Statistical hypothesis testing

EM13-Probability and statistics: Courses 11-12

September 2014

Manuel SAMUELIDES[1] Zhigang SU[2]

[1]Professor

Institut Supereur de l'Aeronautique et de l'Espace

[2]Professor

Sino-European Institute of Aviation Engineering

Civil Aviation University of China

Homework–20141009

Hypothesis testing of normal laws: Suppose that nine observations are selected at random from a normal distribution for which both the mean μ and the variance σ^2 are unknown. For these nine observations, the empirical mean and the empirical variance are respectively $\bar{X} = 22$ and $\sum_{i=1}^{n}(x_i - \bar{X})^2 = 72$

1. Carry out a test of the following hypotheses at the level of significance 0.05:

$$H_0 : \quad \mu \leqslant 20, \qquad\qquad H_1 : \quad \mu > 20$$

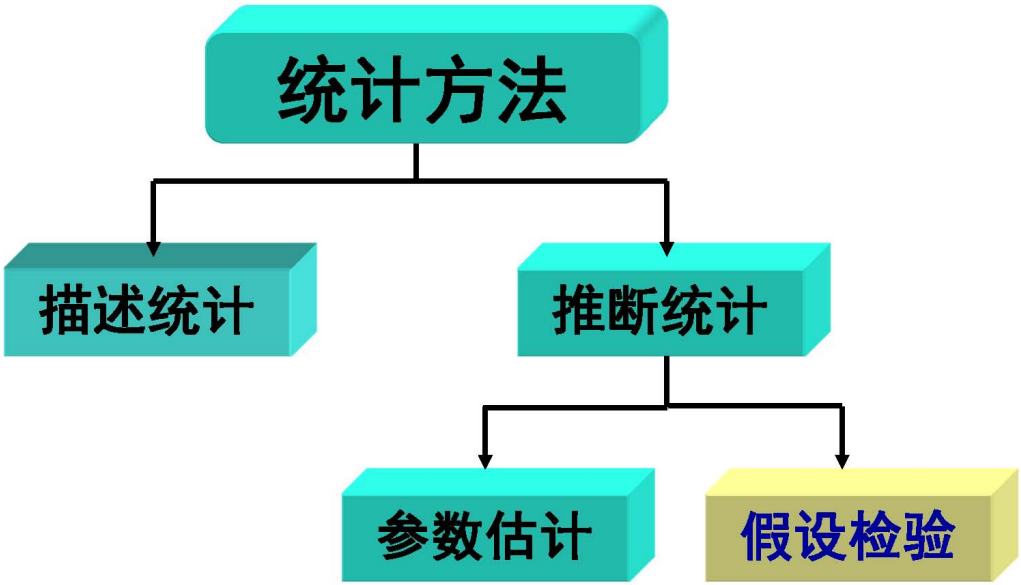
2. Carry out a test the following hypotheses at the level of the significance 0.05 by using the two-sided test:

$$H_0 : \quad \mu = 20, \qquad\qquad H_1 : \quad \mu \neq 20$$

3. From the data, construct a confidence interval for μ with confidence coefficient 0.95.

A **statistical hypothesis test** is a method of statistical inference using data from a scientific study.

Statistical hypothesis testing is a key technique of both Frequentist inference and Bayesian inference although they have notable differences.



假设检验的基本思想是应用小概率原理的反证法

- **小概率原理**：指发生概率很小的随机事件(小概率事件)在一次试验中是几乎不可能发生的。小概率指$p < 5\%$。
- **反证法**：为了检验一个假设是否成立，首先假设它是真的，然后对样本进行观察，如果发现出现了不合理现象(选择某区间值的概率小于显著性水平)，则可以认为假设是不合理的，拒绝假设。否则可以认为假设是合理的，接受假设。

Several concepts — null and alternative hypotheses

Null hypothesis (H_0) / 零假设

A simple hypothesis associated with a contradiction to a theory one would like to prove.

在零假设中，所有因素对变量都不起任何作用。

Alternative hypothesis (H_1) / 备择假设、对立假设

A hypothesis (often composite) associated with a theory one would like to prove.

Remarks

- An important property of a test statistic is that its sampling distribution under the null hypothesis must be calculable, either exactly or approximately, which allows α-values to be calculated.
- 对于任何一个假设检验问题所有可能的结果都应包含在两个假设之内，非此即彼。

假设检验是具有概率性质的反证法

- 所谓假设的不合理不是绝对的，而是基于实践中广泛采用的小概率事件几乎不可能发生的原则。
- 至于事件的概率小到什么程度才算是小概率事件，并没有统一的界定标准，而是必须根据具体问题而定。
- 如果一旦判断失误，错误地拒绝零假设会造成巨大损失，那么拒绝原假设的概率就应定的小一些；如果一旦判断失误，错误地接受原假设会造成巨大损失，那么拒绝原假设的概率就应定的大一些。

	H_0 is true	H_1 is true
Accept Null	Right decision	Wrong decision Type II Error
Reject Null	Wrong decision Type I Error	Right decision

Example

法庭窘境：

- 零假设H_0认为被告是清白的，备择假设H_1认为被告有罪。
- 起诉是因为怀疑被告有罪。
- H_0（现状）与H_1对立并且被认可，除非能够证明其不成立。
- “无法排除H_0”并不能代表被告清白，只是说证据无法将其定罪。
- 陪审团没有必要在H_0“无法推翻”的情况下将其“接受”。
- 当零假设无法被“证明”时，可以通过强度检测判断假设是否近似成立，即进行第二型错误检测。

Several concepts — Likelihood ratio

Likelihood function

Let X be a random variable with a probability distribution with density function f depending on a parameter θ. Then the function

$$\mathcal{L}(\theta|x) = f(x|\theta)$$

considered as a function of θ, is called the **likelihood function**.

The likelihood function is $L(\theta|x) = f(x|\theta)$ (with $f(x|\theta)$ being the pdf or pmf), which is a function of the parameter θ with x held fixed at the value that was actually observed, i.e., the data.

Likelihood ratio

We want to test

- the null hypothesis $\theta \in H_0$
- against the alternative hypothesis $\theta \in H_1$.
- the test is based on the sample $\mathbf{X} = (X_1, \ldots, X_n) \in \mathcal{X}$ with likelihood $f_{\mathbf{X}}(\mathbf{x}|\theta)$.

The likelihood ratio is defined by on $\mathcal{X} \times H_0 \times H_1$ by

$$R(\mathbf{x}, \theta_1, \theta_0) = \frac{f_{\mathbf{X}}(\mathbf{x}|\theta_1)}{f_{\mathbf{X}}(\mathbf{x}|\theta_0)}$$

Example

Mean of normal law $N(\theta, \sigma^2)$. With n-samples of the normal law $N(\theta, \sigma^2)$, a hypothesis test is performing between two point hypotheses

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1$$

then the likelihood-ratio test which rejects H_0 in favour of H_1 when

$$R(x, \theta_0, \theta_1) = \frac{L(\theta_0 \mid x)}{L(\theta_1 \mid x)} = \frac{f(x \mid \theta_0)}{f(x \mid \theta_1)} \leqslant \eta$$

Example

$$f(x \mid \theta_0) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{k=1}^{n} (x_k - \theta_0)^2 \right\}$$

$$f(x \mid \theta_1) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{k=1}^{n} (x_k - \theta_1)^2 \right\}$$

So

$$R(x, \theta_0, \theta_1) = \exp\left\{ \frac{1}{2\sigma^2} \left[-\sum_{i=1}^{n} (x_i - \theta_0)^2 + \sum_{i=1}^{n} (x_i - \theta_1)^2 \right] \right\}$$

$$= \exp\left\{ \frac{1}{\sigma^2} \left[(\theta_0 - \theta_1)(x_1 + \cdots + x_n) - \frac{\theta_0^2 - \theta_1^2}{2/n} \right] \right\}$$

Several concepts — test statistics / 检测统计量

Definition

A **test statistic** is considered as a numerical summary of a data-set that reduces the data to one value that can be used to perform the hypothesis test.

Monotonous likelihood ratio parametric model (MLR)

Suppose that $T(x)$ is an exhaustive statistics (完全统计量) for the parametric model $f(x, \theta)$, the model is said to be MLR iff the likelihood ratio $R(x, \theta_1, \theta_0)$ is a monotonous function of T for every (θ_1, θ_0) such that $\theta_0 < \theta_1$

Example: mean of normal law $N(\theta, \sigma^2), \theta = \mu, \sigma^2$ is known

$$R(\mathbf{X}, \theta_1, \theta_0) = \exp\left\{\frac{1}{\sigma^2}\left[(\theta_1 - \theta_0)(X_1 + \cdots + X_n) - \frac{\theta_1^2 - \theta_0^2}{2}\right]\right\}$$

is MLR with respect to the exhaustive statistics

$$T(\mathbf{X}) = \sum_{i=1}^{n} X_i.$$

Several concepts — significance level / 显著性水平

Significance level

Significance level $\alpha(>0)$ is the low probability of obtaining at least as extreme results given that the null hypothesis is true.

(保证H_0成立的最低概率)

H_0 is rejected by the test $T(X) \geqslant \xi$ at the significance level $\alpha > 0$ iff

$$Pr\{T(X) \geq \xi|\theta\} \leqslant \alpha \qquad \forall\theta \in H_0$$

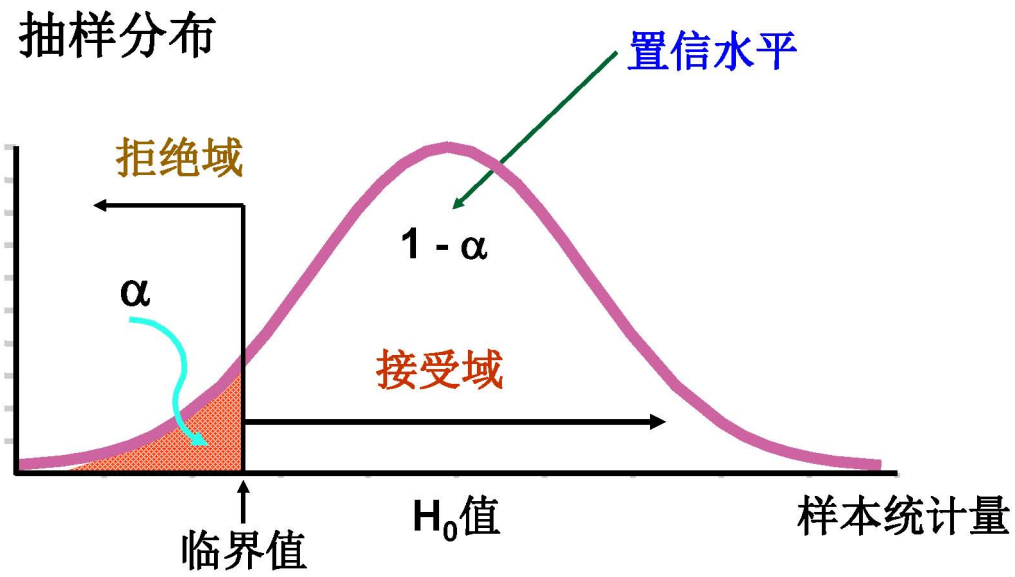
A researcher will often "reject the null hypothesis" when the p-value ($= Pr\{T(X) \geq \xi|\theta\}$) turns out to be less than a certain significance level.
An informal interpretation of a p-value, based on a significance level of about 10%, might be:

- $\alpha \leqslant 0.01$: very strong presumption against H_0
- $0.01 < \alpha \leqslant 0.05$: strong presumption against H_0
- $0.05 < \alpha \leqslant 0.1$: low presumption against H_0
- $\alpha > 0.1$: no presumption against H_0

Several concepts — confidence level / 置信度

Confidence level

If a corresponding hypothesis test is performed, the **confidence level** is the complement of respective level of significance.



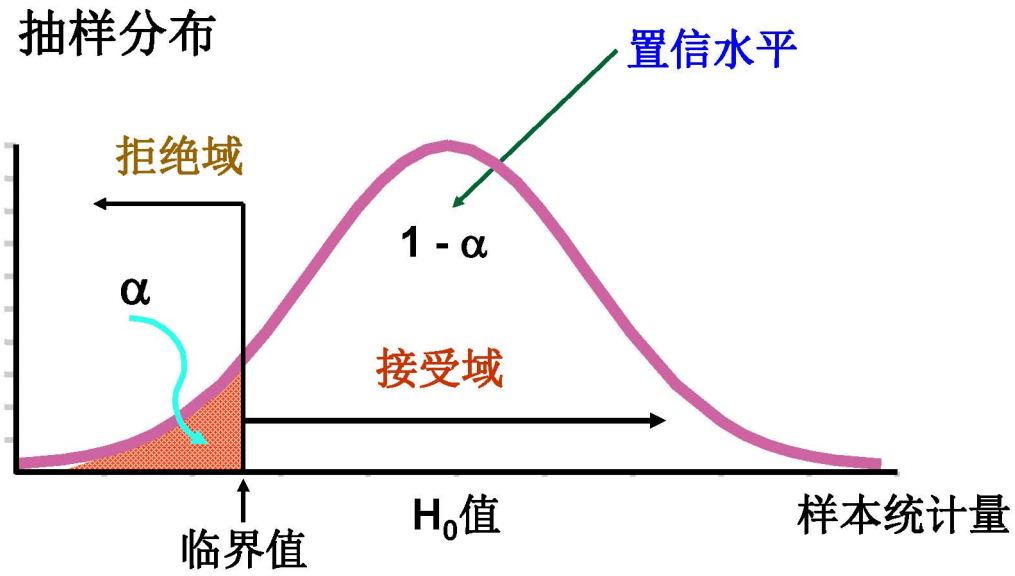
Several concepts — regions of acceptance and rejection

Region of acceptance / 接受区间

The set of values of the test statistic for which we fail to reject the null hypothesis.

Region of rejection / Critical region / 拒绝区间

The set of values of the test statistic for which the null hypothesis is rejected.



Computation of the critical region for a significance level

The likelihood function is

$$f_{\bar{X}}(\bar{x}|\theta) = \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{n}{2\sigma^2}(\bar{x}-\theta)^2\right\}.$$

The obvious test to reject H_0 is $\overline{X} \geq \xi$

$$\begin{aligned} Pr(\overline{X} \geq \xi|\theta) &= 1 - Pr(\overline{X} < \xi|\theta) \\ &= 1 - F(\frac{\sqrt{n}}{\sigma}\xi + \theta) \leqslant \alpha \end{aligned}$$

So

$$\xi \geq sup_{\theta \in H_0} \frac{\sigma}{\sqrt{n}}[F^{-1}(1-\alpha) - \theta] = \frac{\sigma}{\sqrt{n}}F^{-1}(1-\alpha)$$

$F_X(x)$ be the cumulative function of $N(0,1)$.

The threshold of the most powerful test is

$$\xi \geq sup_{\theta \in H_0} \frac{\sigma}{\sqrt{n}}[F^{-1}(1-\alpha) - \theta] = \frac{\sigma}{\sqrt{n}} F^{-1}(1-\alpha)$$

n	1	10	50	100	1000
$\alpha = 0.1$	1.29	0.40	0.18	0.13	0.04
$\alpha = 0.05$	1.65	0.52	0.23	0.16	0.05
$\alpha = 0.01$	2.33	0.74	0.33	0.23	0.07
$\alpha = 0.005$	2.58	0.81	0.36	0.26	0.08

Table:Threshold of the most powerful unilateral test for $\sigma = 1$

Matlab: $xi = norminv(1-\alpha,\ 0,\ 1)./sqrt(n)$
Results of simulation: for $\mu = 0.4$ and $n = 50$, H_0 is generally rejected at an excellent level

The power of test is $\alpha(\theta) = Pr(\bar{X} \geqslant \xi|\theta)$, for $\forall\theta \in H_0$

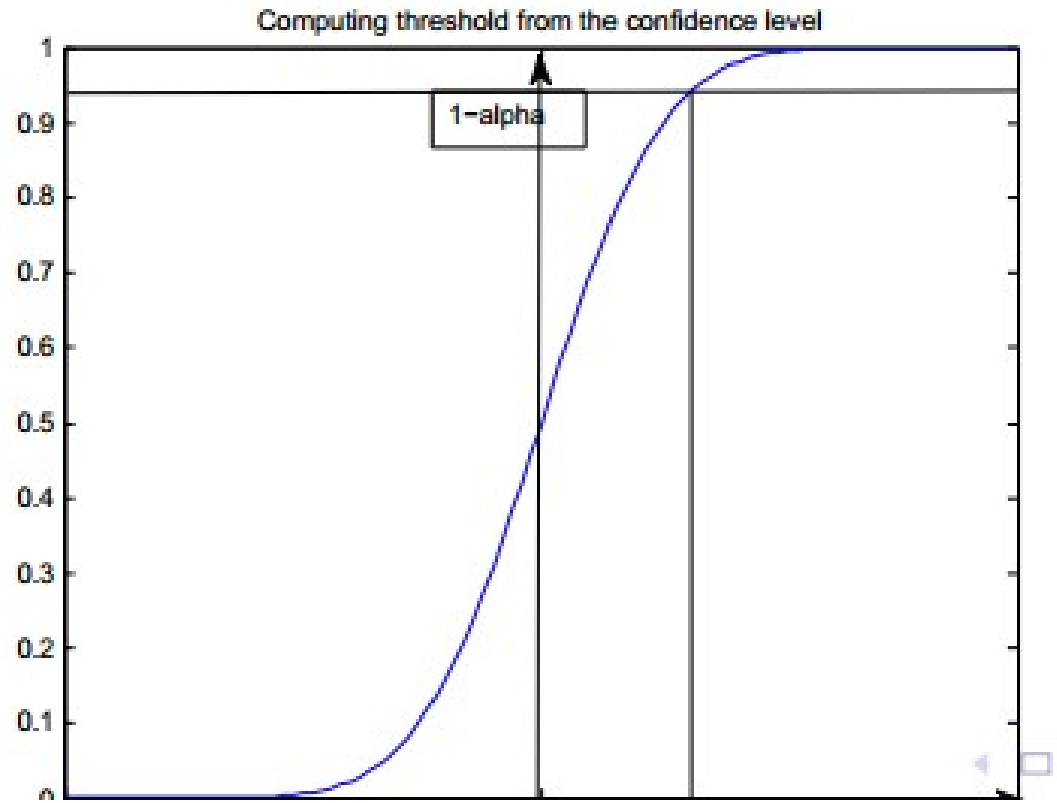
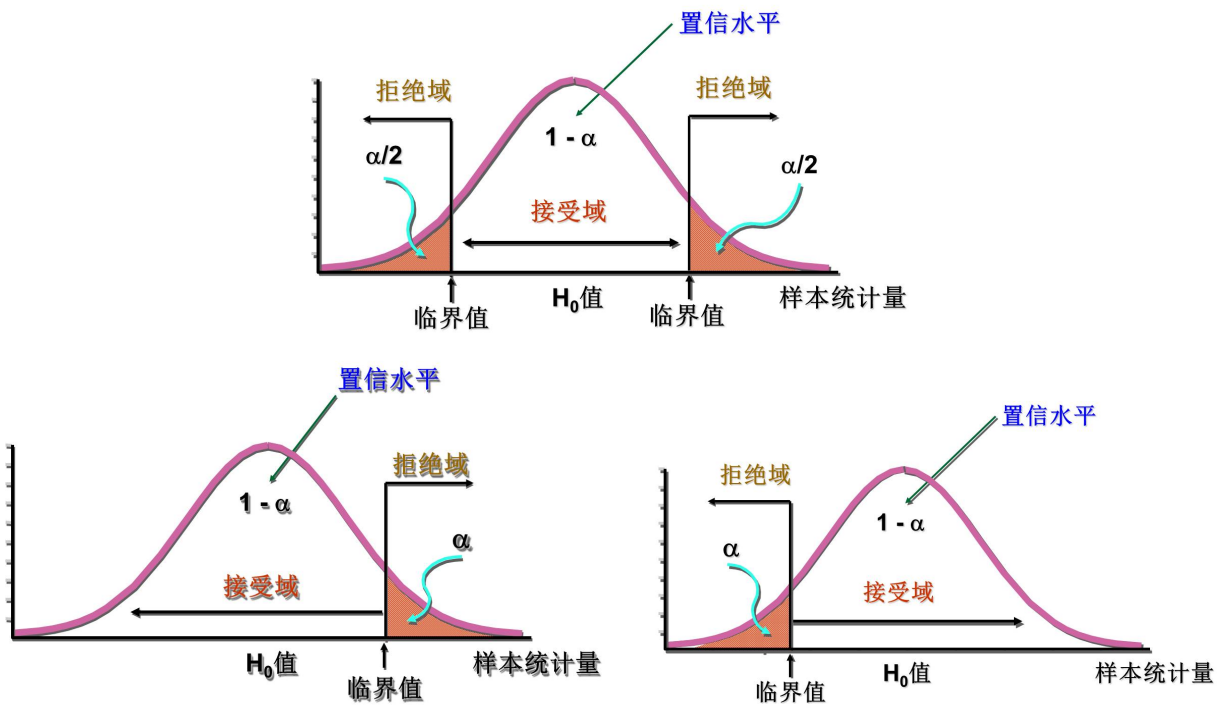


Figure: Computing threshold from the confidence level

Types of hypothesis test

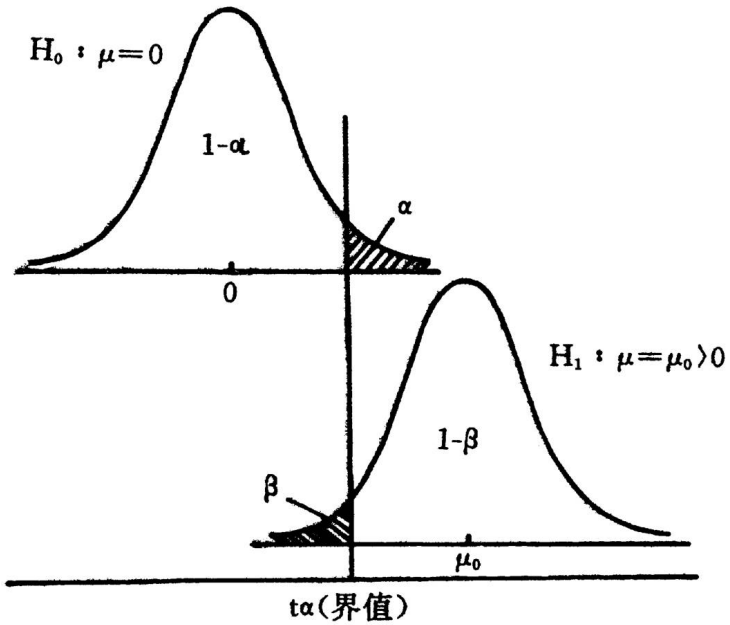
	Task	H_0	H_1
Two-sided (Bilateral)	$\mu \neq \mu_0$	$\mu = \mu_0$	$\mu \neq \mu_0$
Left-sided (Unilateral)	$\mu < \mu_0$	$\mu \geqslant \mu_0$	$\mu < \mu_0$
Right-sided (Unilateral)	$\mu > \mu_0$	$\mu \leqslant \mu_0$	$\mu > \mu_0$



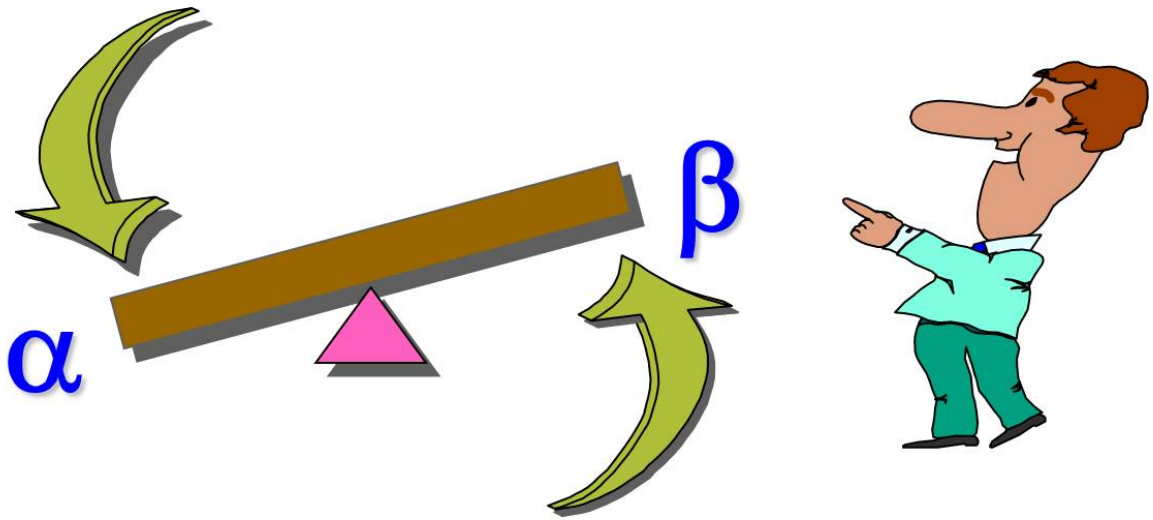
假设检验中的两类错误

假设检验是依据样本提供的信息进行推断的，即由部分来推断总体，因而假设检验不可能绝对准确,是可能犯错误的。

- I型错误：H_0为真时却被拒绝,弃真错误(α错误);
- II型错误：H_0为假时却被接受,取伪错误(β错误);



- α与β是两个前提下的概率，所以$\alpha + \beta \neq 1$
- 对于固定的样本数，α与β一般情况下不能同时减小



Good properties of composite hypothesis testing

- Suppose now we want to test composite hypothesis $\theta \in H_0$ against $\theta \in H_1$.
- Recall that the level of the test with critical region W is $\sup Pr(X \in W_0 | \theta \in H_0)$
- This test is called without bias iff

$$\sup Pr(X \in W_0 | \theta \in H_0) \leqslant \inf Pr(X \in W_0 | \theta \in H_1)$$

- The power of the test is the function γ defined on H_0 by

$$\gamma(\theta_1) = Pr(X \in W_0 | \theta \in H_1)$$

- We are looking for UMP tests, i.e. **uniformly most powerful (UMP / 一致最优)** tests among tests at the same level that are without bias

Neyman – Pearson lemma

Lemma

In statistics, the Neyman – Pearson lemma states that when performing a hypothesis test between two point hypotheses

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1$$

then the likelihood-ratio test which rejects H_0 in favour of H_1 when

$$R(x, \theta_1, \theta_0) = \frac{L(\theta_1 \mid x)}{L(\theta_0 \mid x)} = \frac{f(x \mid \theta_1)}{f(x \mid \theta_0)} \geqslant \eta$$

where

$$Pr(R(x, \theta_1, \theta_0) \geqslant \eta \mid H_0) = \alpha$$

is the most powerful test of size α for a threshold η. If the test is most powerful for all $\theta_1 \in \Theta_1$, it is said to be **uniformly most powerful (UMP / 一致最优)** for alternatives in the set Θ_1.

Remark

- The Neyman – Pearson lemma is named after Jerzy Neyman and Egon Pearson.
- Neyman-Pearson引理说明，似然比检验是所有具有同等显著性差异的检验中最有统计效力的检验.

In practice, the likelihood ratio is often used directly to construct tests. However it can also be used to suggest particular test-statistics that might be of interest or to suggest simplified tests — for this, one considers algebraic manipulation of the ratio to see if there are key statistics in it related to the size of the ratio (i.e. whether a large statistic corresponds to a small ratio or to a large one).

Example(Neyman – Pearson lemma)

Let $X_1, \ldots, X_n$ be a random sample from the $\mathcal{N}(\mu, \sigma^2)$ distribution where the mean μ is known, and suppose that we wish to test for

$$\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 = \sigma_1^2. \end{cases}$$

The likelihood for this set of normally distributed data is

$$f_{\mathbf{X}}\left(\mathbf{X}|\sigma^2\right) \propto \left(\sigma^2\right)^{-n/2} \exp\left\{-\frac{\sum_{i=1}^{n}\left(x_i-\mu\right)^2}{2\sigma^2}\right\}.$$

We can compute the likelihood ratio to find the key statistic in this test and its effect on the test's outcome:

$$\begin{aligned} R(\mathbf{X}, \sigma_0^2, \sigma_1^2) &= \frac{f_{\mathbf{X}}\left(\mathbf{X}|\sigma_0^2\right)}{f_{\mathbf{X}}\left(\mathbf{X}|\sigma_1^2\right)} \\ &= \left(\frac{\sigma_0^2}{\sigma_1^2}\right)^{-n/2} \exp\left\{-\frac{1}{2}(\sigma_0^{-2}-\sigma_1^{-2})\sum_{i=1}^{n}\left(x_i-\mu\right)^2\right\}. \end{aligned}$$

This ratio $R(\mathbf{X}, \sigma_0^2, \sigma_1^2)$ only depends on the data through $\sum_{i=1}^{n} (x_i - \mu)^2$.

Therefore, by the Neyman – Pearson lemma, the most powerful test of this type of hypothesis for this data will depend only on $\sum_{i=1}^{n} (x_i - \mu)^2$.

Also, by inspection, we can see that if $\sigma_1^2 > \sigma_0^2$, then $R(\mathbf{X}, \sigma_0^2, \sigma_1^2)$ is a decreasing function of $\sum_{i=1}^{n} (x_i - \mu)^2$.

So we should reject H_0 if $\sum_{i=1}^{n} (x_i - \mu)^2$ is sufficiently large.

The rejection threshold depends on the size of the test.

In this example, the test statistic can be shown to be a scaled Chi-square distributed random variable and an exact critical value can be obtained.