DATA ANALYSIS

Toni Šćulac
Faculty of Science, University of Split, Croatia
visiting professor at University of Latvia, Latvia

CERN School of Computing 2023, Tartu, Estonia

LECTURES OUTLINE

1) Introduction to Data Analysis
2) Probability density functions and Monte Carlo methods
3) Parameter estimation and Confidence intervals
4) Hypothesis testing and p-value

HYPOTHESIS TESTING AND P-VALUE

GENERAL PICTURE REMINDER

1

Physical phenomena

Described by a theory

Described by PDFs,
depending on unknown parameters
with true values
$\theta^{true}=(m_H^{true},\Gamma_H^{true},\ldots,\sigma^{true})$

2

Sampling reality

EXPERIMENT

3

Data sample

$x = (x_1,x_2,\ldots,x_N)$

x is a multivariate random variable

4

DATA

ANALYSIS

5

Results

- **parameter estimates**
- **confidence limits**
- **hypothesis tests**

BONUS PROBLEM - 4

Some rules to follow:

1. In every lecture there will be one bonus problem presented
2. If you have good knowledge in stats and everything I am presenting is known to you feel free to start working on the problem now!
3. Otherwise, work on the problem after the lectures.
4. Solutions won't be provided, you have to come and talk to me to check if your answer is correct or if you need hints!
5. Google/AI assistance is not allowed. These are problems that I want you to think about on your own

Determine the 90% confidence interval for your b-tagging efficiency if you tag as such 4 b-jets out of 8.

Do even better and draw the Neyman confidence belt for any possible outcome when trying to tag 8 b-jets.

INTRODUCTION TO HYPOTHESIS TESTING

- A key task in most of physics measurements is to discriminate between two or more hypotheses on the basis of the observed experimental data.
  - a new particle called the Higgs boson exists?
  - students cheated on the exam?
- This problem in statistics is known as **hypothesis test**, and methods have been developed to assign an observation considering the predicted probability distributions of the observed quantities under the different possible assumptions.
- A **hypothesis H** specifies the probability for the data, i.e., the outcome of the observation, here symbolically: x
- The probability for x given H is also called the **likelihood of the hypothesis**, written L(x|H).

TEST DEFINITION

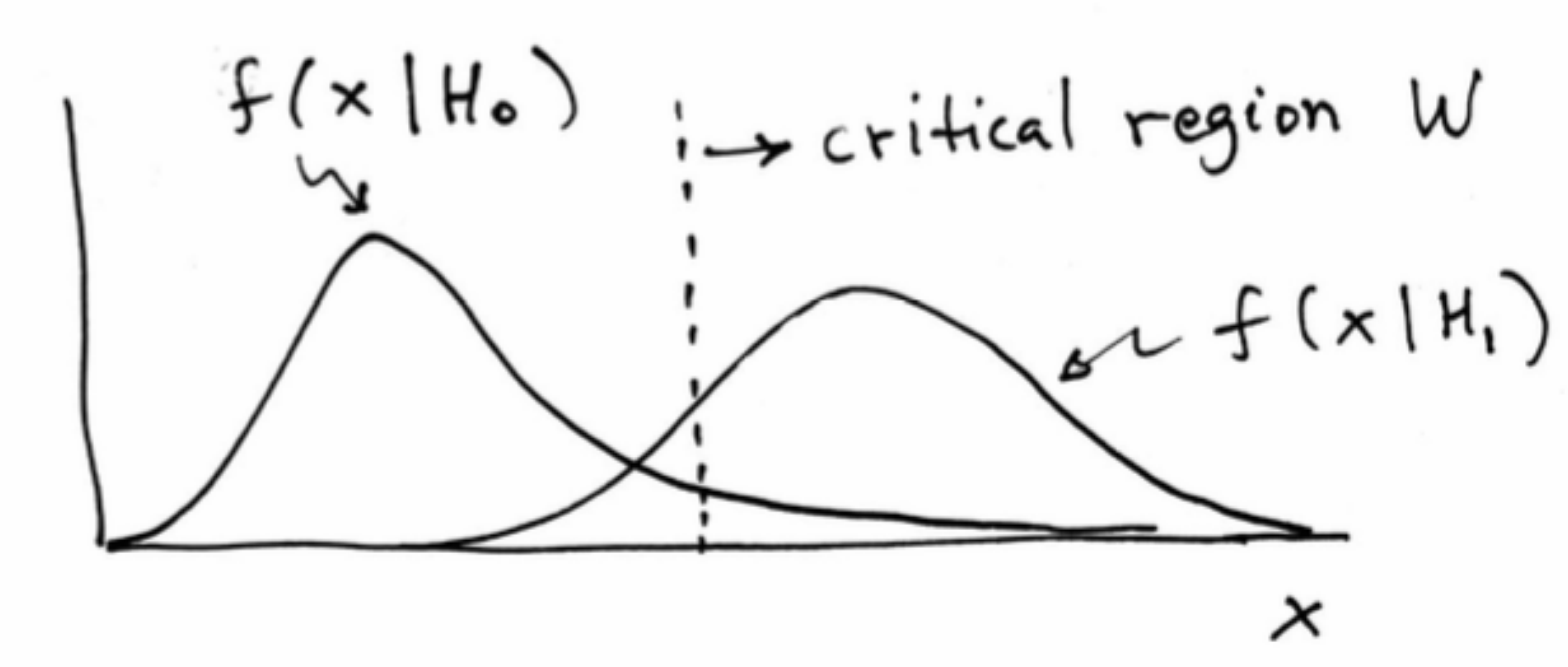
- Goal is to make some statement based on the observed data x as to the validity of the possible hypotheses.
- Consider e.g. a simple hypothesis H_0 and alternative H_1
  - In statistical literature when two hypotheses are present, these are called **null hypothesis** (H_0) and **alternative hypothesis** (H_1)
- A **test** of H_0 is defined by specifying a critical region W of the data space such that there is no more than some (small) probability α, assuming H_0 is correct, to observe the data there, i.e.,

$$P(x \in W | H_0) \leq \alpha$$

  - If x is observed in the critical region, reject H_0.
- α is called the size or **significance level** of the test
- Critical region is also called "rejection" region; complement is acceptance region.

TEST DEFINITION

- In general there are an infinite number of possible critical regions that give the same significance level α
- The choice of the critical region for a test of H_0 needs to take into account the alternative hypothesis H_1
  - Roughly speaking, place the critical region where there is a low probability to be found if H0 is true, but high if H1 is true



COURTROOM TRIAL

- Consider a criminal trial
- There is one simple rule: a defendant is considered **not guilty** as long as his **guilt is not proven**
- The prosecutor tries to prove the guilt of the defendant
  - Only when there is enough **evidence** the defendant is **convicted**
- We start with two hypotheses:
  - H_0: the defendant is not guilty (NULL HYPOTHESIS)
  - H_1: the defendant is guilty (ALTERNATIVE HYPOTHESIS)
- Null hypothesis is considered accepted for time being
- Common sense: the hypothesis of innocence is rejected only if the error is very unlikely
  - We don’t want to convict an innocent person!
  - This is called **Error of the first kind** and we want it to be small
- **Error of the second kind**: liberating someone who indeed committed the crime
  - This one can be large, but we also want it to be small

TYPE-I, TYPE-II ERRORS

- Rejecting the hypothesis H_0 when it is true is a **Type-I error**. The maximum probability for this is the size of the test:

$$P(x \in W | H_0) \leq \alpha$$

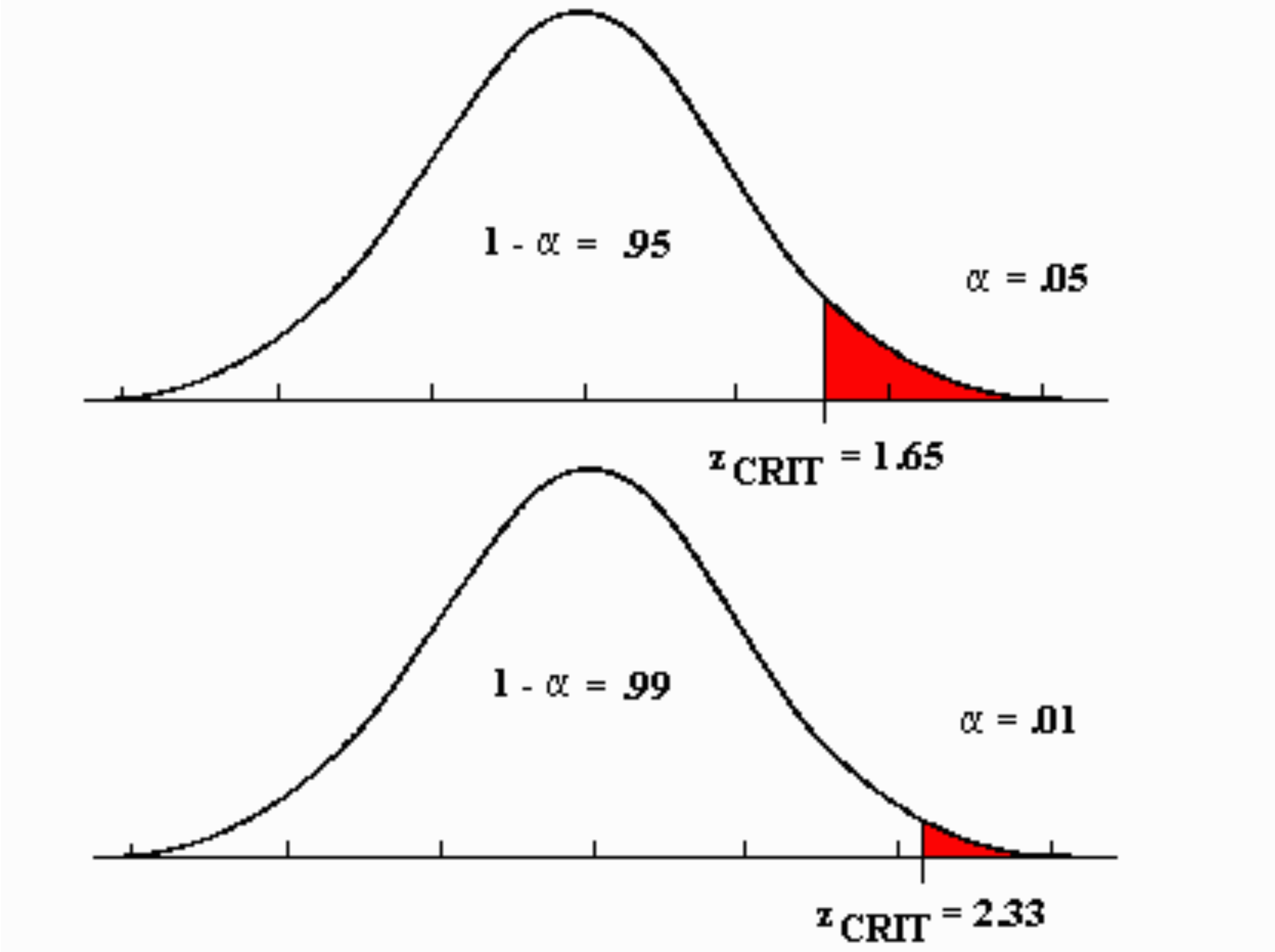
- But we might also accept H_0 when it is false, and an alternative H_1 is true.
- This is called a **Type-II error**, and occurs with probability

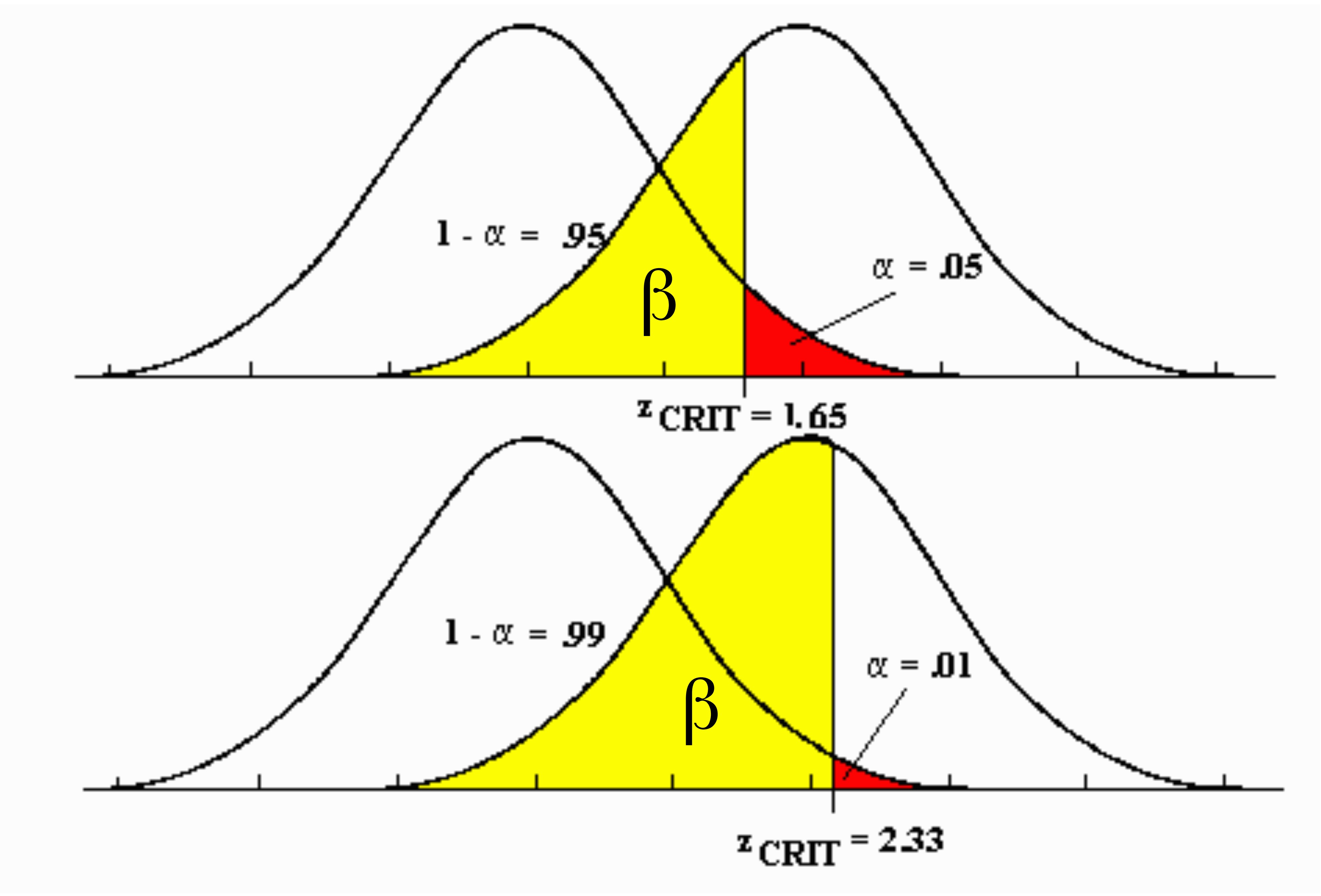
$$P(x \in S - W | H_1) = \beta$$

- 1-β this is called the power of the test with respect to the alternative H_1

ERROR SUMMARY

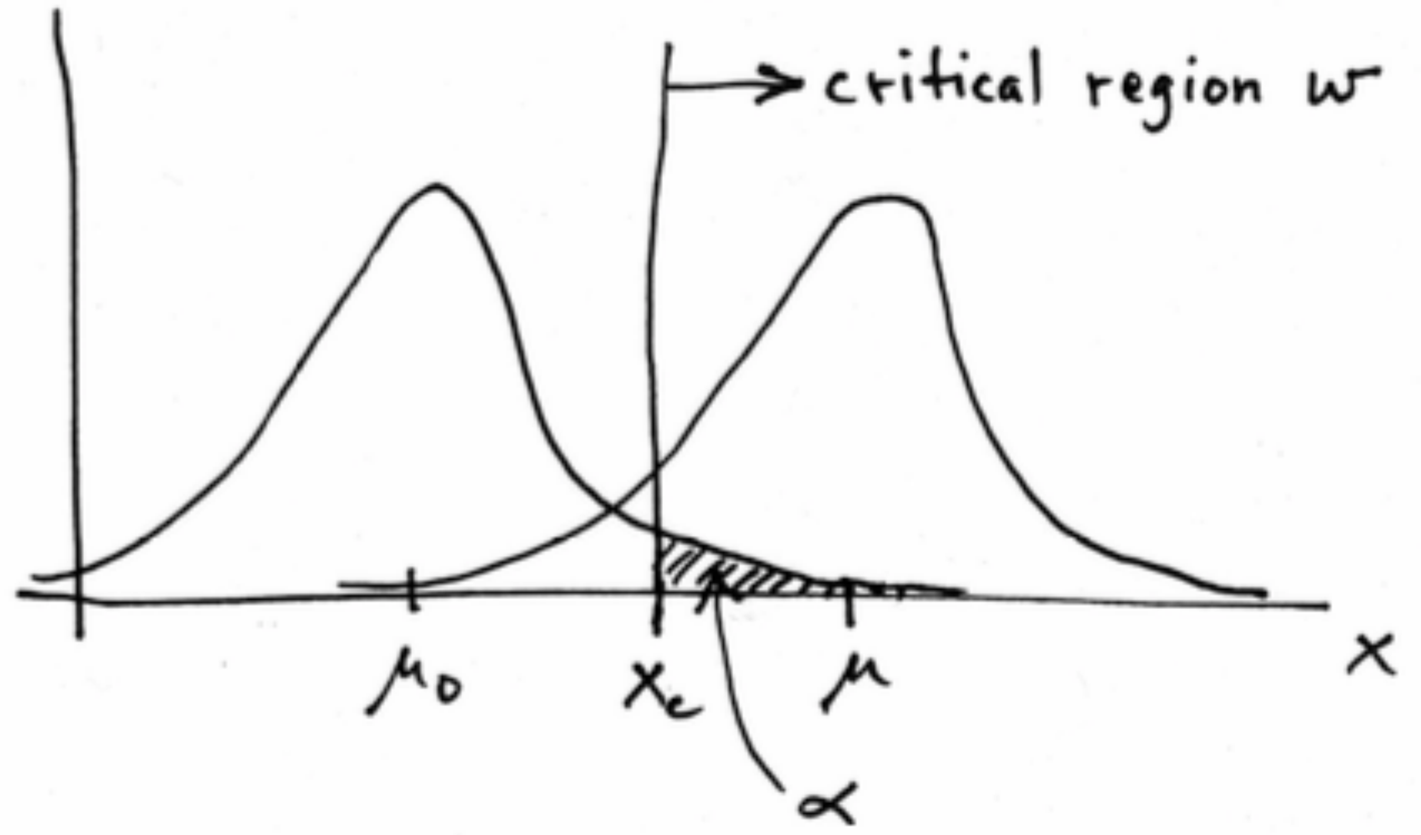
		True state	
		H_0 is true (he is not guilty)	**H_1 is true** (he is guilty)
Decision	**Accept H_0** (acquittal)	Right decision *Probability = 1-α* *(significance level)*	Wrong decison Type II error Probability = β
	Reject H_0 (conviction)	Wrong decision *Type I error* *Probability = α*	Right decision *Probability = 1-β* (power)





CHOOSING A CRITICAL REGION

- To construct a test of a hypothesis H_0, we can ask what are the relevant alternatives for which one would like to have a high power
  - Maximise power wrt H_1 = maximise the probability to reject H_0 if H_1 is true.
- Often such a test has a high power not only with respect to a specific point alternative but for a class of alternatives.
- For example, using a measurement $x \sim$ Gauss (μ, σ) we may test
  - $H_0 : \mu = \mu_0$ versus the composite alternative $H_1 : \mu > \mu_0$
- We get the highest power with respect to any $\mu > \mu_0$ by taking the critical region $x \geq x_c$ where the cut-off xc is determined by the significance level such that $\alpha = P(x \geq x_c \mid \mu_0)$

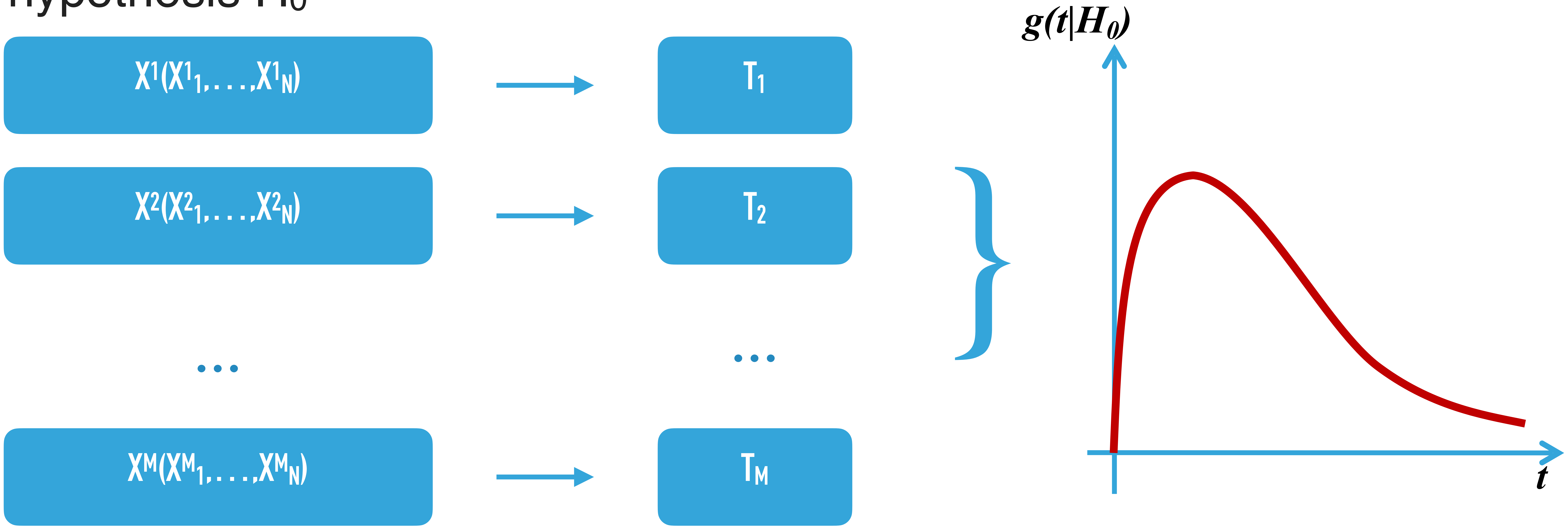


HYPOTHESIS TESTING PROCEDURE

1. CHOOSE NULL (MANDATORY) AND ALTERNATIVE HYPOTHESIS (OPTIONAL)
2. FIND RELEVANT TEST STATISTIC T
3. DERIVE THE DISTRIBUTION OF THE TEST STATISTIC T FOR DIFFERENT HYPOTHESIS
4. DEFINE CONFIDENCE LEVEL AND CRITICAL REGION. DIVIDE THE DISTRIBUTION IN TWO REGIONS
5. COMPUTE THE OBSERVED VALUE OF TEST STATISTIC
6. REJECT OR ACCEPT HYPOTHESIS BASED ON THE OBSERVED TEST STATISTIC

TEST STATISTIC PDF

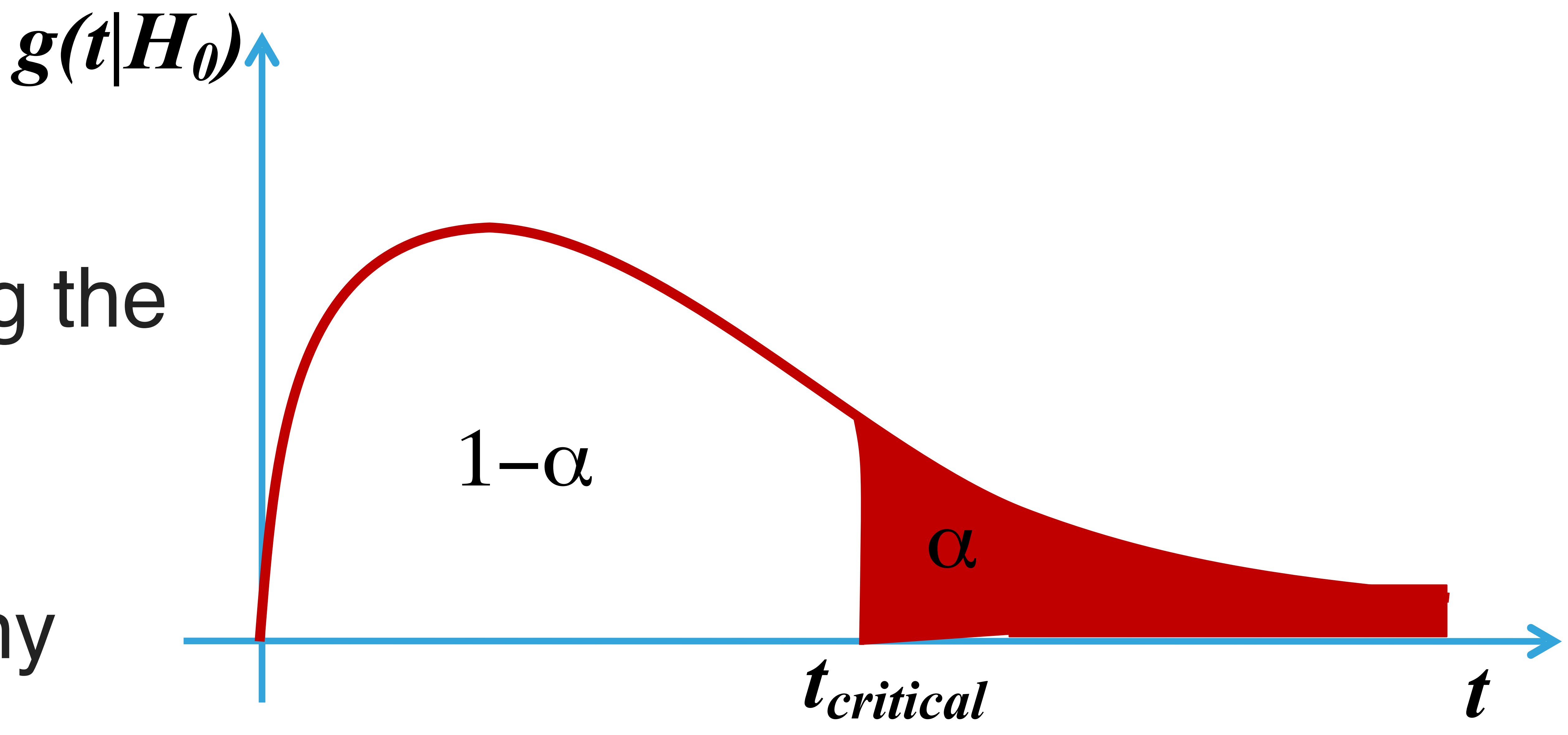
- Using input data define a single **test statistic** $t(x_1,\ldots,x_N)$ whose value reflects the agreement between data and the hypothesis
- Using Monte Carlo simulate many (M) experiments trying to test the null hypothesis H_0



- Obtain a probability density function (PDF) of the test statistic t, given null Hypothesis (H_0) is true, $g(t\,|\,H_0)$

OBSERVED TEST STATISTIC

- Now we have to divide the distribution in two regions:
  - where H_0 is rejected with CL α
  - where H_0 is not rejected with CL 1-α
- $t_{critical}$ is the value of test statistic diving the two regions
- We talk only about rejecting the null hypothesis H_0, not about accepting any other hypothesis



- **We should decide about two regions before looking at the observed value of the test statistics**
- Now we can calculate the observed test statistic t_{obs} and decide:
  - If $t_{obs} > t_{critical}$: reject H_0
  - If $t_{obs} < t_{critical}$: do not reject H_0

WILKS' THEOREM

- When a large number of measurements is available, Wilks' theorem allows to find an approximate asymptotic expression for a test statistic based on a likelihood ratio inspired by the Neyman–Pearson lemma
- **Wilks' Theorem**: if the hypothesised parameters $\theta(\theta_1, \ldots, \theta_N)$ are true then in the large sample limit **test statistic** defined as likelihood ratio

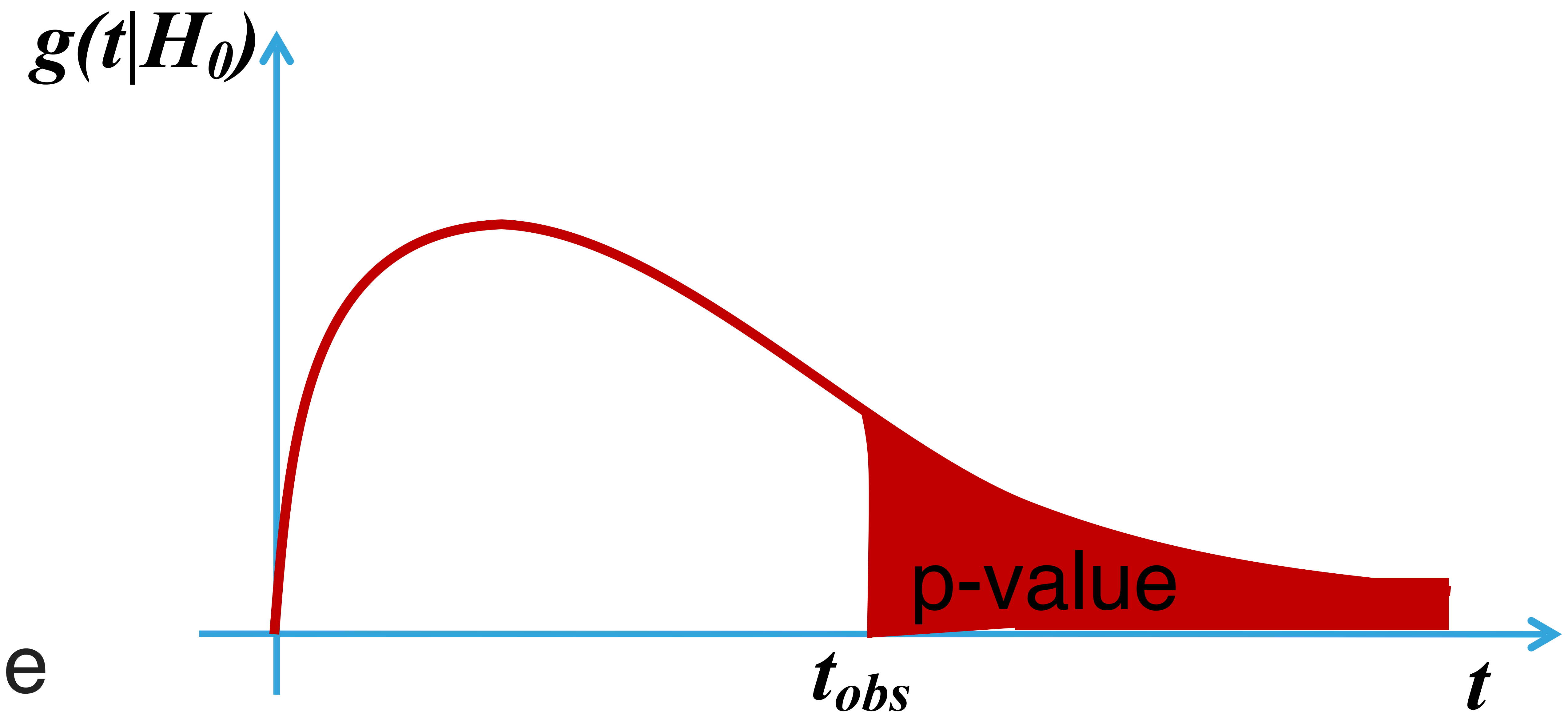
$$\chi^2(\theta) = -2 \ln \frac{L(x; \theta^{true})}{L(x; \hat{\theta})}$$

- **is asymptotically distributed according to the Chi-Square distribution** with k degrees of freedom.

P - VALUE

- Knowing the PDF of our test statistic we can answer one important question:
- What is the probability to obtain the value of t equal or greater than the value t_{obs} we observed?

$$P(t \geq t_{obs}) = \int_{t_{obs}}^{\infty} g(t\,|\,H_0)\,dt$$



- This probability is the so-called p-value
- p-value is defined as the probability to find t in the region of equal and lesser compatibility with H_0 than the level of compatibility observed with actual data

SIGNIFICANCE

- For easier understanding p-values can be converted to **significance**

One tailed p-value	Significance	Gaussian area $\pm n\sigma$	Probability of outcome: 1 in
0.159	1	0.68268949	6.3
0.023	2	0.95449974	44
0.00135	3	0.99730020	740
$3.17 \cdot 10^{-5}$	4	0.99993666	31,574
$2.87 \cdot 10^{-7}$	5	0.99999943	3,488,556

- For example: if you were to measure something with 5σ significance that means that either the null hypothesis is wrong (highly likely) or that due to statistical fluctuations your data sample corresponds to one in 3.5 million and the null hypothesis is correct (possible but extremely unlikely)

SCIENTIFIC DISCOVERY

- Claiming discovery is a serious issue
  - It should stay with us for a long long time (if not forever!)
- So, when do we claim a discovery?
  - When we are sure.
  - But we are never sure!
  - That's right, but we can be pretty sure!
  - 'Pretty' is not a scientific term!?
  - That's right, therefore we developed some kind of a convention in HEP
- Make a hypothesis that the result you obtain is due to the fluctuation of the background (i.e. already know theory)
  - Calculate a probability for that hypothesis
- Reject the hypothesis if that probability is smaller than 0.000000287 (significance $> 5\sigma$)
  - In most other sciences p-value smaller than 0.05 used to reject the null hypothesis!

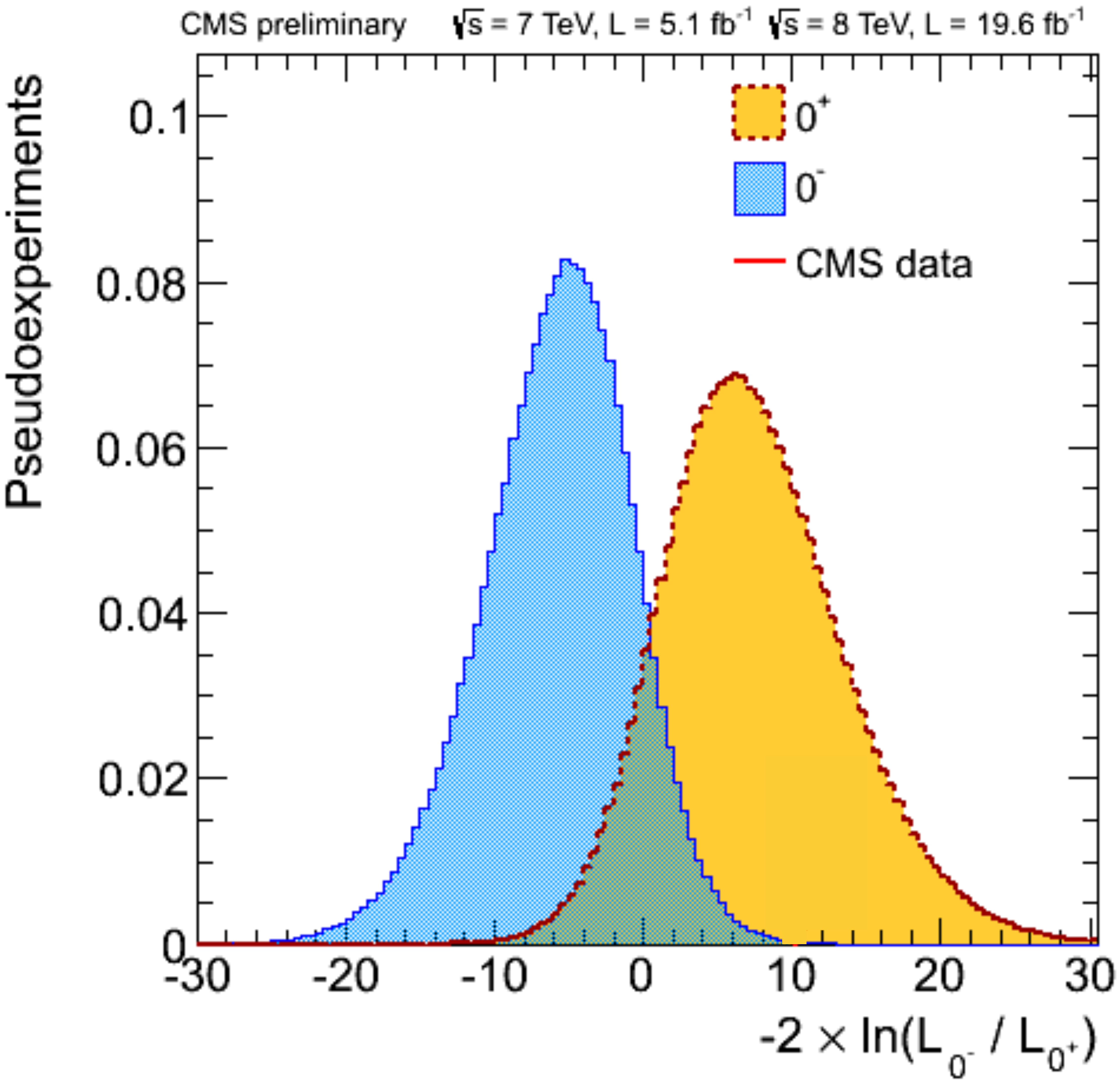
ACCEPTING OR REJECTING THEORIES

- Image we make an experiment and obtain data
  - Theory 1 agrees with data
  - Theory 2 agrees with data
  - Theory N agrees with data
- Then the statement that “Theory 1 is acceptable” is not so strong
  - Not wrong neither
- But imagine this scenario
  - Theory 1 gives precise prediction
  - Experiment doesn’t quite agree with that prediction
  - Than the statement “Theory 1 is not acceptable” is rather strong
  - Therefore we better reject than accept theories

ALTERNATIVE HYPOTHESIS TESTS

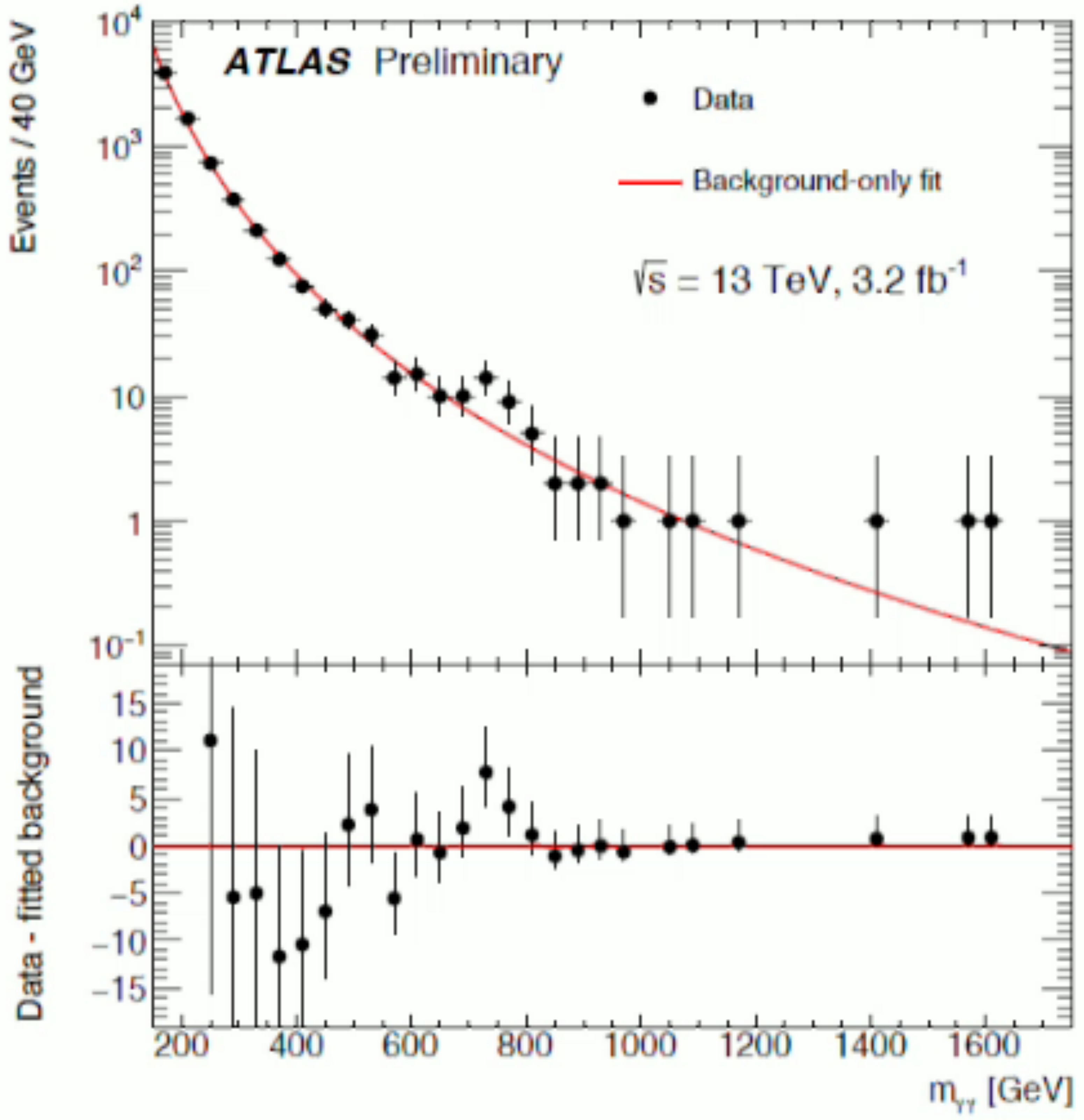
- So far we have only considered null hypothesis
- It is possible to perform hypothesis testing to see if the data favours null or alternative hypothesis
- Build test statistic PDFs for both and measure the observed test statistic
- Use it to exclude alternative hypothesis at a given CL

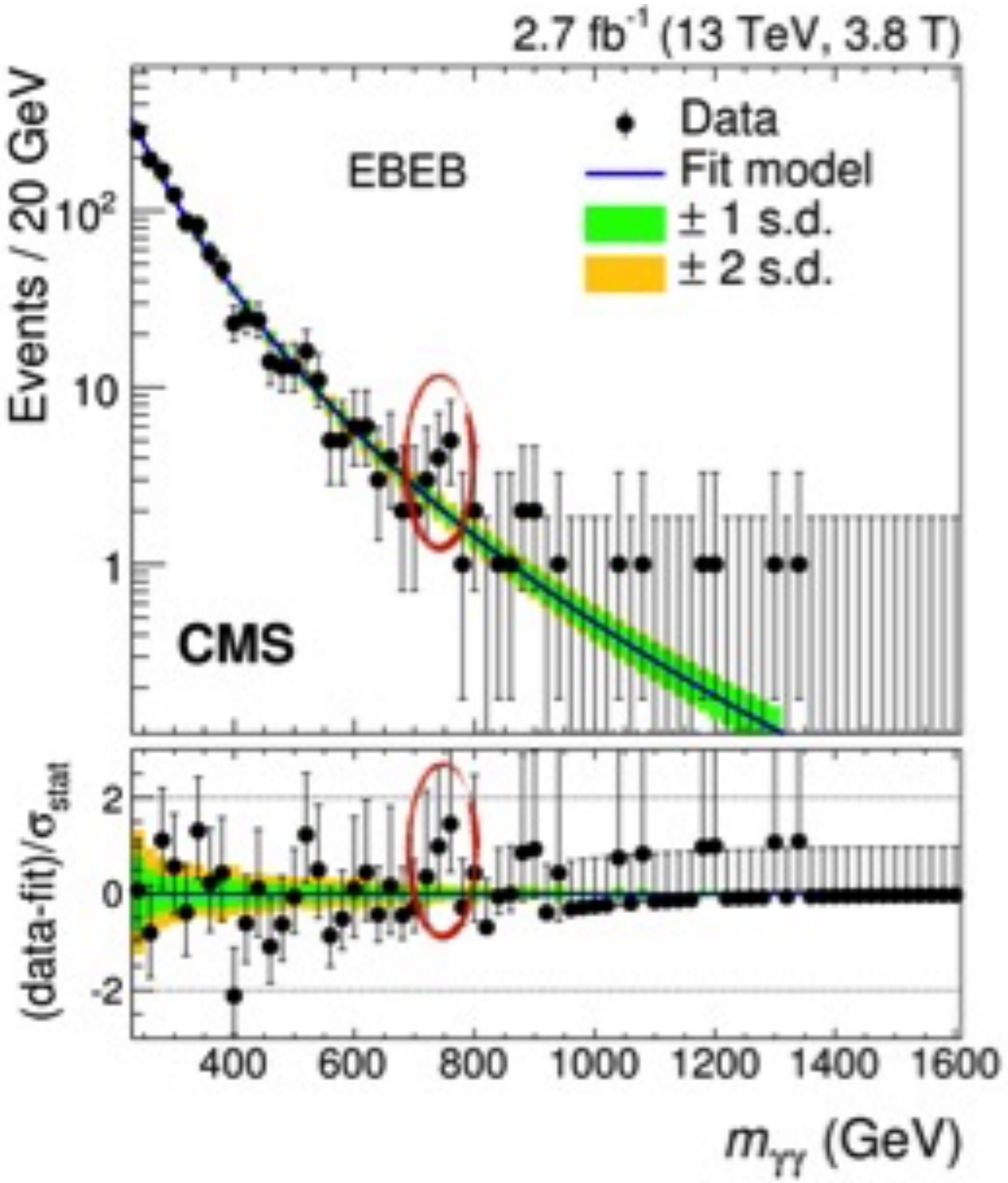
$$\frac{P(t \geq t_{obs} \mid H_1)}{P(t \leq t_{obs} \mid H_0)} < 1 - CL$$



WHY 5 SIGMA?

- In data collected at the Large Hadron Collider (LHC) in 2015 an indication of a new particle or resonance was present
- The statistical significance of the deviation was reported to be 3.9 and 3.4 standard deviations (locally) respectively for each experiment





WHY 5 SIGMA?

- In the interval between the December 2015 and August 2016 results, the anomaly generated considerable interest in the scientific community, including about 500 theoretical studies.
- The anomaly was absent in data collected in 2016, suggesting that the diphoton excess was a statistical fluctuation.
    - The data, however, were always less than five standard deviations