Chap. 6 Hypothesis testing

Introduction

□ Often interested in questions such as:

- Are my data well described by a linear function ?
- Is the average age of people in europe the same as the one in south america ?
- Is this patient affected by cancer ?
- Are my data in favor of the standard model of particle physics or some other model ?
- Is the particle I detect a photon or an electron ?
- ...

→ In order to answer, one has to invoke statistical techniques gathered under the name "**hypothesis testing**"

Introduction

□ **Possible outcomes of hypothesis testing**

		The truth	
		1	**0**
Your conclusion	**1**	true positive	false positive
	0	false negative	true negative

□ **General objective in hypothesis testing:** derive conclusions that are as certain as possible

□ Of course, impossible to be 100% sure of your conclusions

→ Always a chance that you're wrong

→ You need to decide what risk you're ready to take

Workflow of hypothesis testing

1. Clearly state what hypothesis you want to test (considered as true by default)

   → "null" hypothesis H_0

   And, if needed, other alternative hypothesis you also want to confront the data to

   → "alternative" hypothesis H_1

2. Choose one (or more) variable on which the hypothesis test will be based

   → so-called "test statistic" (often written t)

3. Determine what values the test statistic typically get under the null and alternative hypotheses
4. Make your measurement and determine the observed value of the test statistic
5. Compare observed value to typical values under null and alternative hypotheses and conclude

→ **Purpose of this chapter:** detail all these steps !

Central notion: **test statistic**

- □ Word "**test statistic**" may seem strange and difficult to understand at first glance
- □ Not as difficult as it seems
  - → As first approximation: "test statistic" $\overset{\text{synonym}}{=}$ "random variable"
  - → But not any random variable: one that is suited for hypothesis testing
- □ **What is a r.v. suited for hyp. testing ?**
  - → It is a r.v. sensitive to the hypotheses under test (H_0 and H_1)
  - → It behaves differently under H_0 and H_1

$$\boxed{\text{Test statistic} = \textbf{discriminating random variable}}$$

  - → Test statistics chosen so as to offer **discrimination power** as good as possible

Test statistic: examples

□ **Example 1:** Are my data distributed according to the normal distribution with mean $= 0$?

→ Use sample mean as test statistic

□ **Example 2:** Is the activity of a given radioactive source greater than 10 MBq ?

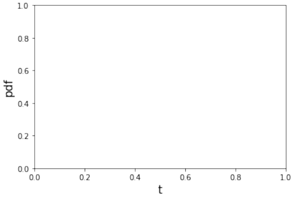
→ Use number of counts as test statistic

□ Test statistic can be either:

- An observable (as in example 2)
- A function of the observables (as in example 1)

Workflow of hyp. testing in pictures

1. Clearly state what H_0 and H_1 are

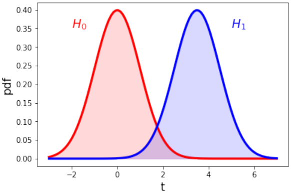


2. Choose test statistics t

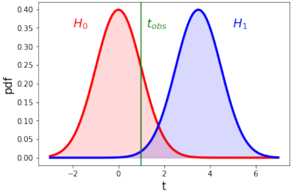
$$t(x_1, \ldots, x_n)$$

3. Determine distributions of t under H_0 and H_1

$$f(t|H_0) \quad \text{and} \quad f(t|H_1)$$

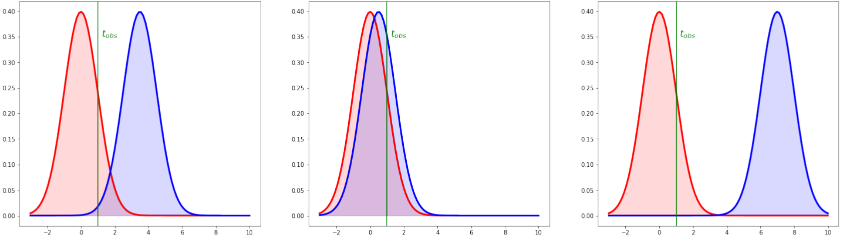


4. Make measurement and calculate t_{obs}



5. Conclude

Discrimination power: illustration



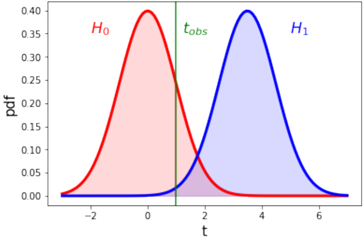
A word of caution

- Word "test statistic" often shortened to "test" or "statistic"
- You may hear/read formulations such as:
  - For my test I use the sample mean as test.
  - For my test I use the sample mean as statistic.
  - The statistic is the sample mean.
  - The distribution of the test is ...
  - The distribution of the statistic is ...

⚠

Don't be confused ! Make sure you understand what this means !

□ **How do we conclude once we have the following plot ?**



→ Need to have a quantitative measure of agreement between observation (t_{obs}) and hypotheses

Measure of agreement between observation and hypotheses

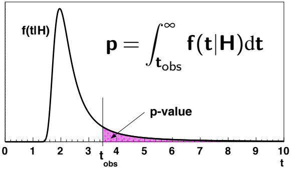
□ Agreement measured with ***p-value***

□ **Definition:**

p-value = probability to observe what you observed in measurement or "more extreme" values

□ Meaning of "more extreme" depends on context:

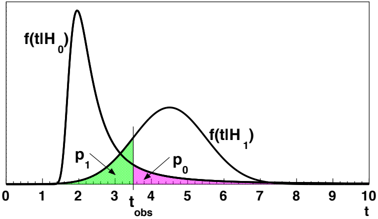
- If only large values are considered a sign of disagreement:



- If only low values: $p = \int_{-\infty}^{t_{\text{obs}}} f(t|H)\mathrm{d}t$
- If both: two-sided definition of *p-value*

More on *p-values*

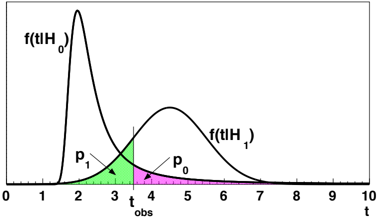
- □ With two hypotheses, the *p-values* look like



- □ *p-values* are random variables

$\rightarrow$ How are they distributed ?

Distribution of *p-values*



□ Suppose $t \sim H_0$ and let $g(p_0|H_0)$ be the distribution of p_0 under H_0:

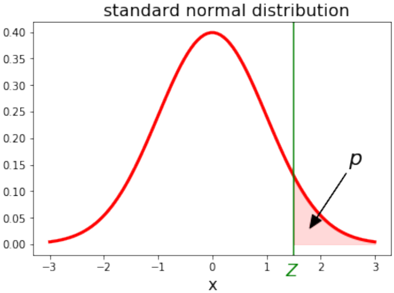
$$g(p_0|H_0) = f(t_{\rm obs}|H_0)\frac{1}{|\mathrm{d}p_0/\mathrm{d}t_{\rm obs}|}$$

□ $p_0 = \int_{t_{\rm obs}}^{\infty} f(t|H_0)\mathrm{d}t = 1 - F(t_{\rm obs}|H_0)$
$\Rightarrow \mathrm{d}p_0 = -dF(t_{\rm obs}|H_0) = -f(t_{\rm obs}|H_0)\mathrm{d}t_{\rm obs}$

□ **Conclusion:** $g(p_0|H_0) = 1$!!!

Significance

- Rather than *p-value*, one sometimes uses the **significance Z**

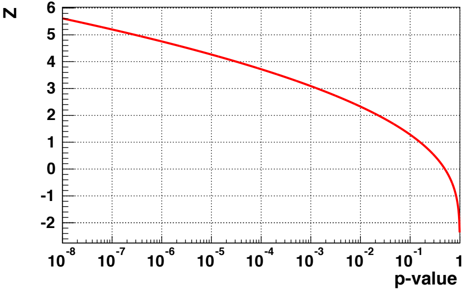


$$Z = \Phi^{-1}(1 - p)$$

- **Important remark**: using Z rather than *p-value* to report values doesn't mean that the test follows normal distribution
  - → Test distribution can be anything
  - → Using Z just means performing a simple change of variable

Significance

□ Why is this change of variable interesting ?



p-value	Z
0.05	1,64
0.00135	3
$2,87 \times 10^{-7}$	5

8 orders of magnitude in p ⇔ 1 order of magnitude in Z

□ Terminology: when $Z = X$, we say that the "significance is $X\sigma$"

Exercice

Let's consider a Poisson counting experiment with one signal and one or more background processes. Let b be the total number of expected background events and N_{obs} the observed number of events. We consider the case $N_{\text{obs}} > b$. Show that, in the asymptotic limit and under the background only hypothesis, the significance of the observation is

$$Z \simeq \frac{\hat{s}}{\sqrt{b}}$$

where $\hat{s} = N_{\text{obs}} - b$ is the estimator of the number of signal events.

How do we conclude eventually ?

□ We choose *a priori* a threshold value for the *p-value*: α

- **If p** $< \alpha$: observation considered too extreme to be compatible with hypothesis

  $\rightarrow$ **Hypothesis is rejected**

- **If p** $> \alpha$: observation considered compatible with hypothesis

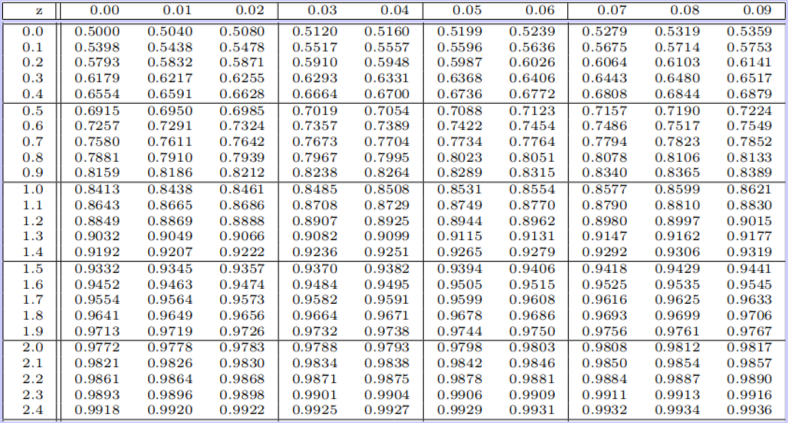
  $\rightarrow$ **Hypothesis is accepted**

□ Terminology: α called the **size** or **significance level** of the test

Exercice

Suppose that, during a certain period of time in a certain population, 49581 boys and 48870 girls were born. Are these observations in favor of the hypothesis according to which the fractions of male birth and female birth are equal to 50% at the 5% level ?

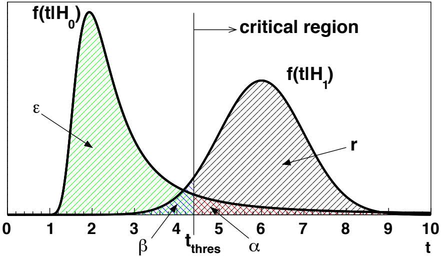
Standard normal distribution cdf values



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936

Equivalent approach

□ Rather than using the *p-value* to conclude, we can use the value of the test t_{obs} itself



- If observation in critical region ($p < \alpha$): null hypothesis rejected
- If observation not in critical region ($p > \alpha$): null hypothesis accepted

How is the size of the test α chosen ?

□ Complex question:

No general answer $\rightarrow$ very problem dependent

□ It depends on:

- The nature of the test you're carrying, of the null and alternative hypothesis
- The consequences of the different possible conclusions

  $\rightarrow$ In particular the consequences of deriving false conclusions (what happens in case of false negative or false positive ?)
- The risk you're ready to take to suffer the consequences

Choice of α: example 1

□ **Testing well established hypothesis** (e.g. Einstein's theory of relativity):

- *A priori* very confident in hypothesis

  → Rejecting it would be a major event

- Need very strong evidence to reject hypothesis

  $\Rightarrow$ Choose small size

- Typical value is $\alpha = 2.87 \times 10^{-7}$

  → Requires "5σ observation" to reject hypothesis

Choice of α: example 2

□ **Cancer diagnosis:**

- If you conclude that your patient is not affected by cancer while he is (false negative)

  → Enormous consequences

- If you conclude that your patient is affected by cancer while he isn't (false positive)

  → Important consequences too

⇒ **Goal:** minimize false negative and false positive probabilities

Problem: not possible to lower false negative and false positive probabilities to arbitrary low values at the same time

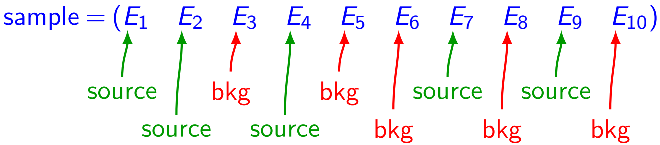
→ Always a trade-off between the two

→ Do you prefer minimizing false negatives or false positives ?

Choice of α: example 3

□ **Selection of pure samples**:

- Occurs when you have composite samples and want to measure some properties of only one type of event appearing in the mixture
- Example: measurement of radioactive source:



$$\text{sample} = (E_1 \quad E_2 \quad E_3 \quad E_4 \quad E_5 \quad E_6 \quad E_7 \quad E_8 \quad E_9 \quad E_{10})$$

source bkg bkg source source

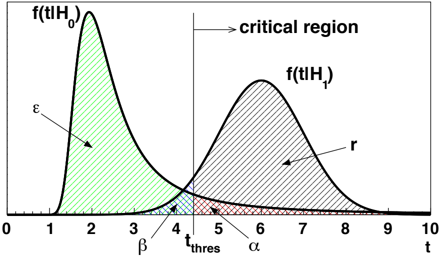
source source bkg bkg bkg

→ From this sample, how to build a sample enriched in events from the source ?

In other words: How to reject the background as much as possible while keeping signal events ?

□ In next slides, will describe how to maximize purity

Type-I and type-II errors (and related quantities)

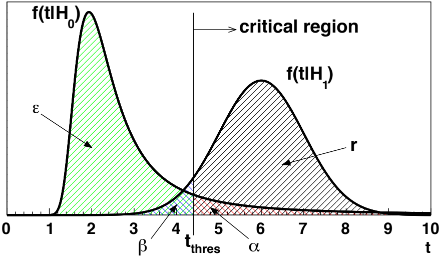


- □ **Type-I error (error of the first kind)**=reject of H_0 if it's true
  - Probability of type-I error $= \alpha$
- □ **Type-II error (error of the second kind)**=accept H_0 if it's false
  - Probability of type-II error $= \beta$
- □ **Related quantities:**
  - Efficiency: $\varepsilon = 1 - \alpha$
  - Power (or rejection): $r = 1 - \beta$

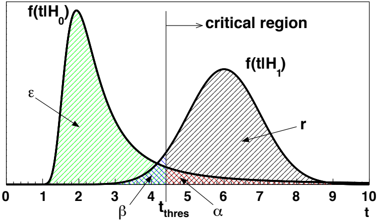
Purity

□ By definition:

$$\boxed{\begin{aligned} \text{purity} &= \text{fraction of events from } H_0 \text{ in selected sample} \\ &= P(H_0 | t \notin \mathscr{C}) \quad (\mathscr{C} = \text{critical region}) \end{aligned}}$$



Purity calculation



□ Notations:

- $c = P(H_0)$
- $1 - c = P(H_1)$

$$f(t) = c \times f(t|H_0) + (1-c) \times f(t|H_1)$$

□ Purity:

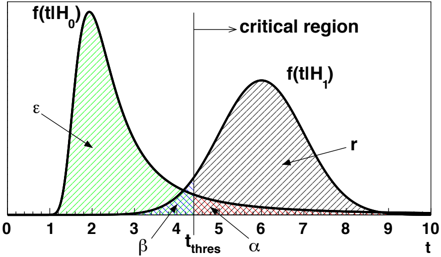
$$P(H_0|t \notin \mathscr{C}) = \frac{P(t \notin \mathscr{C}|H_0)P(H_0)}{P(t \notin \mathscr{C})} = \ldots = \frac{1}{1 + \frac{\beta}{\varepsilon}\frac{1-c}{c}}$$

□ **Conclusion:** purity maximized by maximizing the ratio ε/β

→ For fixed ε (α), must maximize power

Purity

□ **Conclusion:** for fixed ε (α), must maximize power



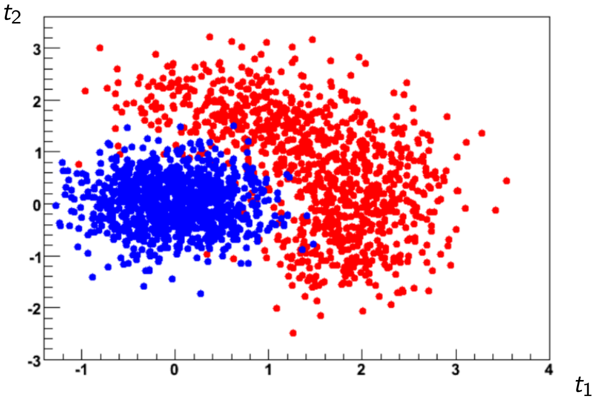
□ **1D case:**

- No degrees of freedom once size fixed
- If varying the size is acceptable, maximize ratio ε/β

□ **2D (or higher) case:**

- More complex (see next slide)

Purity: 2D (or higher) case



- □ No unique critical region for a given size α
- □ Must select critical region that maximizes power

→ How do we do that ?

Neyman-Pearson lemma

□ Critical region that maximizes power for fixed α given by:

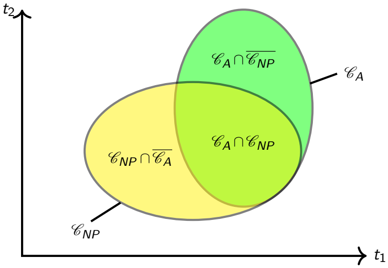
$$\boxed{\frac{f(t|H_0)}{f(t|H_1)} \leq k_\alpha}$$

□ **Notes:**

- t is a multidimensional object: $t = (t_1, t_2, \ldots)$
- $\frac{f(t|H_0)}{f(t|H_1)}$ called **likelihood ratio**
  $\rightarrow$ Traditionally noted Λ
- k_α is a function of α

□ Neyman-Pearson critical region denoted as $\mathscr{C}_{NP}$ and referred to as the Best Critical Region (BCR)

Neyman-Pearson lemma: proof


t_2
$\mathscr{C}_A \cap \overline{\mathscr{C}_{NP}}$
$\mathscr{C}_A$
$\mathscr{C}_A \cap \mathscr{C}_{NP}$
$\mathscr{C}_{NP} \cap \overline{\mathscr{C}_A}$
$\mathscr{C}_{NP}$
t_1

- □ $\mathscr{C}_A$ and $\mathscr{C}_{NP}$ have same size

$$P(C_A|H_0) = P(C_{NP}|H_0) = \alpha$$

- □ Must prove that

$$P(C_A|H_1) \leq P(C_{NP}|H_1) \quad \text{when} \quad \Lambda \leq k_\alpha$$

- □ **Proof:**

$$P(C_A|H_1) \leq P(C_{NP}|H_1) \Leftrightarrow P(C_A \cap \overline{C_{NP}}|H_1) \leq P(C_{NP} \cap \overline{C_A}|H_1)$$

If lemma true:

$$P(C_{NP} \cap \overline{C_A}|H_1) \geq \frac{1}{k_\alpha} P(C_{NP} \cap \overline{C_A}|H_0) = \frac{1}{k_\alpha} P(C_A \cap \overline{C_{NP}}|H_0) \geq P(C_A \cap \overline{C_{NP}}|H_1)$$

Thus $P(C_A|H_1) \leq P(C_{NP}|H_1)$

Neyman-Pearson lemma

$$\boxed{\frac{f(t|H_0)}{f(t|H_1)} \leq k_\alpha}$$

□ Neyman-Pearson lemma allows reformulation of multidimensional problems in simpler terms

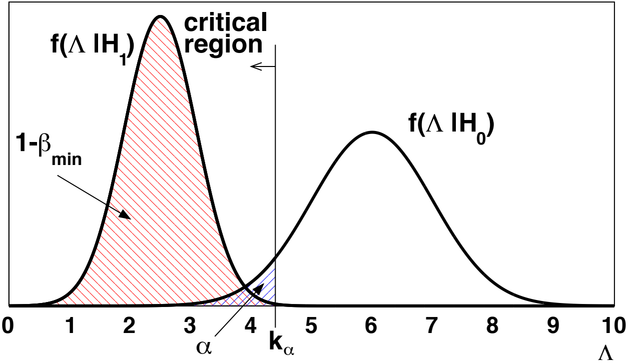
$$\begin{aligned} ND \quad &\longrightarrow \quad 1D \\ t = (t_1, t_2, \ldots) \quad &\longrightarrow \quad \Lambda = \frac{f(t|H_0)}{f(t|H_1)} \end{aligned}$$

□ $1D$ machinery described previously can be employed using Λ as test statistics

→ Automatically leads to "optimal" hypothesis tests

→ Λ is the most discriminating variable

Neyman-Pearson lemma



Exercice

Let $(X_1, X_2, \ldots, X_n)$ be a set of n i.i.d. variables distributed according to a gaussian distribution with mean equal to μ and variance equal to 1. We consider the two following hypotheses:

- $H_0 : \mu = \mu_0$
- $H_1 : \mu = \mu_1 > \mu_0$

Show that the BCR region is given by

$$\bar{x} \geq \frac{\mu_0 + \mu_1}{2} + \frac{1}{n} \frac{\ln k_\alpha}{\mu_0 - \mu_1}$$

where $\bar{x}$ is the sample mean and k_α a constant depending only of the size of the test α.

Link between confidence interval building and hypothesis testing

□ Confidence interval building can be done using hypothesis testing language

□ **Example**: CI for mean of normal distribution with known variance

→ Sample: $(x_1, \dots, x_n)$

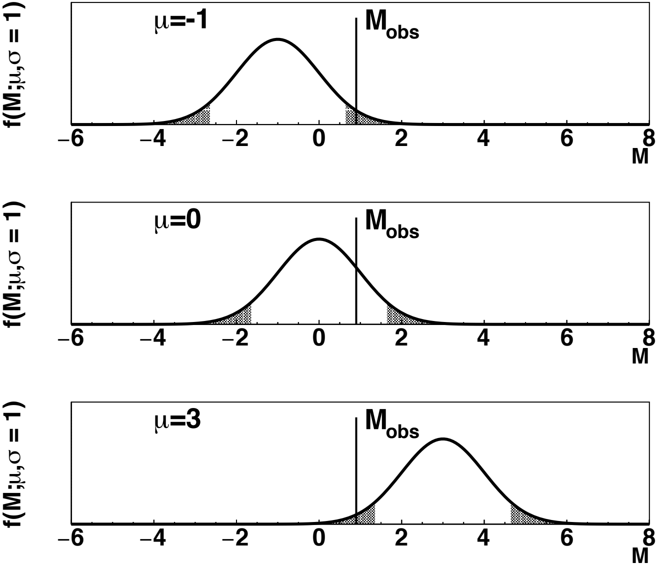
→ Test statistic: sample mean $M = \frac{1}{n} \sum x_i$

→ $M \sim \mathcal{N}(\mu, \sigma/\sqrt{n})$

→ Perform hypothesis test for each value of μ and, based on the observed value of M, accept or reject these values

CI made of all values of μ not rejected in hypothesis test

Link between confidence interval building and hypothesis testing



Link between confidence interval building and hypothesis testing

- □ This way of building confidence intervals sometimes called **"hypothesis test inversion"**
- □ It is strictly equivalent to Neyman construction with

$$\boxed{\text{confidence level} = 1 - \text{ size of test}}$$

→ Hyp. test inversion thus leads to good coverage properties

- □ Note that notations can be confusing
  - In confidence interval chapter: α=confidence level
  - In this chapter: α=size of test