

Lecture 1: Chi-Squared & Some Basics

D. Jason Koskinen
koskinen@nbi.ku.dk

Advanced Methods in Applied Statistics
Feb - Apr 2024

Variance

- Because it's something we all should know

$$\sigma^2 \equiv \langle (X - \mu)^2 \rangle$$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

σ^2 is the variance

μ is the expected value, which can sometimes also be the mean

N is the number of data points

x_i is the individual observed data points

Unbiased Variance

- Just because it's something we all should know

$$S_{N-1} \equiv \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2$$

S_{N-1} is the 'unbiased' estimator of the variance

\bar{x} is the mean calculated from the data itself

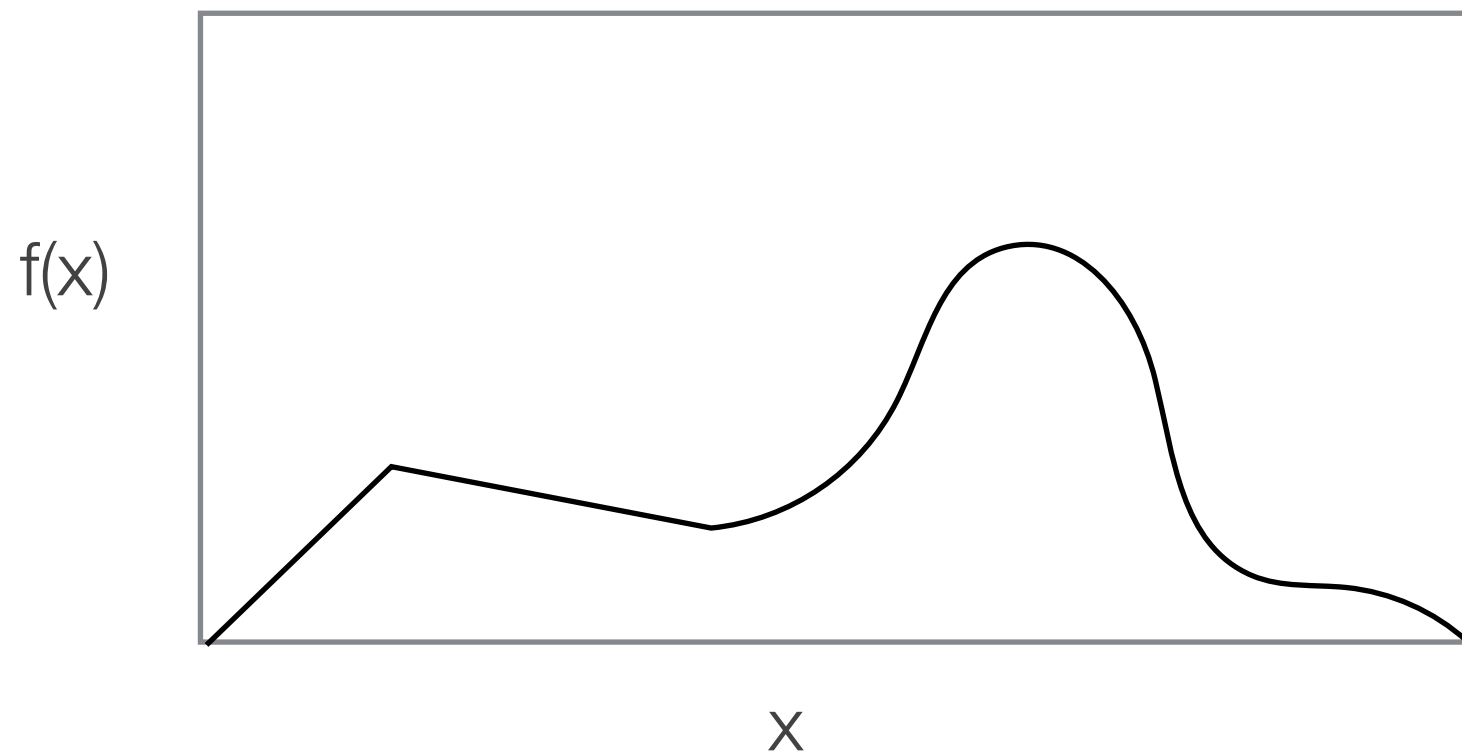
N is the number of data points

x_i is the individual observed data points

For further information on $1/(N-1)$ see Bessel's correction [wikipedia](https://en.wikipedia.org/wiki/Bessel%27s_correction)

Probability Distribution Function

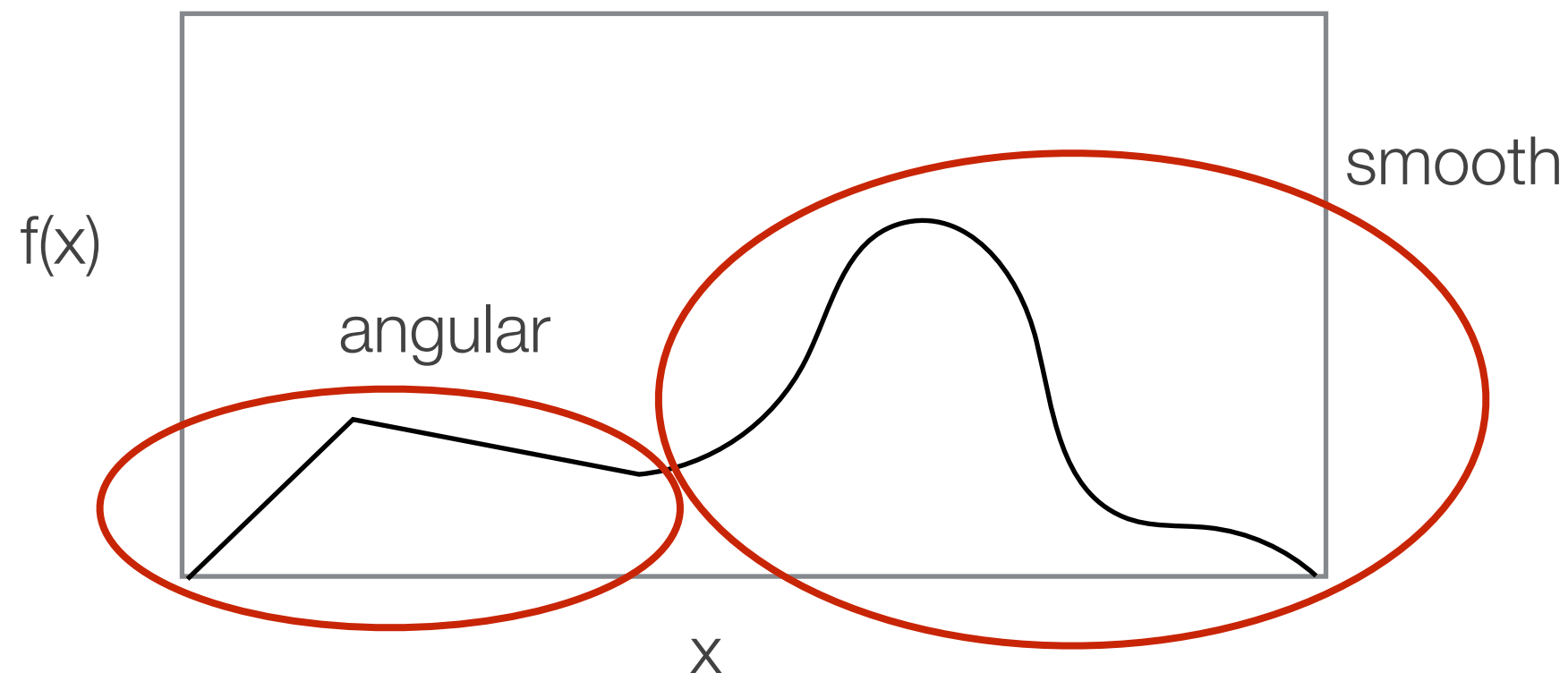
- **Probability Distribution Functions (PDF)**, where sometimes the "D" is density, is the probability of an outcome or value given a certain variable range



- The PDF does not have to be nicely described by a single continuous equation

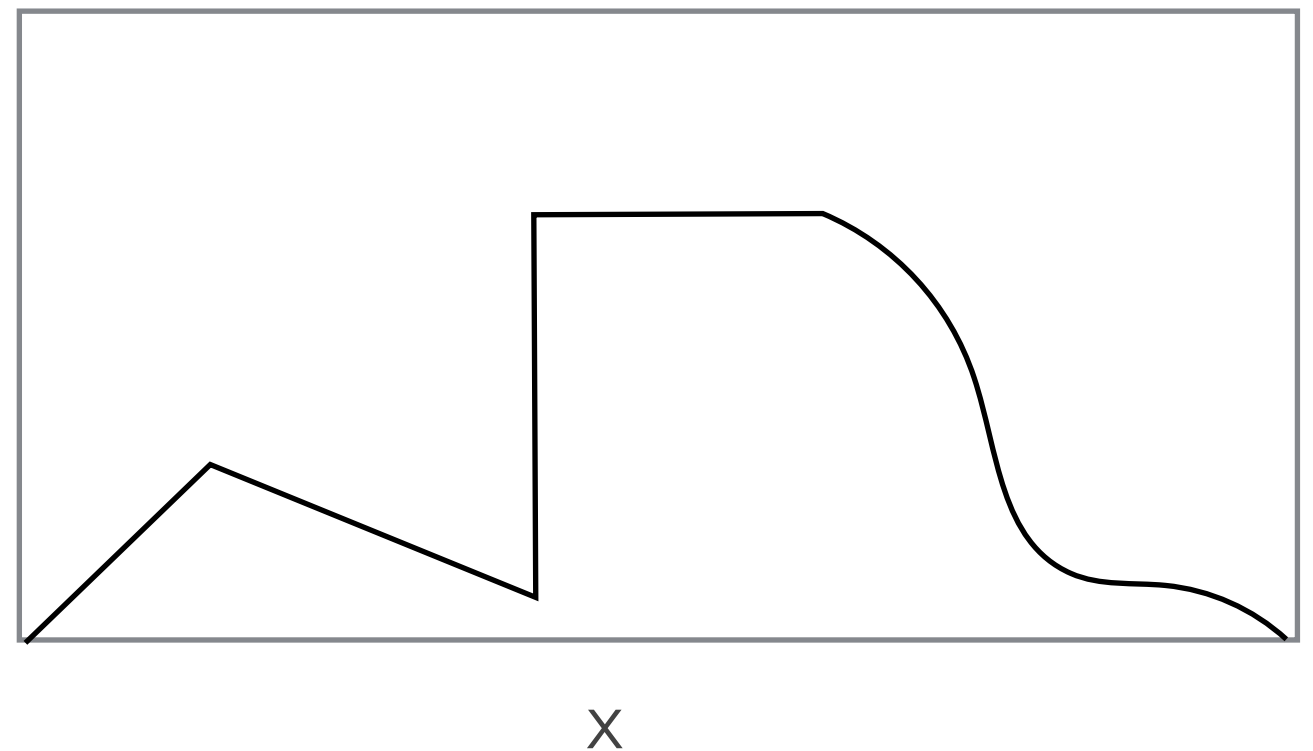
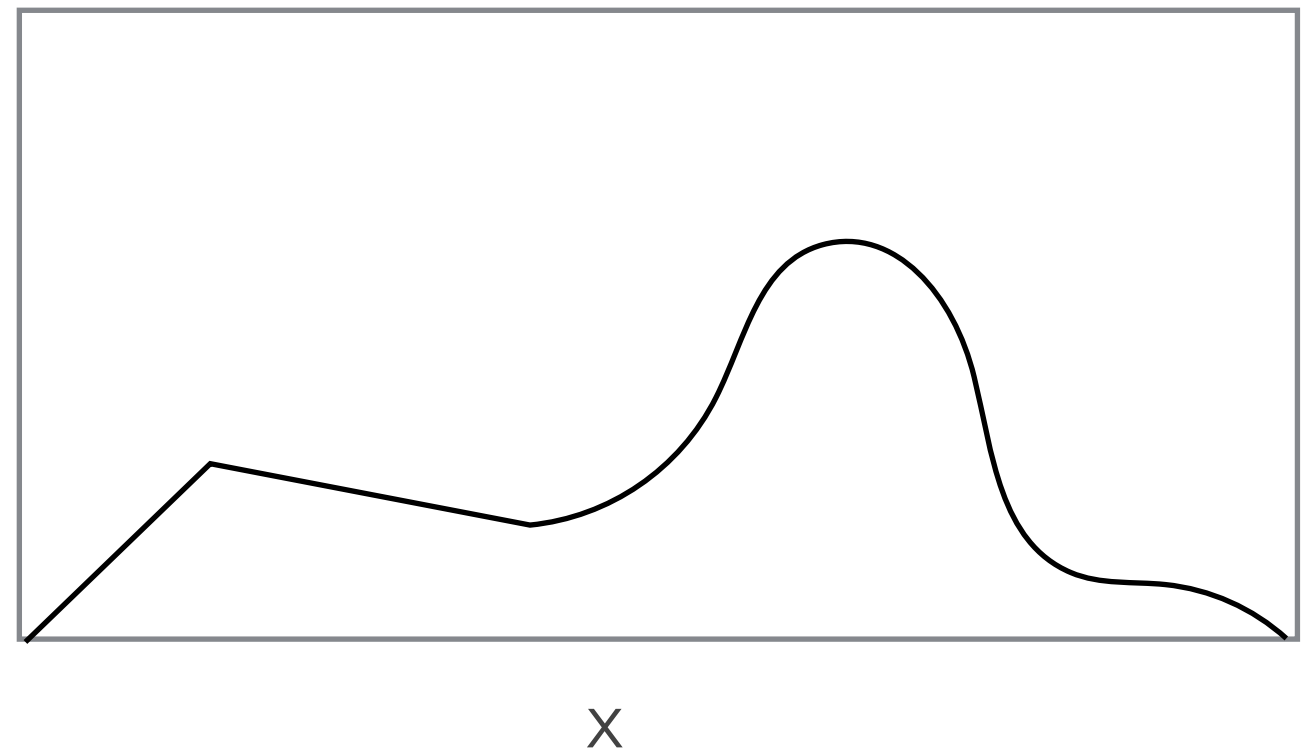
Probability Distribution Function

- The PDF does not have to be nicely described w/ equations, and sometimes cannot be



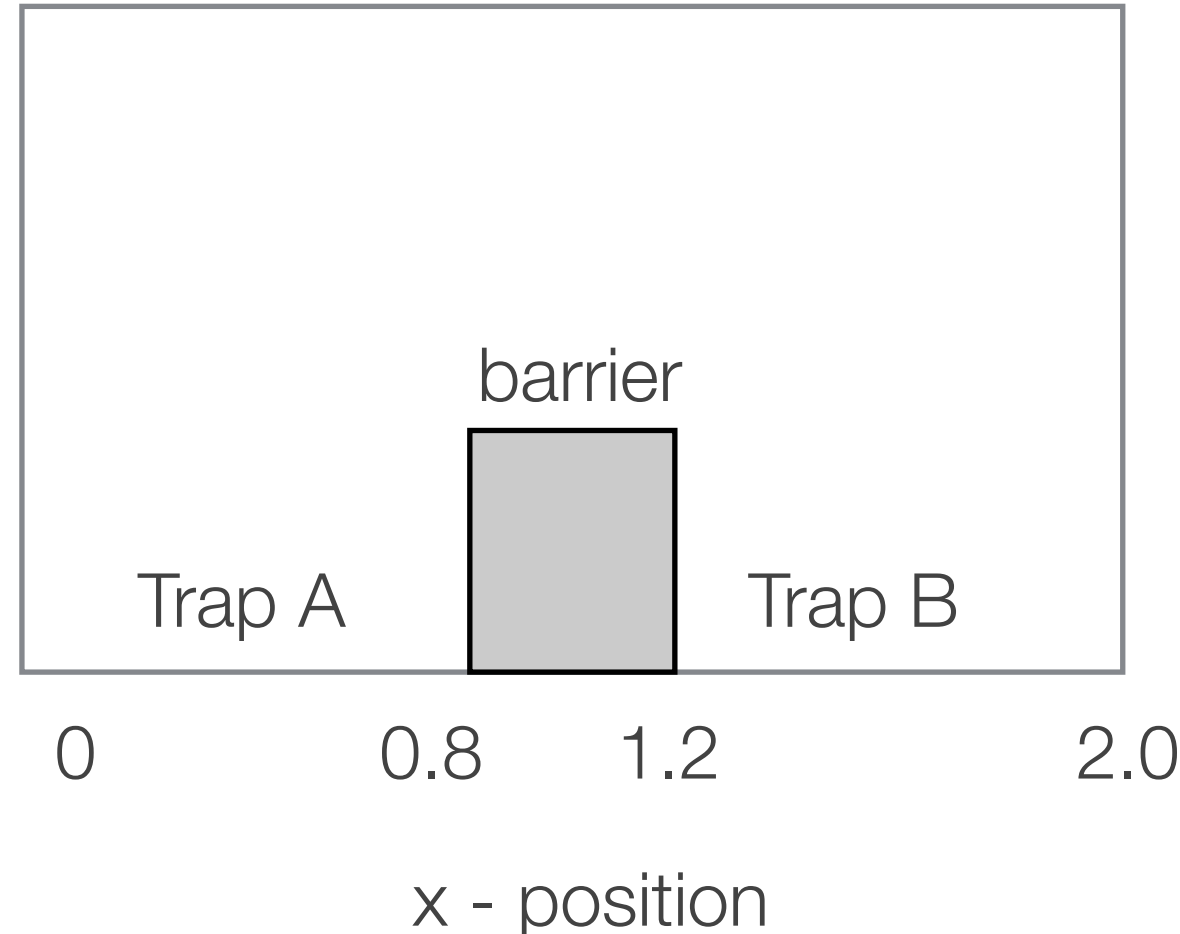
PDFs

- They can be discrete, $f(x)$ continuous, or a combination
- They often have an implied conditionality
 - “What is the energy of an outgoing electron from nuclear beta-decay?” implies beta-decay
 - PDF must always be normalized to **one**



PDF Possibility

- Let's imagine an experiment which has two identical electron traps (A & B) separated by a finite barrier. An electron w/ energy below the barrier threshold is deposited in trap A. Sketch out the PDF of the x position after a very short time.

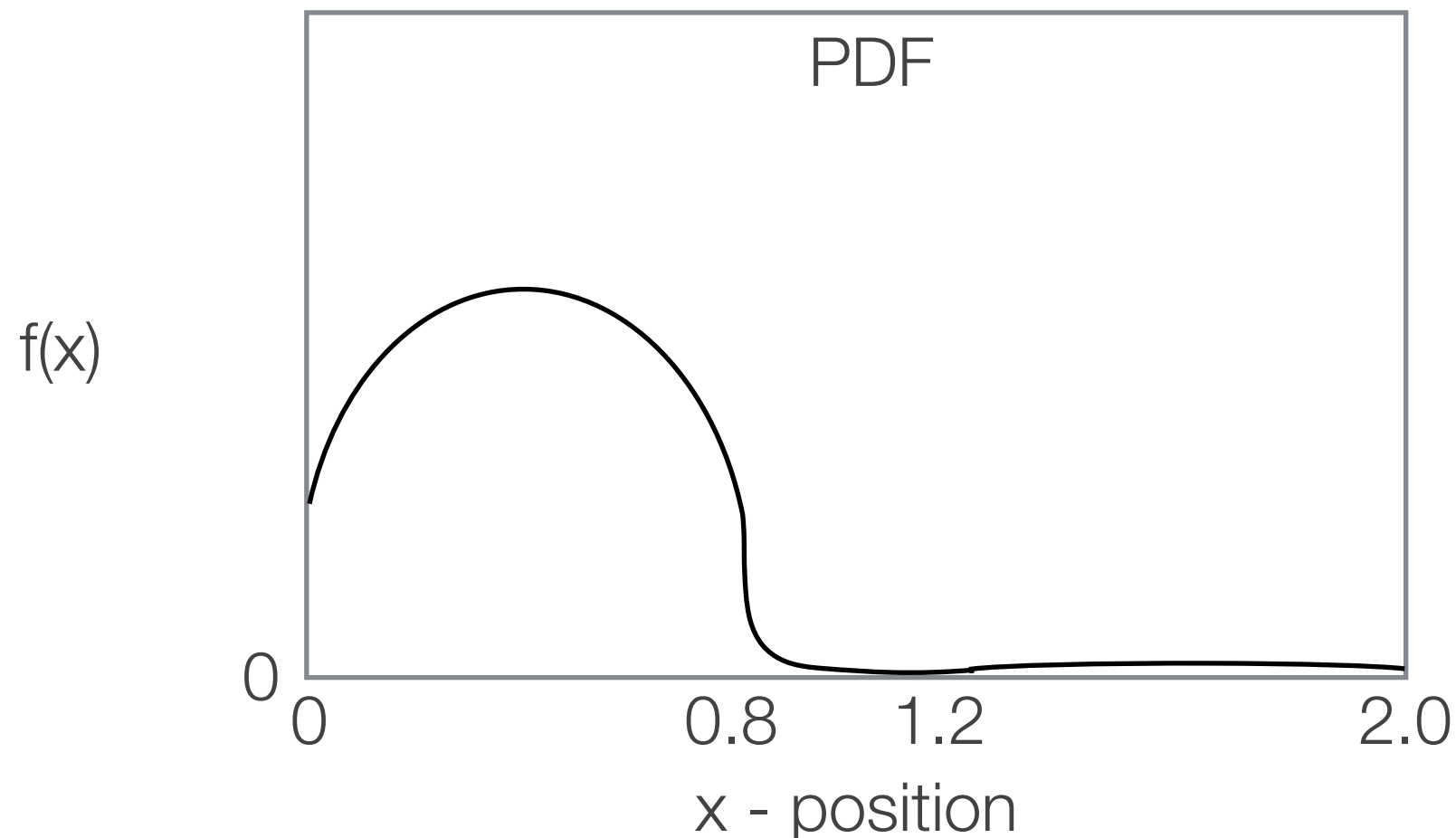
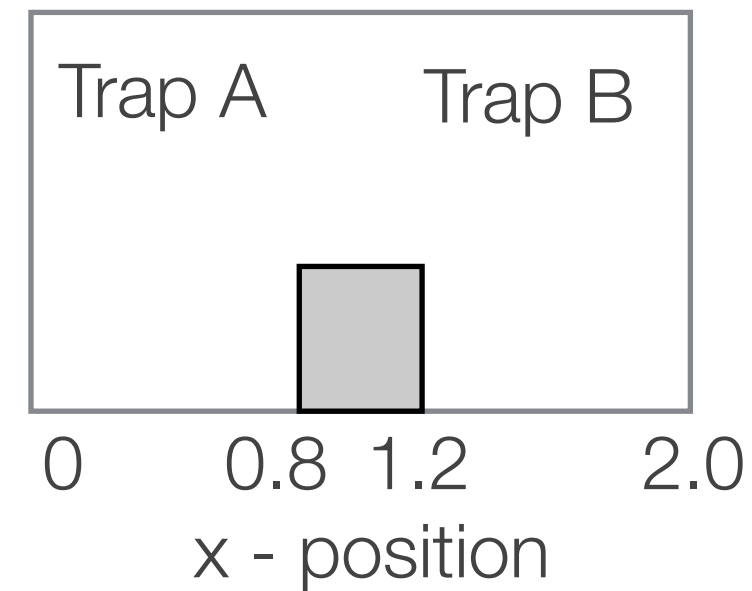


$$\text{time} \approx \frac{1}{\infty}$$

*rough sketch,
don't take it too literal

PDF Possibility

- Sketch out the PDF of the x position after a very short time.
- My trap has a potential which keeps it mostly in the middle of the trap, and it's mostly in trap A because it hasn't had time to tunnel.



$$\text{time} \approx \frac{1}{\infty}$$

*rough sketch,
don't take it too literal

PDF Possibility

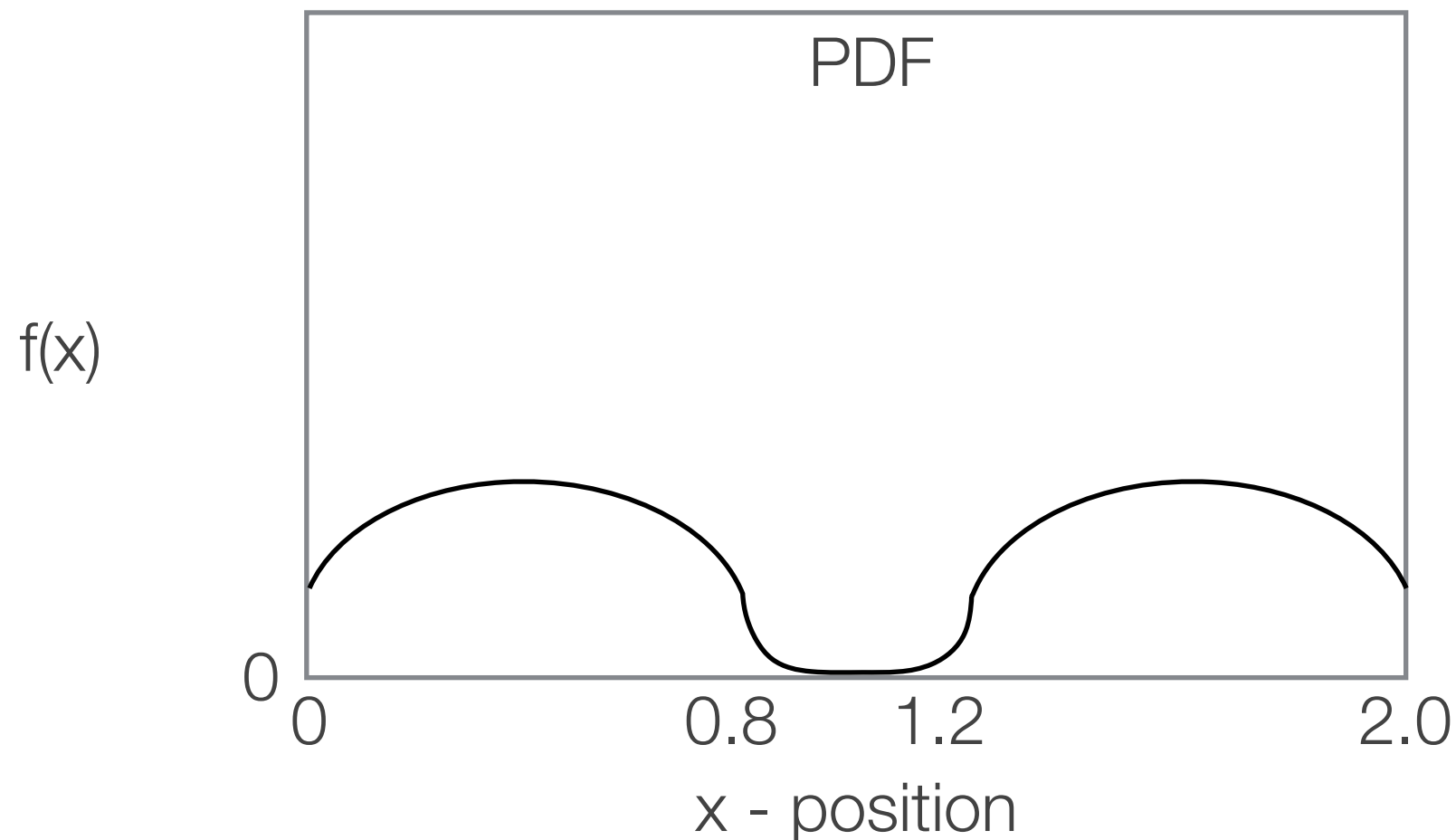
- Sketch out the PDF of the x position after a near infinitely long time.

time $\approx \infty$

PDF Possibility

- Sketch out the PDF of the x position after a near infinitely long time.
 - Same distribution shape as before, but now the probability of being in trap A and trap B are equal.
 - The PDF is still normalized

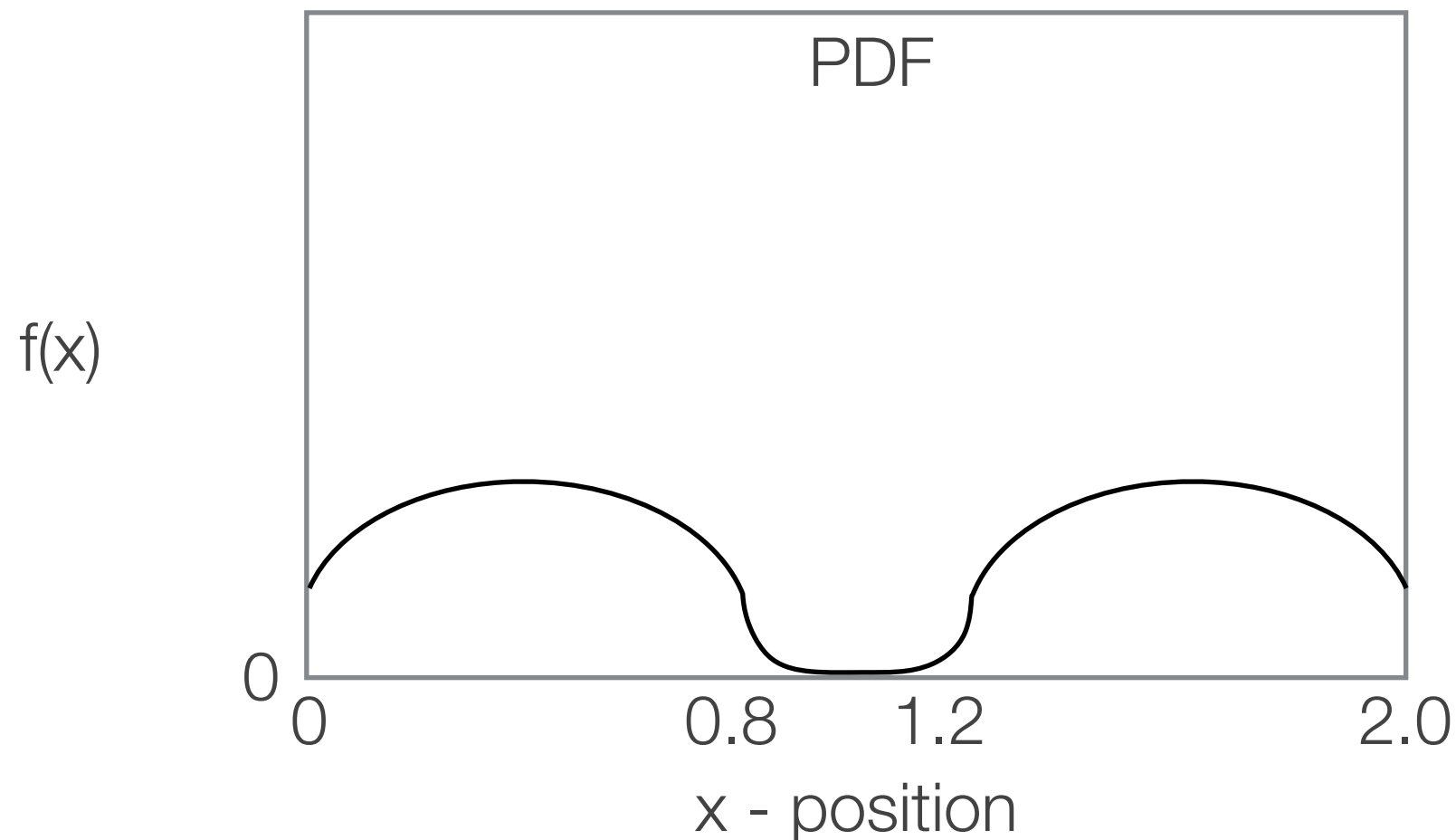
time $\approx \infty$



PDF Possibility

- Notice that there are discontinuities in the PDF, which is not uncommon in experimental PDFs due to boundary conditions. How many discontinuities as a function of x ?

time $\approx \infty$



Some PDF Remarks

- Previous examples are univariate PDFs, i.e. probability only as a function of a single variable (x), but the PDF comes from a multivariate situation
 - Multivariate, because the PDF doesn't just depend on x , but also the time of the measurement, energy of the electron, barrier height, etc.
 - We'll stick with univariate (or at least 1-dimensional unchanging PDFs) initially, before moving onto more complex situations later in the course
- Probability distribution functions can be used to not only derive the most likely outcome, but having recorded the outcome, figure out the mostly likely situation. For example, if we record a single electron at a position in trap B, it is more likely that the data was taken at $t = \infty$ versus $t = 1/\infty$

Cumulative Distribution Function

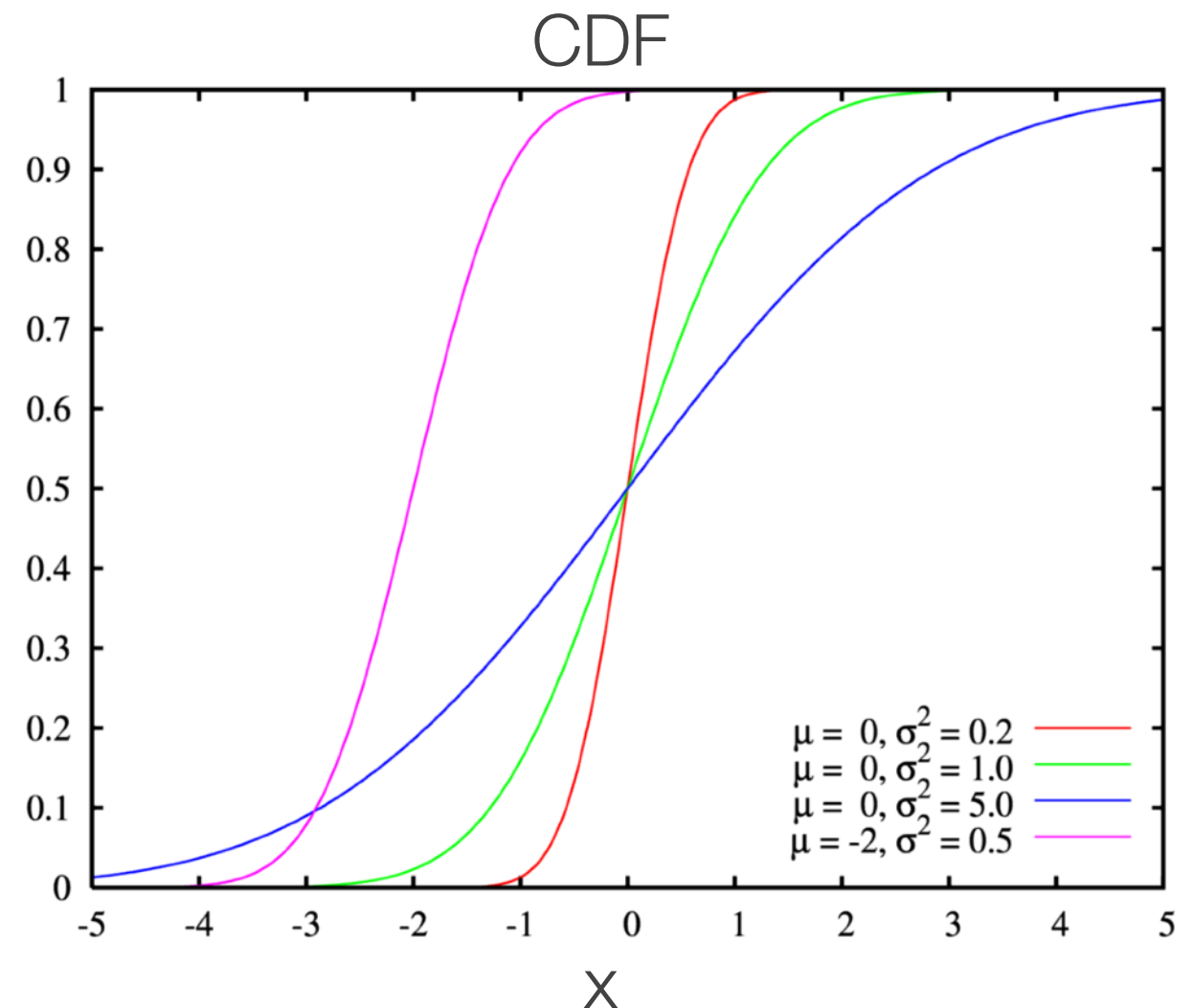
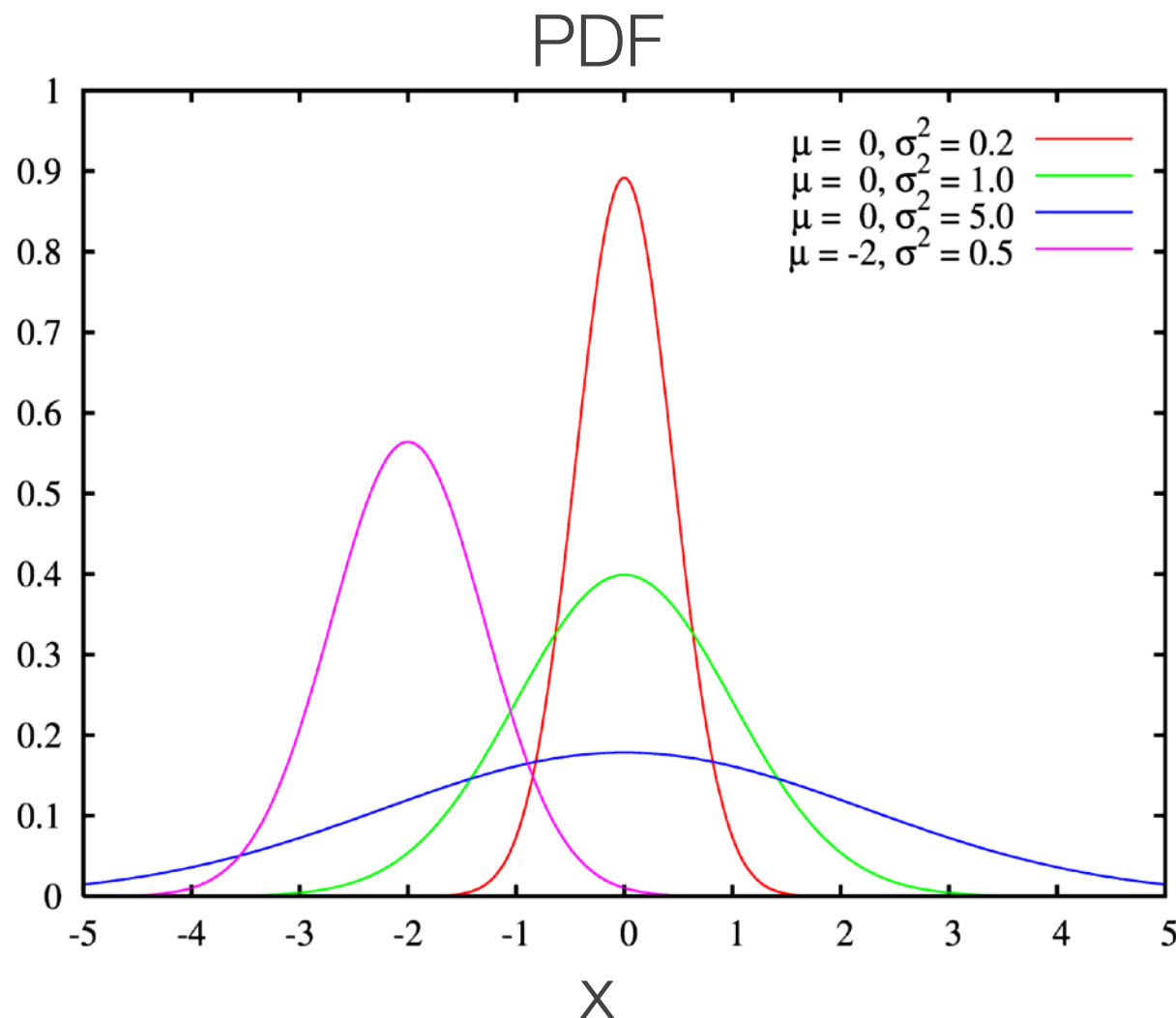
- The Cumulative Distribution Function (CDF) is related to the PDF and gives the probability that a variable (x) is less than some value x_0
- Basically, the integral or sum from $-\infty$ to x_0

$$CDF(x_0) = F(x_0) = \int_{-\infty}^{x_0} f(x)dx$$

where $f(x)$ is the PDF

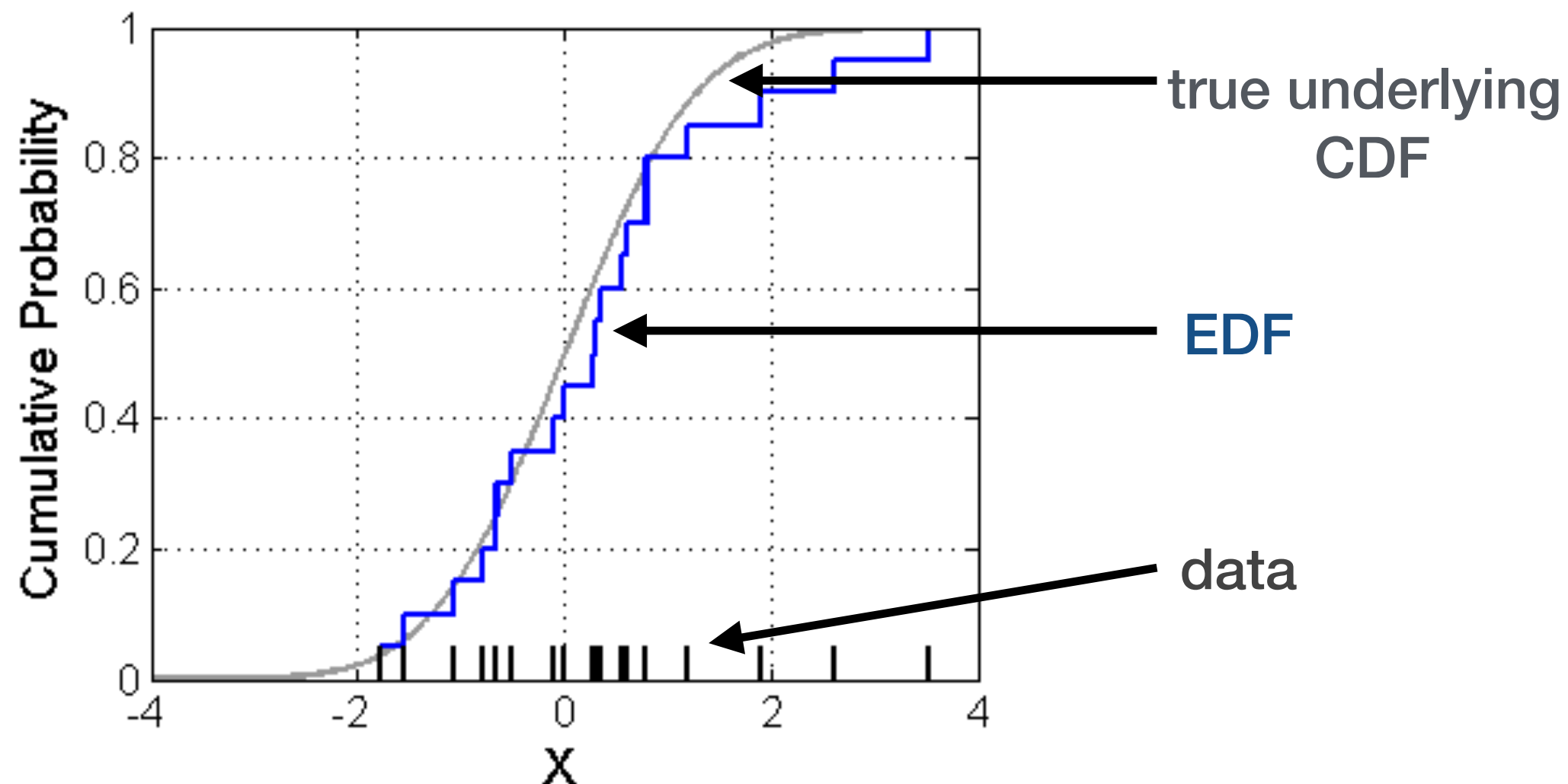
Cumulative Distribution Function

- The Cumulative Distribution Function (CDF) is related to the PDF and gives the probability that a variable (x) is less than some value x_0



Empirical Distribution Function

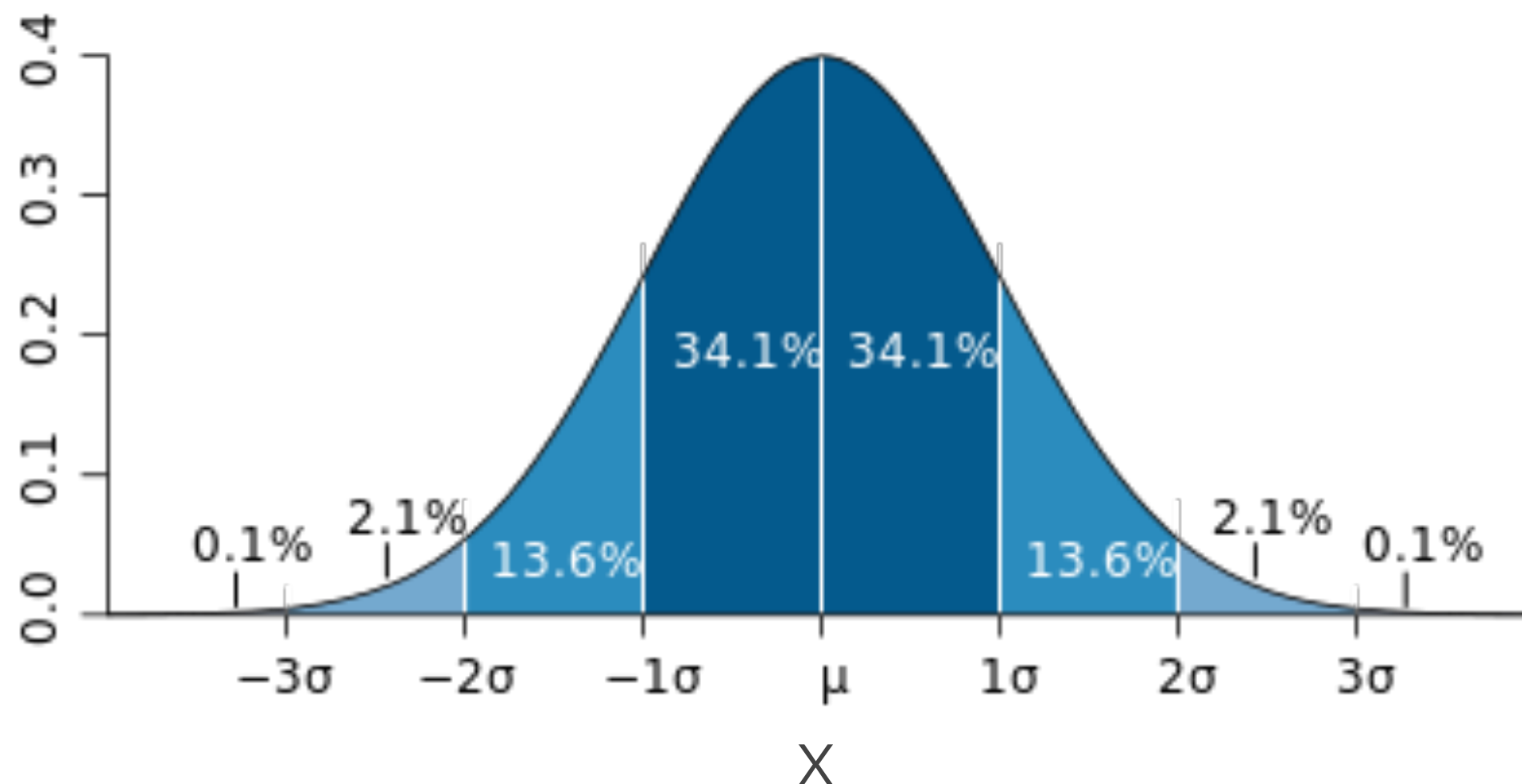
- The Empirical Distribution Function (EDF) is similar to the CDF, but constructed from data
 - Used in methods we'll cover later, e.g. the Kolmogorov-Smirnov test
 - Much less common than the CDF or PDF



Gaussian PDF

- Gaussian Probability Distribution Function (PDF) only relies on the expectation or mean (μ) and the standard deviation (σ) of a sample

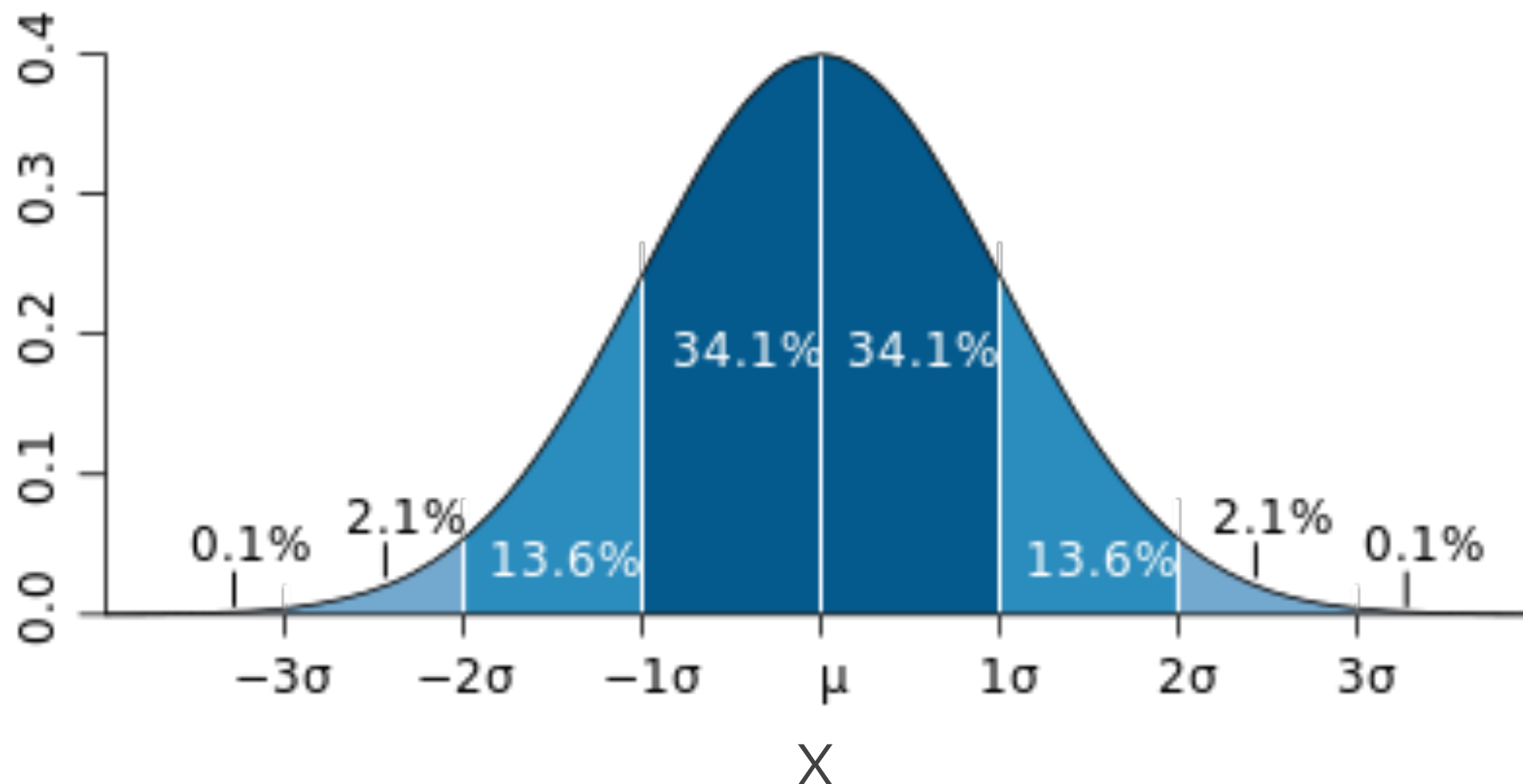
$$f(X; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$



Gaussian PDF

- Gaussian is one of the single most common PDFs, in part because of the Central Limit Theorem (CLT)

$$f(X; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$



Central Limit Theorem

- Because the central aim is the practical application of analyses techniques, we will not be focused on theorems, math proofs, and theoretical derivations. This is an ***applied*** methods course.
- In loose terms, the CLT says that for a large number of measurements of a continuous variable X done in batches*, the distribution of the batch means \bar{X} will be approximately gaussian.
 - Even if the underlying PDF (or joint PDFs) of X are not themselves gaussian

*As a rule of thumb, the batch size should be ≥ 30

Statistical Tests

- Chi-squared test

$$\chi^2 = \sum \frac{(\textit{Observed} - \textit{Expected})^2}{(\textit{Expected Uncertainty})^2}$$

- Often, the χ^2 is shown assuming N observations across some range of values (i), for example

$$\chi^2 = \sum_i \frac{(N_{i,obs} - N_{i,exp})^2}{\sigma_{i,exp}^2}$$

- If the uncertainties are only statistical, and N is large enough that $\sigma_{i,exp} = \sqrt{N_{i,exp}}$, then we get the conventional

$$\chi^2 = \sum_i \frac{(N_{i,obs} - N_{i,exp})^2}{N_{i,exp}}$$

Chi-Squared

- The Chi-squared lets us know how far away our observed data is from our expectation(s)
 - The denominator is the *uncertainty*², so the entire χ^2 is always calculated relative to the total uncertainty
 - The total uncertainty is a combination of the statistical uncertainty **AND** any systematic uncertainty

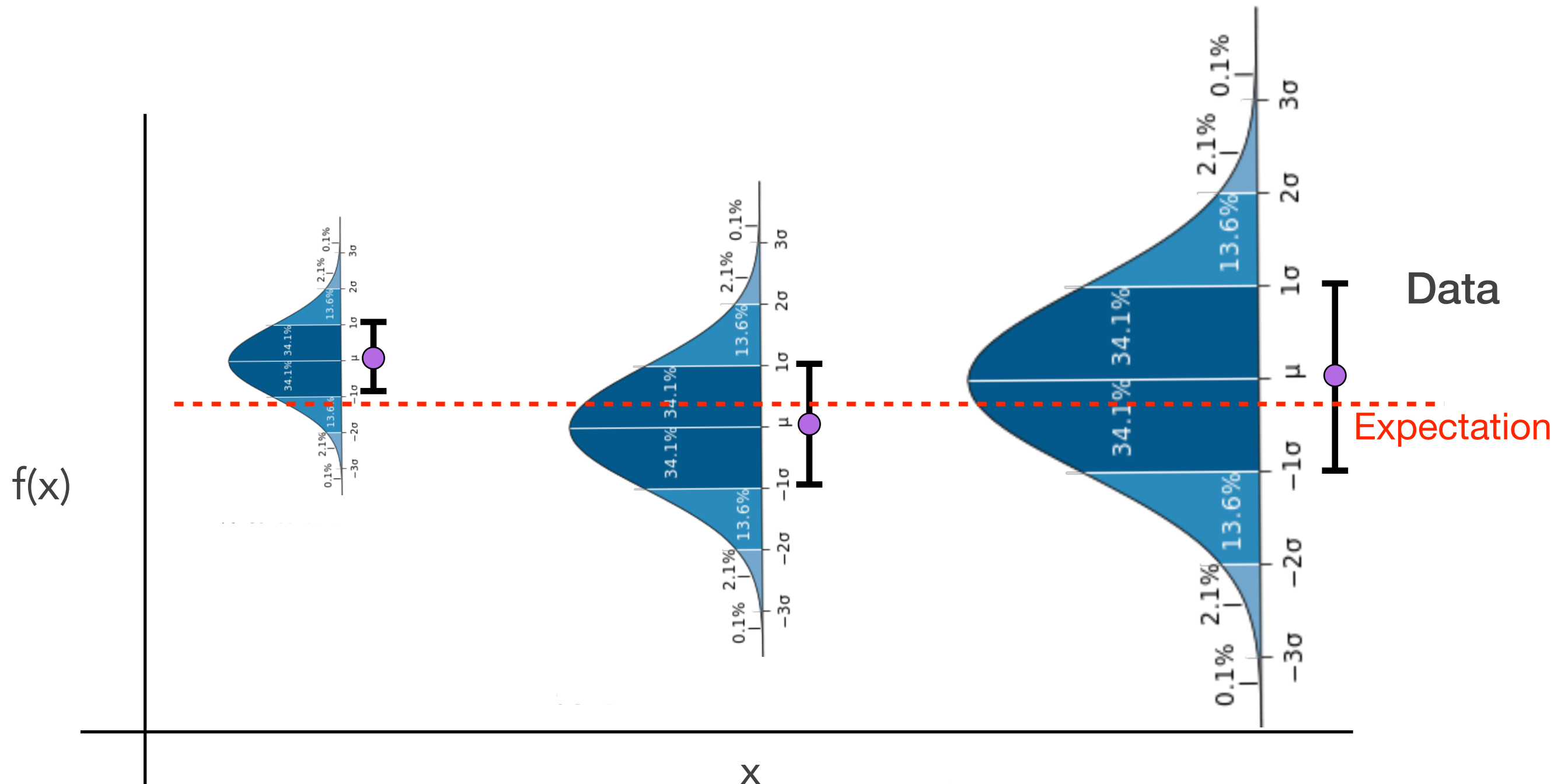
Basic Reduced Chi-Square

$$\chi_{reduced}^2 = \chi^2 / D.O.F.$$

$$\chi_{reduced}^2 \ll 1$$

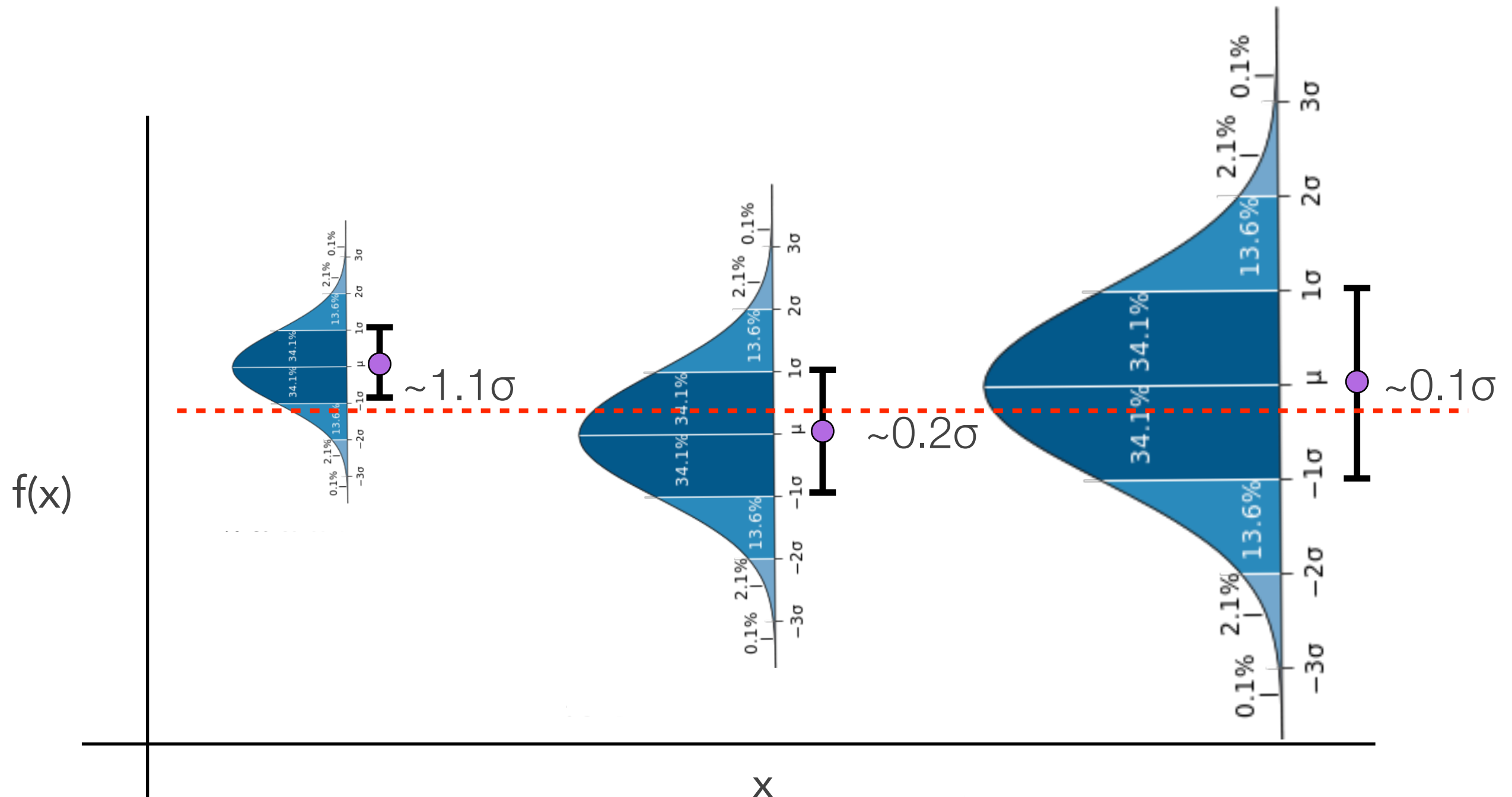
$$\chi_{reduced}^2 \approx 1$$

$$\chi_{reduced}^2 \gg 1$$

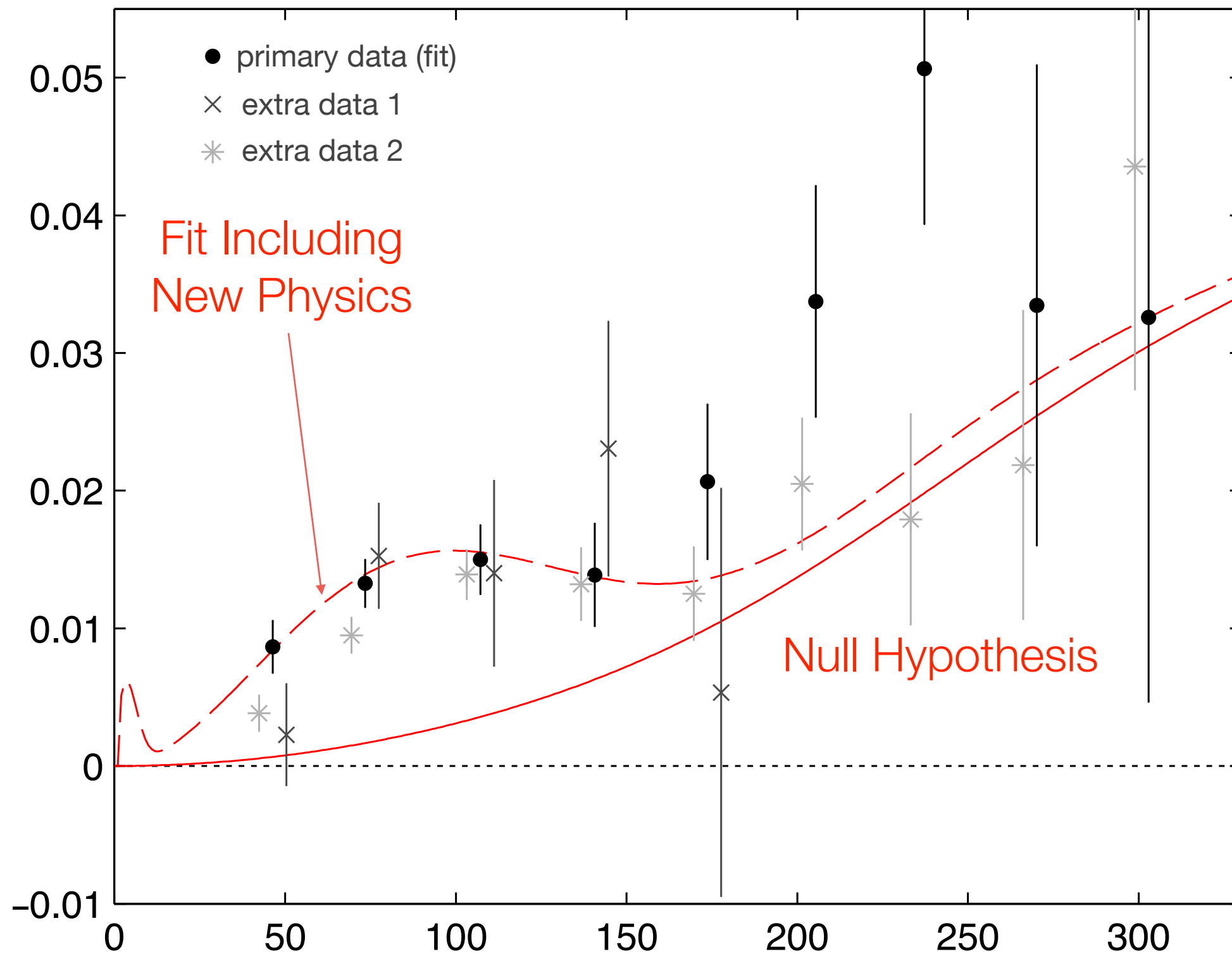


Basic Reduced Chi-Square

- Each data point has an associated approximate difference to the expectation of: 1.1σ , 0.25σ , and 0.1σ . So the total is 1.35 and with 3 data points, we get an approximate reduced chi-square of $\sim 0.4-0.5$.



Chi-By-Eye

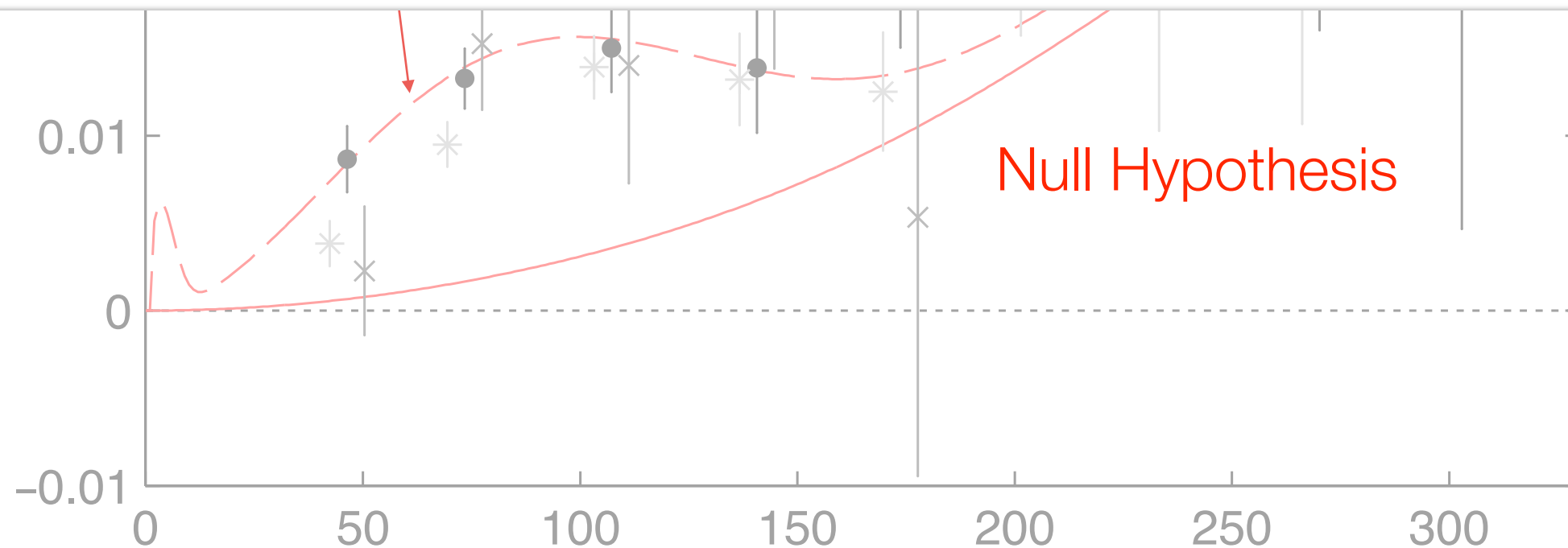


Chi-By-Eye

doi:10.1103/PhysRevLett.112.241101

Detection of *B*-Mode Polarization at Degree Angular Scales by BICEP2

P. A. R. Ade,¹ R. W. Aikin,² D. Barkats,³ S. J. Benton,⁴ C. A. Bischoff,⁵ J. J. Bock,^{2,6} J. A. Brevik,² I. Buder,⁵ E. Bullock,⁷ C. D. Dowell,⁶ L. Duband,⁸ J. P. Filippini,² S. Fliescher,⁹ S. R. Golwala,² M. Halpern,¹⁰ M. Hasselfield,¹⁰ S. R. Hildebrandt,^{2,6} G. C. Hilton,¹¹ V. V. Hristov,² K. D. Irwin,^{12,13,11} K. S. Karkare,⁵ J. P. Kaufman,¹⁴ B. G. Keating,¹⁴ S. A. Kernasovskiy,¹² J. M. Kovac,^{5,*} C. L. Kuo,^{12,13} E. M. Leitch,¹⁵ M. Lueker,² P. Mason,² C. B. Netterfield,^{4,16} H. T. Nguyen,⁶ R. O'Brient,⁶ R. W. Ogburn IV,^{12,13} A. Orlando,¹⁴ C. Pryke,^{9,7,†} C. D. Reintsema,¹¹ S. Richter,⁵ R. Schwarz,⁹ C. D. Sheehy,^{9,15} Z. K. Staniszewski,^{2,6} R. V. Sudiwala,¹ G. P. Teply,² J. E. Tolan,¹² A. D. Turner,⁶ A. G. Vieregg,^{5,15} C. L. Wong,⁵ and K. W. Yoon^{12,13}
(BICEP2 Collaboration)



Gaussian/Poisson Uncertainty is Everywhere

- Thanks to basic statistics, and Siméon Poisson, an estimate of the uncertainty on data points is generically $\sqrt{\text{number of events}}$. It works because almost all data is at some level a collection of discrete events.
 - Does not include the impact of systematic uncertainties
 - Does not include the impact of any biases either
 - Works better for larger number of events than smaller

Exercise 1

- Read in data from “FranksNumbers.txt”
 - Link is on the course webpage
 - There is some non-numeric text in the file, so data parsing is important
 - Use any methods and/or combinations of coding languages which work(s) for you
 - Parse data in python, analyze in MatLab
 - Parse data and analyze in Rust or R
 - Parse data in C, analyze in Fortran (not recommended, but possible)
 - Copy/paste using spreadsheets (Excel, OpenOffice, etc.) is discouraged because the data is already in .txt files, and reading in .txt files is a very important skill
 - Note that a future data set has 1.28M entries, which will kill a spreadsheet
- Calculate the mean and variance for each data set in the file
 - There should be 5 unique data sets

Exercise 1 pt.2

- Using the eq. $y = x * 0.48 + 3.02$, calculate the Pearson's χ^2 for each data set assuming that the uncertainty on the 'y-value' is \sqrt{value}
 - Write your own method
 - Bonus: use a class or external package to get value
- Using the same equation, calculate a χ^2 where the uncertainty on each data point is ± 1.22
- From the two χ^2 , which uncertainty shows better agreement with the data?
 - ± 1.22 or \sqrt{events} ?

Discussion/Comments

- What values did people get for the χ^2 ?
- Note: do **not** use a test statistic, such as the χ^2 , to estimate or assign statical uncertainties or systematic uncertainties on experimental derived data

Some chi-squared Remarks

- A chi-squared distribution is based on gaussian uncertainties, so beware when errors/uncertainties are not gaussian. For example:
 - Low statistics
 - Biases in the data can also produce non-gaussianity
- The concept that a reduced chi-squared near 1 is 'good' depends strongly on the degrees of freedom (DoF) and/or data
 - A reduced chi-squared of 1.2 w/ 20 DoF is not a cause for concern
 - A reduced chi-squared of 1.2 w/ 1000 DoF is very, very bad and incredibly unlikely

Conclusion

- Know your distribution functions (probability, cumulative, and empirical)
- Central Limit Theorem says that means of most variables will produce a gaussian distribution of the mean value for a large numbers of measurements
- Chi-square(d) calculation is a frequent metric for goodness-of-fit and quantitative data/hypothesis matching
- Very light load this week, so try and get your software working
 - If you have problems 'ask' classmates who have similar computer setups
 - If you have solutions help your classmates
- First problem set should be available now in Absalon
- Read "Not Normal: the uncertainties of scientific measurements", there will be a discussion this afternoon

Extra

Distribution Functions

- Many nice illustrations for different functions at https://commons.wikimedia.org/wiki/Probability_distribution
- Many of the plots used in the lecture notes come from wikipedia (because it's a great resource)