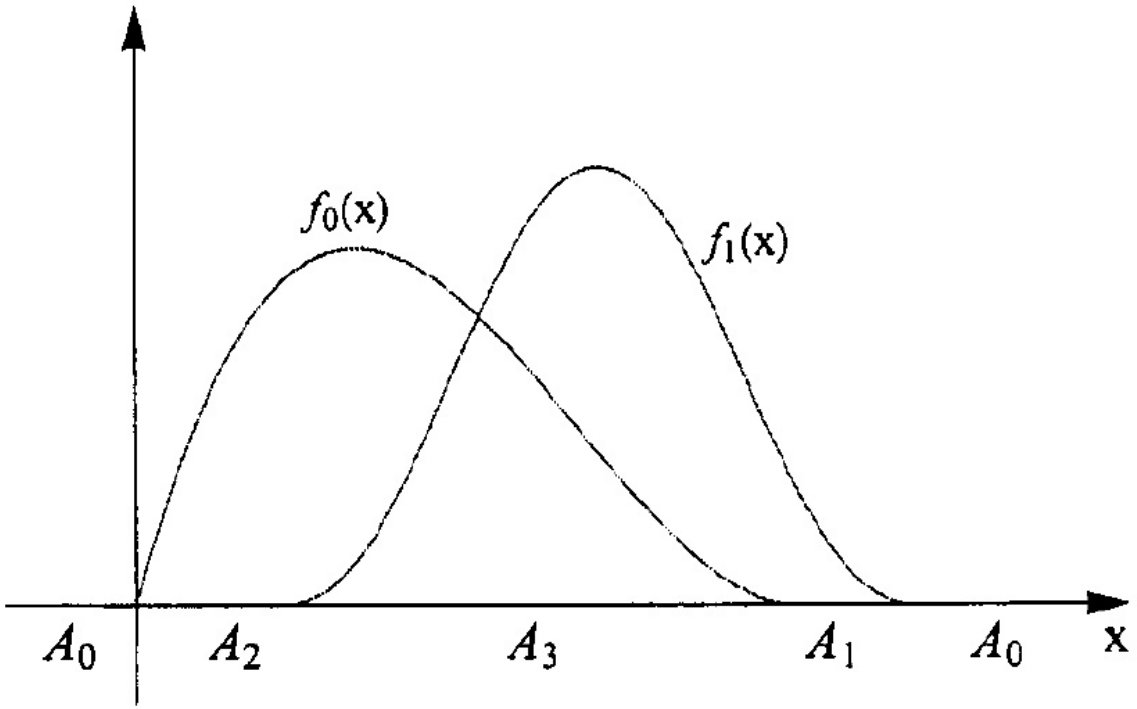
$$\alpha^* = \int_{C^*} f(x,\theta_0)\mathrm{d}x = \int_C f(x,\theta_0)\mathrm{d}x = \alpha.$$

$$\beta^* = 1 - \int_{C^*} f(x,\theta_1)\mathrm{d}x = 1 - \int_C f(x,\theta_1)\mathrm{d}x - \int_{C^c \cap A_1} f(x,\theta_1)\mathrm{d}x \leq$$

$$\leq 1 - \int_C f(x,\theta_1)\mathrm{d}x = \beta.$$

Set $A_2 = \{x : f(x,\theta_0) > 0, f(x,\theta_1) = 0\}$, then for $C^{**} = C \setminus A_2$ we have $\alpha^{**} \leq \alpha$ and $\beta^{**} = \beta$.
Set $A_0 = \{x : f(x,\theta_0) = 0, f(x,\theta_1) = 0\}$ plays no role in minimization of α and β.
Set $A_3 = \{x : f(x,\theta_0) > 0, f(x,\theta_1) > 0\}$.
The problem now reduces to partitioning A_3 to C and its complement in order to improve the critical region.

Partition of values of x



Reduction principle and Neyman-Pearson lemma

- Reduction principle: in choosing critical regions, one should restrict the consideration to the set based on the likelihood ratio

$$C_K = \left\{ x : \frac{f(x, \theta_1)}{f(x, \theta_0)} \geq K \right\}.$$

Most important result in statistical testing:

Theorem 62: (Neyman-Pearson lemma)

Let $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$. Let C^* be a critical region such that there exists a constant $K > 0$ for which

a) if $\frac{f(x,\theta_1)}{f(x,\theta_0)} > K$ then $x \in C^*$;

b) if $\frac{f(x,\theta_1)}{f(x,\theta_0)} < K$ then $x \notin C^*$.

Let C be any critical region. Then $\alpha_C \leq \alpha_{C^*}$ implies $\beta_C \geq \beta_{C^*}$. Moreover, if $\alpha_C < \alpha_{C^*}$ then $\beta_C > \beta_{C^*}$.

Interpretation of N-P lemma

- Idea: Impose an upper bound on one of the probabilities of error and minimize the probability of the other kind of error.
- Intepretation: The test C^* has the minimum error of type II of all tests with fixed error of type I, i.e., C^* is the most powerful test
- Note that Neyman-Pearson lemma dost not restrict us to suppose only the same distributions of sample in H_0 and H_1.

Size of the test, level of the test, significance level

Definition 49: The size of the test C is defined as

$$\bar{\alpha}(C) = \sup_{\theta \in \Theta_0} \pi_C(\theta).$$

- the size of the test of a simple null hypothesis corresponds to the probability of a type I error

Definition 50: Any number $\alpha \geq \bar{\alpha}(C)$ is called the level of the test C. A test C satisfying $\bar{\alpha}(C) \leq \alpha$ is called α-level test.

- E.g., a test with size 0.01 is a 1%-level test, as well as 5%-level test, etc.

Definition 51: If only tests C with size $\bar{\alpha}(C) \leq \alpha_0$ are considered then α_0 is called the significance level.

- in the discrete case we may not be able to find a test C with $\alpha(C)$ equal the desired significance level α_0. In continuous case we can always find such a test!

p-value

- statement that the null hypothesis was rejected at the significance level, e.g., $\alpha = 0.05$ is often not informative enough; we would not know if it is still rejected at some lower level, e.g. $\alpha = 0.01$ or even lower

Definition 52: The lowest significance level at which H_0 is rejected is called p-value.

- the smaller the p-value the stronger the evidence for rejecting H_0
- it can be calculated as a probability that a random variable with a distribution of a test statistics has values beyond the value of the test statistic (depends on the type of alternative hypothesis), e.g. for Z normal and $H_1 : \mu > \mu_0$, p-value equals $P(Z \geq z)$, where z is the value of the test statistics for the particular data set; for Z normal and $H_1 : \mu \neq \mu_0$, p-value equals $P(Z \leq -z) + P(Z \geq z)$. Analogously for other distributions.

Generalization of Neyman-Pearson lemma

- Suppose $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$, possibly both composite
- the likelihood ratio can be generalized to

$$v(X) = \frac{\sup_{\theta \in \Theta_1} f(X, \theta)}{\sup_{\theta \in \Theta_0} f(X, \theta)}$$

  and we reject the null hypothesis if $v(X)$ has high value
- unfortunately, that is hard to compute

Definition 53: The ratio

$$\lambda(X) = \frac{\sup_{\theta \in \Theta_0} f(X, \theta)}{\sup_{\theta \in \Theta} f(X, \theta)}$$

is called the generalized likelihood ratio statistic.

Generalization of Neyman-Pearson lemma

- if Θ_0 is a closed set then $\lambda(X)$ is well defined, since $\sup_{\theta\in\Theta} f(X,\theta)$ is attained for MLE
- $\lambda(X) \leq 1$ and

$$v(X) \leq \frac{\max\{\sup_{\theta\in\Theta_1} f(X,\theta), \sup_{\theta\in\Theta_0} f(X,\theta)\}}{\sup_{\theta\in\Theta_0} f(X,\theta)} = \frac{1}{\lambda(X)}$$

  and hence $v(X)\lambda(X) \leq 1$.
- also, sometimes $\lambda(X)$ does not depend on any parameters, then α-size critical region can be derived from

$$P(\lambda(X) \leq K | H_0) = \alpha.$$

Construction of critical region

Example 98: Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ be random sample, where both μ and $\sigma^2 > 0$ are unknown. Using the generalized likelihood ratio, derive the α-size critical region for test $H_0 : \mu = \mu_0, \sigma^2 > 0$ arbitrary, against $H_1 : \mu \neq \mu_0, \sigma^2 > 0$ arbitrary.

Tests vs. confidence intervals

- Let $[L(X), U(X)]$ be $(1-\alpha)$-level confidence interval of $\theta \in \Theta$, i.e. for every $\theta \in \Theta$

$$P_\theta\left(L(X) \leq \theta \leq U(X)\right) = 1 - \alpha.$$

- Now, suppose test $H_0 : \theta = \theta_0$ against alternative $H_1 : \theta \neq \theta_0$.
- Defining the set

$$A(\theta) = \{x = (x_1, \ldots, x_n) : L(x) \leq \theta \leq U(x)\},$$

we can construct the critical region of the α-level test as $C = [A(\theta_0)]^c$. Then

$$P(H_0 \text{ is rejected } | H_0 \text{ is true }) = P_\theta(X \in [A(\theta_0)]^c | \theta = \theta_0) = \\ = P_{\theta_0}(\theta_0 \notin [L(X), U(X)]) = \alpha.$$

Recall of notation

Recall of notation:

- α .. significance level
- $1-\beta$ .. power
- z_p .. upper pth quantile of $N(0,1)$
- $t_{p,\nu}$ .. upper pth quantile of t_ν
- $\chi^2_{p,\nu}$ .. upper pth quantile of χ^2_ν
- F_{p,ν_1,ν_2} .. upper pth quantile of F_{ν_1,ν_2}

Overview of one-sample tests: σ^2 known

$X_1, \ldots, X_n \sim N(\mu, \sigma^2)$

i) $H_0 : \mu = \mu_0, H_1 : \mu > \mu_0 (\mu < \mu_0)$
test statistics: $U = \frac{\bar{X} - \mu_0}{\sigma}\sqrt{n}$
critical region: $U \geq z_\alpha \quad (U \leq z_{1-\alpha})$

ii) $H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$
test statistics: $U = \frac{\bar{X} - \mu_0}{\sigma}\sqrt{n}$
critical region: $|U| \geq z_{\frac{\alpha}{2}}$

Overview of one-sample tests: μ, σ^2 unknown

iii) $H_0 : \mu = \mu_0, H_1 : \mu > \mu_0 (\mu < \mu_0)$
test statistics: $t = \frac{\bar{X} - \mu_0}{\sqrt{\frac{1}{n}\sum (X_i - \bar{X})^2}} \sqrt{n-1}$
critical region: $t \geq t_{\alpha, n-1} \quad (t \leq -t_{\alpha, n-1})$

iv) $H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$
critical region: $|t| \geq t_{\frac{\alpha}{2}, n-1}$

v) $H_0 : \sigma^2 = \sigma_0^2, H_1 : \sigma^2 > \sigma_0^2 \ (\sigma^2 < \sigma_0^2)$
test statistics: $\chi^2 = \frac{\sum (X_i - \bar{X})^2}{\sigma_0^2}$
critical region: $\chi^2 \geq \chi^2_{\alpha, n-1} \quad (\chi^2 \leq \chi^2_{1-\alpha, n-1})$

vi) $H_0 : \sigma^2 = \sigma_0^2, H_1 : \sigma^2 \neq \sigma_0^2$
critical region: $\chi^2 \leq \chi^2_{1-\frac{\alpha}{2}, n-1}$ or $\chi^2 \geq \chi^2_{\frac{\alpha}{2}, n-1}$

Overview of two-sample tests

$X_1, \ldots, X_m \sim N(\mu_1, \sigma_1^2), Y_1, \ldots, Y_n \sim N(\mu_2, \sigma_2^2)$

σ^2 known

vii) $H_0 : \mu_1 = \mu_2, H_1 : \mu_1 > \mu_2$

test statistics: $U = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$

critical region: $U \geq z_\alpha$

viii) $H_0 : \mu_1 = \mu_2, H_1 : \mu_1 \neq \mu_2$

critical region: $|U| \geq z_{\frac{\alpha}{2}}$

σ^2 unknown, $\sigma_1^2 = \sigma_2^2$

ix) $H_0 : \mu_1 = \mu_2, H_1 : \mu_1 > \mu_2$

test statistics: $T = \frac{\bar{X} - \bar{Y}}{\sqrt{(m-1)S_1^2 + (n-1)S_2^2}} \sqrt{\frac{m+n-2}{\frac{1}{m} + \frac{1}{n}}}$

critical region: $T \geq t_{\alpha, m+n-2}$

x) $H_0 : \mu_1 = \mu_2, H_1 : \mu_1 \neq \mu_2$

critical region: $|T| \geq t_{\frac{\alpha}{2}, m+n-2}$

Overview of two-sample tests

μ unknown

xi) $H_0 : \sigma_1^2 = k\sigma_2^2, H_1 : \sigma_1^2 > k\sigma_2^2$

test statistics: $F = \frac{\frac{\sum(X_i - \bar{X})^2}{m-1}}{\frac{\sum(Y_i - \bar{Y})^2}{k(n-1)}}$

critical region: $F \geq F_{\alpha,m-1,n-1}$

xii) $H_0 : \sigma_1^2 = k\sigma_2^2, H_1 : \sigma_1^2 \neq k\sigma_2^2$

critical region: $F \leq F_{1-\frac{\alpha}{2},m-1,n-1}$ or $F \geq F_{\frac{\alpha}{2},m-1,n-1}$

Overview of tests based on CLTs

$X \sim BIN(n,p)$, n large but neither np nor $n(1-p)$ is small
hence approximately $\frac{X}{n} \sim N(p, \frac{p(1-p)}{n})$
to test $H_0 : p = p_0$ we can use

$$Z = \frac{\frac{X}{n} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} \sim_{\text{as.}} N(0,1)$$

For two sample situations: $X \sim BIN(n_1,p_1), Y \sim BIN(n_2,p_2)$, n_1, n_2 large

$$\frac{X}{n_1} - \frac{Y}{n_2} \sim_{\text{as.}} N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

and under $H_0 : p_1 = p_2 = p$ using $\hat{p}_{MLE} = \frac{X+Y}{n_1+n_2}$ we can use

$$Z = \frac{\frac{X}{n_1} - \frac{Y}{n_2}}{\sqrt{\frac{X+Y}{n_1+n_2}\left(1 - \frac{X+Y}{n_1+n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim_{\text{as.}} N(0,1).$$

Example

Example 99: All philogaps presently on the market have an average concentration of muzzz of at least 3.7 mg per philogap. A company claims to have discovered a new method of production that will decrease the average muzzz content to the level below 3.7 mg. To test this claim, the muzzz content of 15 philogaps of this company are analyzed, and their average muzzz content is found to be 3.52 mg. It is known that the standard deviation of the muzzz content in a philogap does not depend on the production process and equals 0.35 mg. It is also known that the muzzz content is normally distributed. At the significance level $\alpha = 0.01$, does this finding indicate that the new production process decreases the concentration of muzzz in philogaps?

Examples

Example 100: It is claimed that sports-car owners drive on the average 18000 miles per year. A consumer firm believes that the average mileage is lower and obtained information from 40 randomly selected sports-car owners that resulted in a sample mean of 17463 miles with sample standard deviation of 1348 miles. What can we conclude about this claim? Use $\alpha = 0.01$.

Example 101: The intelligence quotients (IQs) of 17 students from one area of a city showed a sample mean of 106 with a sample standard deviation of 10, whereas the IQs of 14 students from another area chosen independently showed a sample mean of 109 with sample standard deviation of 7. Is there a significant difference between the IQs of the two groups at $\alpha = 0.02$? Assume the population variances are equal.

Example

Example 102: In primary elections, 28% of the Republicans in New Hampshire voted for candidate A. A poll of 180 Republicans in Iowa show that 41 of them will vote for candidate A. Does this result indicate that the Republican support of candidate A is lower in Iowa than in New Hampshire?

Pair test versus two-sample test

- rather then two samples we have a random sample of pairs $(X_i, Y_i), i = 1, \ldots, n$ (independent as couples). So our data are couples of measurements, usually on the same subject. Test: μ_X equals to μ_Y; similarly about variances.
- define $Z_i = X_i - Y_i$, thus get a (one-dimensional) random sample and e.g. test $\mu_Z = \delta$ where $\mu_Z = \mu_X - \mu_Y$ by one-sample t-test. δ can be 0 but it can also be any real.
- How to decide which test to apply?:
  - $n_1 \neq n_2$ – apply two-sample test
  - $n_1 = n_2$ but samples are independent – apply two-sample test
  - $n_1 = n_2$ and two samples are dependent e.g. because the measurement is taken on same subjects – apply pair test

Applying the other type than you should may lead to incorrect result of the test! - for two-sample design, a much larger sample is needed to achieve statistical significance for a given difference and variability in the data

Example of paired test

Example 103: Say that we observe patients with trouble falling asleep and we measure time in minutes, once without getting a certain drug inducing sleep and after taking the drug. Let us denote it X_i and Y_i respectively. Let the observed data be given by the table

Subject	No pill (X_i)	Pill (Y_i)
1	65	45
2	35	5
3	80	61
4	40	31
5	50	20

Test the effectiveness of the sleeping pills based on observed time needed before falling asleep. Test at level $\alpha = 0.05$.