USTH MASTER SPACE

Earth Observation - Astrophysics - Satellite Technologies

USTH

VIETNAM FRANCE UNIVERSITY

Lecture 3: Hypothesis testing, and model fitting

Hoàng Đức Thường

Department of Space and Applications (DSA), USTH

Hanoi, 2021

The process of science

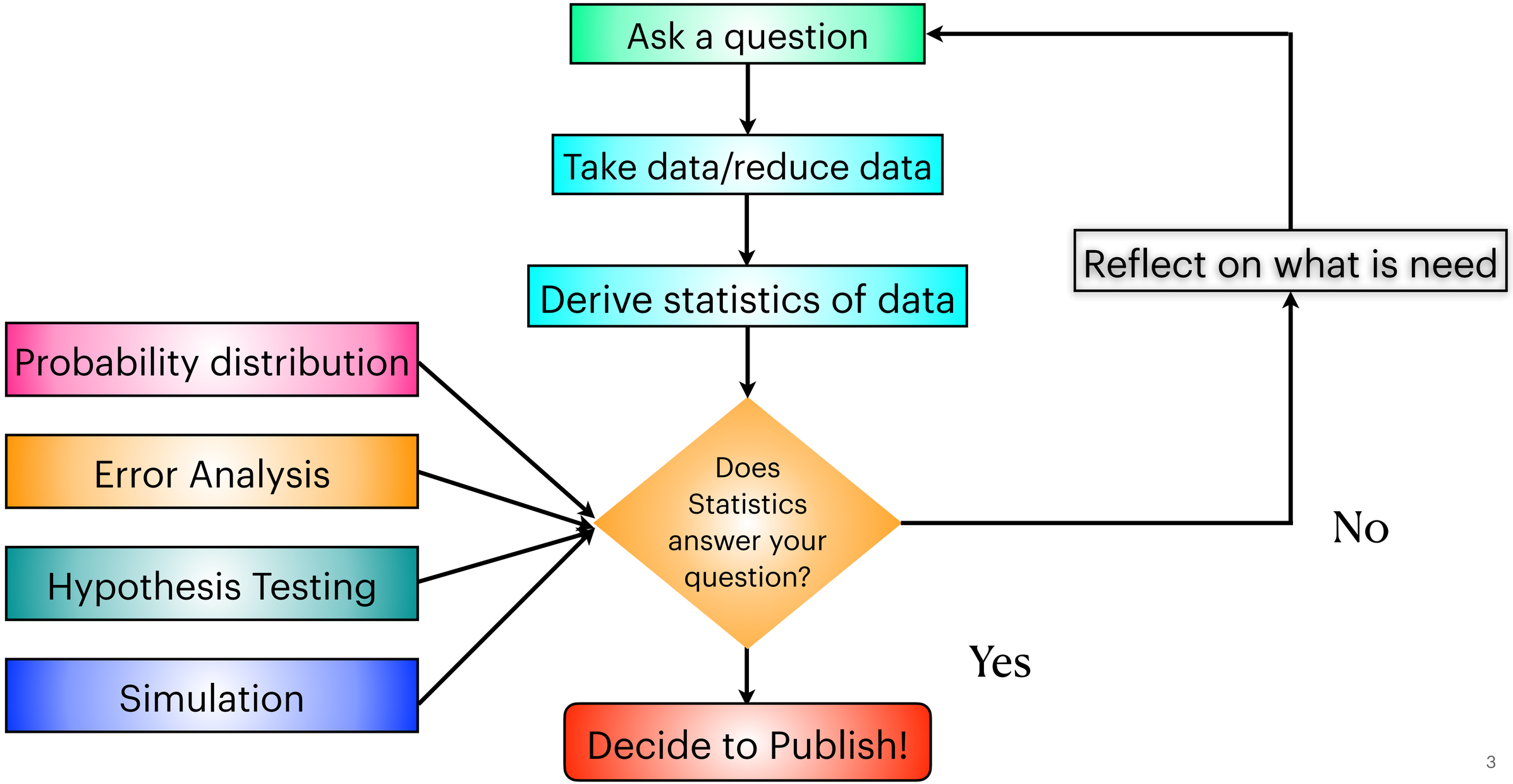
Observation/
Measurements

Analyze data

Conclusion
(Test hypothesis,
probabilities)

Scientific questions

The process of decision making



Hypothesis testing, and model fitting

In this class we will study the use of the χ^2 statistic as a (1)hypothesis test of a model and (2) estimate the best fit parameter estimation.

Hypothesis testing, and model fitting

At the end of this class, you should be able to do:

- Apply the χ^2 as a hypothesis test, the goodness-of-fit of the data to the model.
- Apply the χ^2 statistic in parameter fitting.
- Determine parameter errors.

Hypothesis testing, and model fitting

It is often the case that we need to do sample comparison:

We have someone else's data to compare with ours; or someone else's model to compare with our data; or even our data to compare with our model.

We need to make the comparison and to decide something. We are doing hypothesis testing, are our data consistent with a model, with somebody else's data?

Comparing data and models

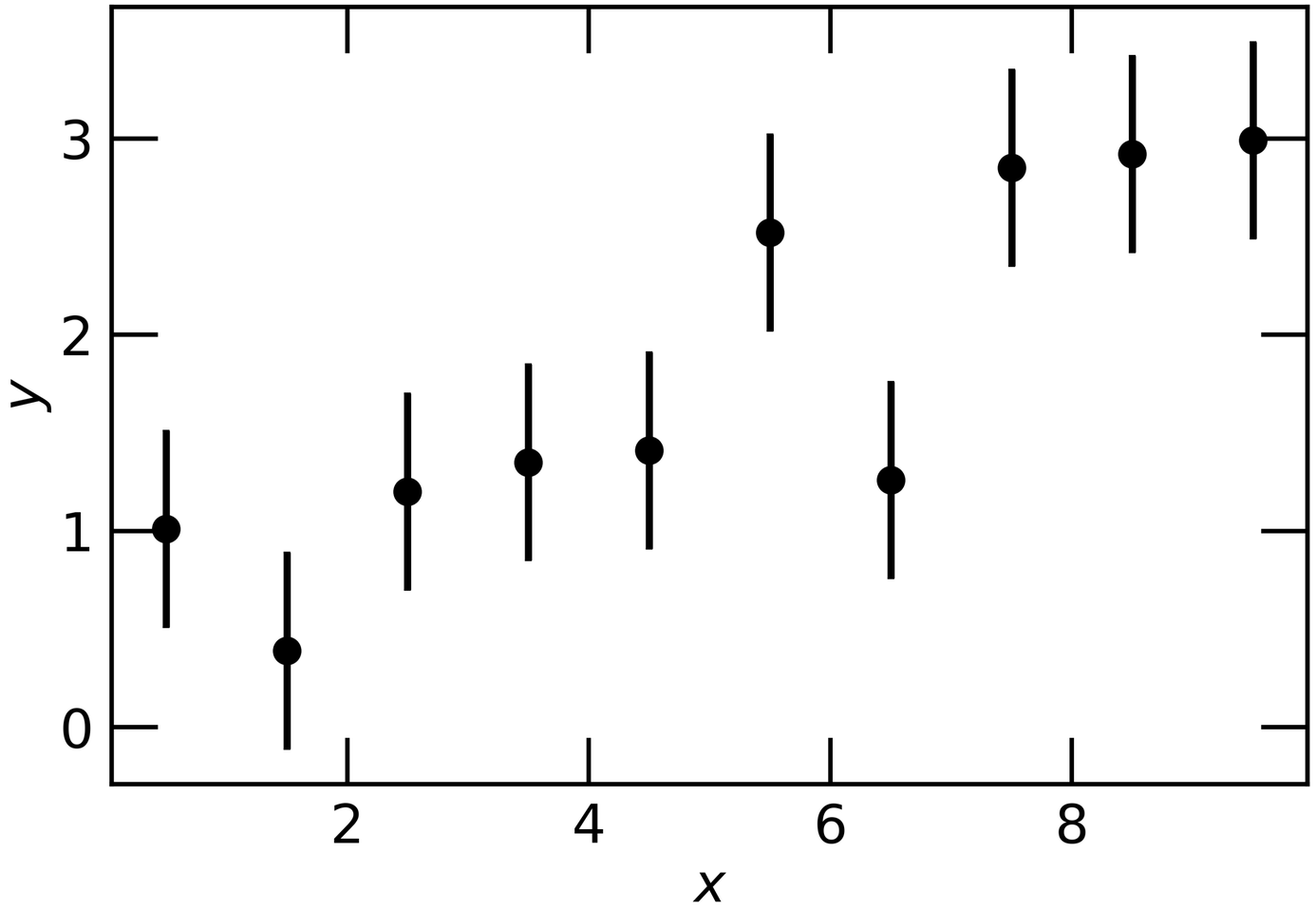
- *When comparing our data to a model, we use to do:*

1. **Hypothesis testing:** we have a set of N measurements with error $x_i \pm \sigma_i$, which a theorist says that it should have values of mean μ_i. How probable is it that these measurements would have been obtained, if the theory is correct?

2. **Parameter estimation:** we have a theoretical model that describes the data $y = ax + b$. What are the best fitting parameters and errors in those parameters?

Comparing data and models

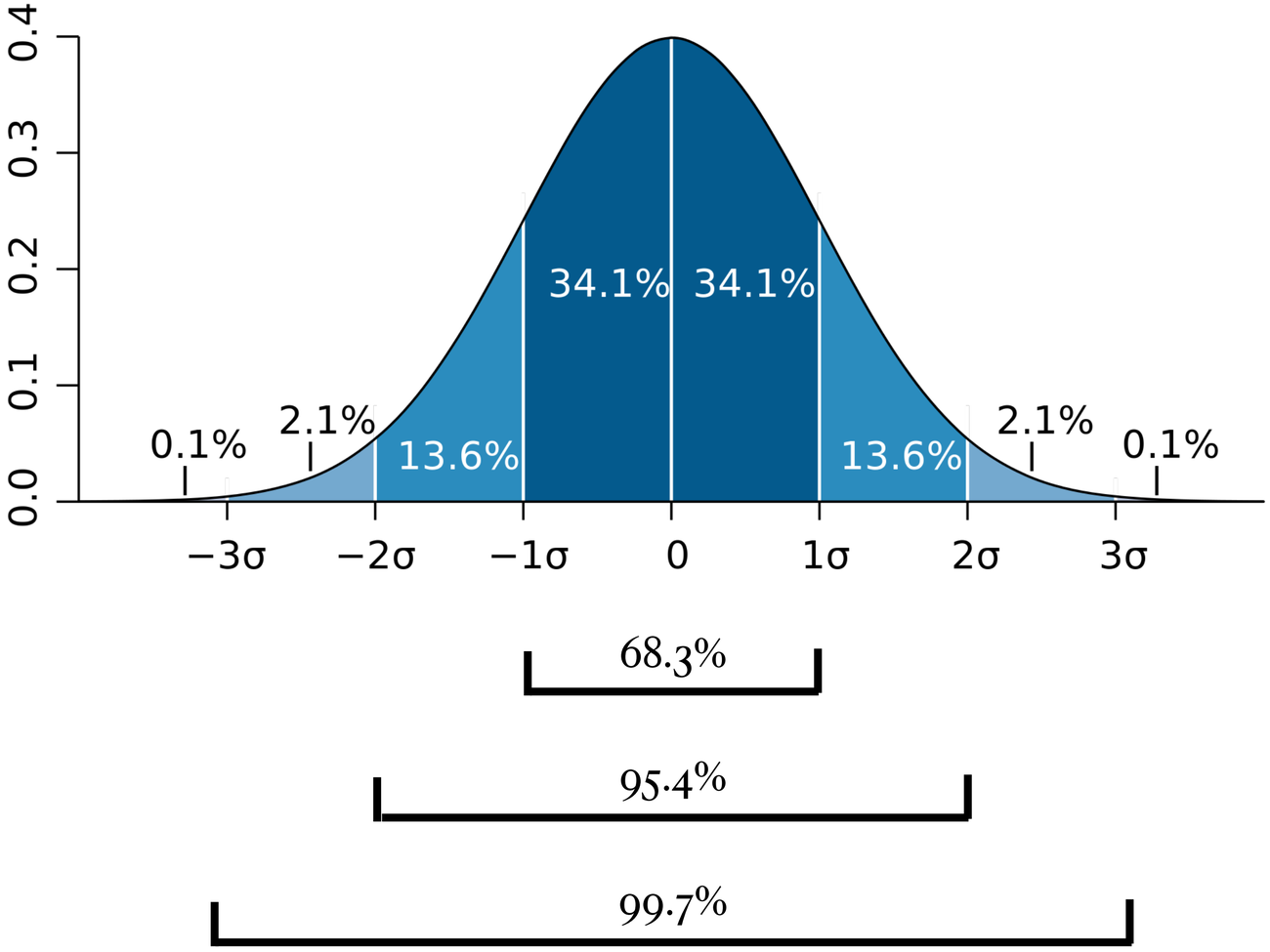
- A key question in statistic is to build theoretical models fit with our experimental data.



- Consider a dataset of N = 10 measurements and its errors.

1. Hypothesis testing

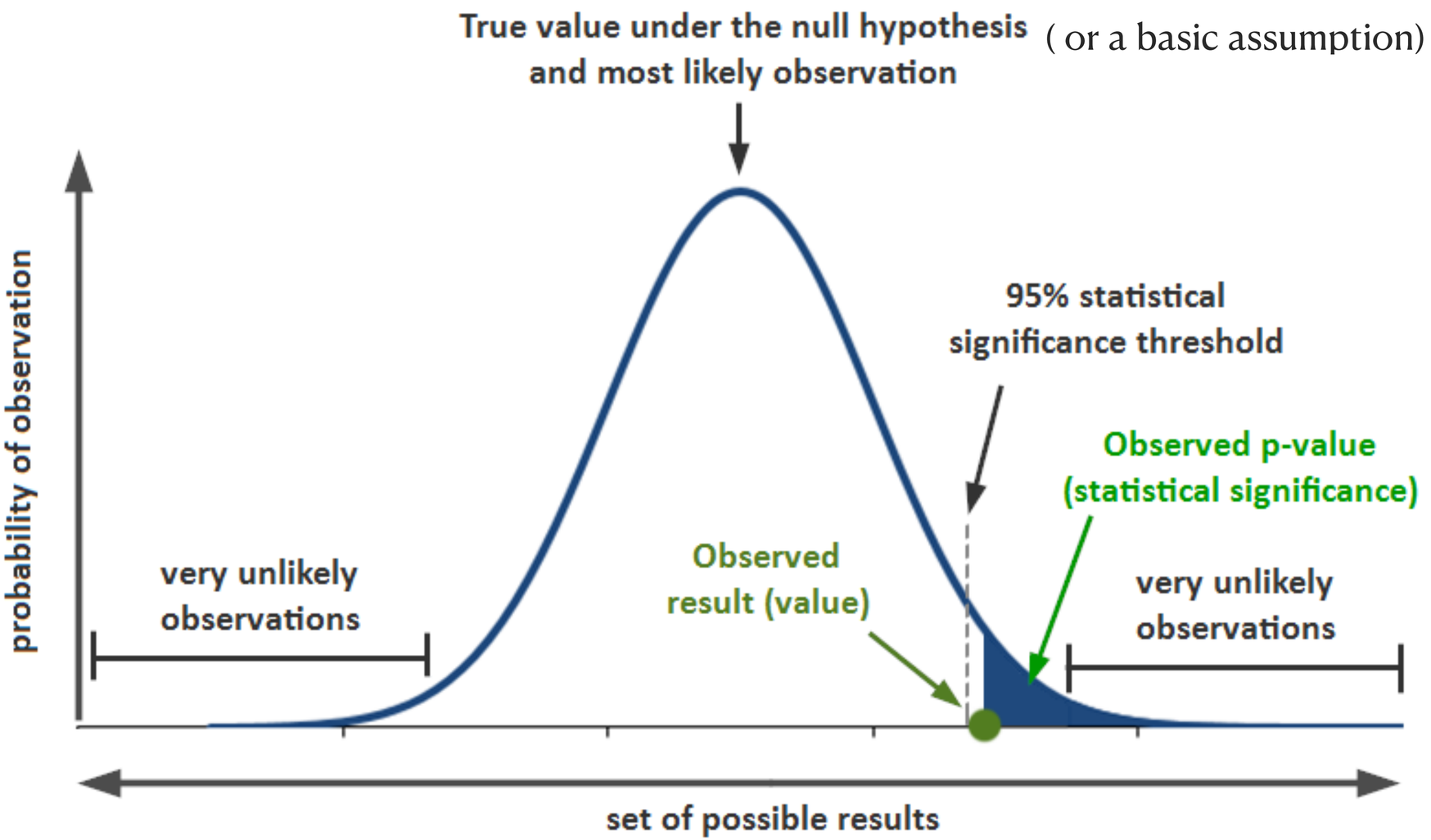
$$N(\mu = 0, \sigma = 1)$$

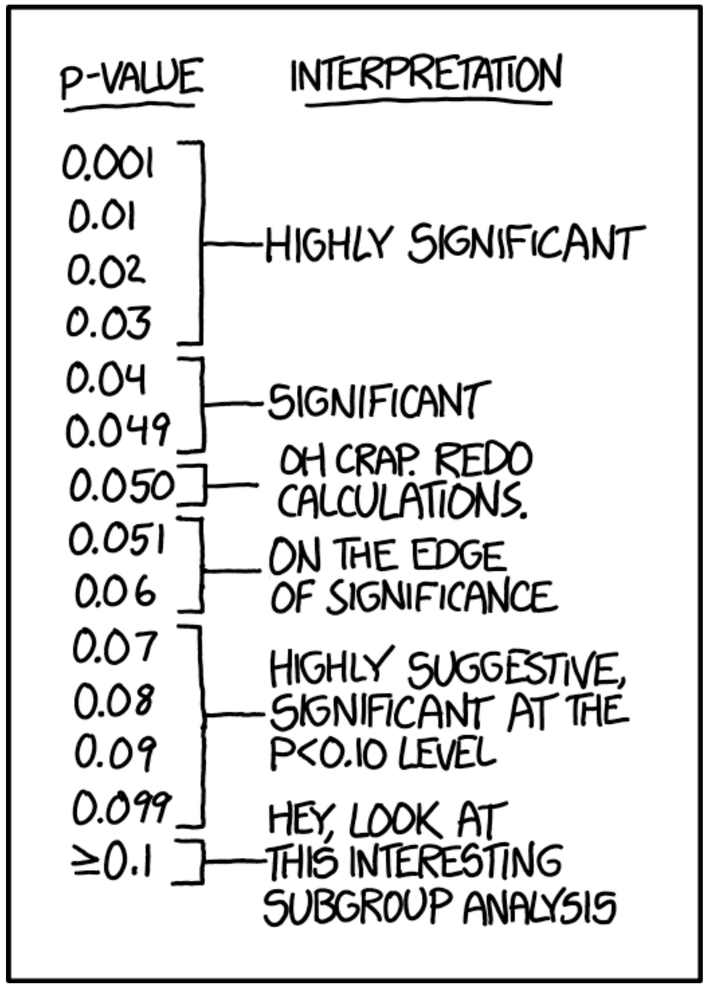


What is hypothesis testing ? why do we use it ?

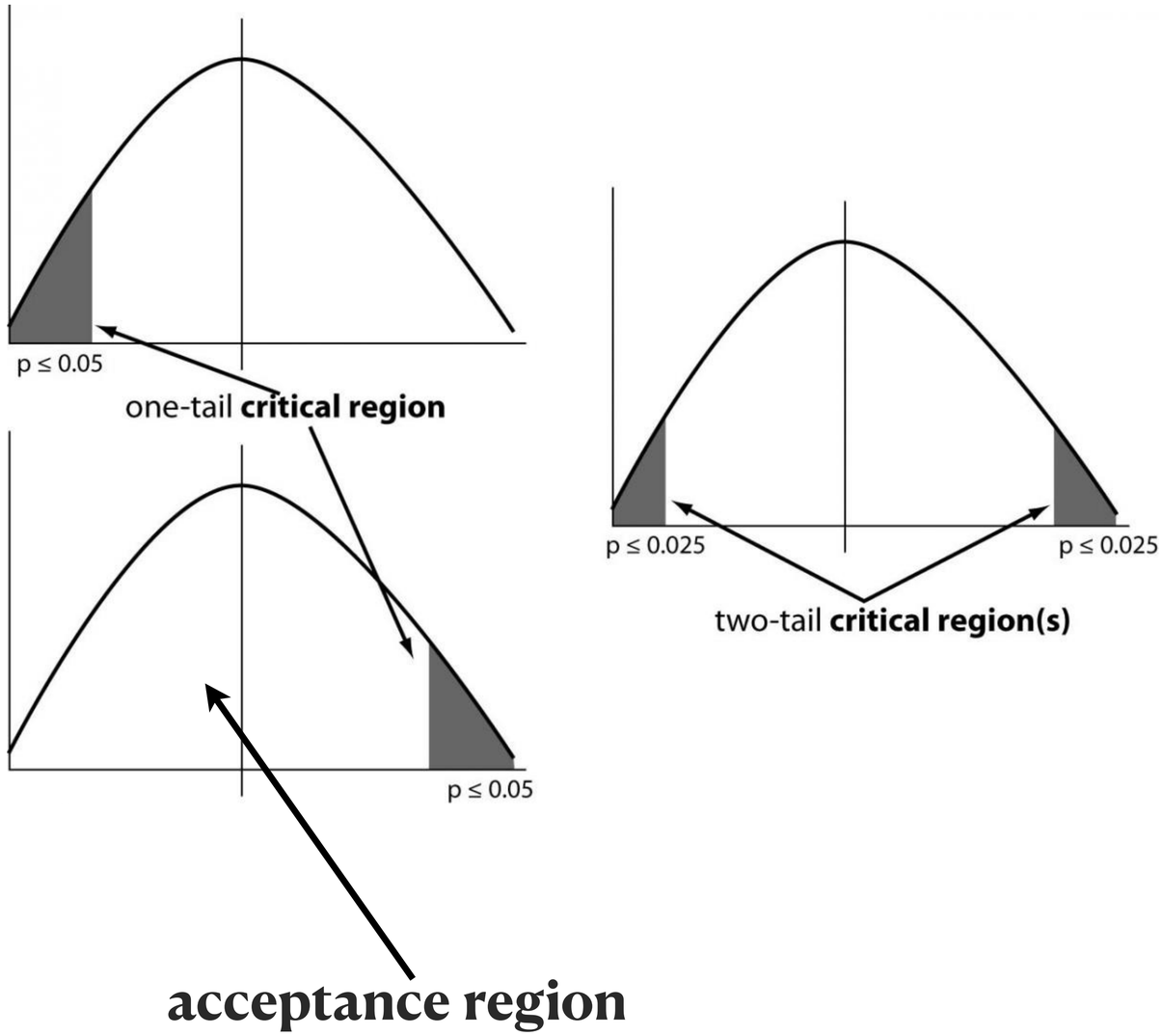
- A statistical method is used in making decision base on experimental data. In simple words: Fit a model to data.
- Determine which statement is the best supported by the sample data. Simply: Choose the best parameters.
- The basic of hypothesis is Normalization or standard normalization.

1. Hypothesis test





1. Hypothesis testing



Which are important parameter of hypothesis testing ?

- The **critical region** is also called **alpha region.** This is Type I error.
- The **acceptance region** is called **beta region.** This is Type II error.
- **p-value:** or calculated probability, is the probability value.
- The degrees of freedom: to evaluate independence of test.

1. Hypothesis testing

Type of tests:

- *Parametric tests*: Compare the values of parameters
  - Example: Does the mass of the photon = mass of the electron?
- *Non-Parametric Tests*: Compare the shapes of probability distributions
  - Example: The number of photons received at a telescope should follow a Poisson distribution.

1. Hypothesis testing

Some widely used hypothesis tests:

- F-Test: Equality of Variances (The probability of χ^2 distribution)
- Student's t-test: Comparison of Means
- Z-test
- Likelihood ratio method: When no uniformly test exist
- Bayesian methods
- Chi-square χ^2 test

1. Chi-Square (χ^2) test

- Pearson's (1900) paper introduced chi-square as a foundation stone of modern statistical analysis. A comprehensive and readable paper is published by Cochran (1952).
- The chi-square statistic describes the goodness-of-fit if the data compare to model.
- The χ^2 statistic follows the chi-square probability distribution.

Chi-Square (χ^2) test

- The most important statistic to help on the hypothesis testing and parameter estimation is the χ^2 **statistic** between the data $x_i \pm \sigma_i$ and model μ_i

$$\chi^2 = \frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} + \ldots + \frac{(x_N - \mu_N)^2}{\sigma_N^2} = \sum_{i=1}^{N} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2$$

Real data

Model (Expected)

sample index

Error / uncertainty in the individual measurements

Chi-Square (χ^2) test

- **Example:** A satellite has 6 channel frequencies (observed frequency O_i). The expected frequency value (E) for all channels is 100. Calculate the χ^2 value.

Satellite sensor	Frequency (GHz)
1	95
2	72
3	103
4	105
5	97
6	128

. The $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$

$$= \frac{(95-100)^2}{100} + \frac{(72-100)^2}{100} + \frac{(103-100)^2}{100} + \frac{(105-100)^2}{100} + \frac{(97-100)^2}{100} + \frac{(128-100)^2}{100}$$

$$= 16.36$$

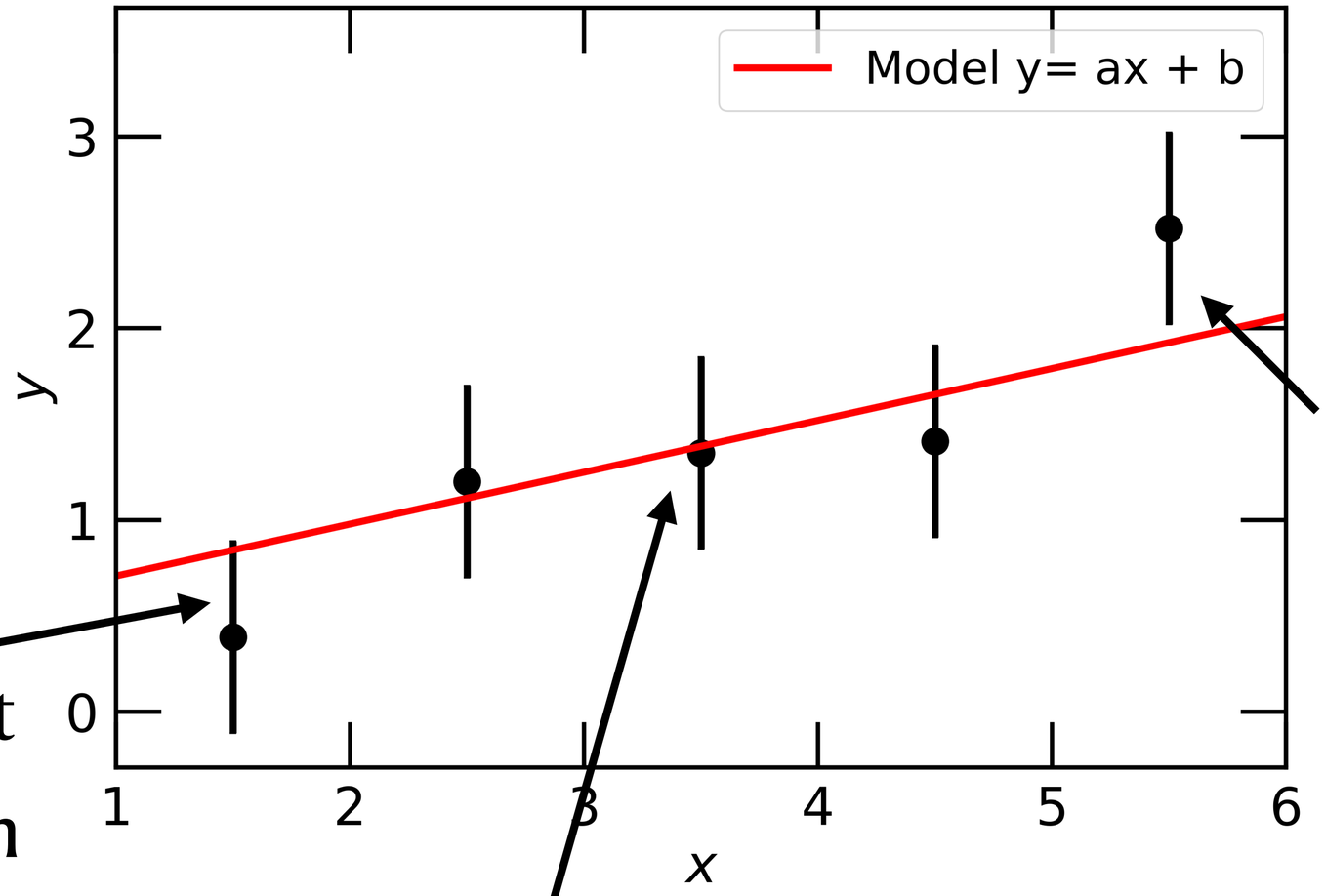
- We do not know the value of $\chi^2 = 16.36$ is significant or not. We need to consider the degrees of freedom. In this example we are simply calculate the different frequency of a single variable, so that:

- The degrees of freedom = number of category - 1 = 6 -1 =5.

- With the degrees of freedom, we can look at the probability distribution to evaluate the value of $\chi^2 = 16.36$.

Chi-Square (χ^2) test

χ^2 is sum over the data points



χ^2 This data point is $1-\sigma$ away from the model, so the $\chi^2 \sim 1.0$

χ^2 This data point is on the model, so the $\chi^2 \sim 0.0$

χ^2 This data point is about $1.5-\sigma$ away from the model, so the $\chi^2 \sim 2.25$

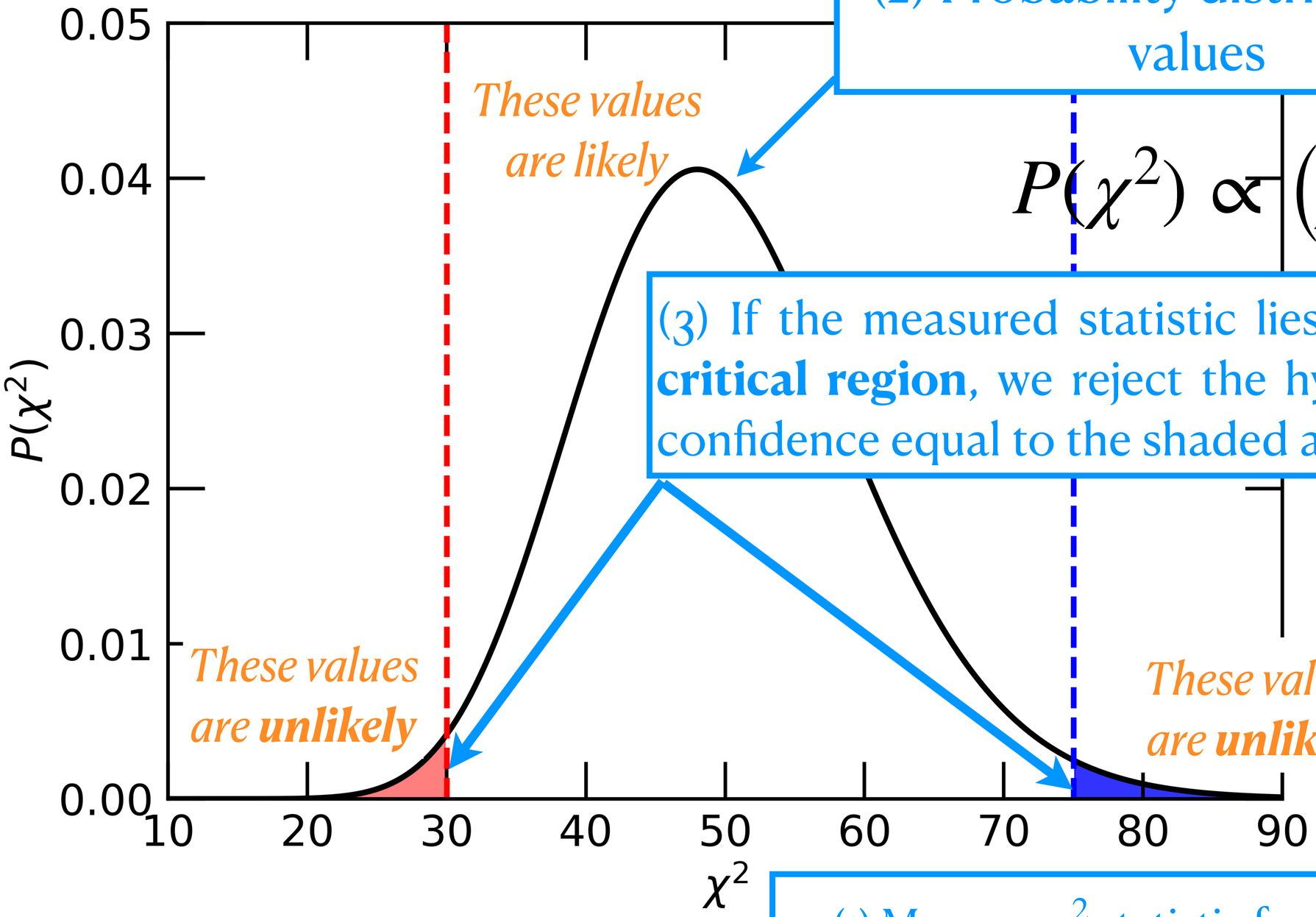
Chi-Square (χ^2) statistic as hypothesis test

- We can use the χ^2 statistic to construct a hypothesis test which estimate the "goodness of fit" between data and model.
- Test statistic with χ^2 value: First, c**alculate the** χ^2 for the dataset.
- Distribution of values: The χ^2 probability distribution
- If the p-value is not low, the data and the model are consistent -> **rule in**
- if the p-value is low, try a new model. -> **rule out**

$$p = \int P(\chi^2) d\chi^2$$

(2) Probability distribution of values

$$P(\chi^2) \propto \left(\chi^2\right)^{\frac{\nu}{2}-1} e^{-\chi^2/2}$$

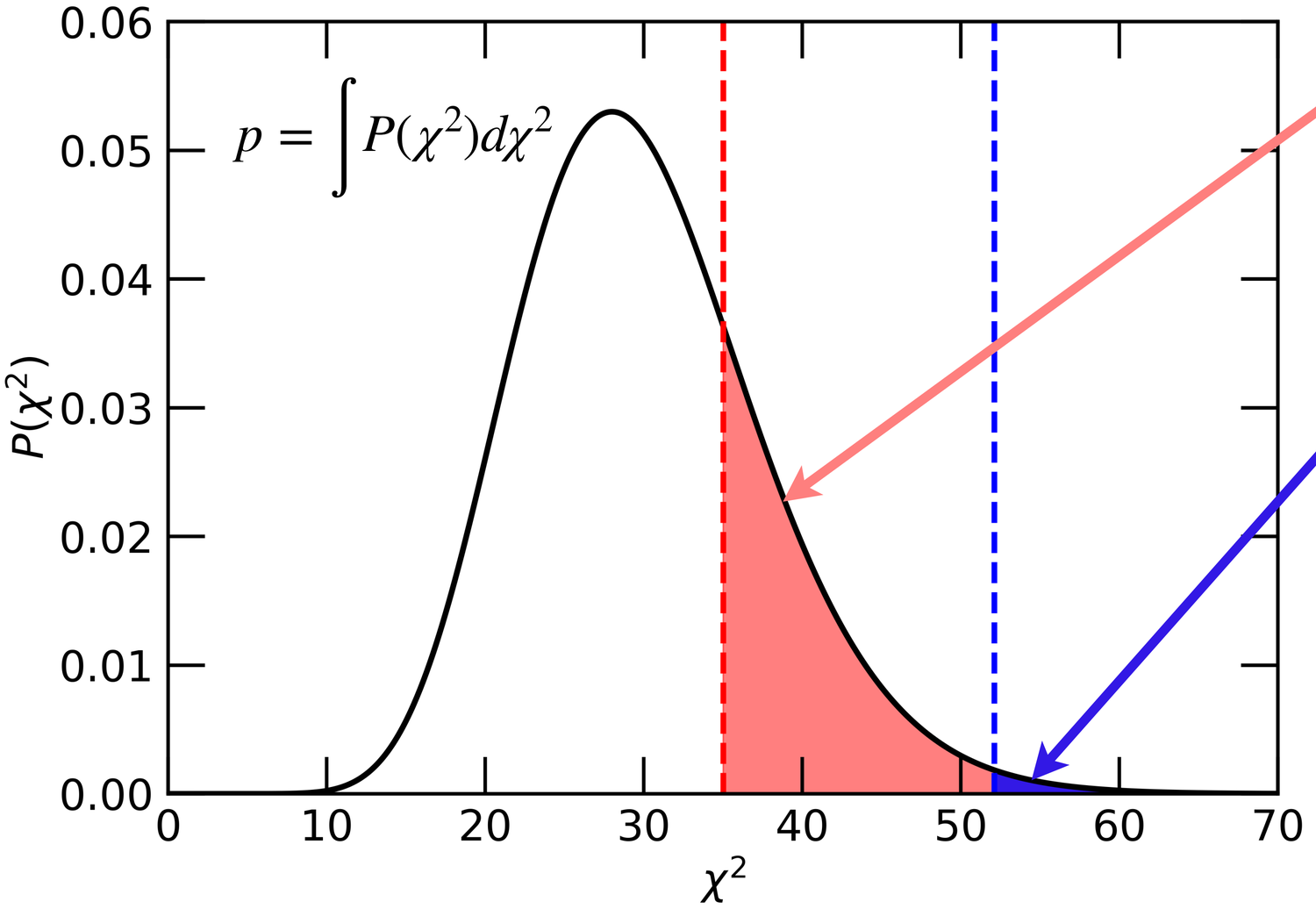


(3) If the measured statistic lies in the **two-tail critical region**, we reject the hypothesis with a confidence equal to the shaded area.

(1) Measure χ^2 statistic from our dataset

Chi-Square (χ^2) statistic as hypothesis test

- Example: Suppose that the mean $k = 30$ and consider datasets with two different values, $\chi^2 = 35$ and $\chi^2 = 51$. We would integrate the following areas:



p=0.24
for $\chi^2 > 35$

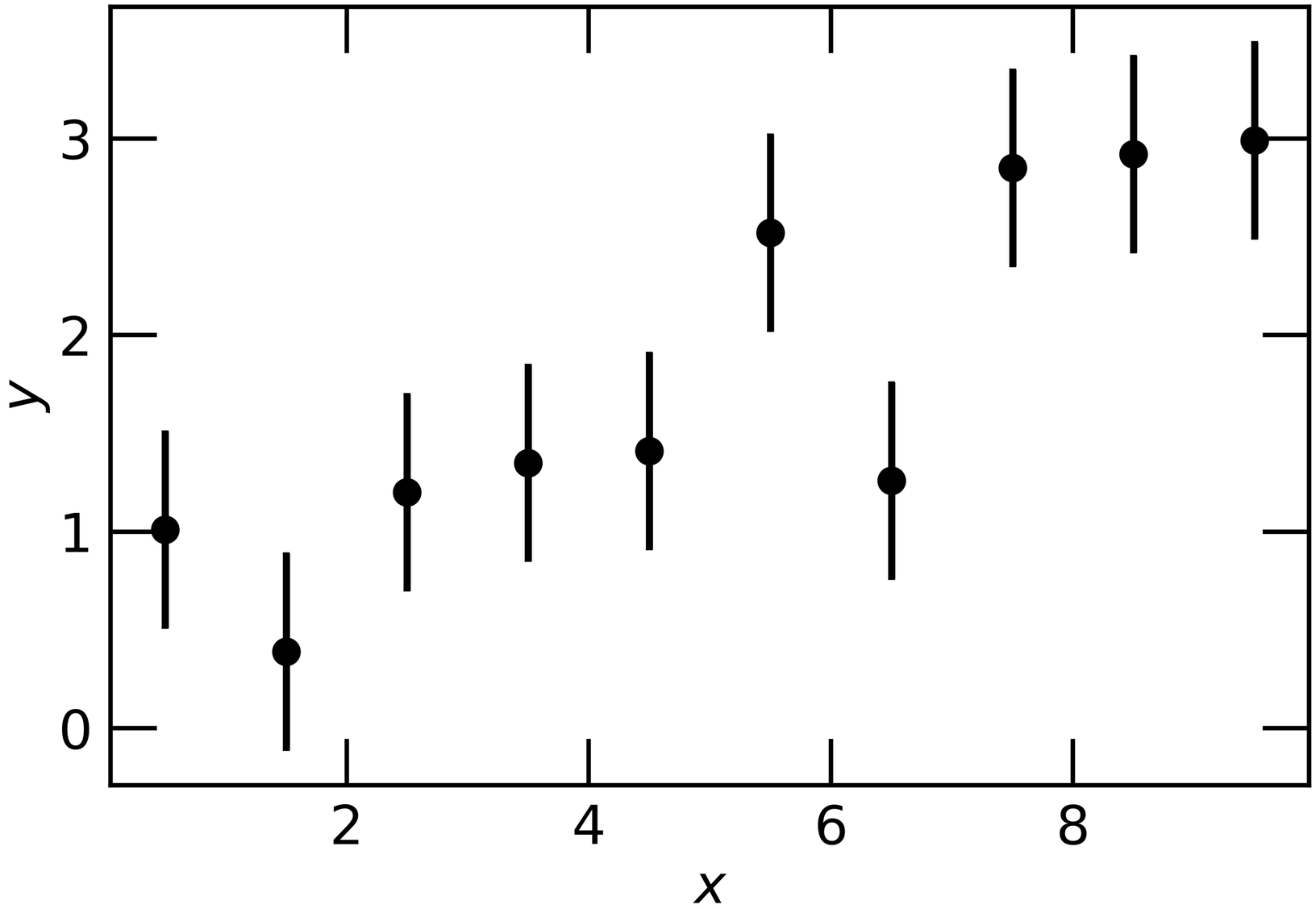
p=0.01
for $\chi^2 > 51$

- 1 % chance of obtaining a set of measurements. We should pull out this hypothesis.

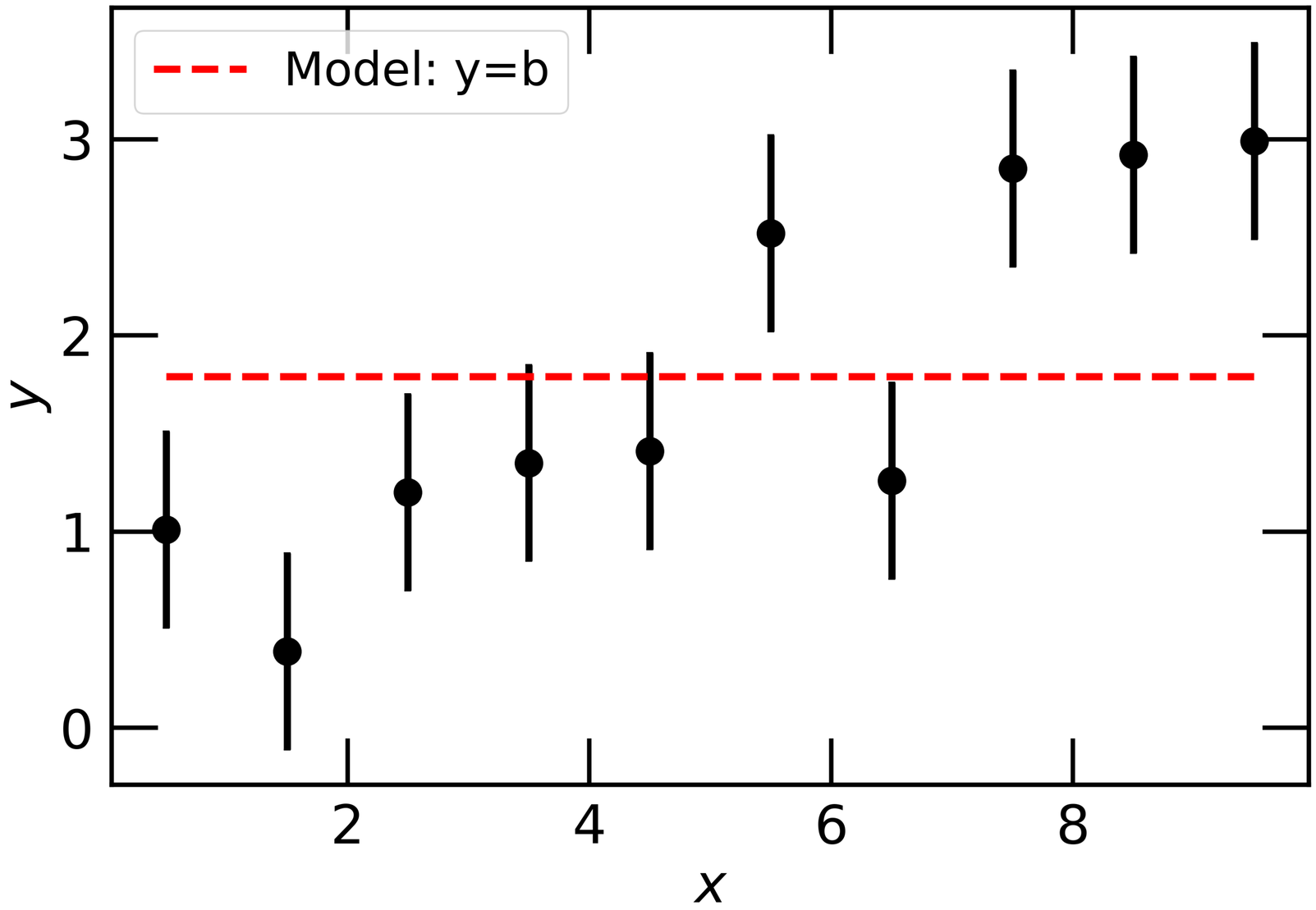
2. Parameter estimation

2. Parameter estimation using minimum Chi-Square (χ^2)

What should be the parameters of a model to fit the example N = 10 points?



2. Parameter estimation using minimum Chi-Square (χ^2)

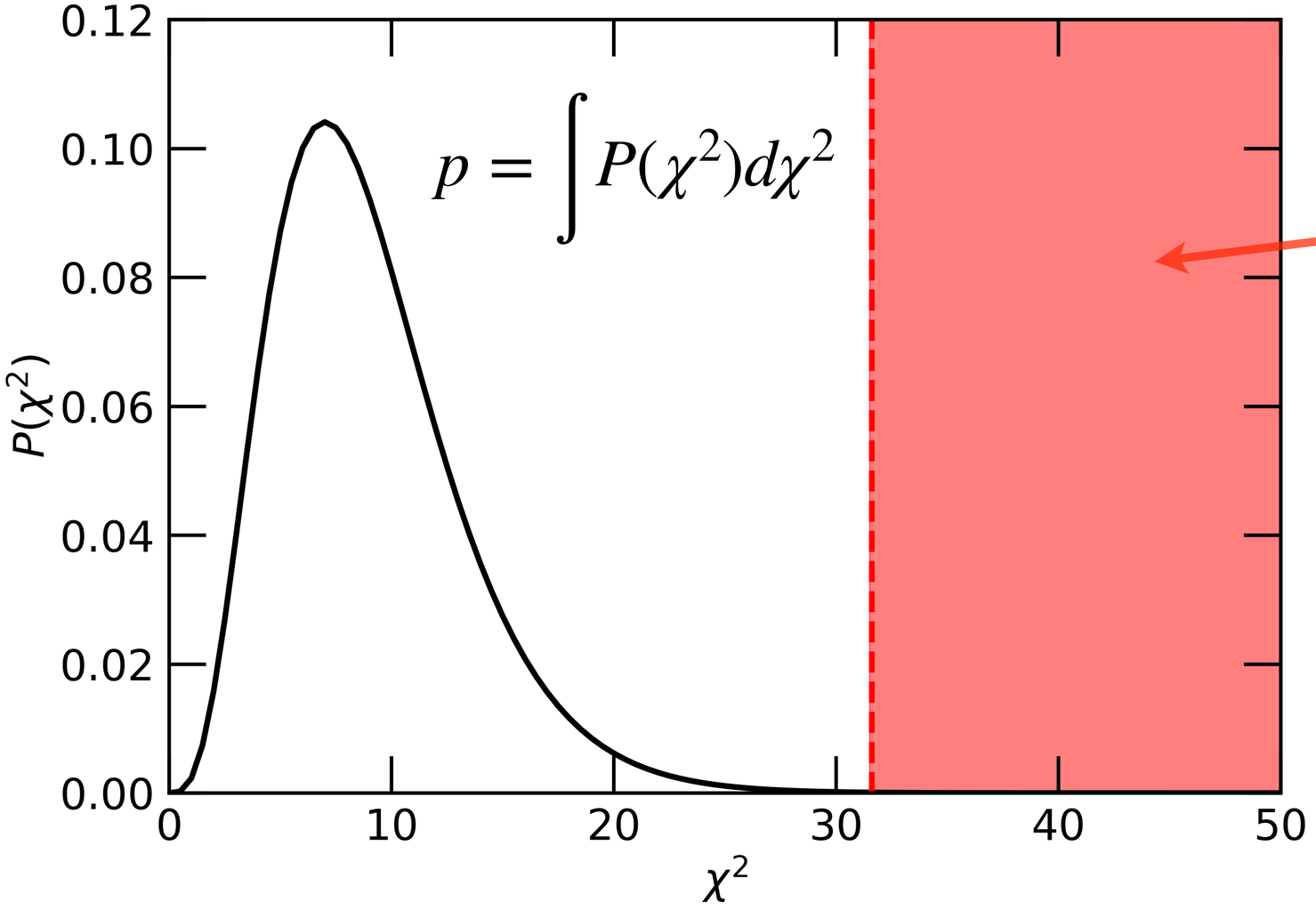


A model y = b can fit the data?

Minimizing χ^2, the value $\chi^2 = 31.6$ for the parameter: b = 1.79

2. Parameter estimation using minimum Chi-Square (χ^2)

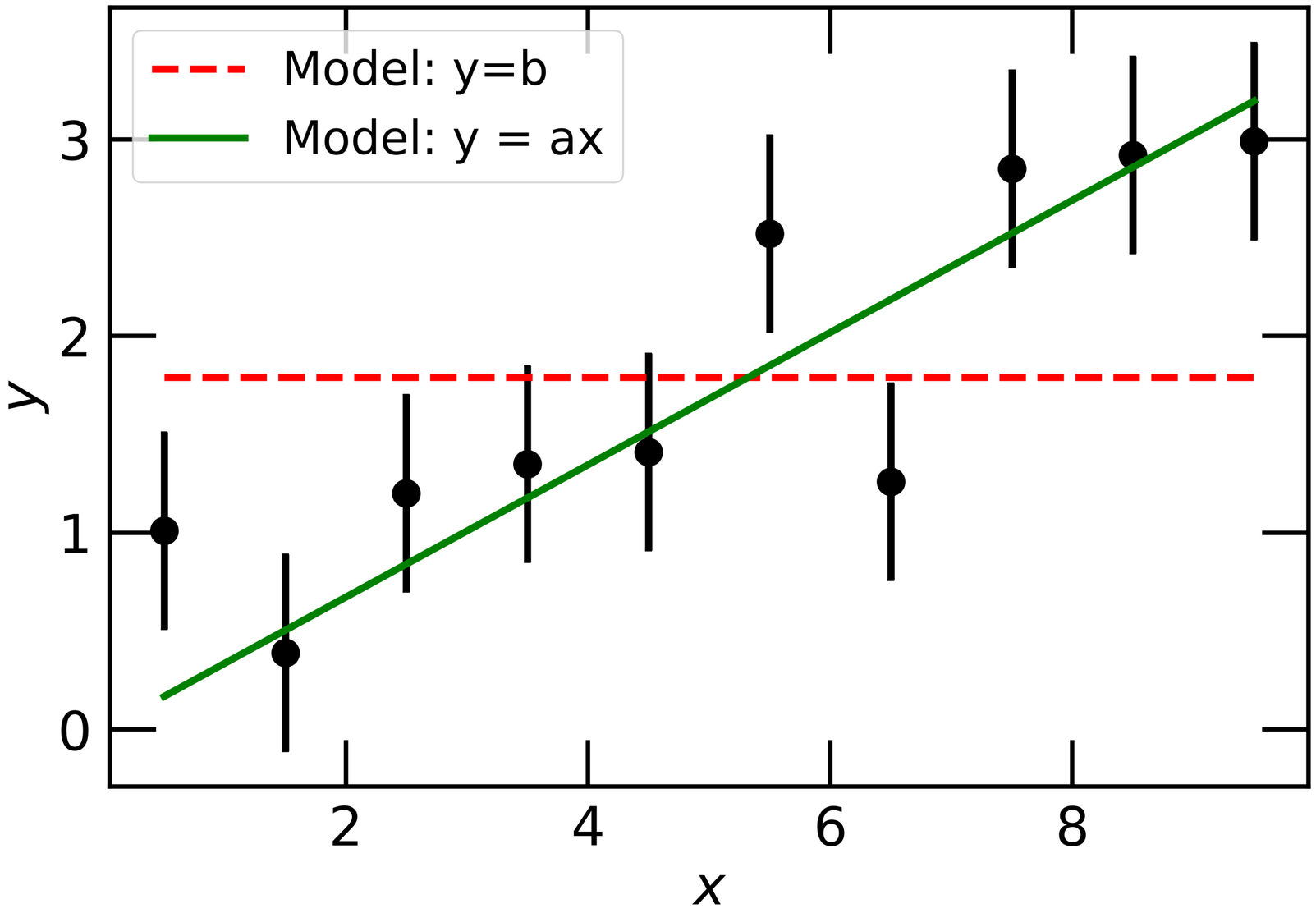
A model y = b can fit the data? Consider the χ^2 probability distribution with the number of degree of freedom (mean) $k = N - 1 = 9$. Calculate p-value for $\chi^2 > 31.6$



p-value = 0.000233

Probability in the tail critical region - **The model is ruled out!**

2. Parameter estimation using minimum Chi-Square (χ^2)

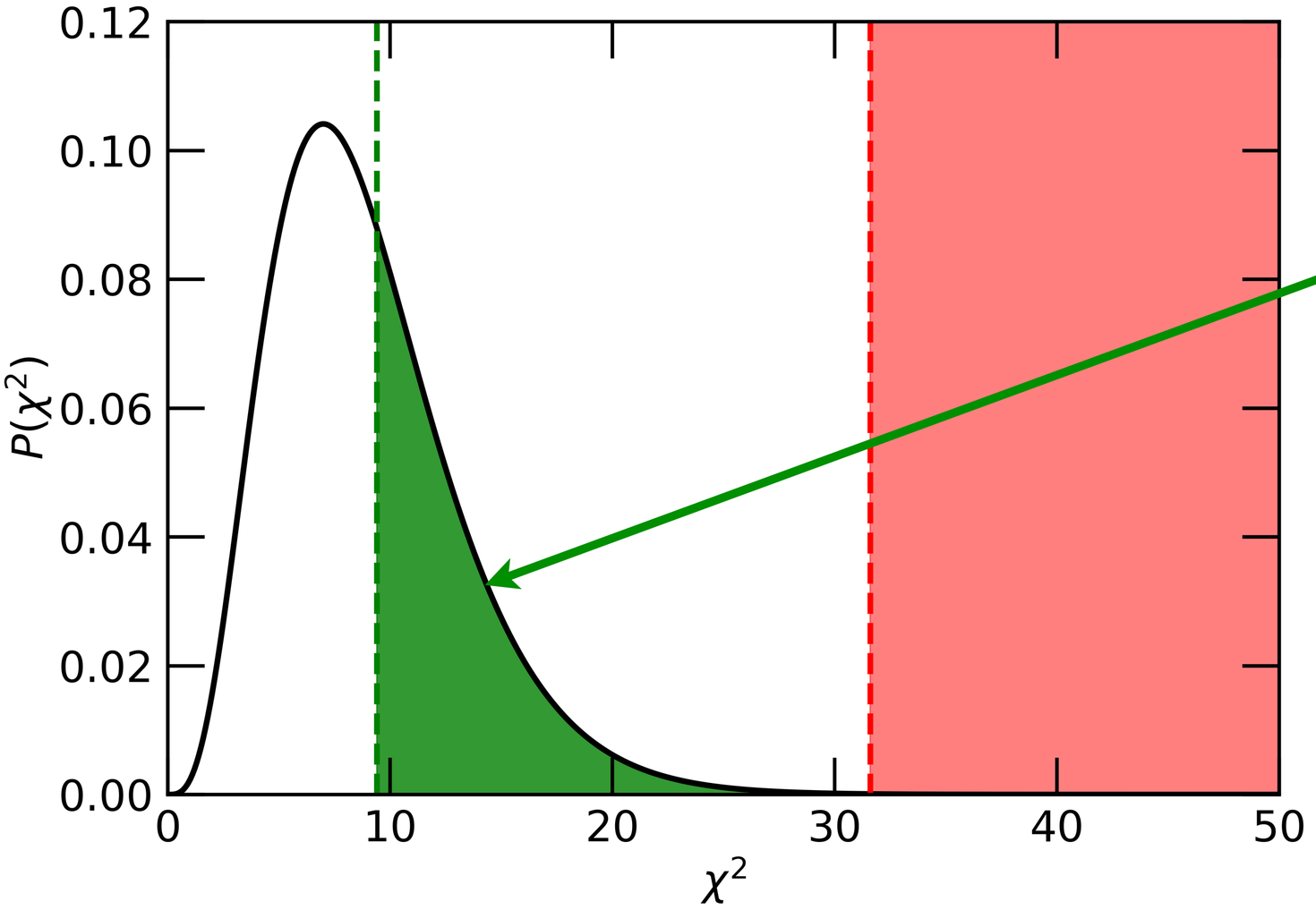


The model y = ax can fit the data?

Minimizing χ^2, the value $\chi^2 = 9.4$ for the parameter: y = 0.336x

2. Parameter estimation using minimum Chi-Square (χ^2)

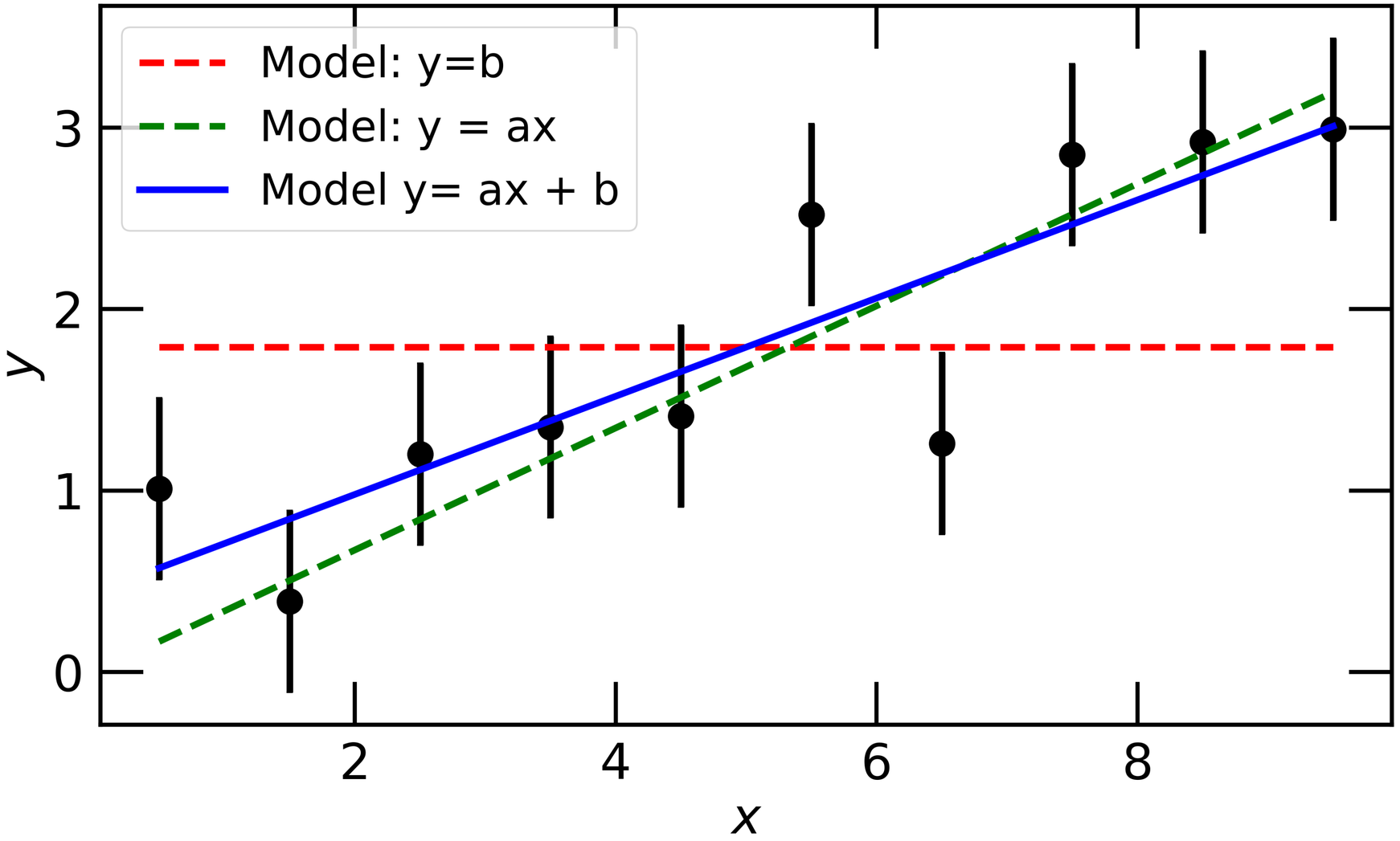
A model y = ax can fit the data? Consider the χ^2 probability distribution with the number of degree of freedom (mean) $k = N - 1 = 9$. Calculate p-value for $\chi^2 > 9.4$



p-value = 0.4

Probability in the true value region - **The model is ruled in!**

2. Parameter estimation using minimum Chi-Square (χ^2)



The model y = ax+b can fit the data?

y=1.79: $\chi^2 = 31.6$

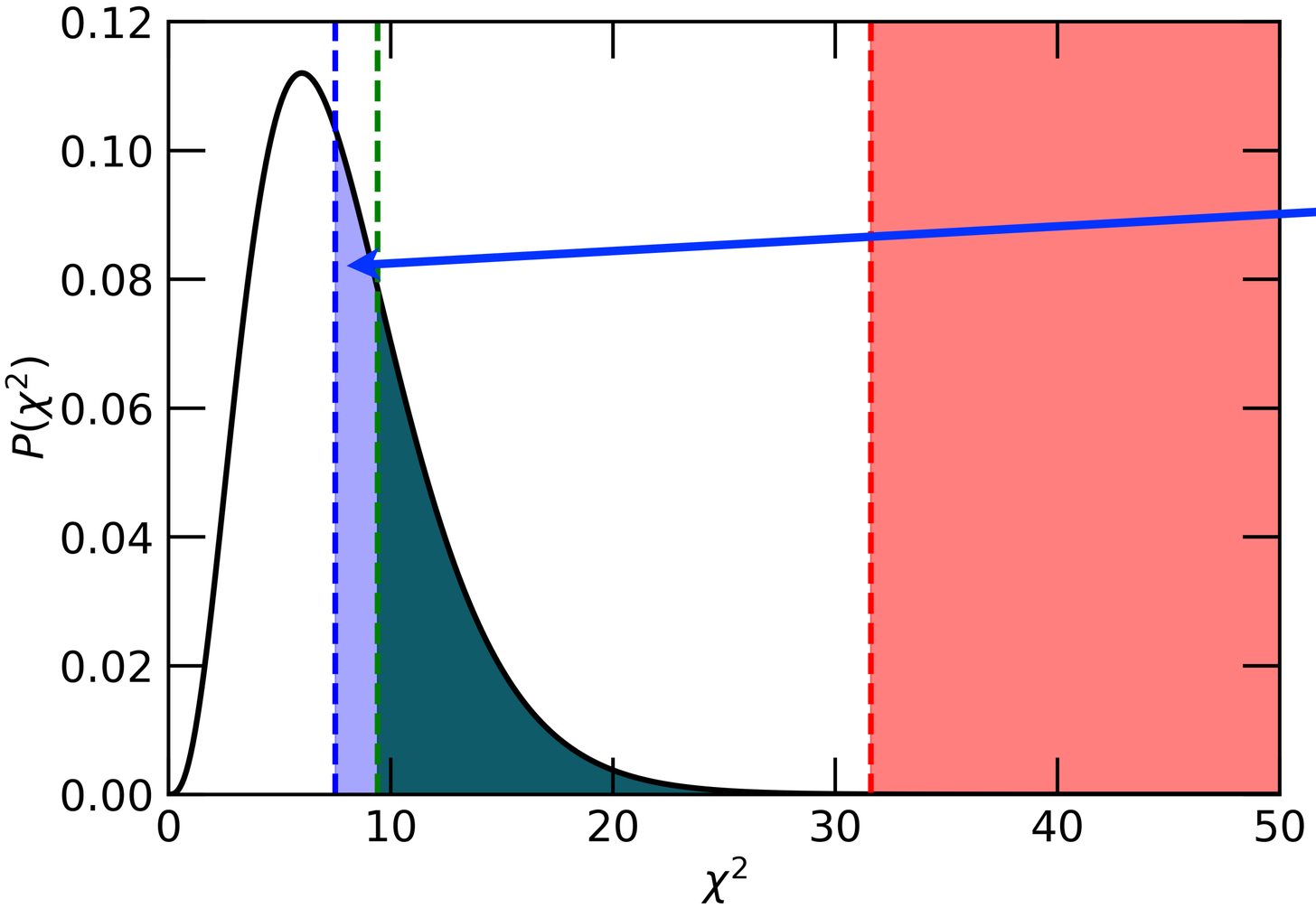
y = ax: $\chi^2 = 9.4$

y = ax +b: $\chi^2 = 7.5$

The best fit model: (y=0.27x+0.43)

2. Parameter estimation using minimum Chi-Square (χ^2)

A model y = ax+b can fit the data? Consider the χ^2 probability distribution with the number of degree of freedom (mean) $k = N - 2 = 8$. Calculate p-value for $\chi^2 > 31.6$



p-value = 0.48

Probability in the true value region - **The model is ruled in!**

Least squares fitting a straight line: $y = ax + b$ (Linear regression)

- Suppose we have N data points with it errors (x_i, y_i, σ_i)
- Assume we know a theoretical model of the relationship of the data points: $y = f(x, a, b, \ldots)$, where a, b, ... are constant parameters.

$$\chi^2(a, b, \ldots) = \sum_{i=1}^{n} \frac{[y_i - f(x_i, a, b, \ldots)]^2}{\sigma_i^2}$$

- This is very similar to the Maximum Likelihood Method.

Least squares fitting a straight line: $y = ax + b$ (Linear regression)

- Lets take a simple linear equation: $y = ax + b$
- How to estimate the parameters a, b?

$$\chi^2(a,b) = \sum_{i=1}^{n} \frac{[y_i - ax_i - b]^2}{\sigma_i^2}$$

- A procedure to obtain a and b is to minimize the following χ^2 with respect to a and b.

Lets calculate on the chalkboard!

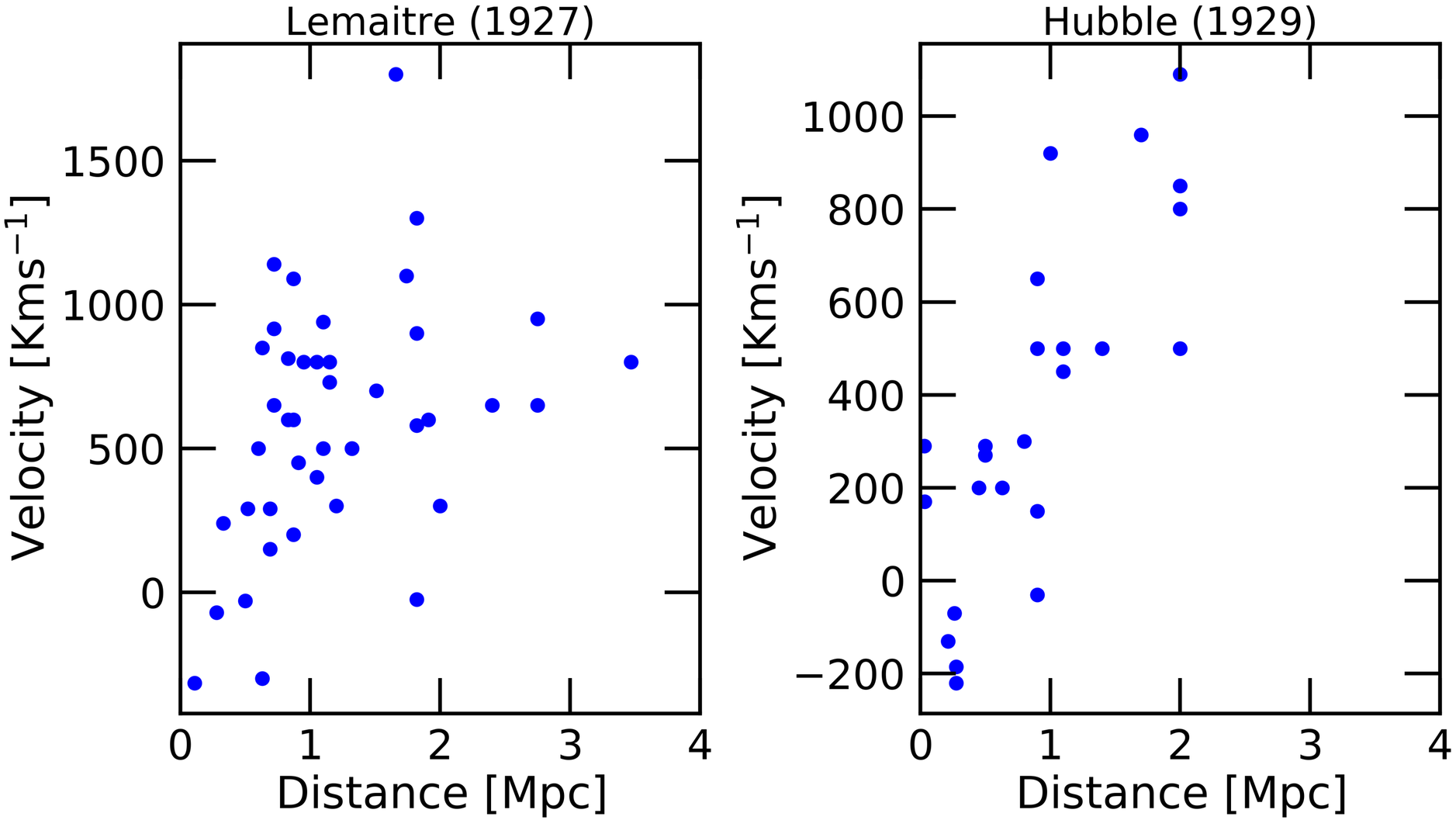
Least squares fitting a straight line: $y = ax + b$ (Linear regression)

- **Example**: Find the parameters a, b for a model y = ax+b, and estimate χ^2 value.

x	1.0	2.0	3.0	4.0	5.0
y	2.2	2.9	4.3	5.2	6.3
σ	0.2	0.4	0.3	0.1	0.35

Chi-Square (χ^2)

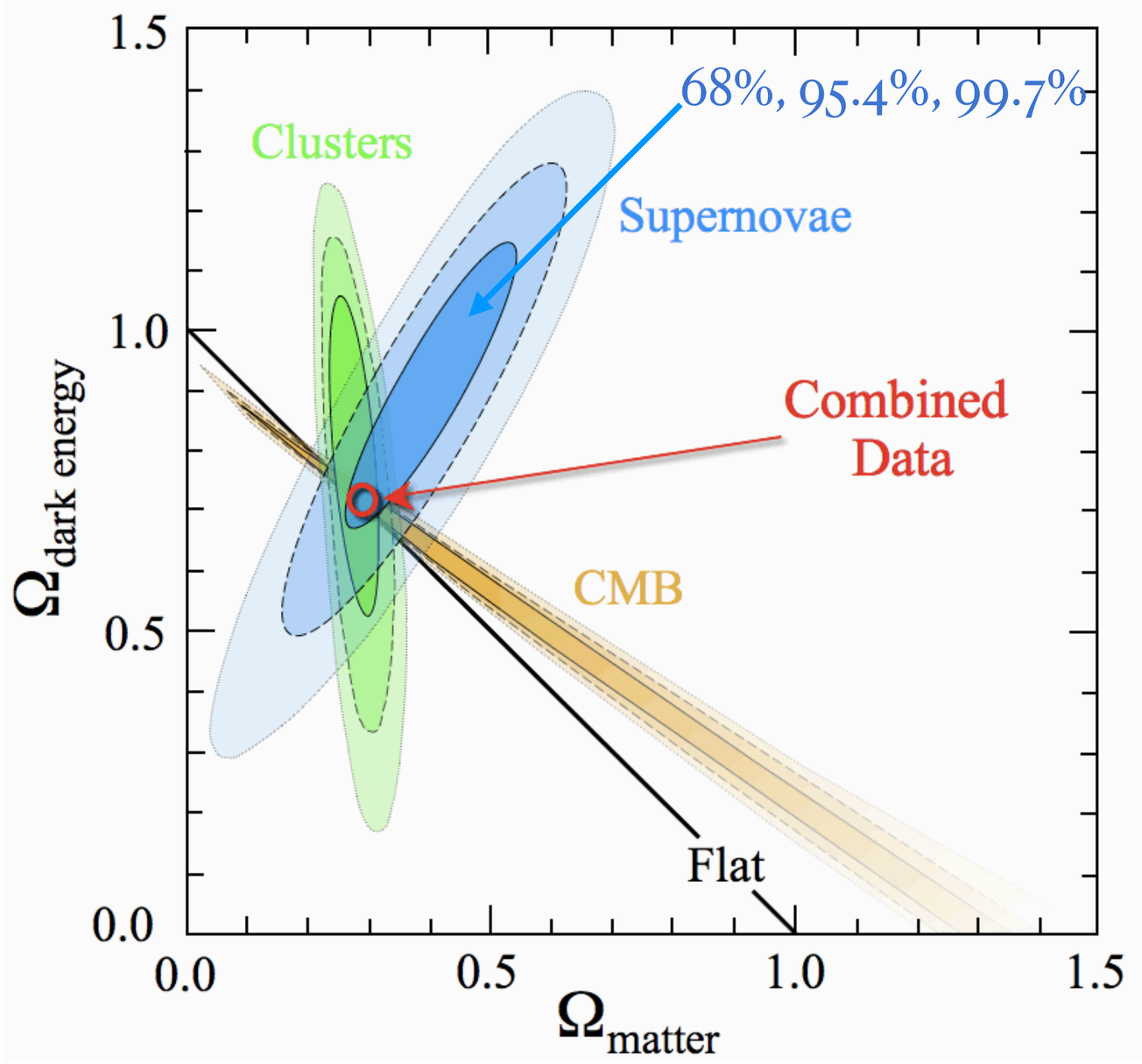
- **Exercise**: For the Hubble, Lemaitre datasets, lets fit a models of linear equation.



Joint confidence regions

- A model has free parameters. Then, how do we determine the most likely values of these parameters and their error ranges?

- **Example**: The join analytical measurements of galaxy clusters, supernovae, and Cosmic microwaves background (CMB) to estimate the dark matter parameter ($\Omega_m \approx 0.3$), and the dark energy ($\Omega_\Lambda \approx 0.7$) in the universe.



Joint confidence regions

- Assuming that we are fitting 2 free parameters (a,b) for the model y = ax+b.
- The “best-fitting” values of (a,b) are found by minimizing the χ^2 statistic.
- The joint error distribution of parameters (a,b) can be found by calculating the values of χ^2 over a grid of (a,b) and enclosing region $\chi^2 < \chi^2_{min} + \Delta\chi^2$.
- It means we can plot 2D contour of $\chi^2 < \chi^2_{min} + \Delta\chi^2$.

Joint confidence regions

- It means we can plot 2D contour of $\chi^2 < \chi^2_{min} + \Delta\chi^2$. We already know the value of minimum χ^2 . The value of $\Delta\chi^2$ of 2 parameters is (2.30, 6.17, 11.8) correspond for (68%, 95.4%, 99.7%) confident level or (1σ, 2σ, 3σ).

2 parameters model

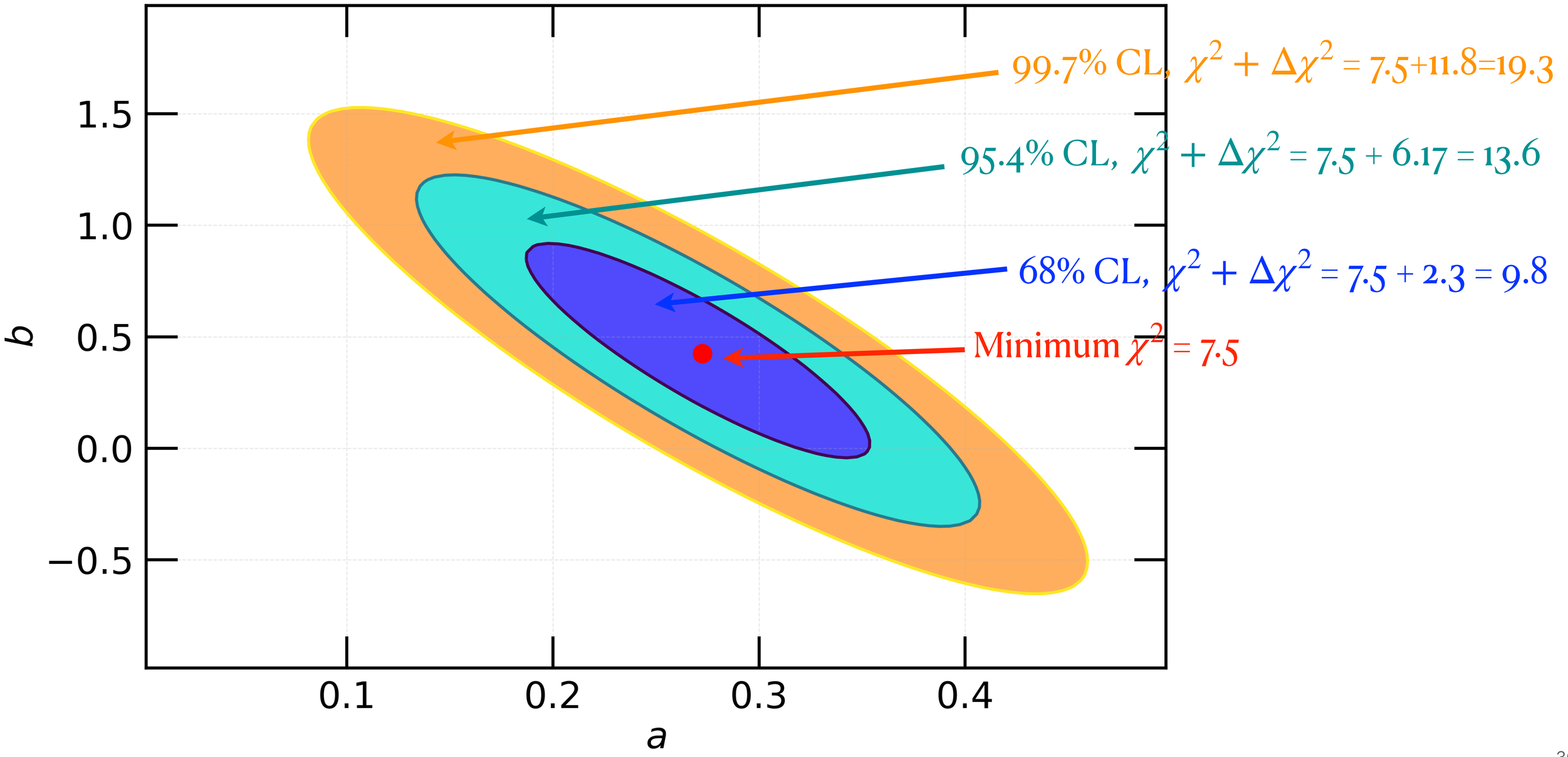
$\Delta\chi^2$ as a Function of Confidence Level and Degrees of Freedom

	ν					
p	1	2	3	4	5	6
68.3%	1.00	2.30	3.53	4.72	5.89	7.04
90%	2.71	4.61	6.25	7.78	9.24	10.6
95.4%	4.00	6.17	8.02	9.70	11.3	12.8
99%	6.63	9.21	11.3	13.3	15.1	16.8
99.73%	9.00	11.8	14.2	16.3	18.2	20.1
99.99%	15.1	18.4	21.1	23.5	25.7	27.8

1σ

Source: Numerical Recipes, chapter 15.

Joint confidence regions



Modified Chi-Square (χ^2) for correlated data

- If the data points are correlated, the χ^2 equation must be modified:

$$\chi^2 = \sum_i \sum_j (d_i - m_i)\left(C^{-1}\right)_{ij}(d_j - m_j) = (\mathbf{d} - \mathbf{m})^T \mathbf{C}^{-1}(\mathbf{d} - \mathbf{m})$$

- Here, the co-variance matrix $C_{ij} = \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle$

- Note that $C_{ii} = \langle x_i^2 \rangle - \langle x_i \rangle^2 = \sigma^2$ (variance)

- In fact, the data used to correlate, so that the equation above is very popular in data analysis of minimum χ^2 for Maximum Likelihook.

- Beside the linear regression, we can use **principal component analysis**, **interpolation, Gaussian process**, **Bayesian likelihood** methods to model a dataset.
- There are several methods to estimate the errors of parameters: Jack-knife, bootstrap (re-sampling), Fisher matrix (estimate covariance), Monte Carlo simulations.

Practical work

Practical work

- A telescope receives photons from the sky/a satellite receive emitted light from the ground/ a satellite at Lagrange point 2 observes Cosmic Microwave Background (CMB) photons. The basically physical process is that a satellite sensor is hit by particles.

- The process can be described by a Poisson statistic. Assuming the average particles coming to the satellite (mirrors, sensor system) is 400 particles per second per centimeter cube (In fact, this is an example of the number of CMB photons).

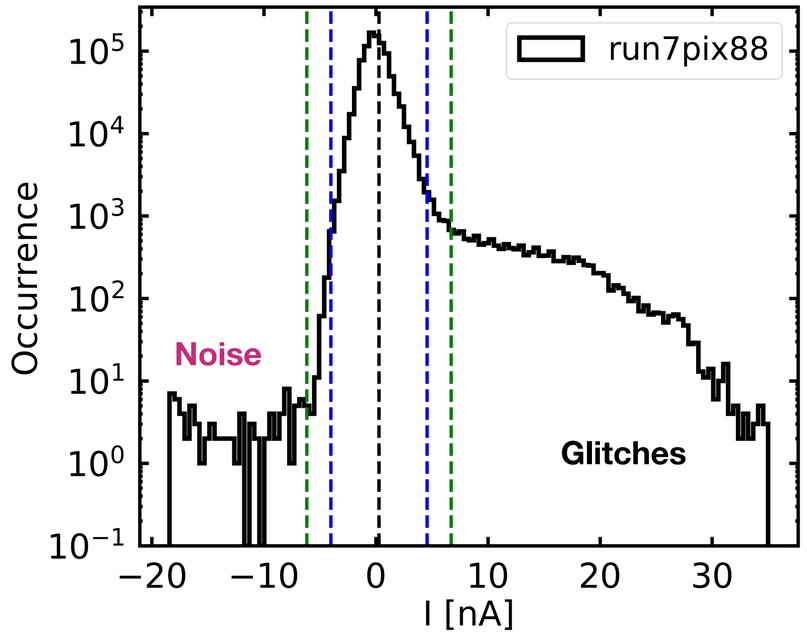
Practical work

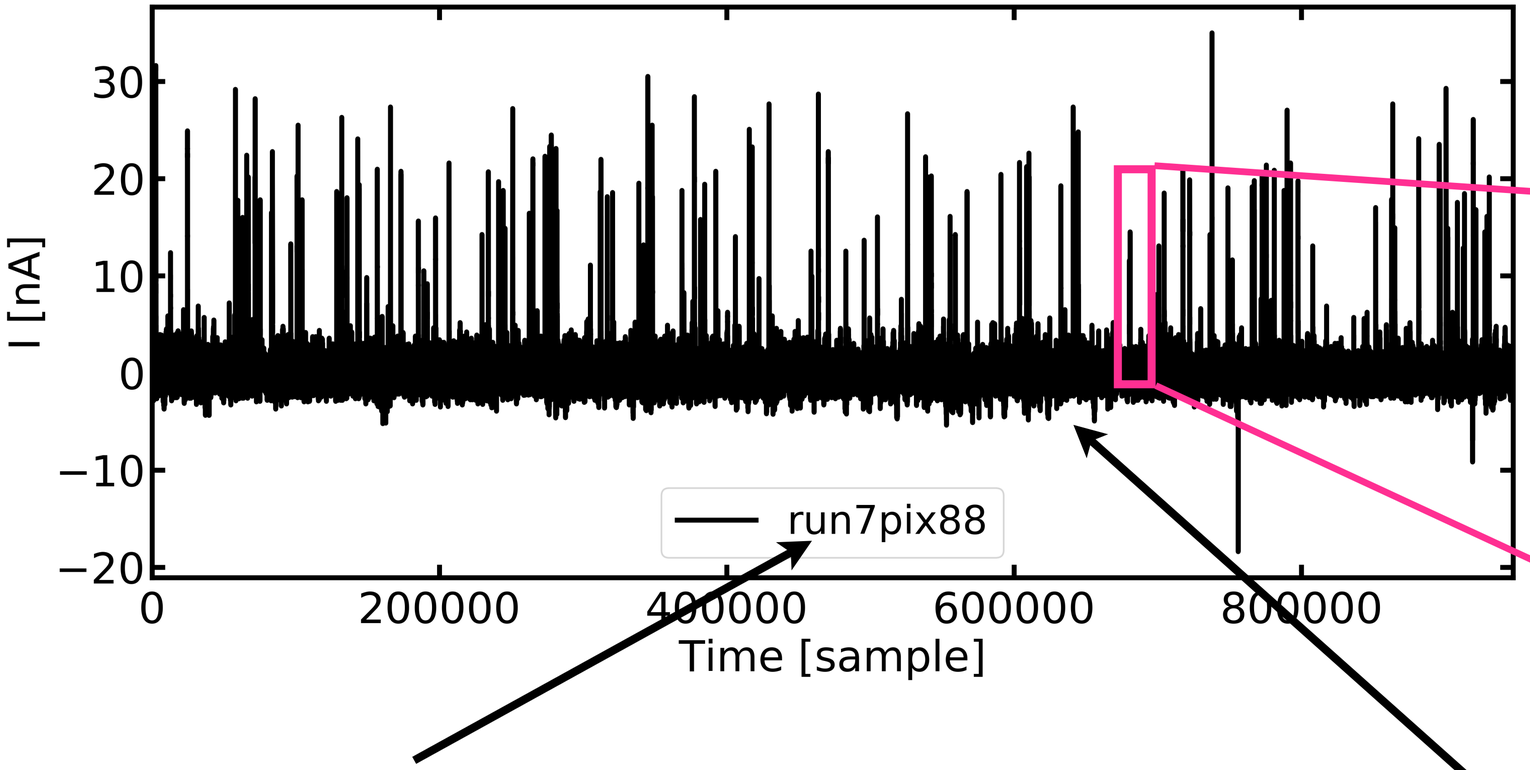
- **Problem 1**: Draw the Poisson probability distribution of N events with the mean = 400 photons per second per centimeter cube.

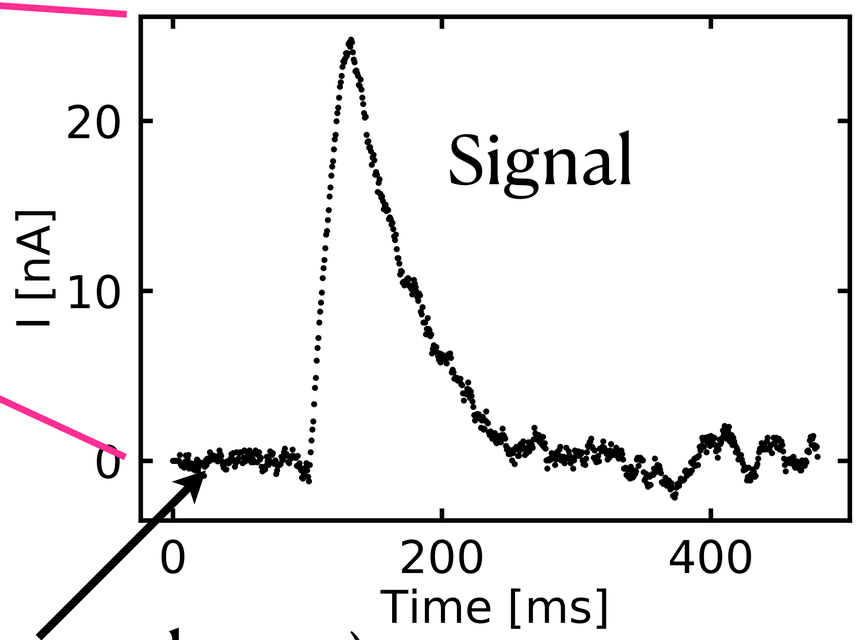
Practical work

Example:

~ 10 minutes of the data when particles hit the sensor.





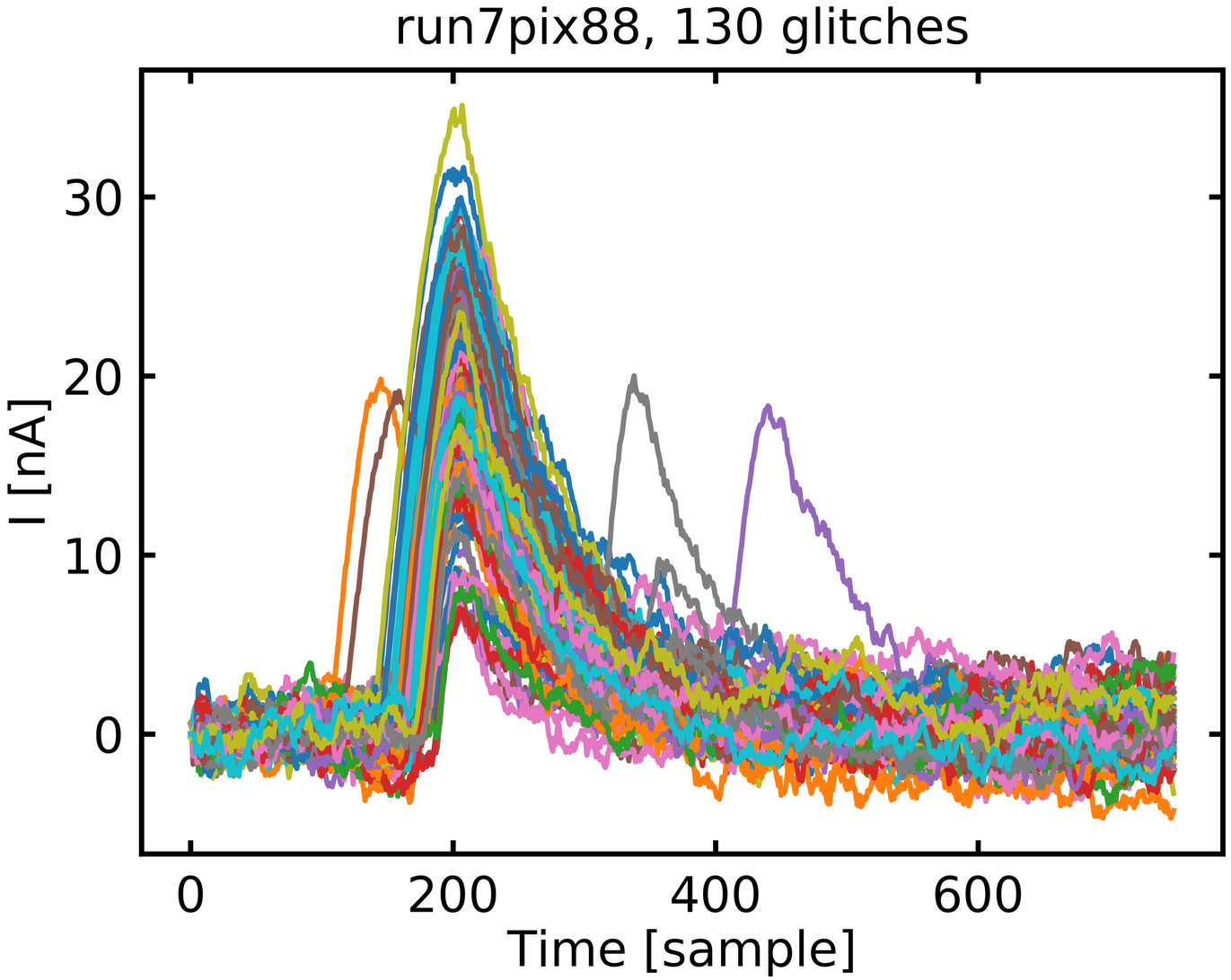


Name of data collection and sensor.

Noise (normal state)

Practical work

- The total number of event in the time order data.
- Sample rate is 6.4 ms. (This number is used to convert x-axis to time unit.)
- This data will be given to you!



Practical work

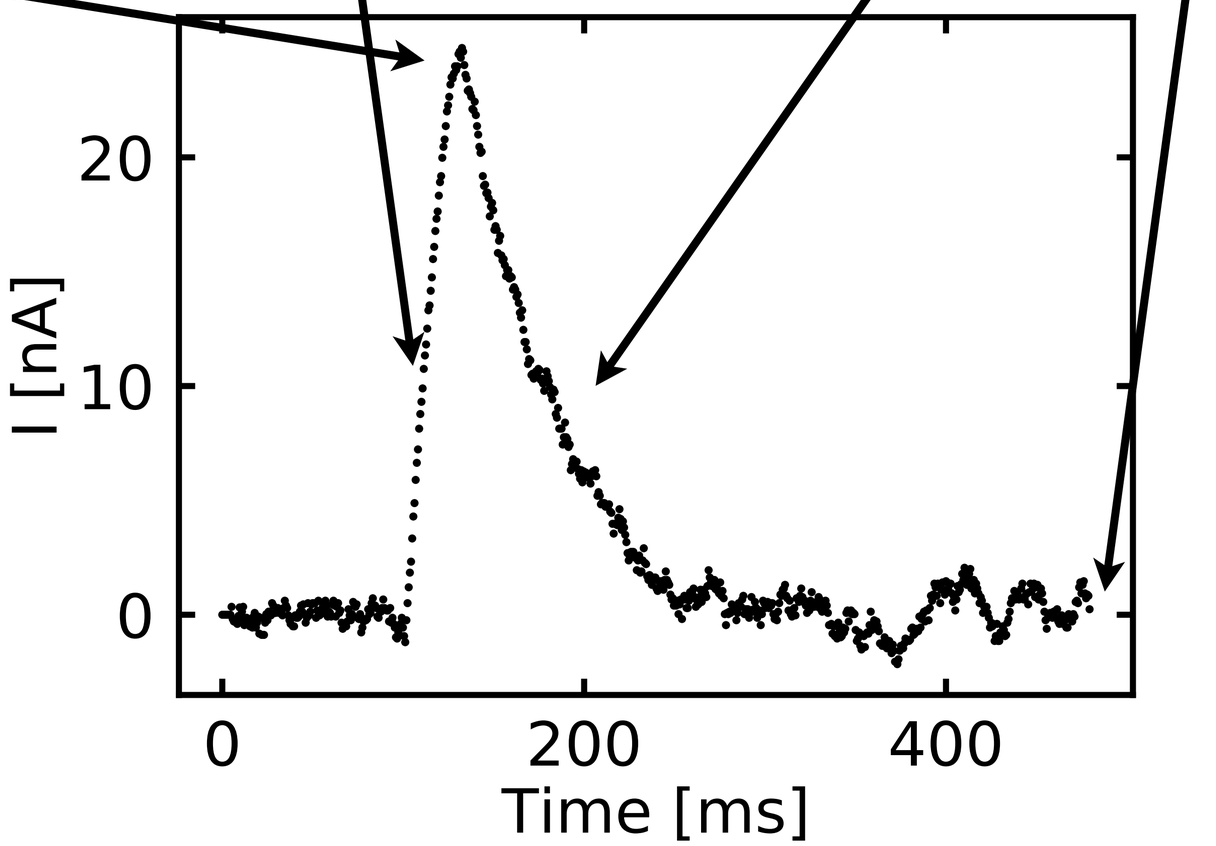
Exponential Model: $$S(t) = a\left(1 - \exp^{-(t-t_0)/\tau_0}\right)\exp^{-(t-t_0)/\tau_1} + c$$

Amplitude **Rising time** **Decay time**

τ_0 : Rising time constant

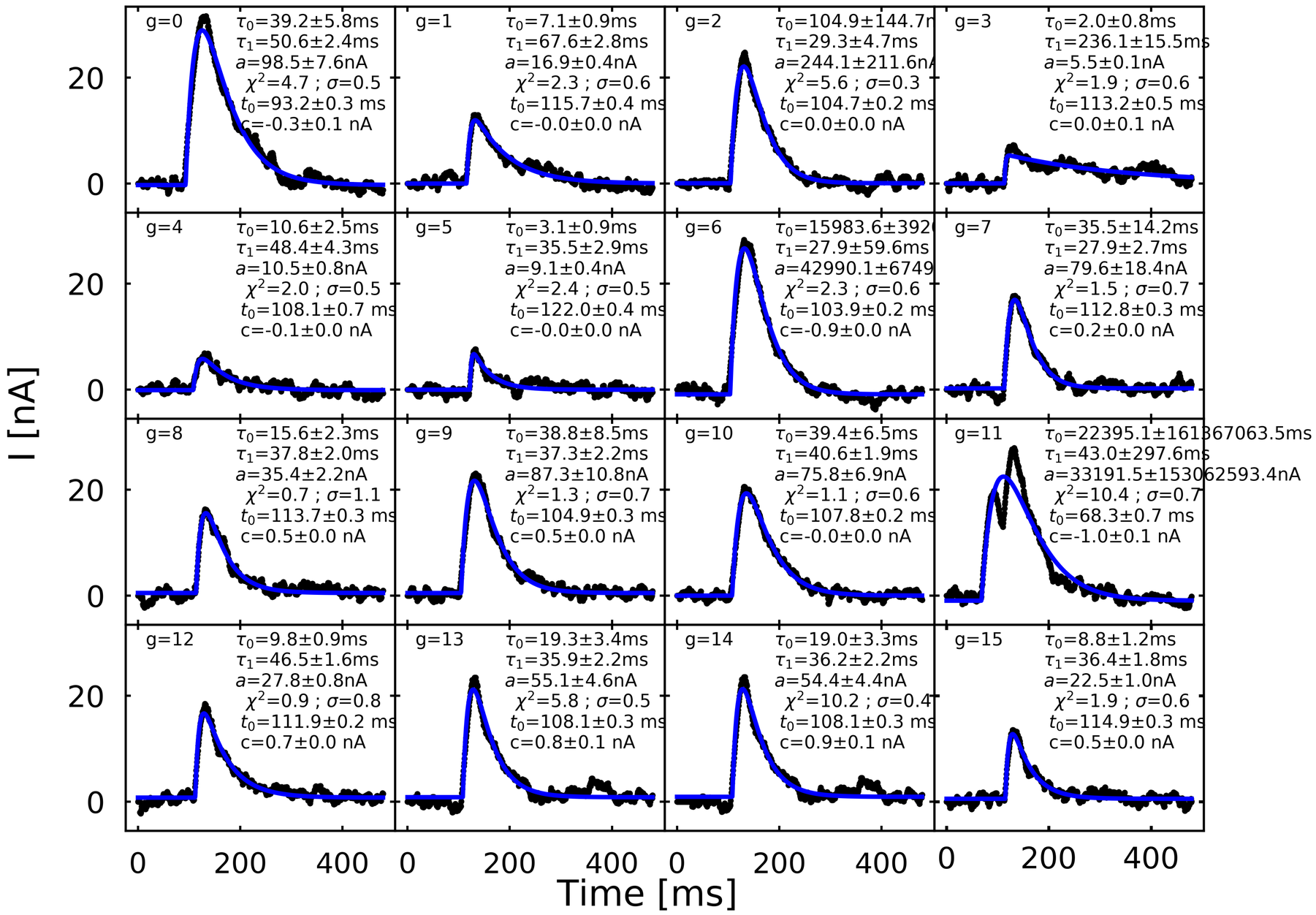
τ_1 : Decay time constant

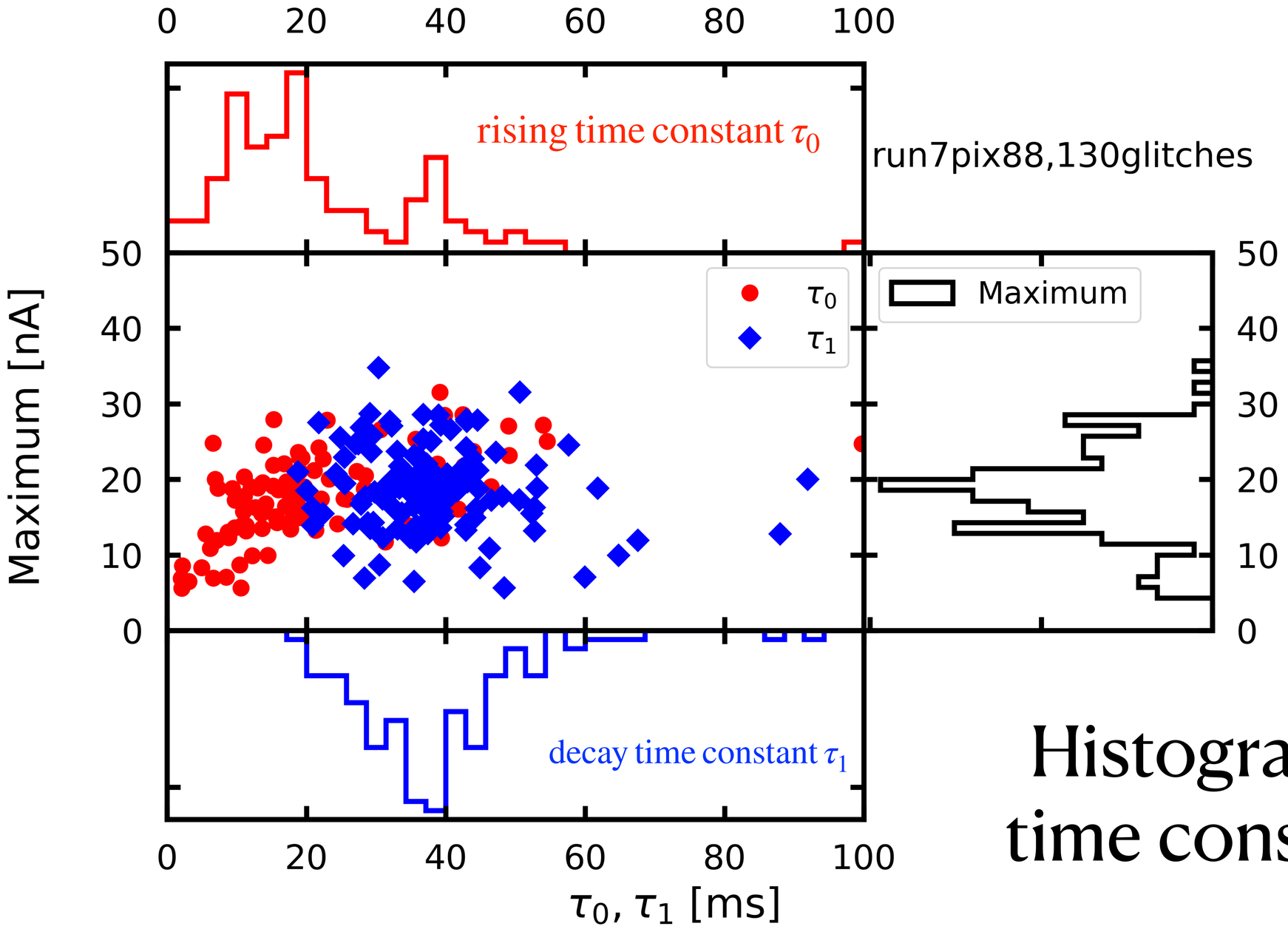
c : Offset



USTH MASTER SPACE
Earth Observation - Astrophysics - Satellite Technologies

This is example of fitting model and chi-square estimation, parameters estimation.





Histogram of time constants

Practical work

Problem 2: Summary practical work requirements

- First read the data and plot
- Fit the data with the exponential model
- Calculate chi-square for every events
- Draw a chi-square distribution base on the degree of freedom = size(data) - number of parameters.
- Select good dataset using chi-square distribution evaluation
- Plot the histogram of estimated parameters of good data
- Plot 2D join confidence region of τ_0, τ_1

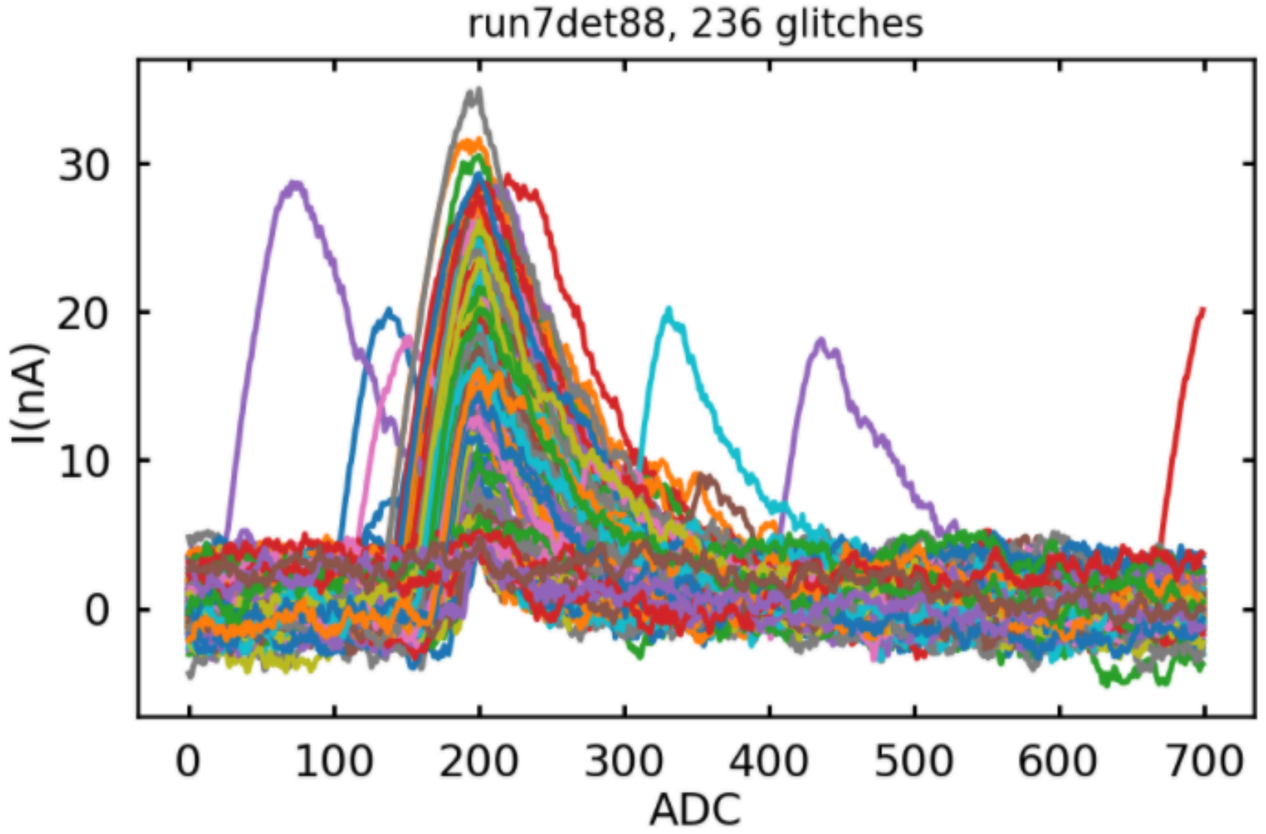
Practical work

- First read the data and plot

```
[12]: signal = np.load('signal_glitches.npy')
```

```
[15]: fig,ax=plt.subplots(figsize=(12,8))
for i in range(len(signal[:,0])):
    ax.plot(signal[i,:])
ax.set_title(label = labeltxt +', '+str(len(signal[:,0]))+' glitches',y=1.02, fontsize=26)

plt.xlabel('ADC',fontsize=30) ;   plt.ylabel('I(nA)',fontsize=30)   ; plt.tight_layout()
plt.tick_params(axis='x', labelsize=30, which='major', pad=18); plt.tick_params(axis='y', labelsize=30, which='major', pad=18)
plt.tick_params(which='both', width=3) ; plt.tick_params(which='major', length=8) ; plt.tick_params(which='minor', length=4)
```



Practical work

- **Requirement**:
  - You will submit your work using a Jupyter Notebook file.
  - Format of the file name: practice_[your_full_name]