Data Analysis

Kevin Nelson, Jianming Qian, Alexander Takla
Michigan Math and Science Scholars

Data Analysis

- **We collect data in the lab.**
  - Resonant frequency of waves on a string
  - Number of radioactive decays in a given time period
  - Time between light pulses to measure the muon decay time

- **We analyze data to draw conclusions**
  - Find relationship between resonant frequency and number of nodes in the string
  - Determine the probabilistic nature of quantum mechanics
  - Find the lifetime of the muon

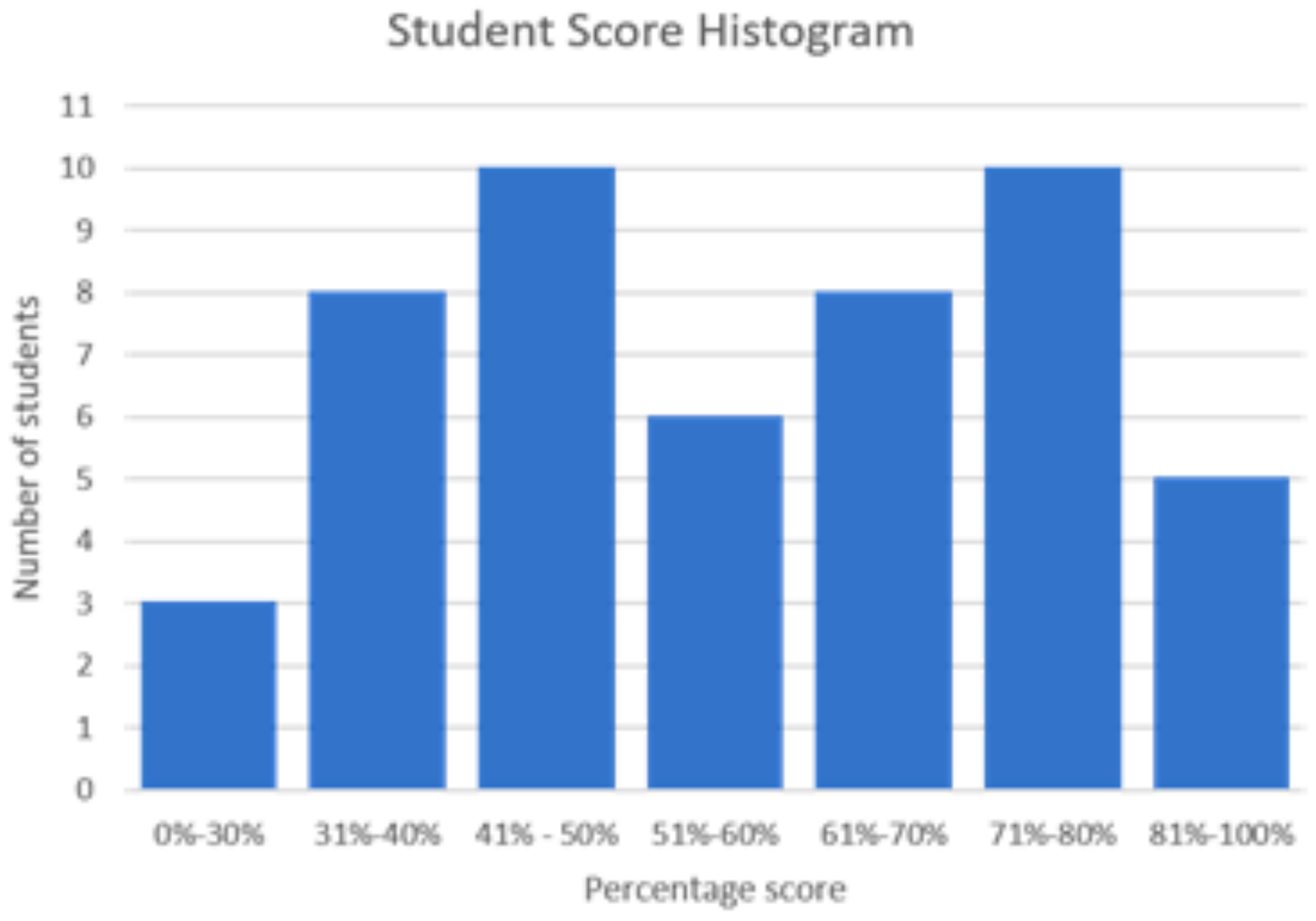
- **Python is one way that physicists analyze data**
  - Today we will discuss concepts in statistics to give formal mathematical understanding to our data analysis.

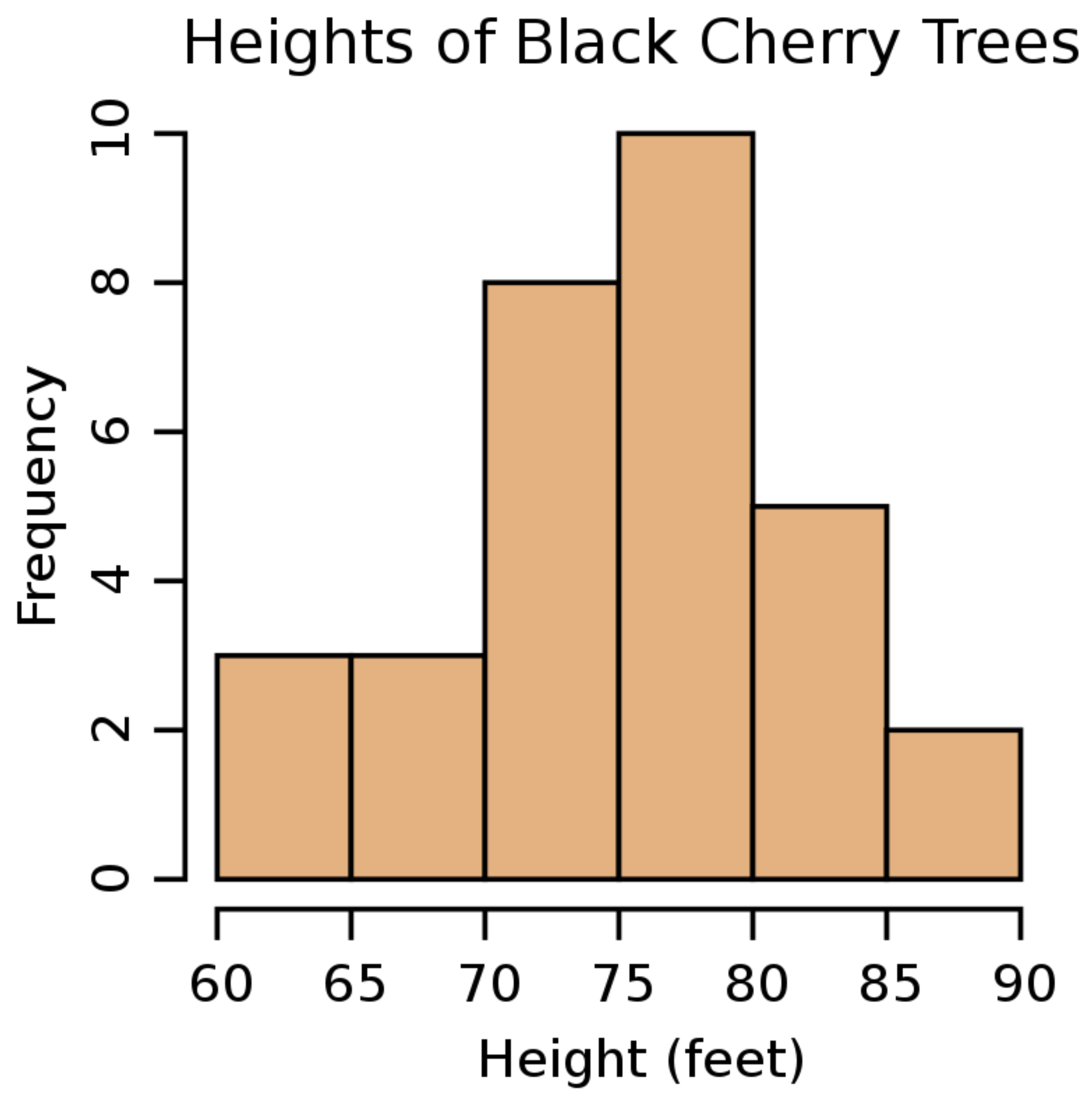
  - Answer questions like:
    - What is the probability that this data was observed?
    - How do scientists discover something new?

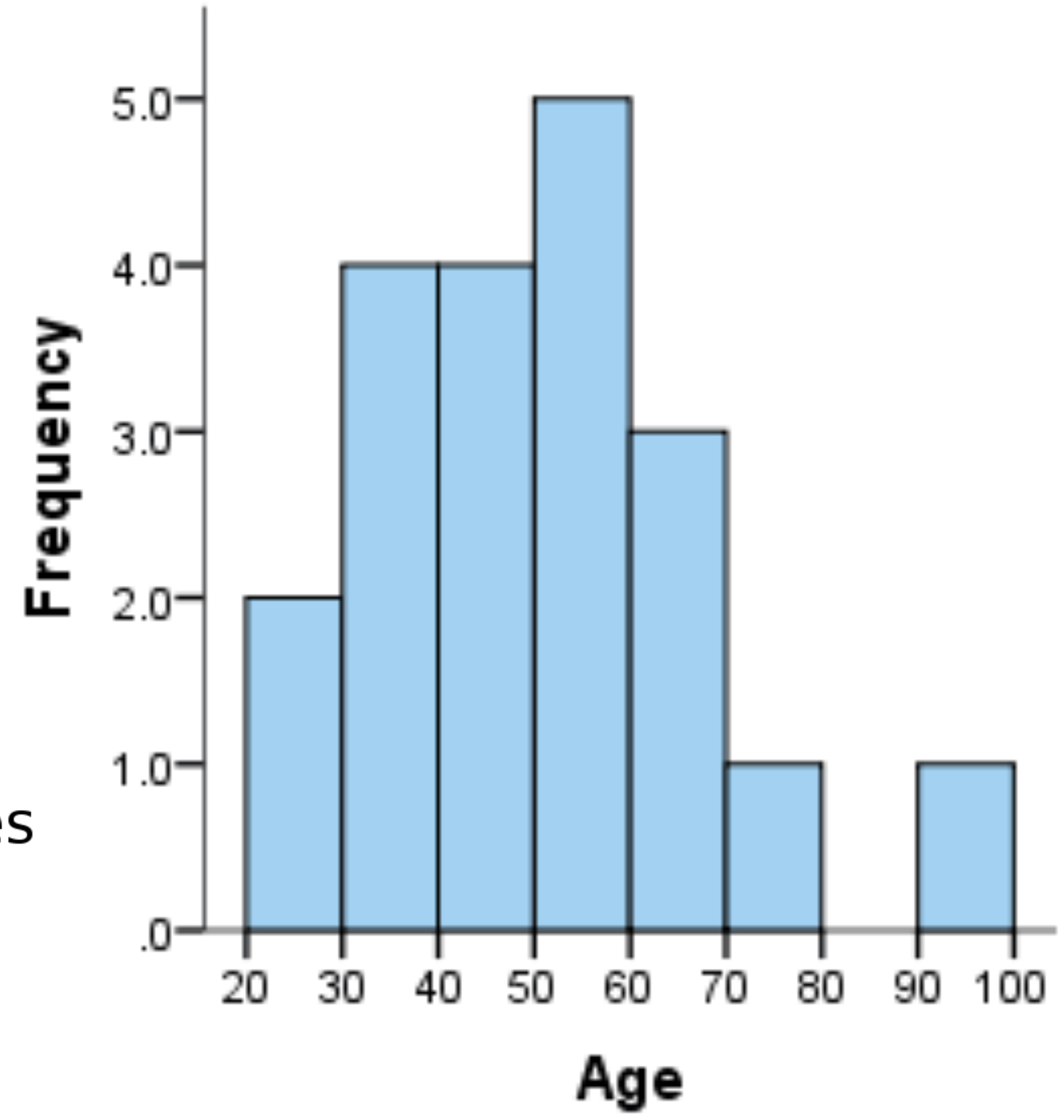
Histograms

- **When you measure some quantity over and over again one way to visualize it is with a histogram**

- **For example:**
  - Survey 20 people and make a histogram of their ages
  - Measure the height of 31 cherry trees in an orchard
  - After a test, your teacher might give you a histogram showing the distribution of scores on the test



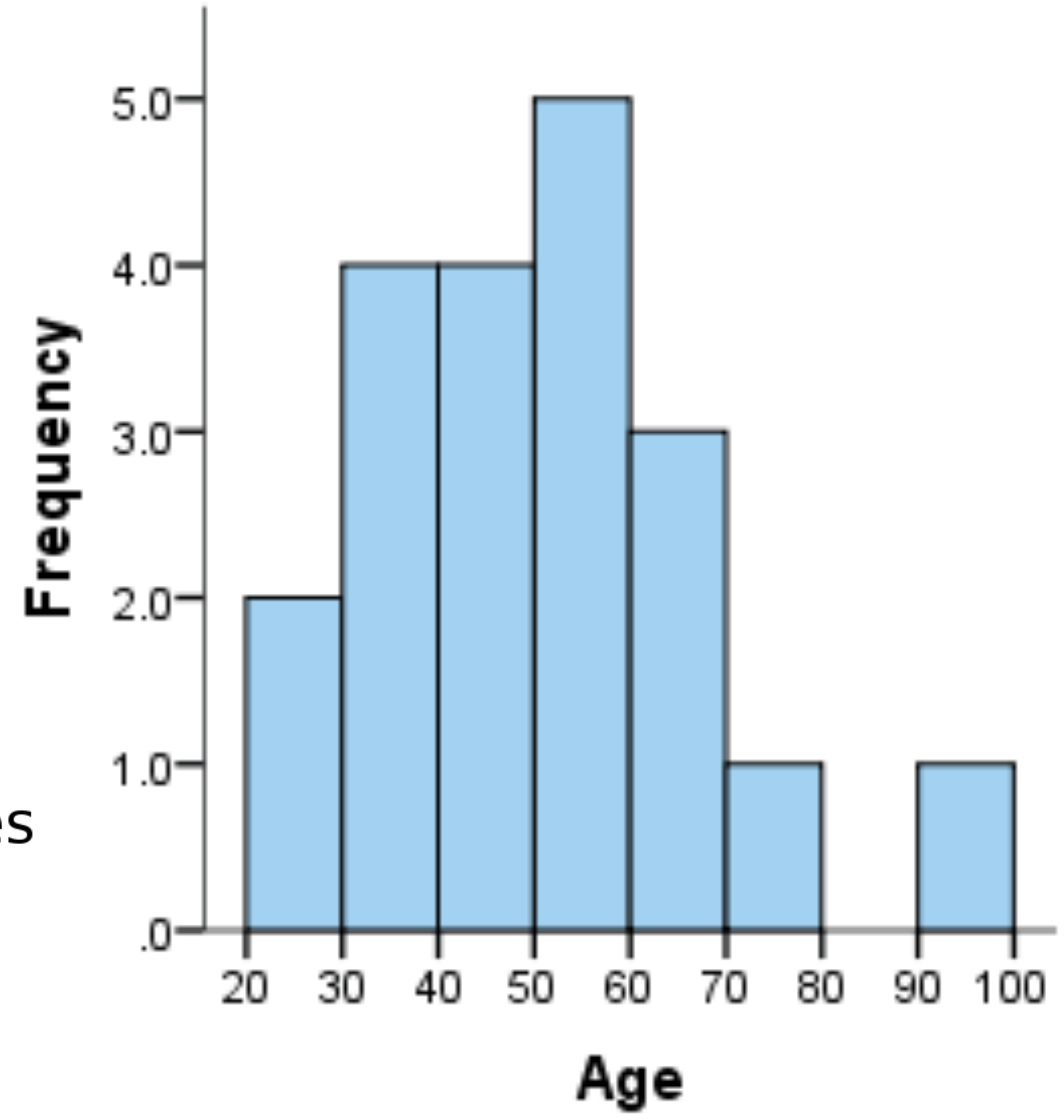


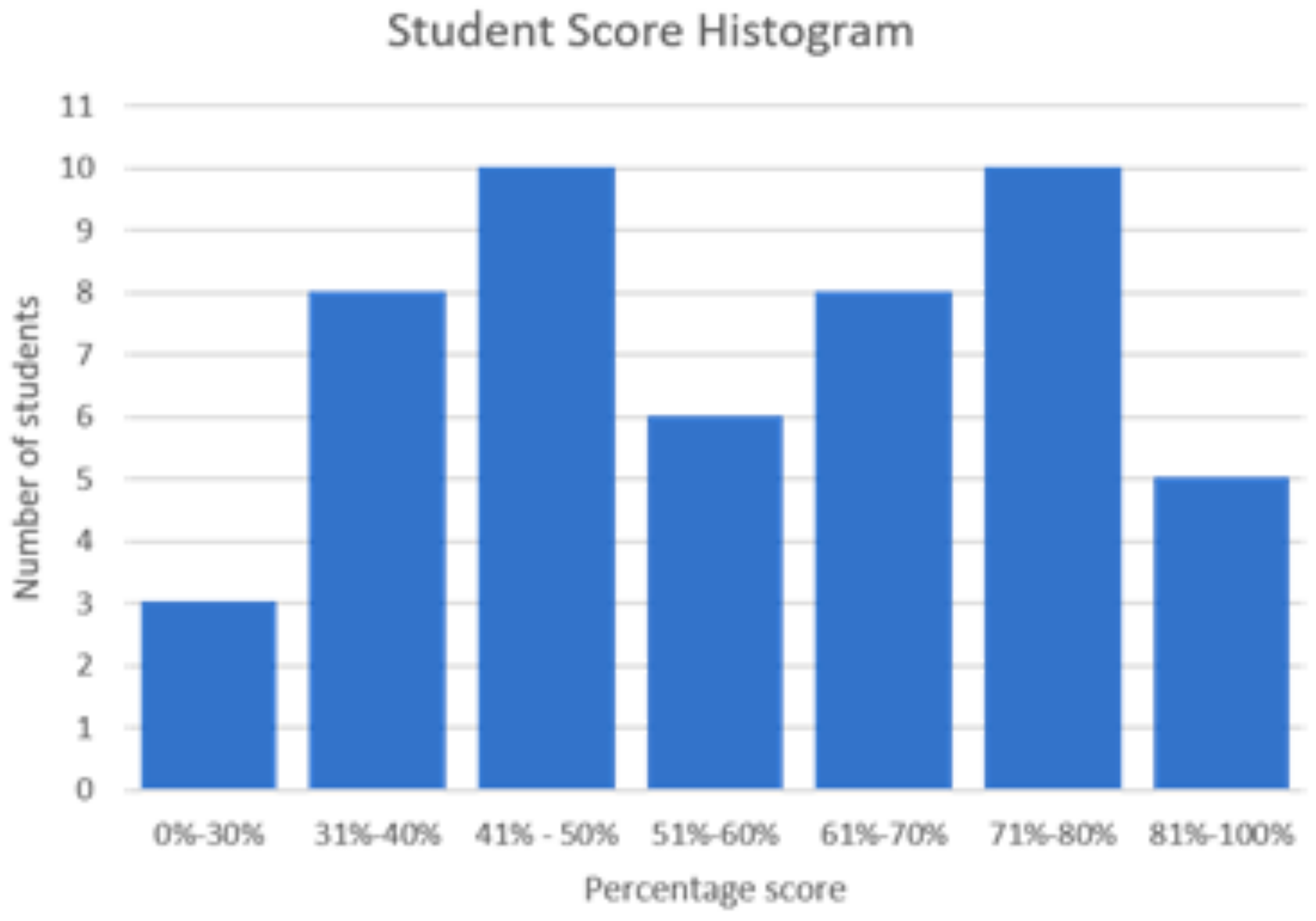


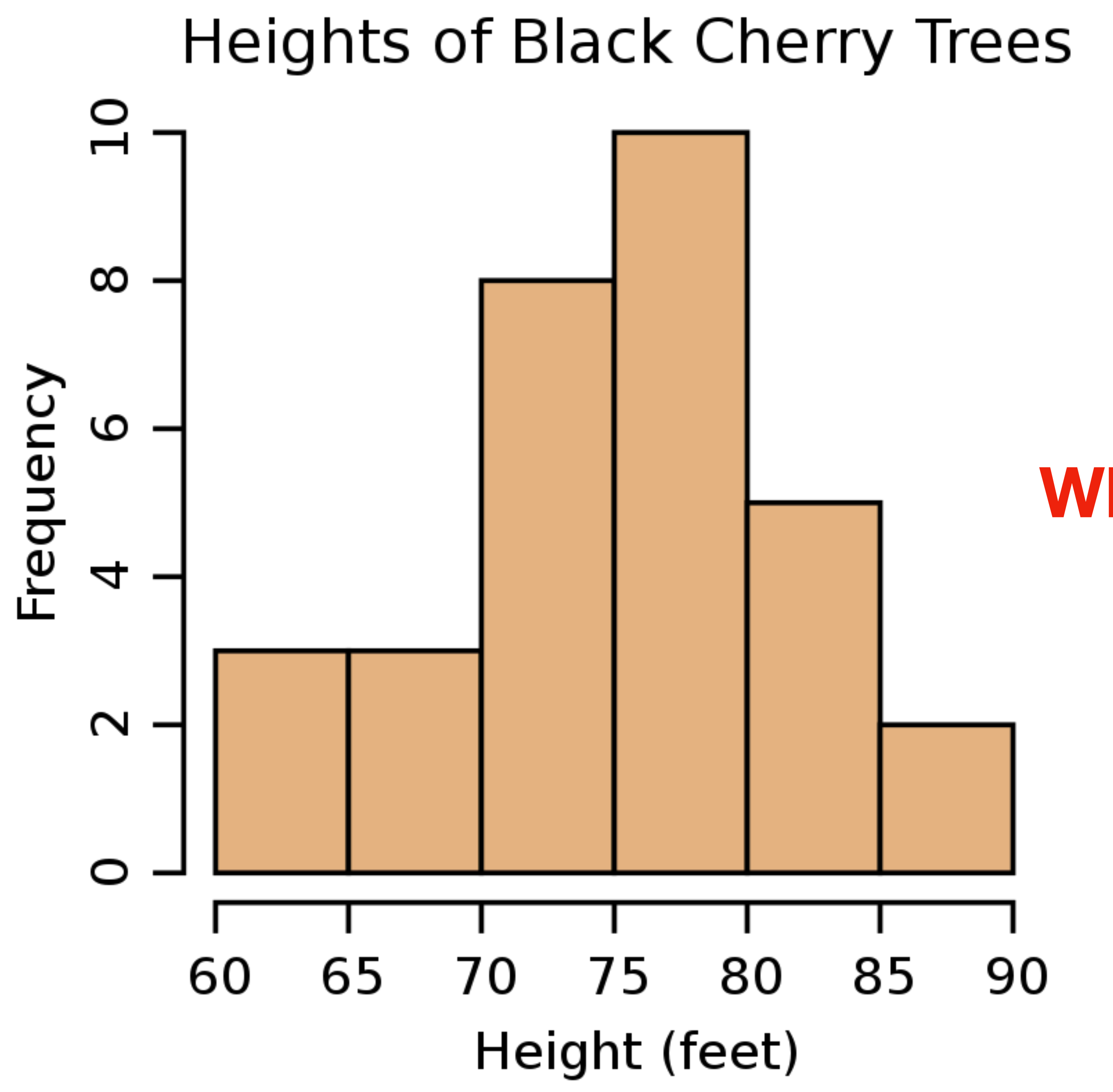
Histograms

- **When you measure some quantity over and over again one way to visualize it is with a histogram**

- **For example:**
  - Survey 20 people and make a histogram of their ages
  - Measure the height of 31 cherry trees in an orchard
  - After a test, your teacher might give you a histogram showing the distribution of scores on the test





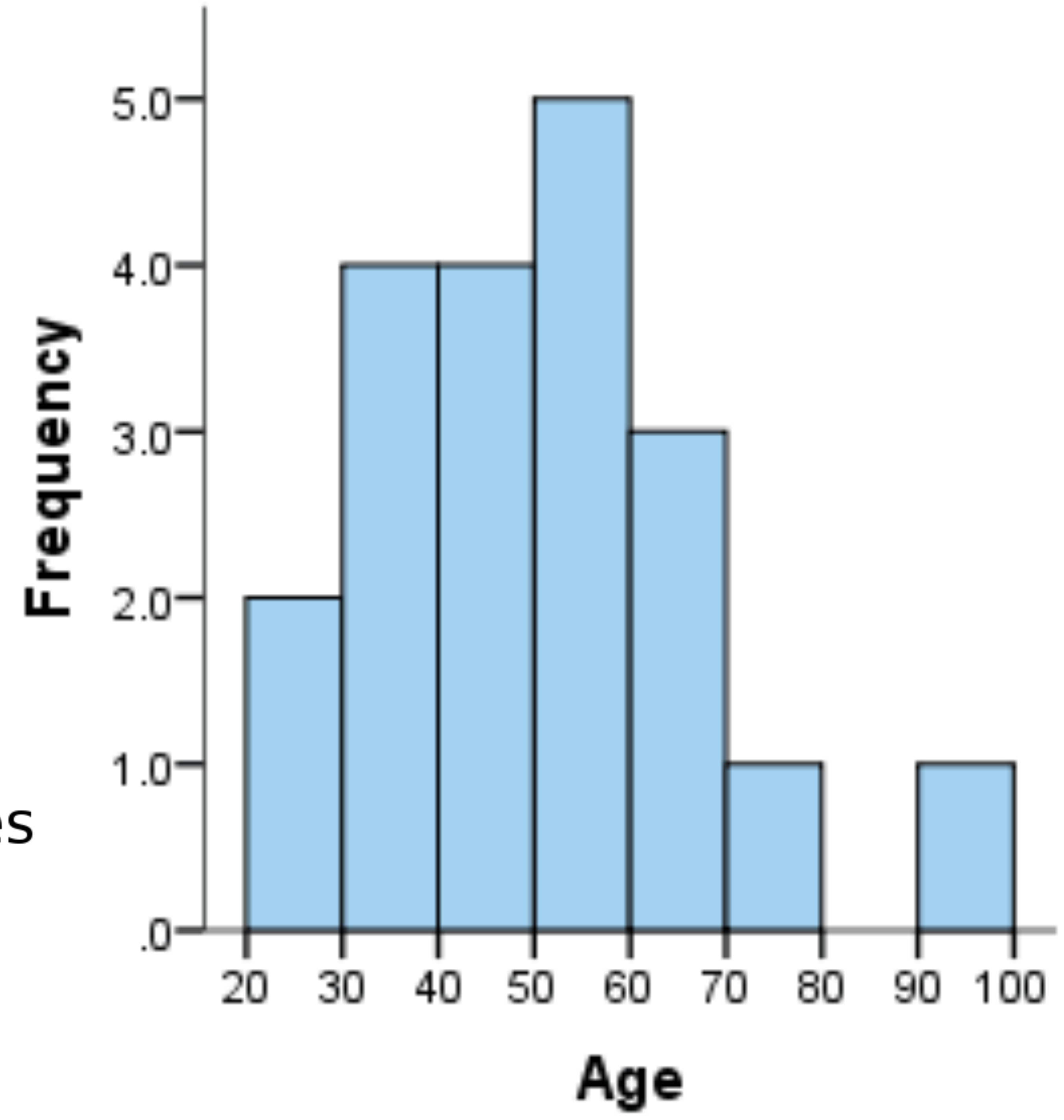


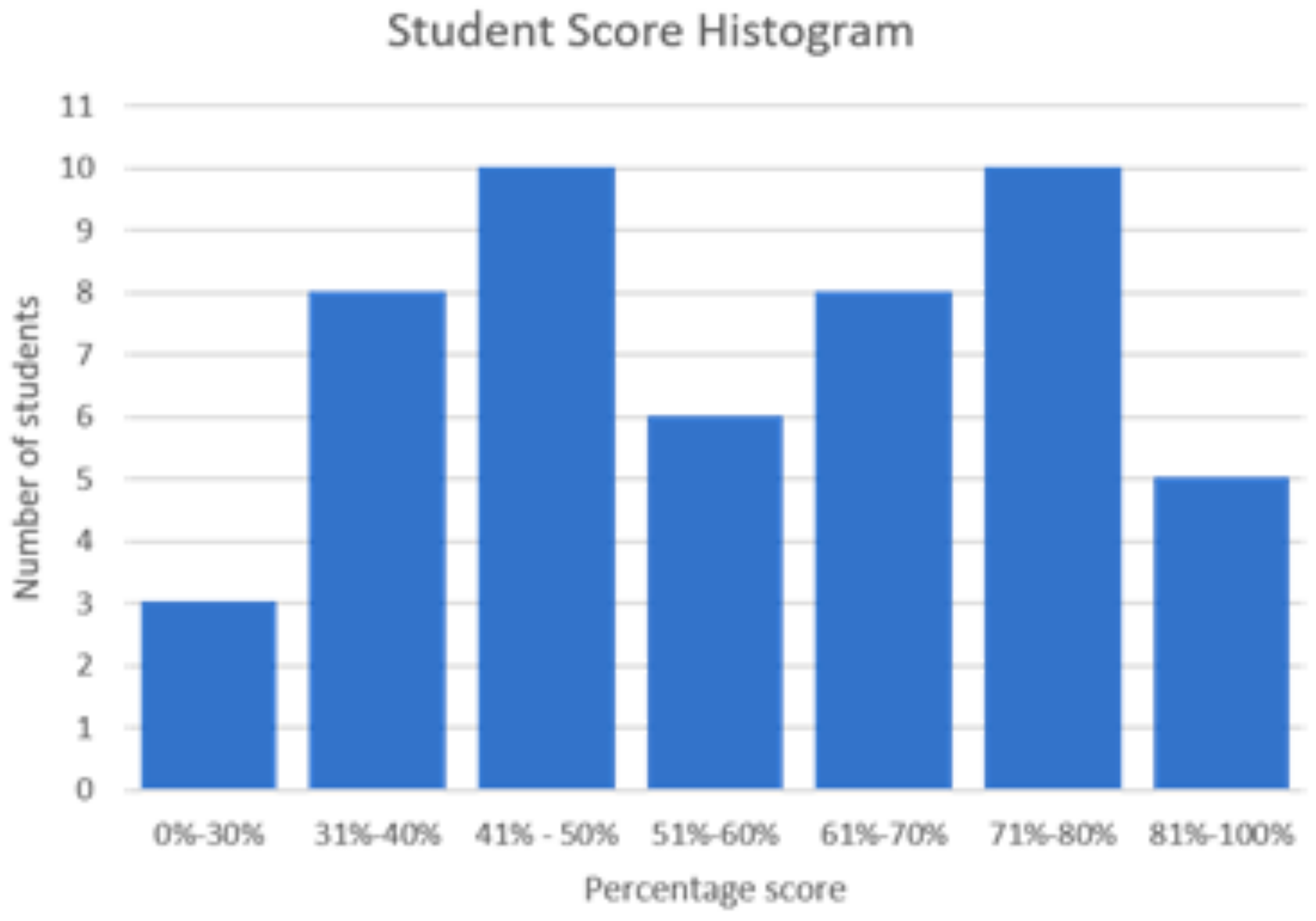
What is missing from all these Histograms?

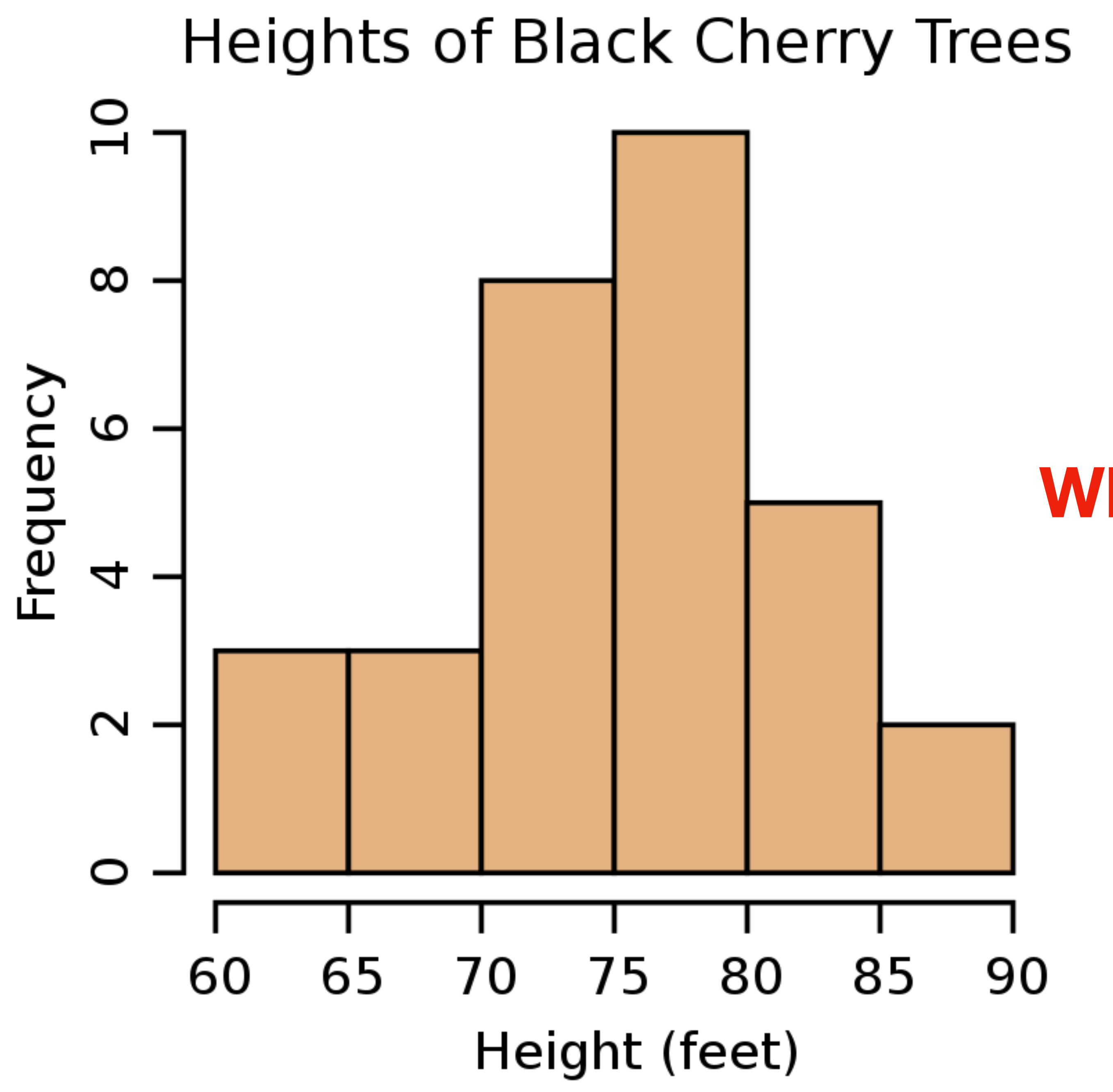
Histograms

- **When you measure some quantity over and over again one way to visualize it is with a histogram**

- **For example:**
  - Survey 20 people and make a histogram of their ages
  - Measure the height of 31 cherry trees in an orchard
  - After a test, your teacher might give you a histogram showing the distribution of scores on the test







What is missing from all these Histograms?

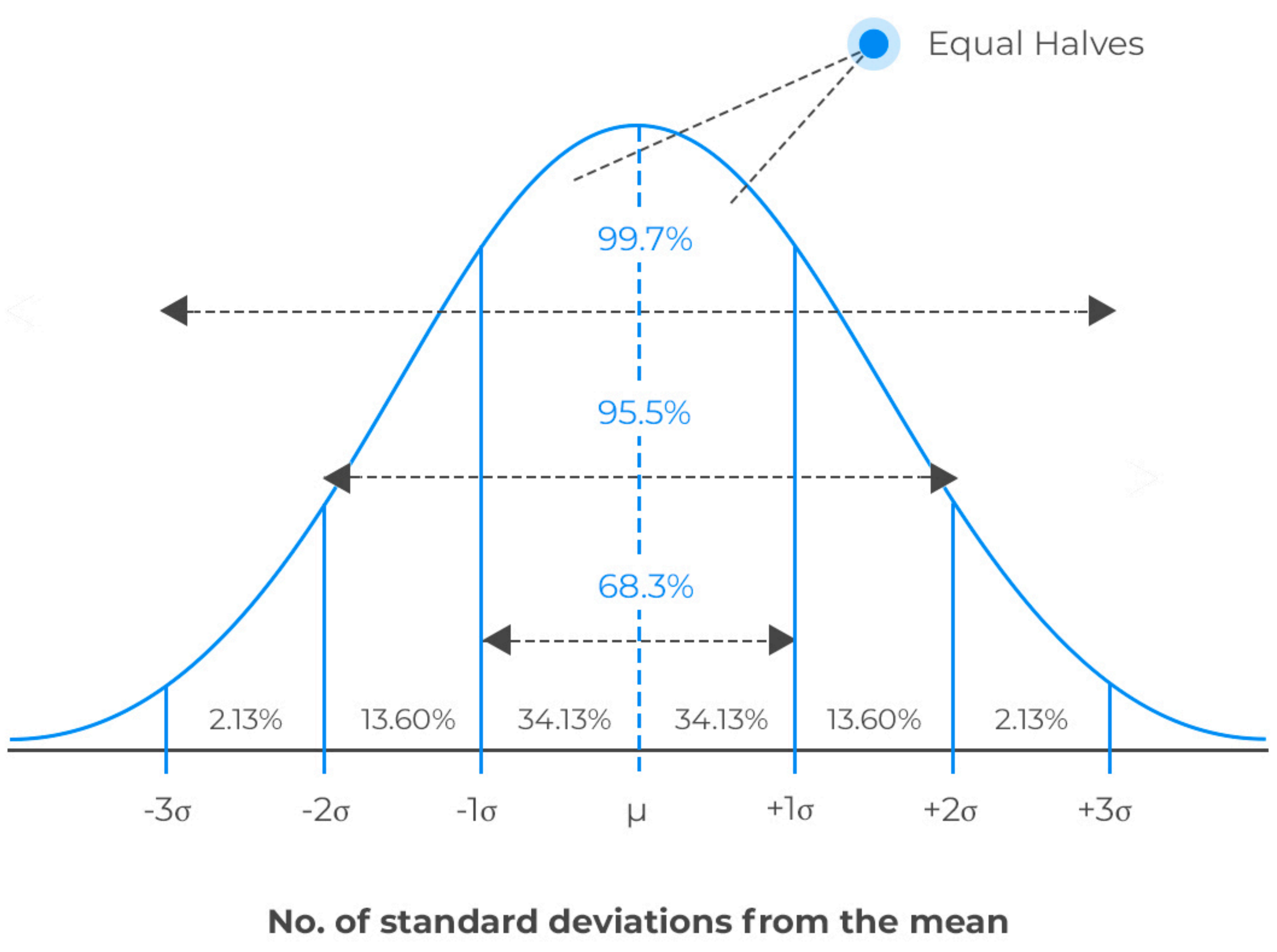
Error bars!!!!

Uncertainty

- **Scientists need to estimate the uncertainty on their measurements.**

- **In particular physics "discovery" actually has very precise statistical meaning**
  - Probability less than 3×10^{-7} or 5 standard deviations on a gaussian distribution

- **When you count the number of entries (like in a histogram) the uncertainty on the measurement is $\sqrt{N}$ for N measurements**



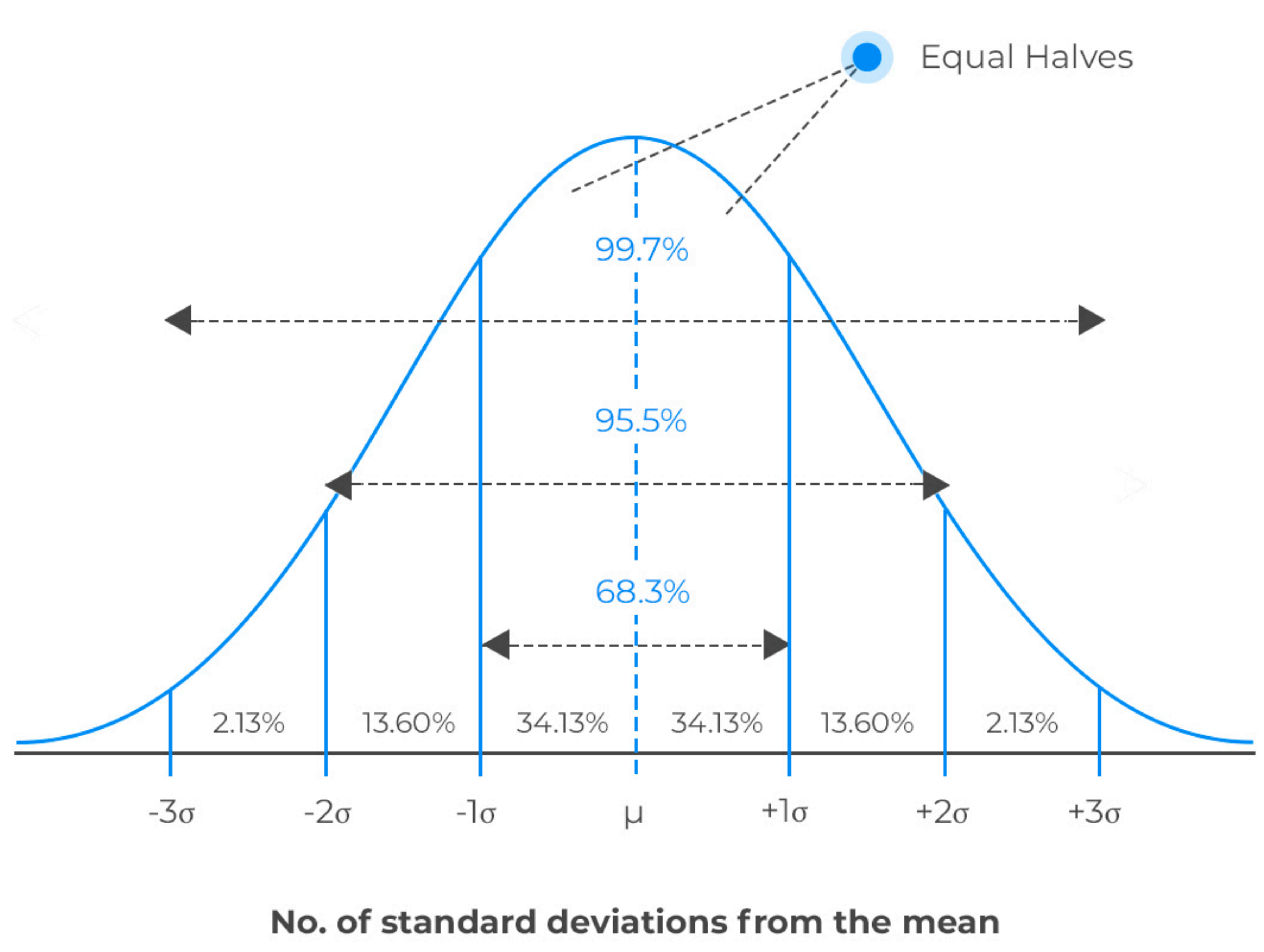
Kevin Nelson

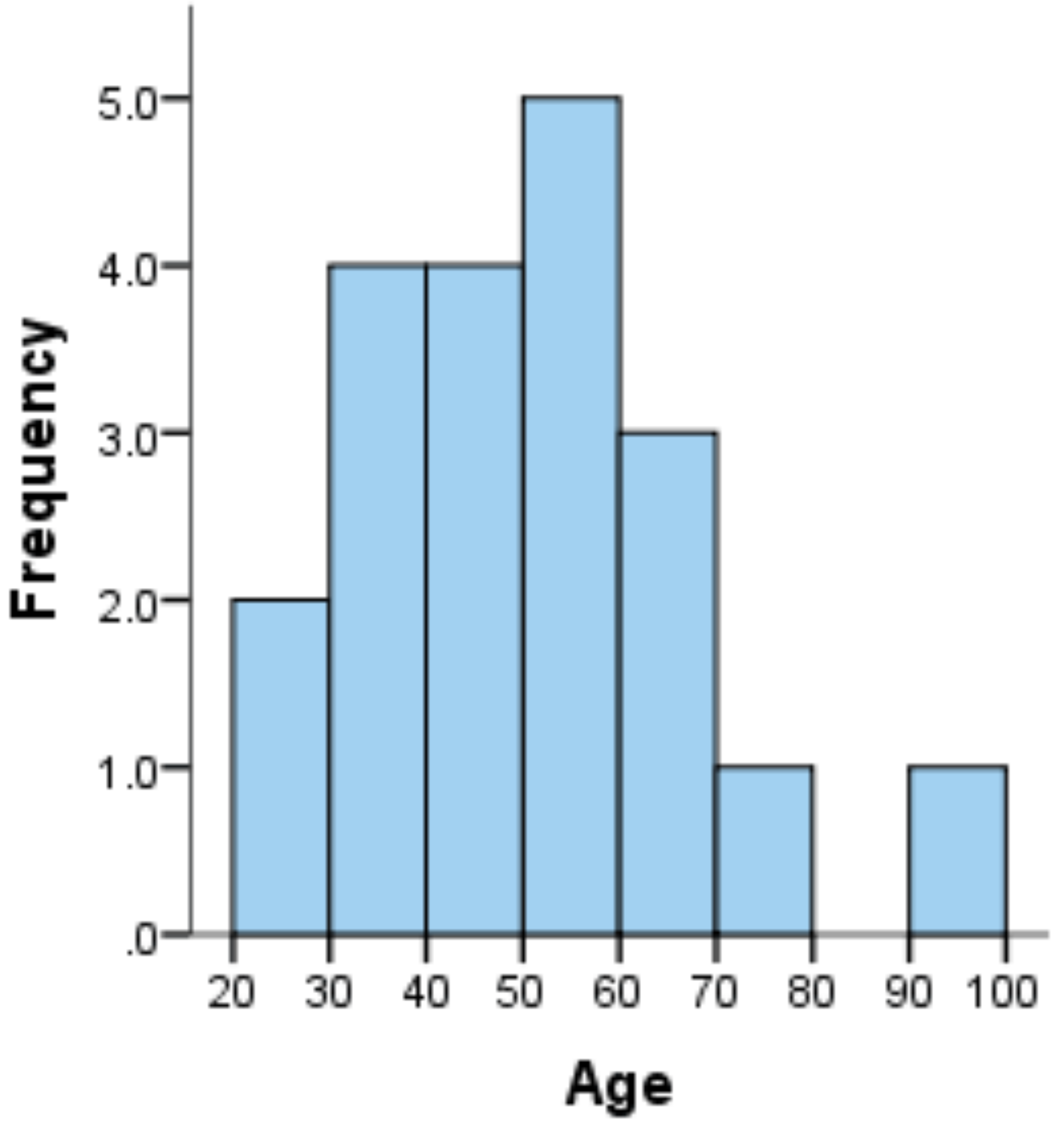
Uncertainty

- **Scientists need to estimate the uncertainty on their measurements.**

- **In particle physics "discovery" actually has very precise statistical meaning**
  - Probability less than 3×10^{-7} or 5 standard deviations on a gaussian distribution

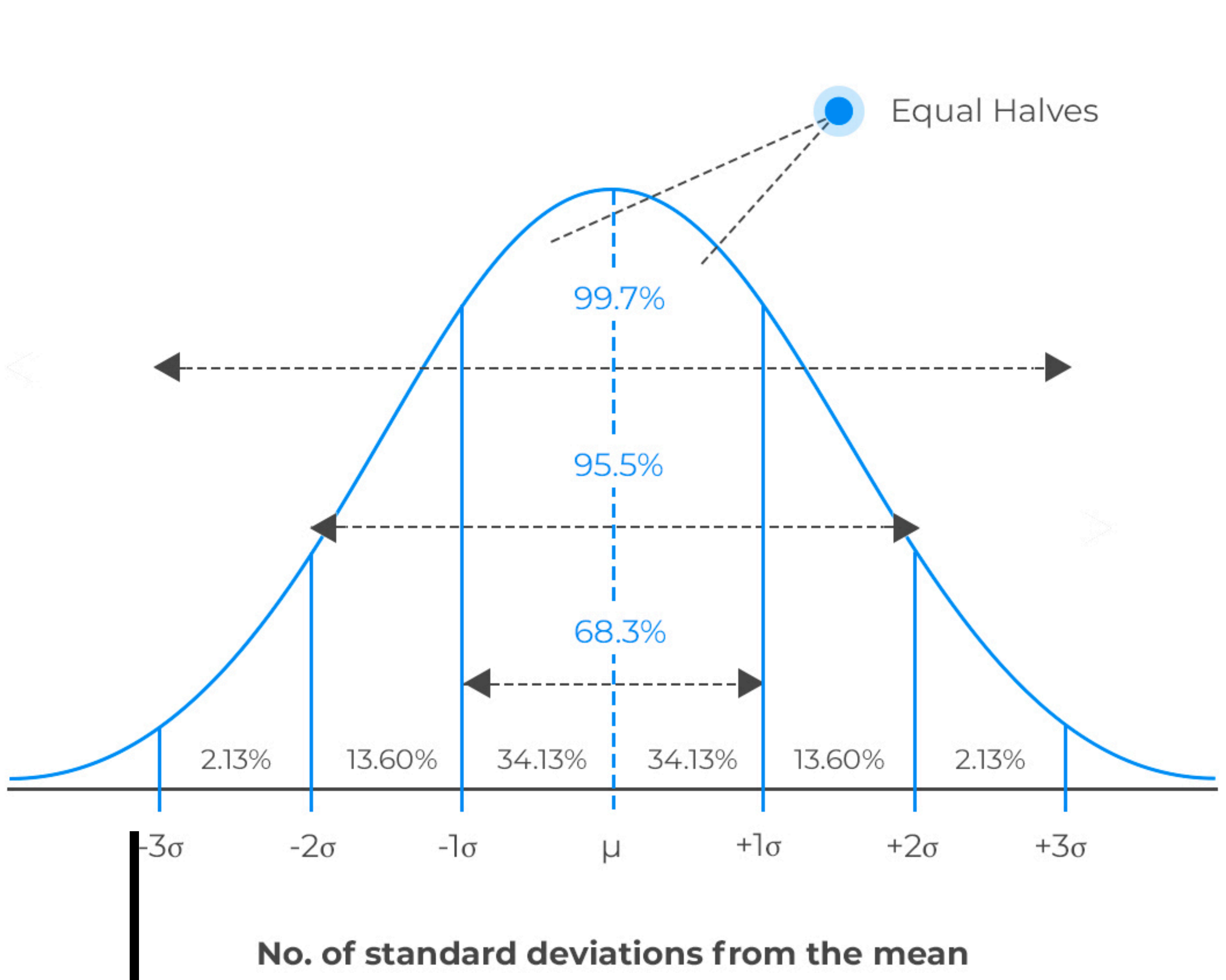
- **When you count the number of entries (like in a histogram) the uncertainty on the measurement is $\sqrt{N}$ for N measurements**

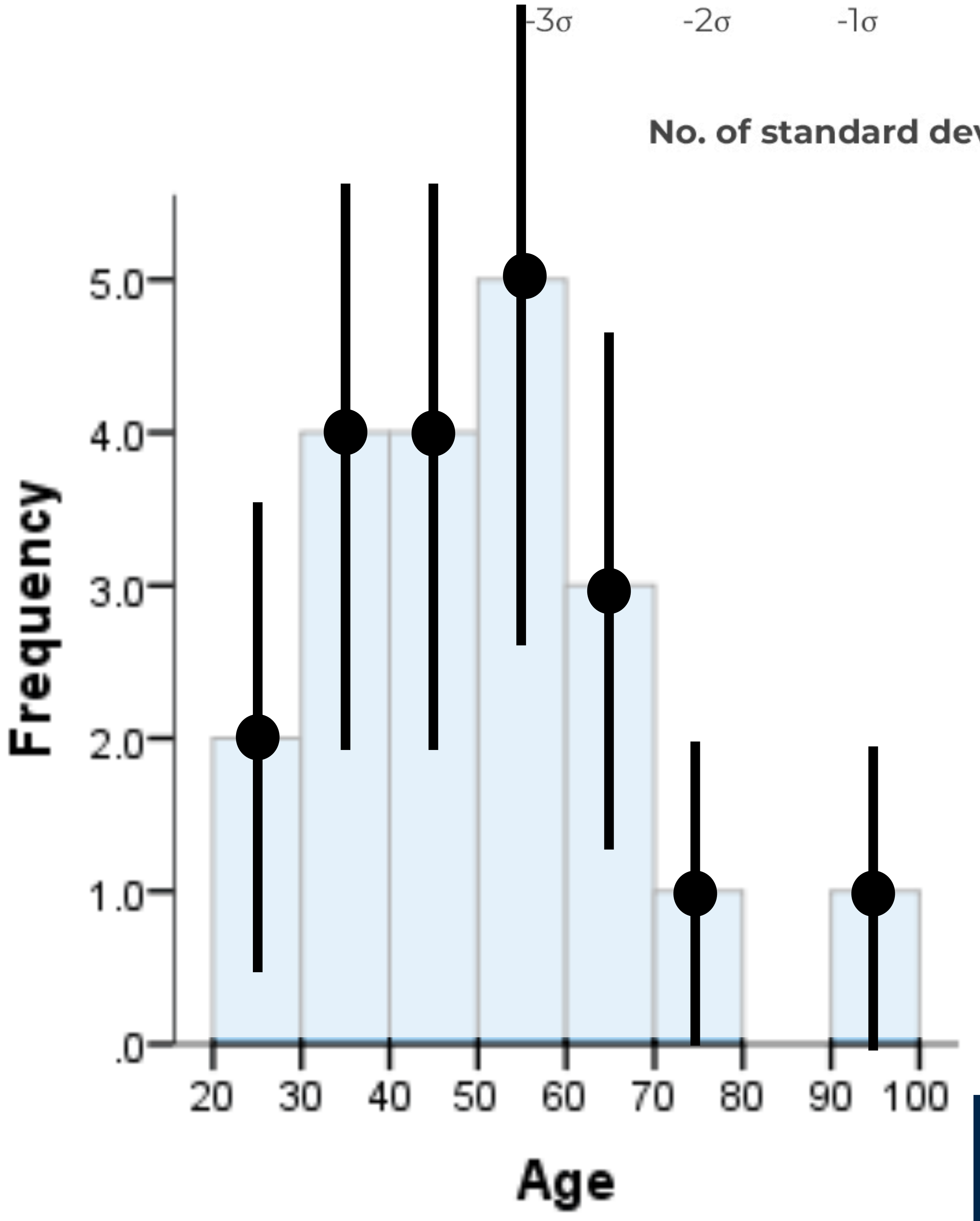




Uncertainty

- **Scientists need to estimate the uncertainty on their measurements.**
- **In particular physics "discovery" actually has very precise statistical meaning**
  - Probability less than 3×10^{-7} or 5 standard deviations on a gaussian distribution
- **When you count the number of entries (like in a histogram) the uncertainty on the measurement is $\sqrt{N}$ for N measurements**
- **We often use circles with lines through them like this to show the data and the error**





Chi-Square

- What if we have a prediction of the data? How do we compare the prediction to the observed data?

$$\chi^2 = \Sigma \left(\frac{Exp - Obs}{\sigma} \right)^2$$

- Chi-Square is a sum over all the data points the value of the expectation minus the observed divided by the error

- If chi-square is low, then there is good agreement between expected and observed $Exp - Obs \to 0$

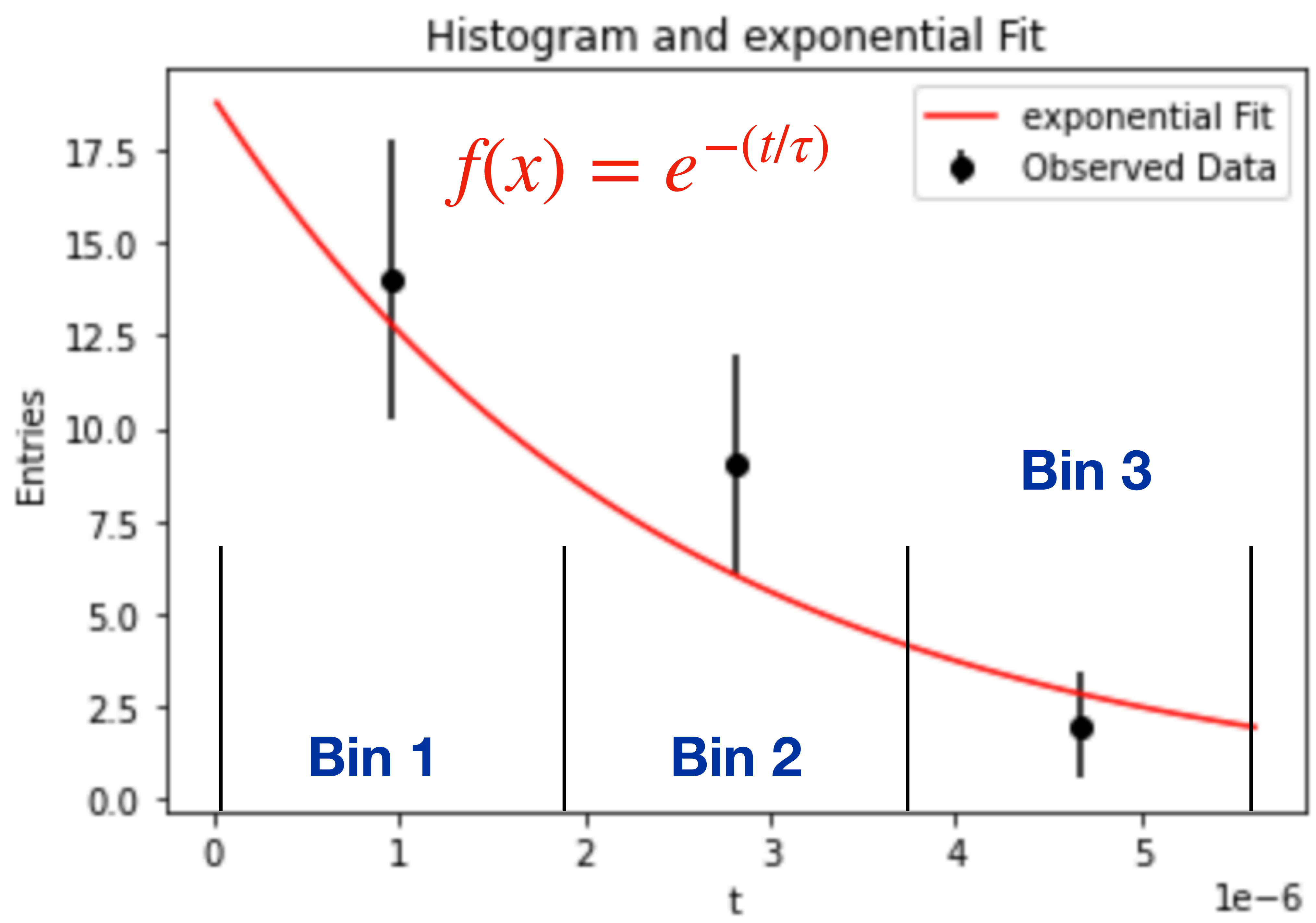
- If chi-square is large, then there is bad agreement between expected and observed $Exp - Obs >> 0$

- Remember, σ is the error on the observation: $\sqrt{N}$

Example: Calculating Chi-Square

- **We have a histogram with 3 bins. We count the number of observed entries in each bin and plot the data with error bars**

- **The theoretical prediction is an exponential decay curve. We plot the theoretical prediction which is a smooth curve**

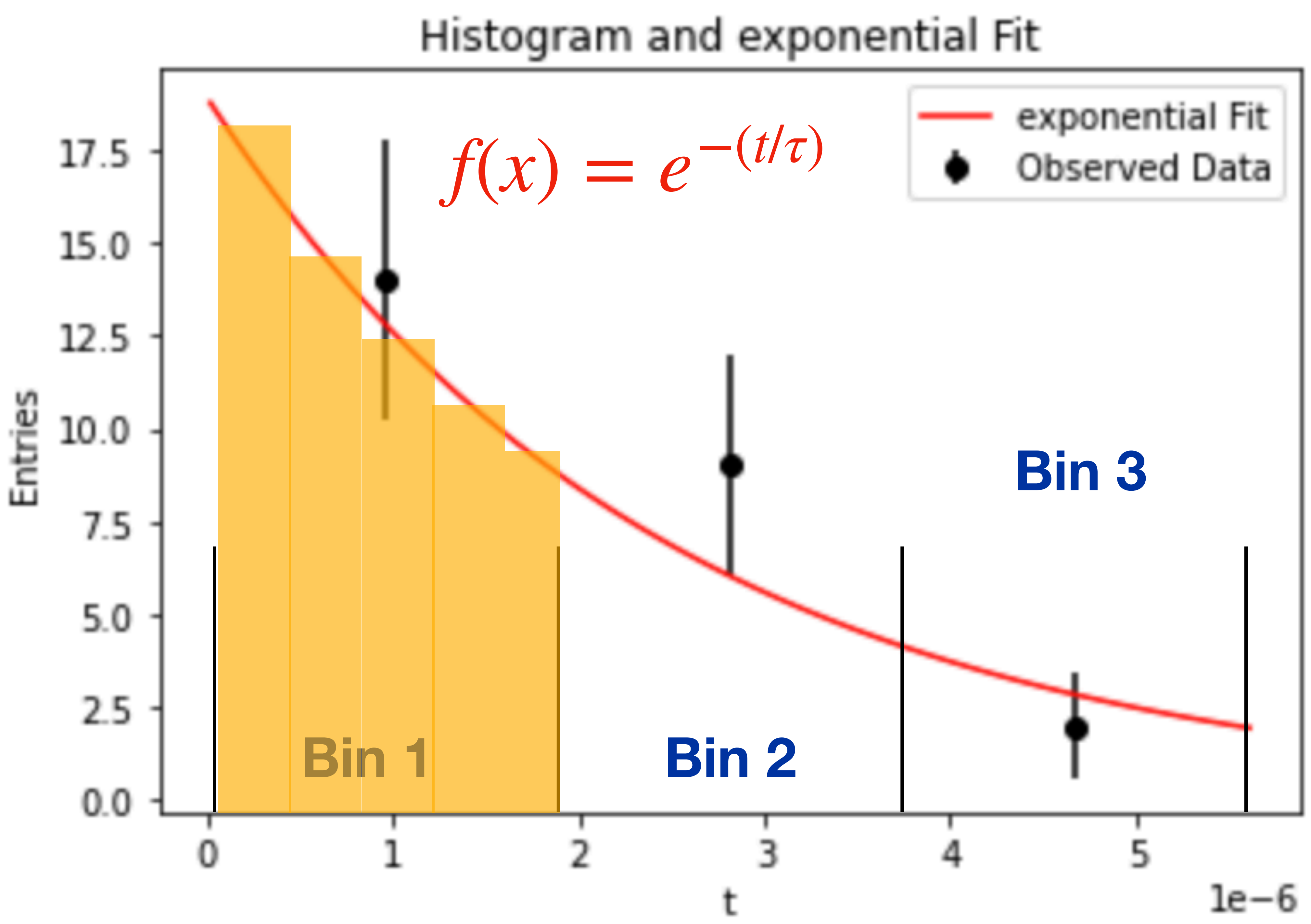


Example: Calculating Chi-Square

- **We have a histogram with 3 bins. We count the number of observed entries in each bin and plot the data with error bars**

- **The theoretical prediction is an exponential decay curve. We plot the theoretical prediction which is a smooth curve**

- **To get the expected number of entries in each bin, we calculate the integral in each bin**



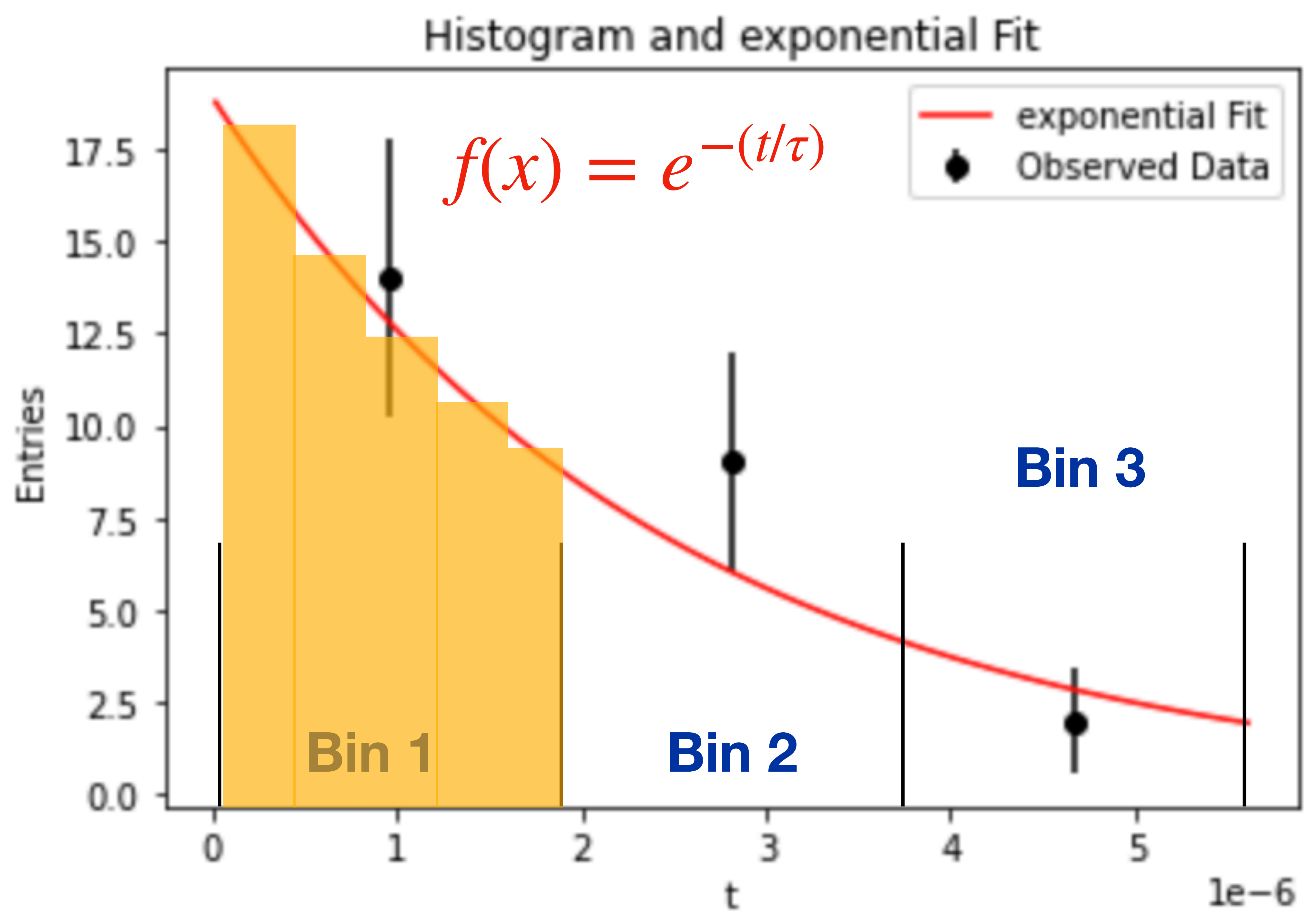
Approximate the integral as a Reimann sum

Example: Calculating Chi-Square

- We have a histogram with 3 bins. We count the number of observed entries in each bin and plot the data with error bars
- The theoretical prediction is an exponential decay curve. We plot the theoretical prediction which is a smooth curve
- To get the expected number of entries in each bin, we calculate the integral in each bin

$$\chi^2 = \Sigma \left(\frac{Exp - Obs}{\sigma} \right)^2$$

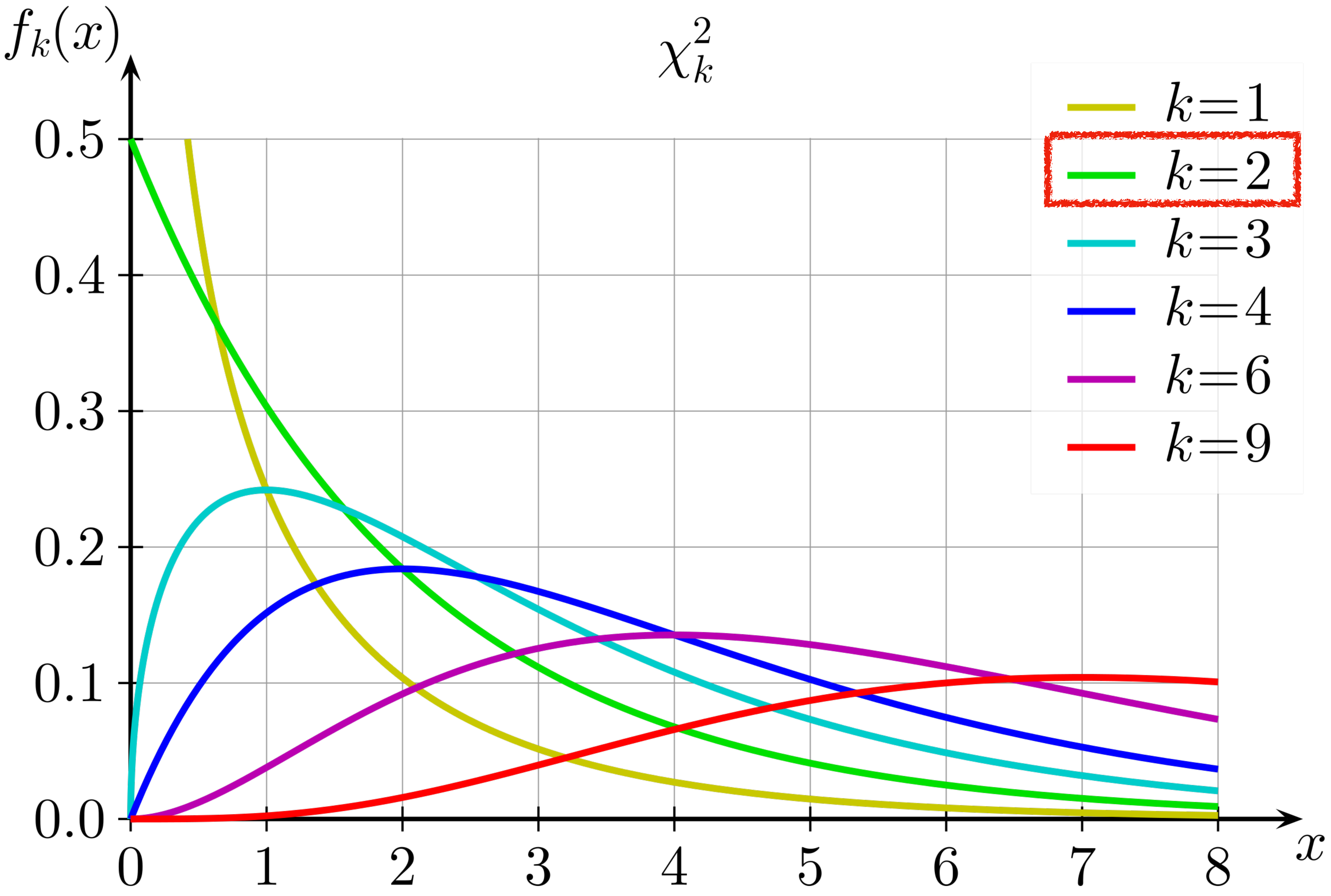
- Then add up the contribution for each bin
- In this case, $\chi^2 = 1.14$



Approximate the integral as a Reimann sum

Chi-Square Distribution

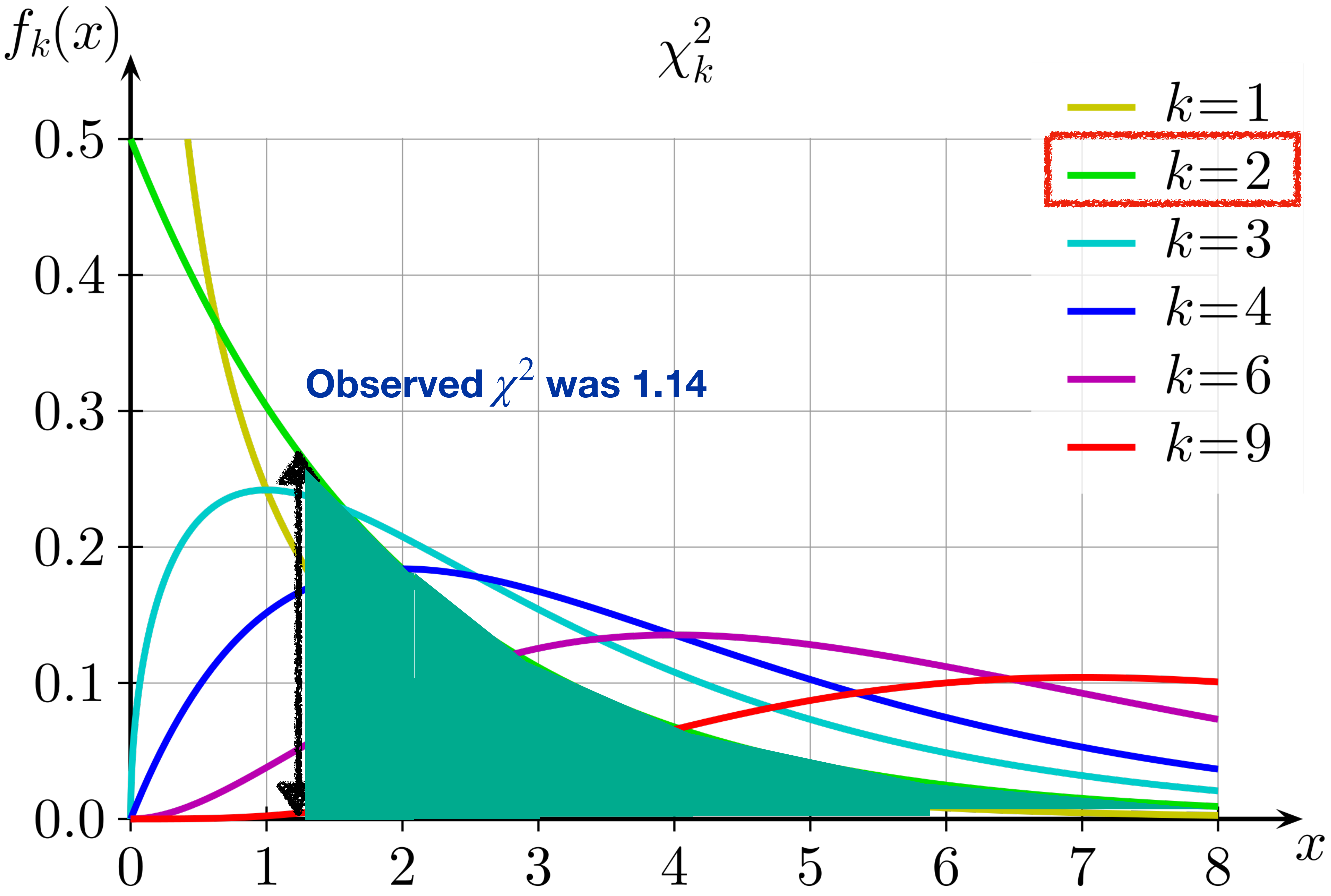
- **Why is computing the chi-square for a set of data and the prediction useful?**
- **We check the observed chi-square against the predicted distribution with the right number of degrees of freedom.**
- **In the previous example, there were 2 degrees of freedom (3 bins - 1 free parameter)**



Chi-Square Distribution

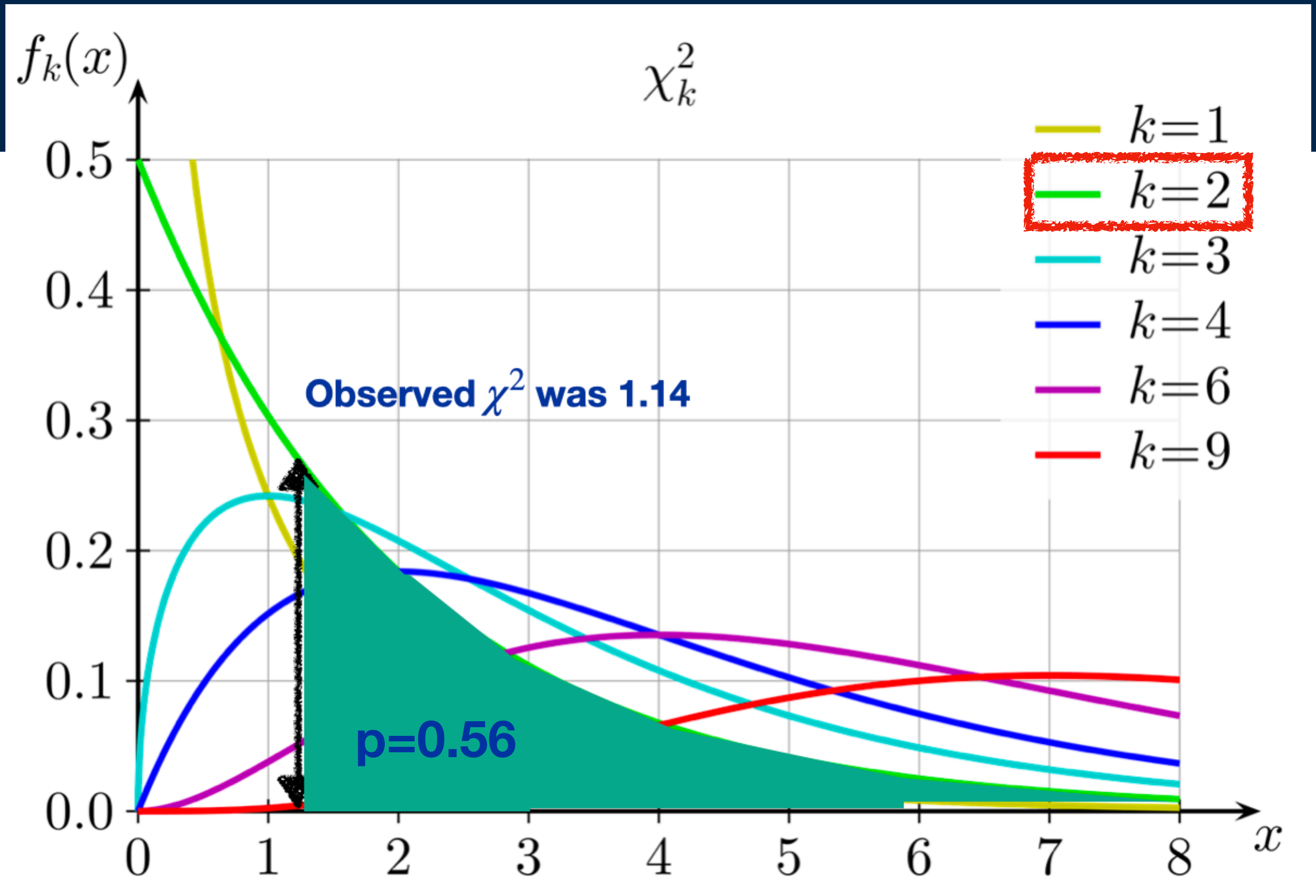
- Why is computing the chi-square for a set of data and the prediction useful?
- We check the observed chi-square against the predicted distribution with the right number of degrees of freedom.
- In the previous example, there were 2 degrees of freedom (3 bins - 1 free parameter)
- The probability of observing data in this level of agreement with the theoretical prediction is the integral

$$\int_{\chi^2}^{\infty} f_k(x)dx = 0.56$$



Interpreting p-values

- The integral described is a χ^2 probability or p-value
- A p-value greater than 0.05 is “good agreement”
- If the p-value is very small, we discover something new
- What if we can get better agreement? We could try changing the parameter and recomputing χ^2

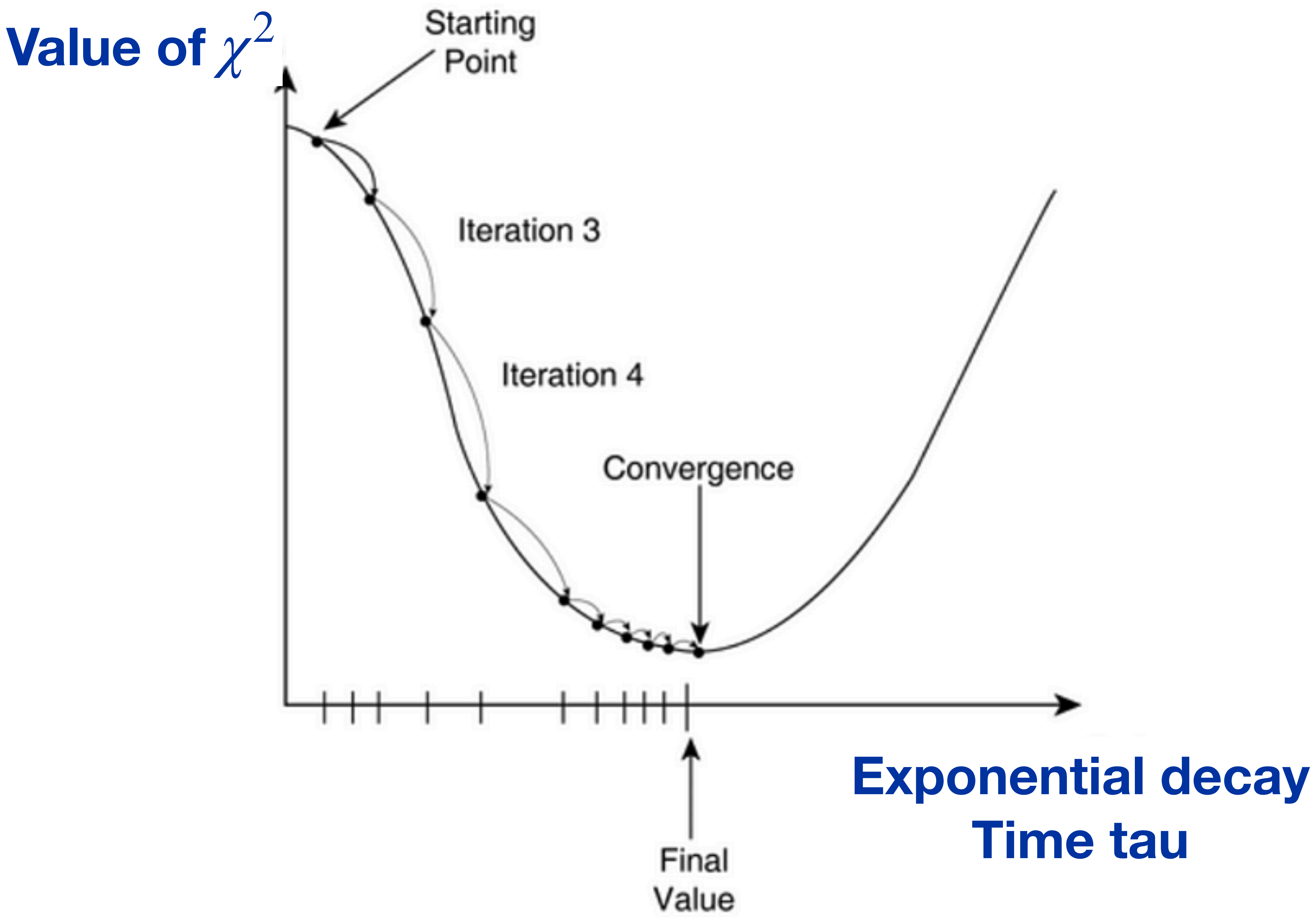


Fitting

- The theoretical description has *parameters* that need to be fit

- We take small steps in each direction to see if the value of χ^2 gets better or worse

- If we find a place where the value of χ^2 doesn't change, then we are at the local minimum



Example: Calculating Chi-Square

- We have a histogram with 3 bins. We count the number of observed entries in each bin and plot the data with error bars
- The theoretical prediction is an exponential decay curve. We plot the theoretical prediction which is a smooth curve
- We try increasing or decreasing the length scale tau. This is represented by the blue and green dashed lines.
- We are done fitting when changing the parameter up or down gives us a larger value of χ^2