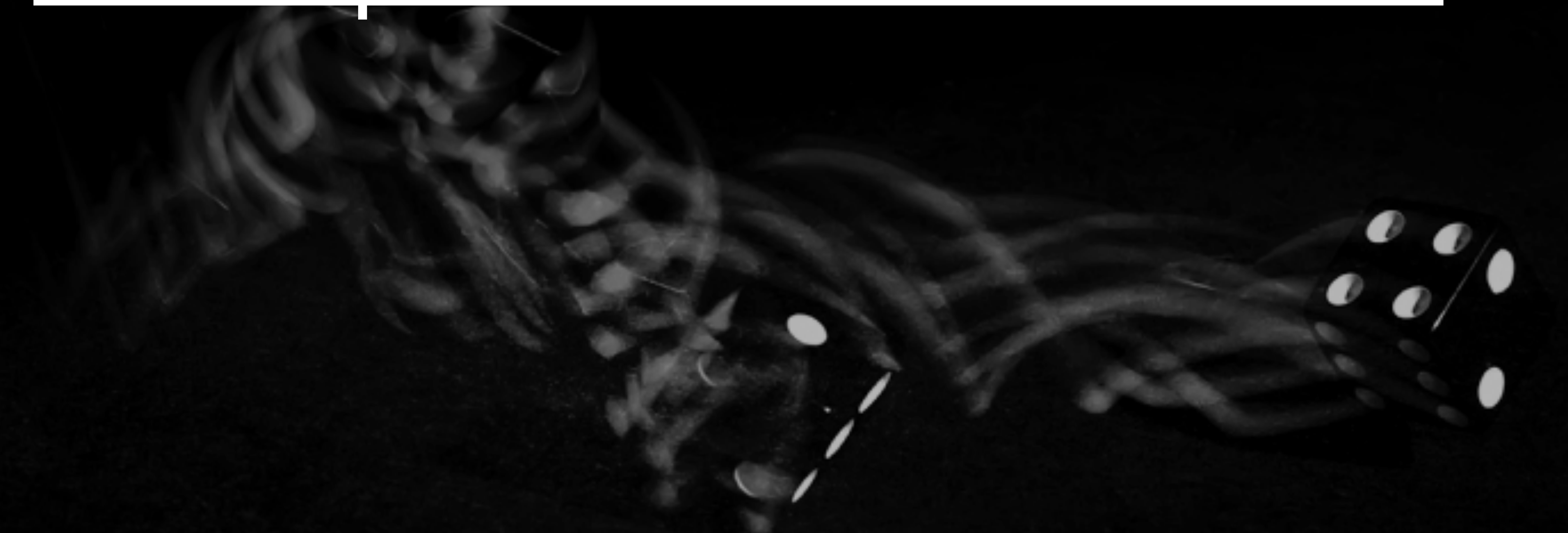
Lecture 1: Chi-Squared & Some Basics



D. Jason Koskinen
koskinen@nbi.ku.dk

Advanced Methods in Applied Statistics
Feb - Apr 2019

Photo by Howard Jackman

Variance

- Because it's something we all should know

$$\sigma^2 \equiv \langle (X - \mu)^2 \rangle \qquad \sigma^2 = \frac{1}{N}\sum_{i=0}^{N}(x_i - \bar{x})^2$$

σ^2 is the variance

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

N is the number of data points

x_i is the individual observed data points

*mathworld.Wolfram

Unbiased Variance

- Just because it's something we all should know

$$S_{N-1} \equiv \frac{1}{N-1}\sum_{i=0}^{N}(x_i - \bar{x})^2$$

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

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

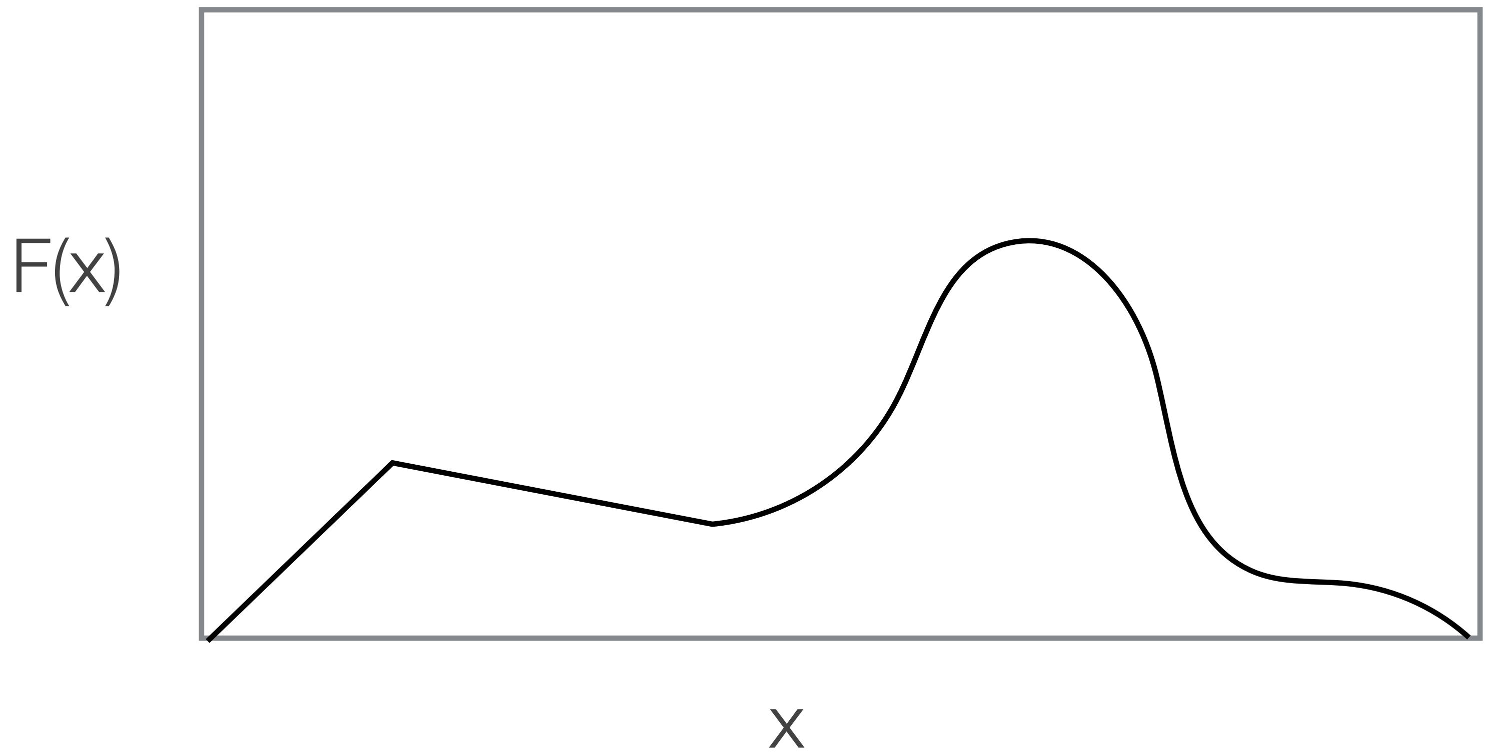
N is the number of data points

x_i is the individual observed data points

For further information on 1/(N-1) see Bessel's correction wikipedia

Probability Distribution Function

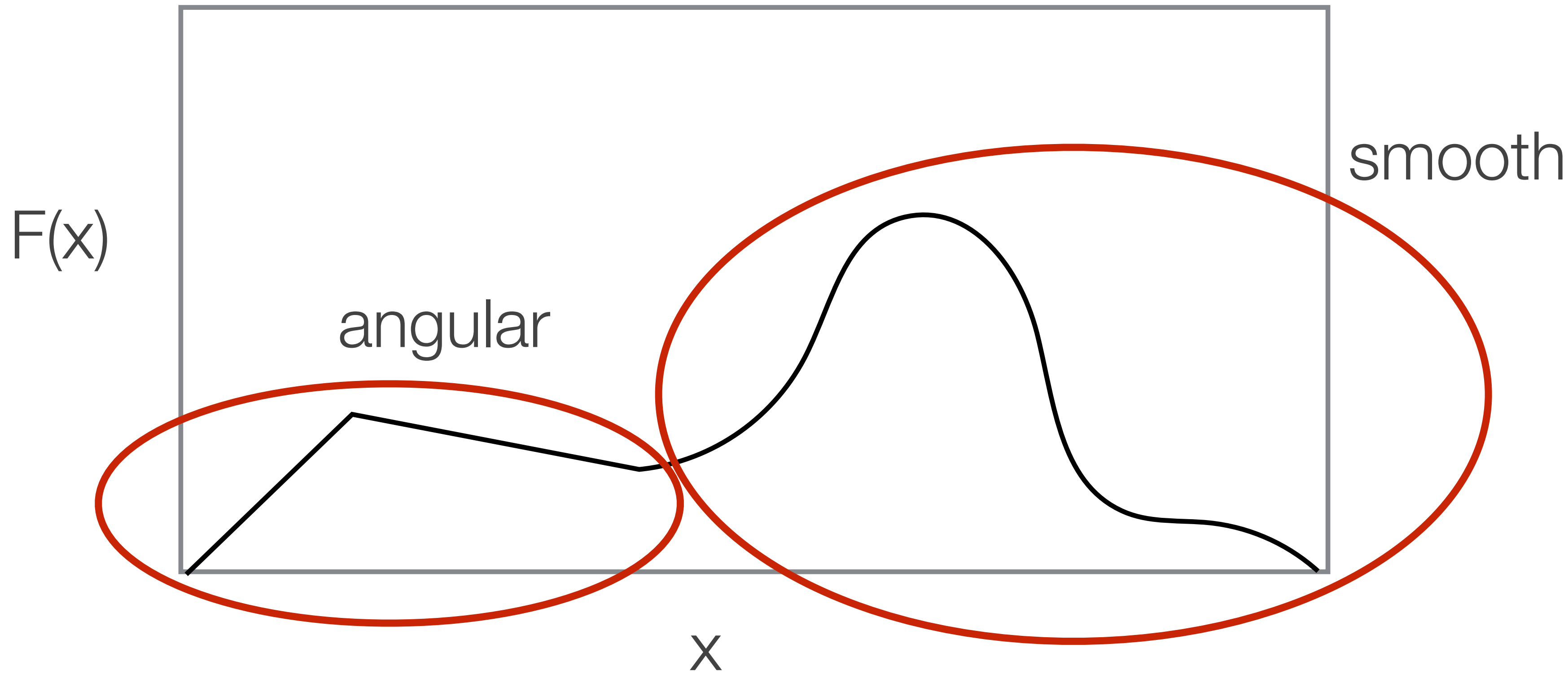
- **P**robability **D**istribution **F**unctions (PDF), where sometimes the "D" is density, is the probability of an outcome or value given a certain variable range



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

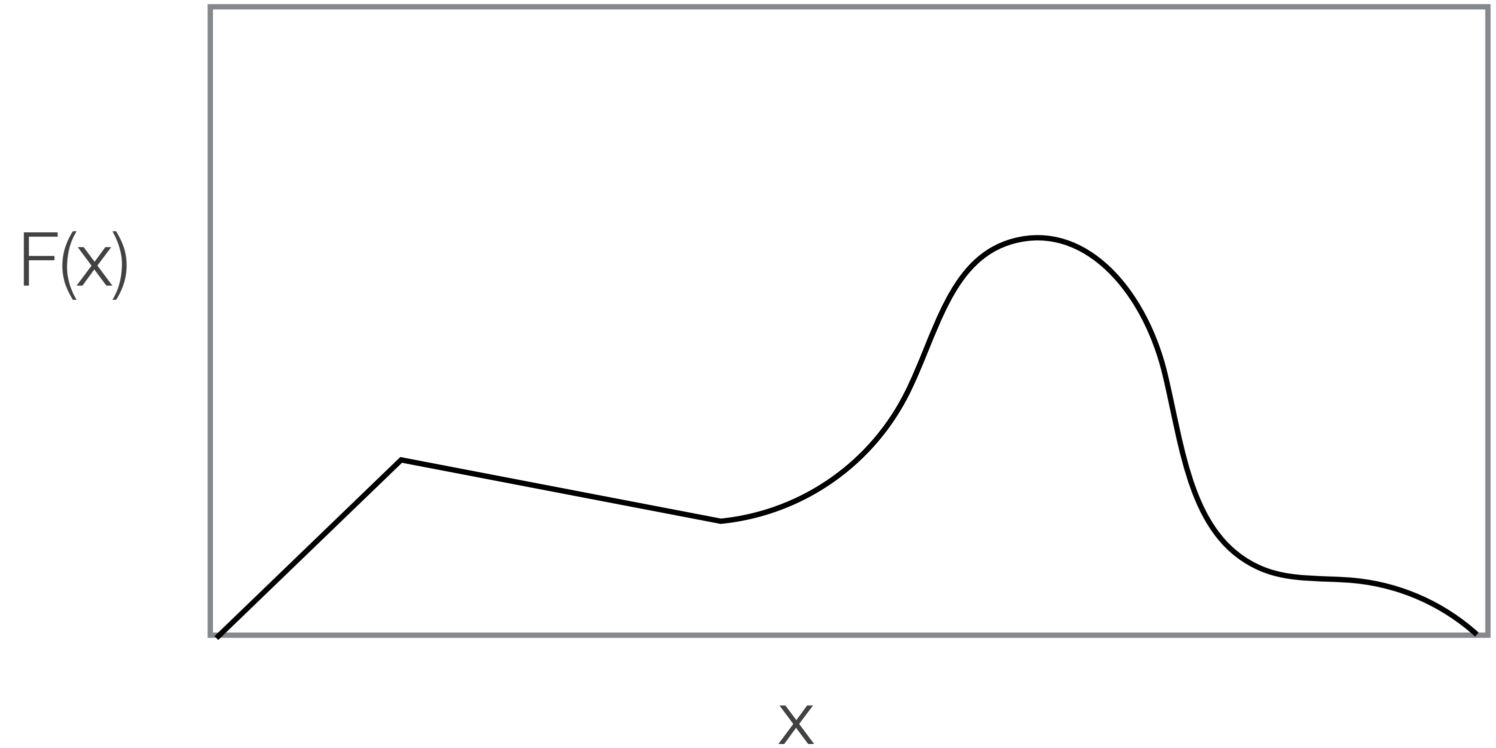
Probability Distribution Function

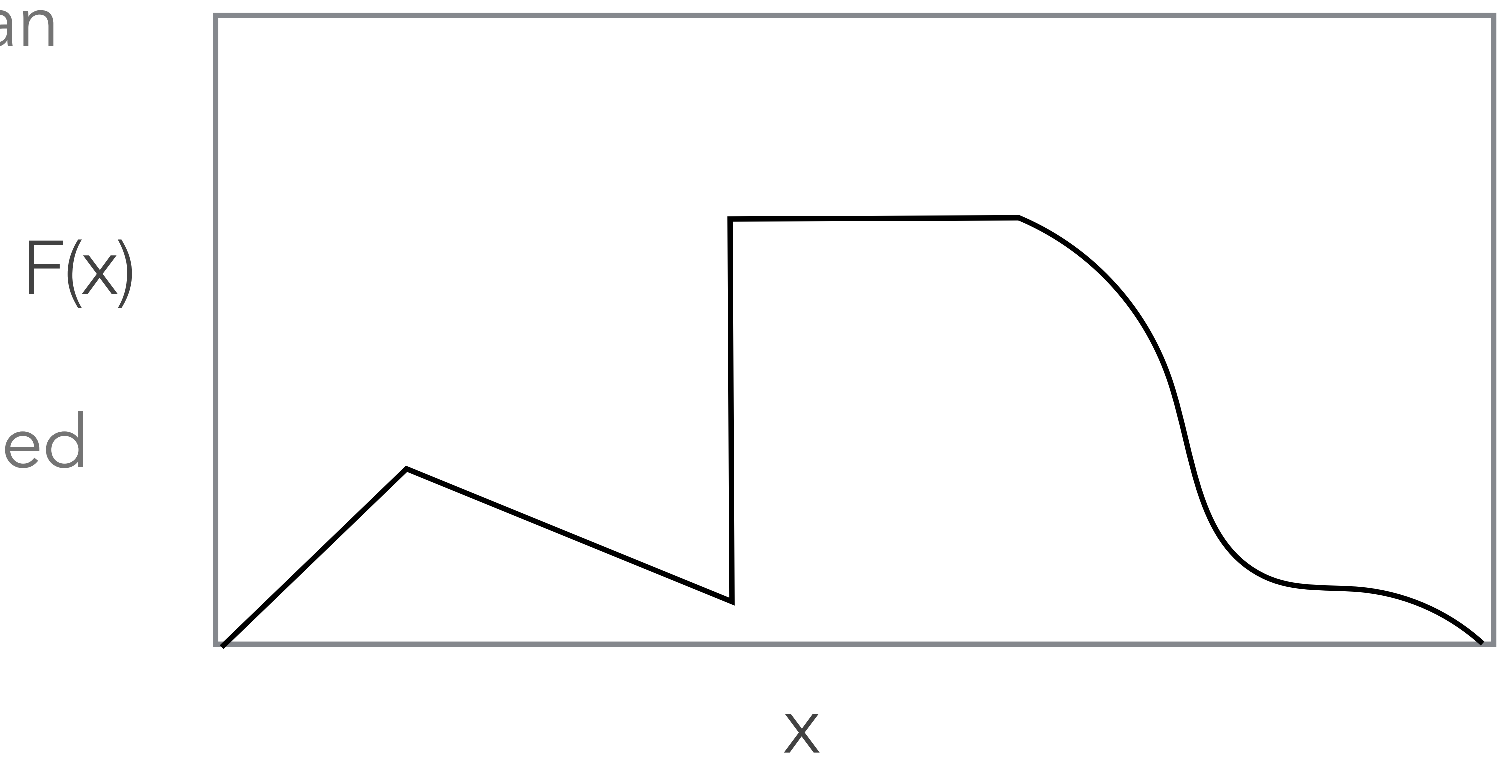
- The PDF does not have be nicely described w/ equations, and sometimes cannot be



PDFs

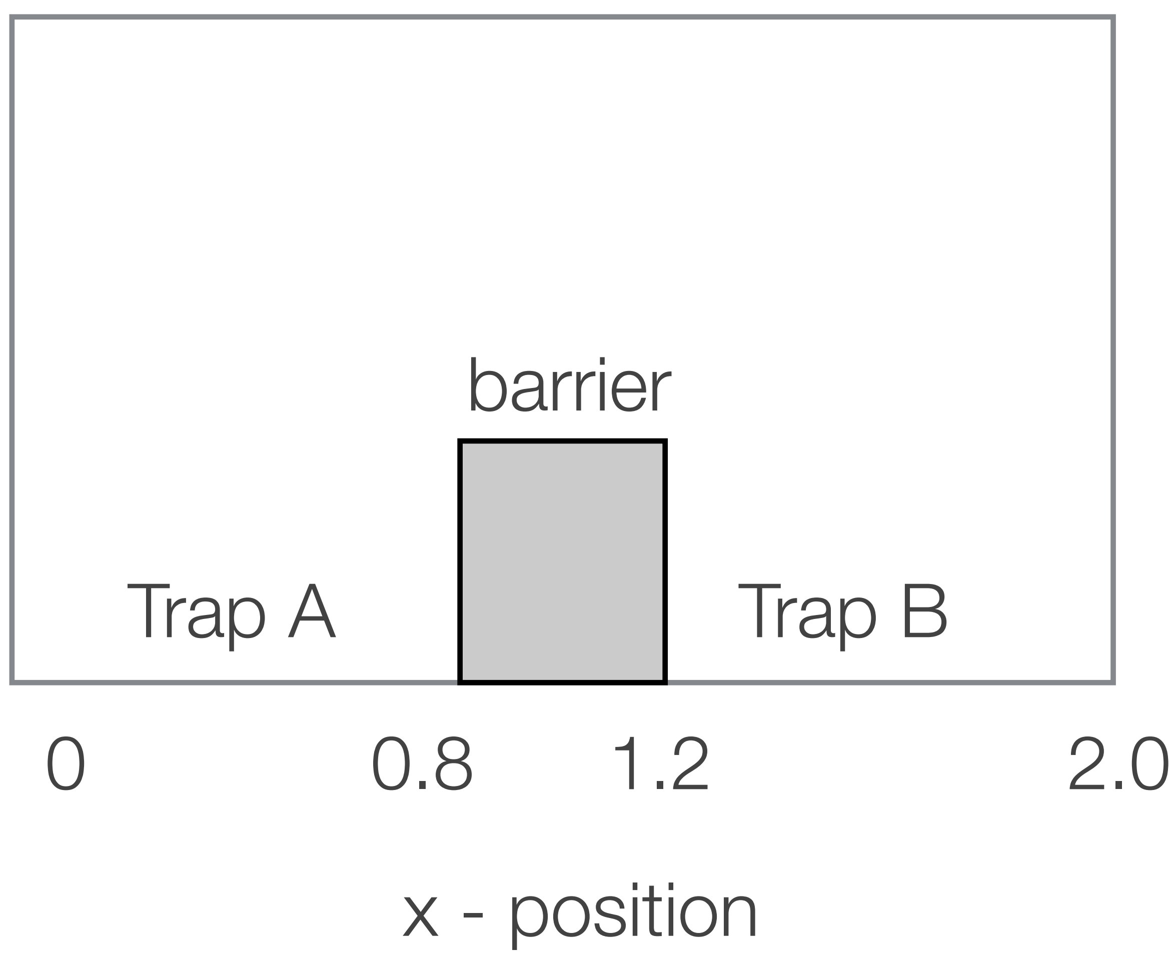
- They can be discrete, continuous, or a combination
- They often have an implied conditionality
  - "What is the energy of an outgoing electron from nuclear beta-decay?" implies beta-decay
  - PDF should be normalized to 'one'





PDF Possibility

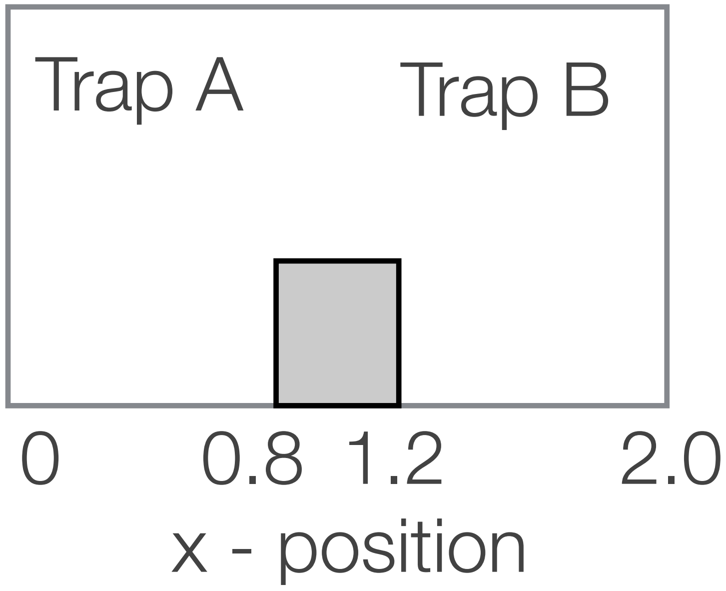
- Let's imagine an experiment which has two identical electron traps (A & B) separated by a finite barrier. An electron w/ energy below the barrier threshold is deposited in trap A. Sketch out the PDF of the x position after a very short time.



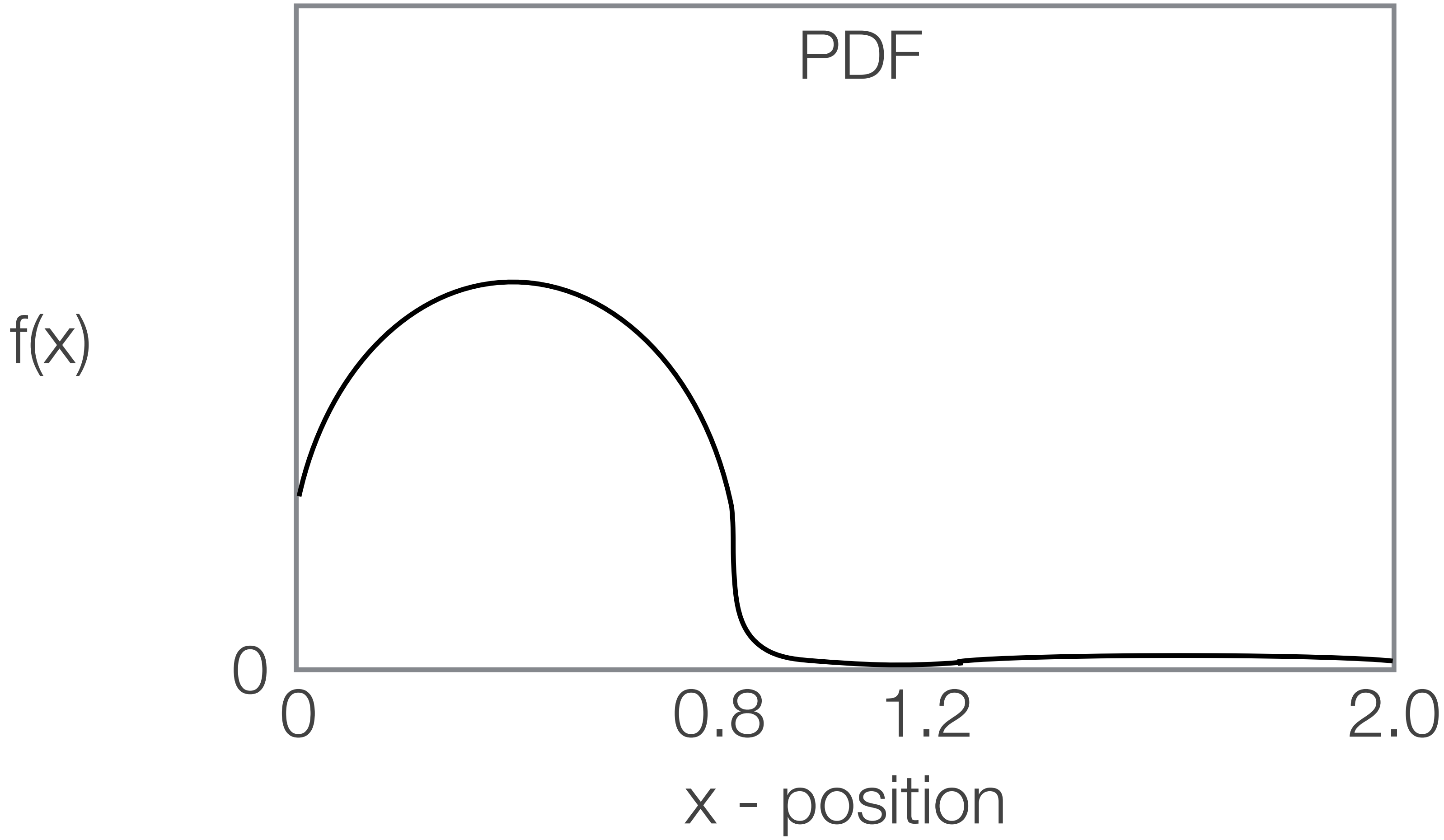
$$\text{time} \approx \frac{1}{\infty}$$

*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

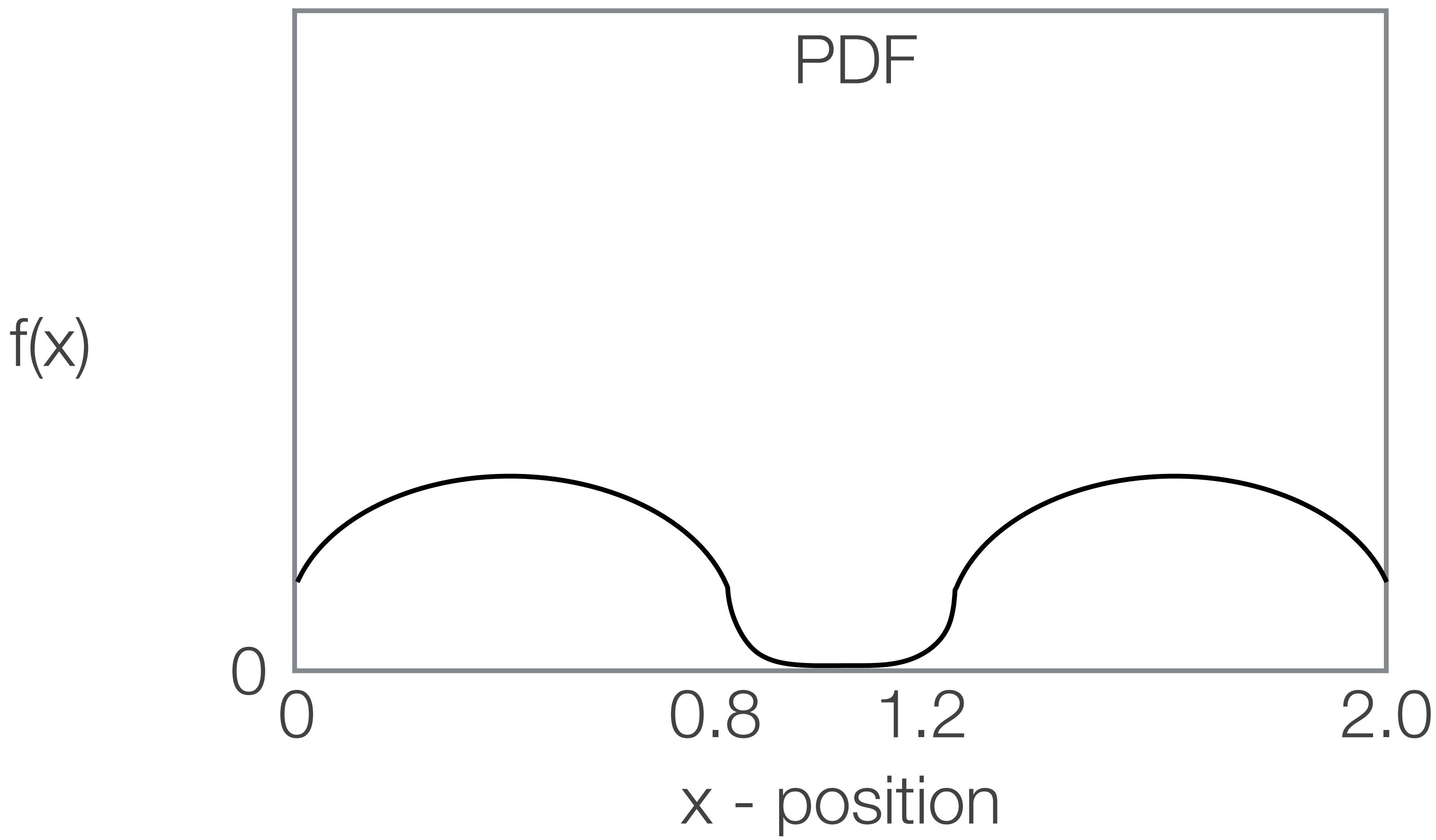
- Sketch out the PDF of the x position after a near infinitely long time.

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

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.
  - Had to renormalize the PDF

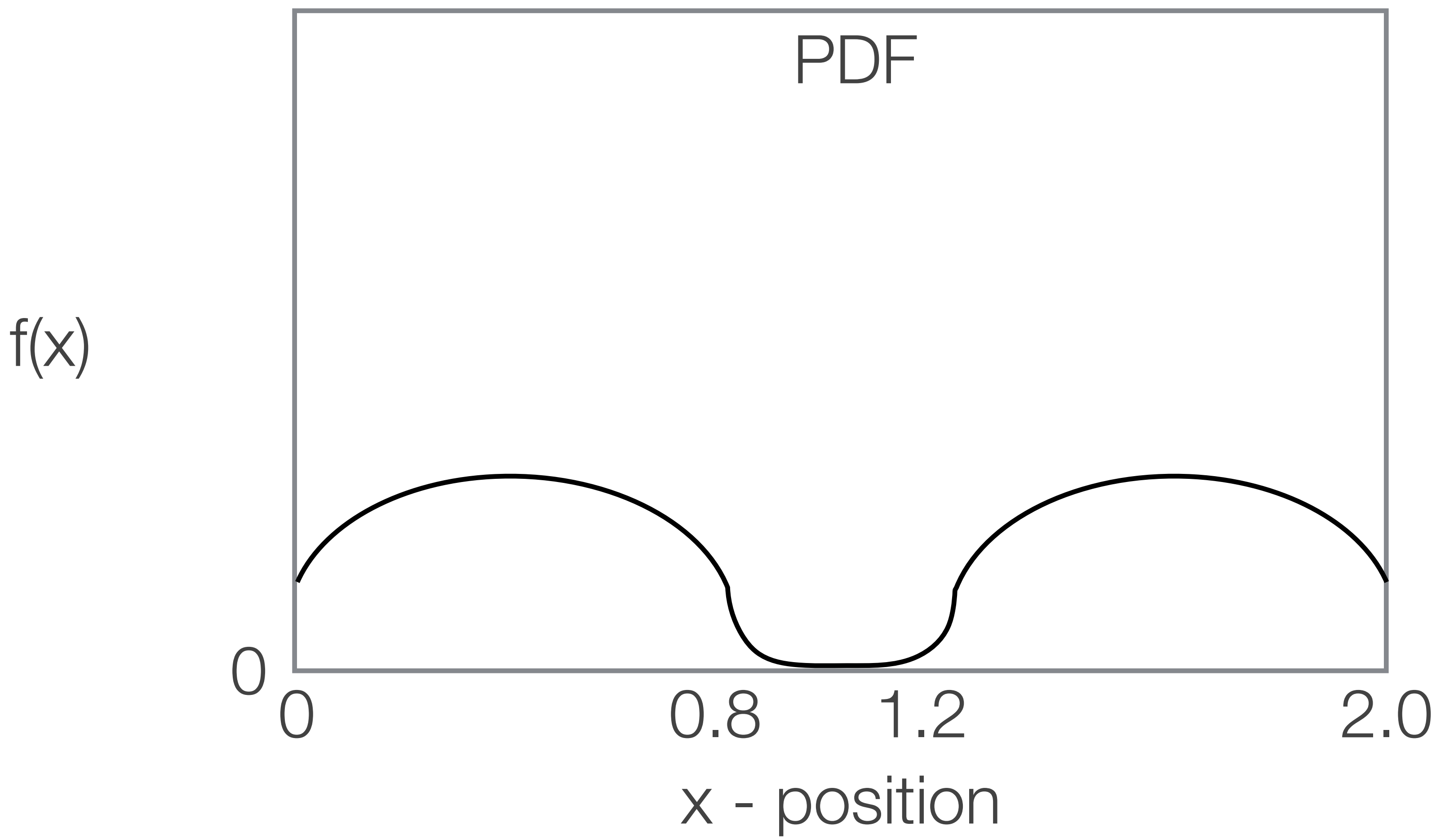
$$\text{time} \approx \infty$$



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?

$\text{time} \approx \infty$



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

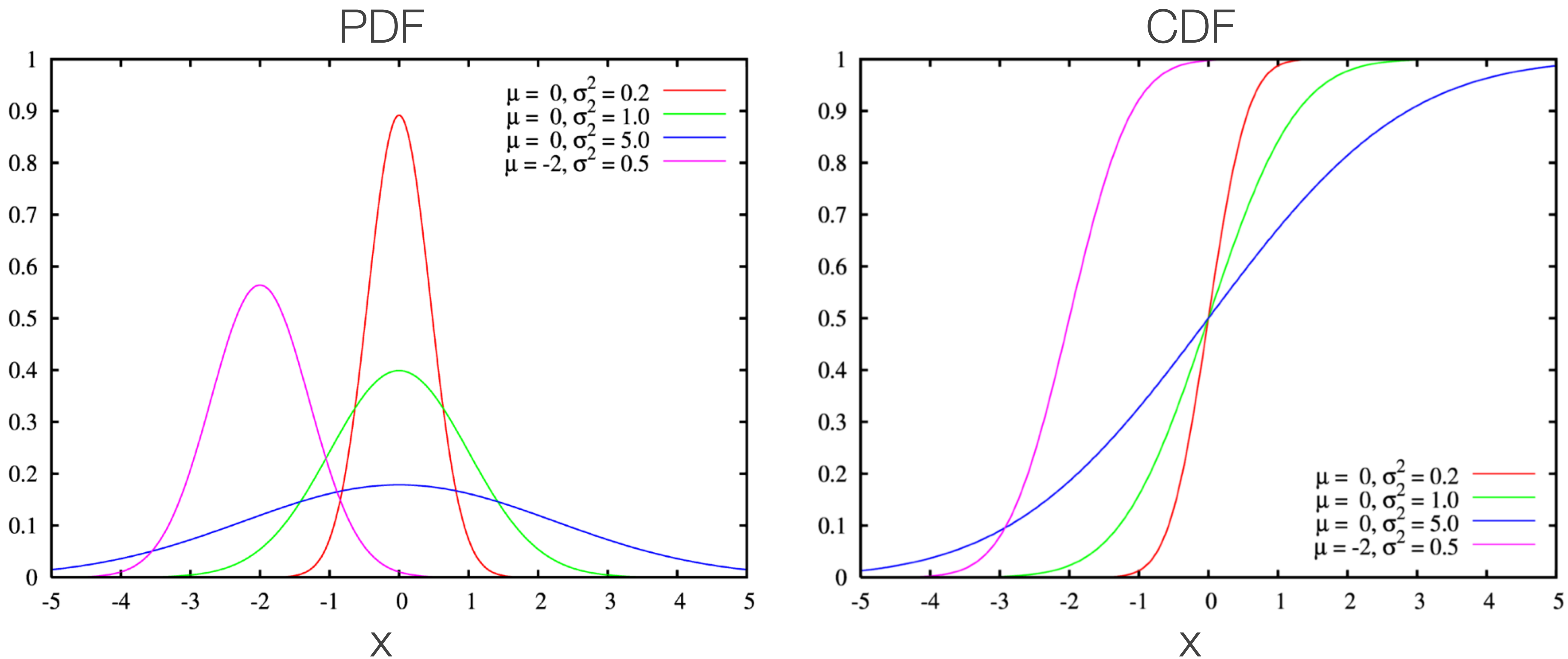
- The Cumulative Distribution Function (CDF) is related to the PDF and gives the probability that a variable (x) is less than some value x_0
- Basically, the integral or sum from -infinity to x_0

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

where $f(x)$ is the PDF

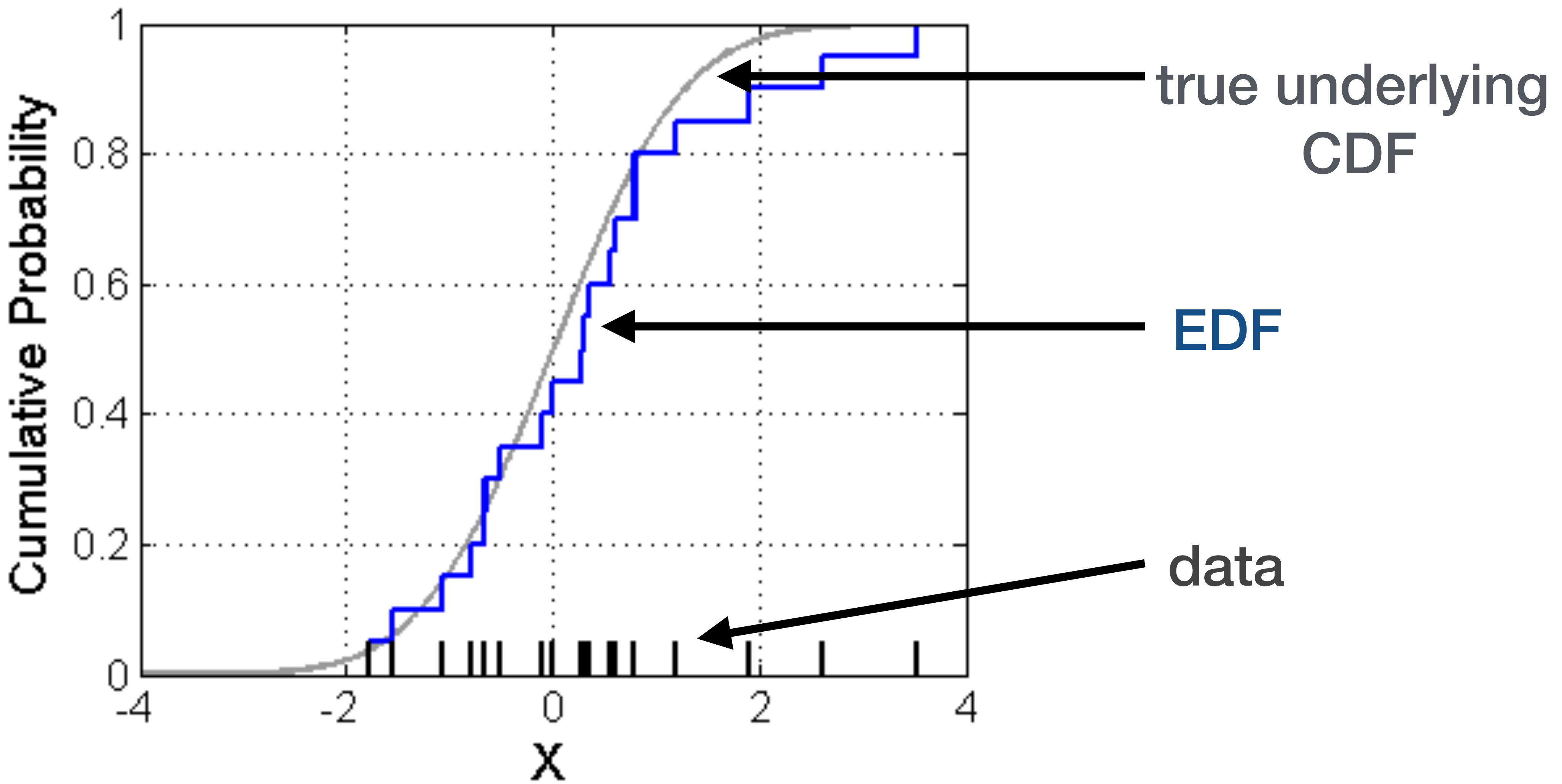
Cumulative Distribution Function

- 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

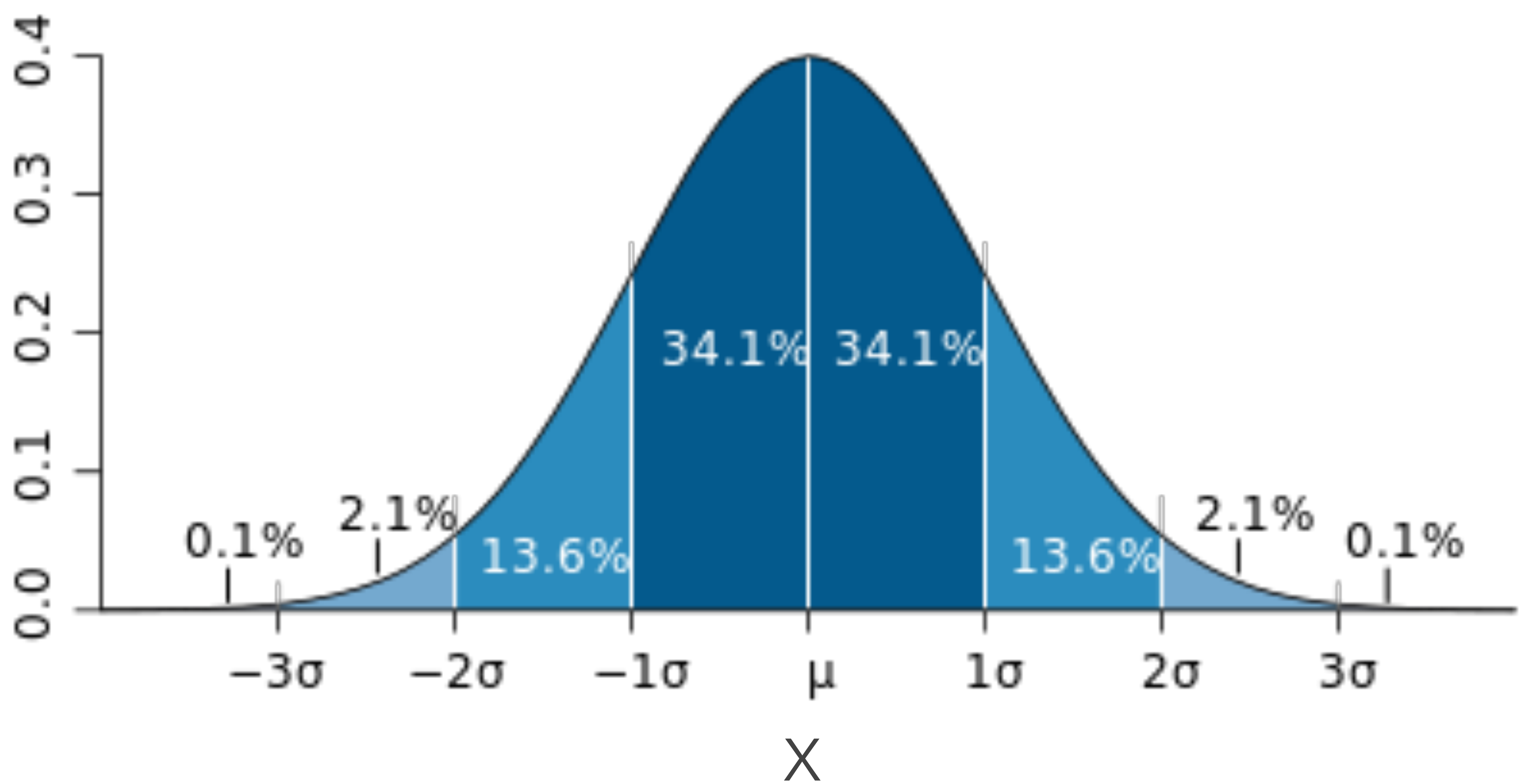
- The Empirical Distribution Function (EDF) is similar to the CDF, but constructed from data
  - Used in methods we'll cover later, e.g. the Kolmogrov-Smirnov test
  - Much less common than the CDF or PDF



Gaussian PDF

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

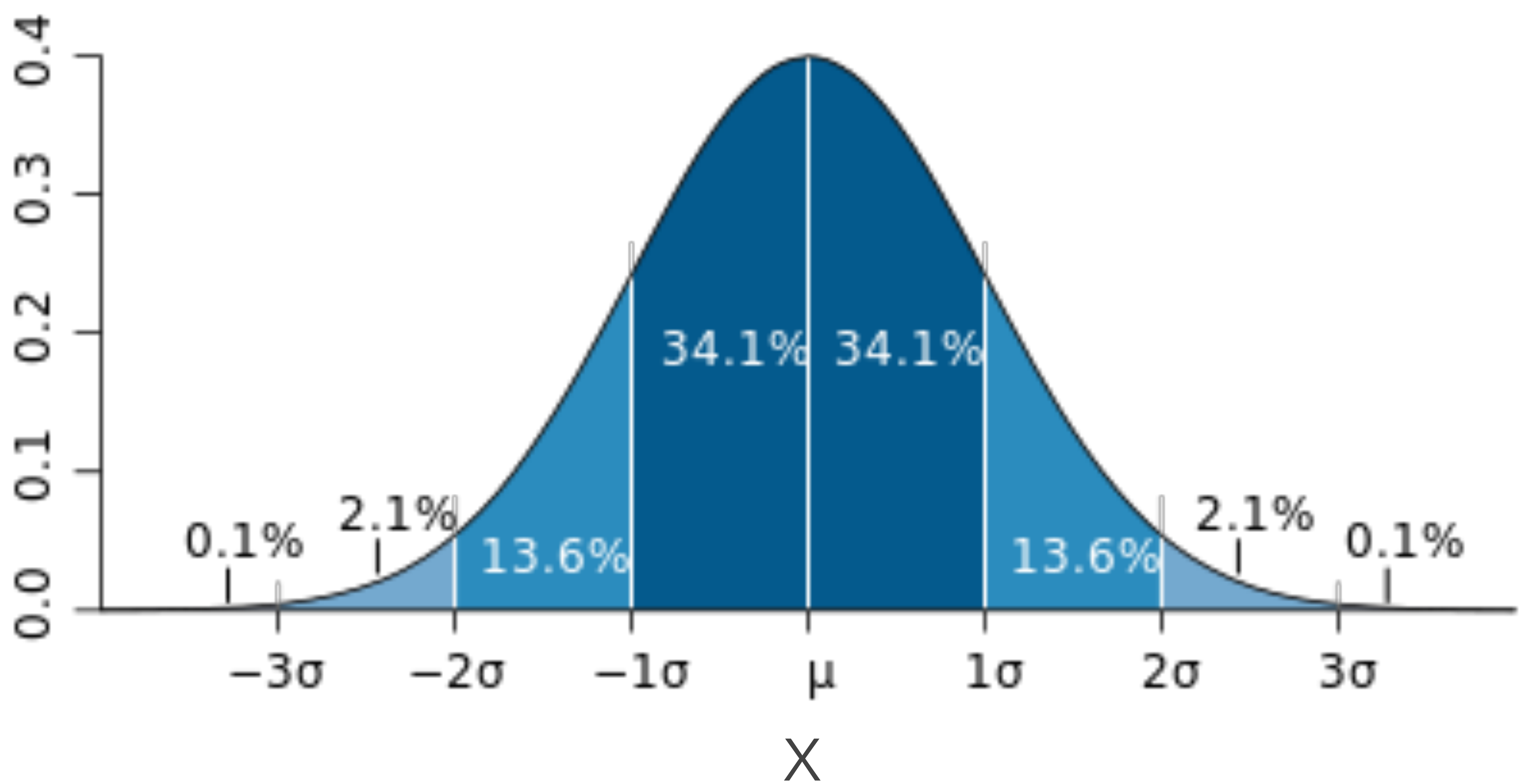
$$f(X;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$



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 overly concerned with theorems, math proofs, and theoretical derivations. This is an ***applied*** methods course after all.
- In loose terms, the CLT says that for a large number of measurements of continuous variables (or combinations thereof) the outcome approaches a gaussian distribution.
  - Even if the underlying PDF (or joint PDFs) are not themselves gaussian

Statistical Tests

- Pearson's Chi-squared test

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

- Many different permutations for a Figure of Merit (FOM), and a quick modification of χ^2 is a nice tool to have when seeing new results

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected + \sigma^2_{expected}}$$

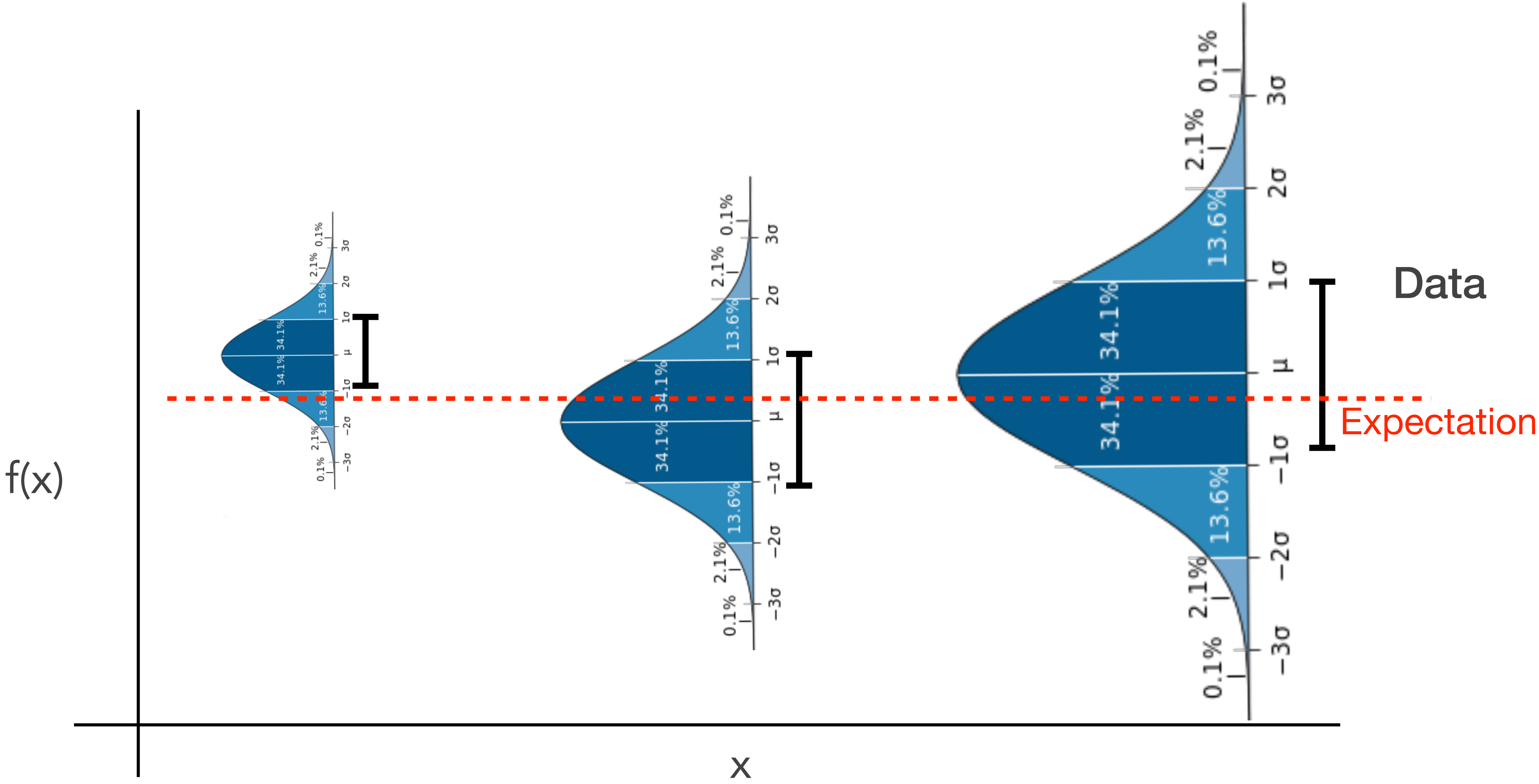
Basic Reduced Chi-Square

$$\chi^2_{reduced} \ll 1$$

$$\chi^2_{reduced} \approx 1$$

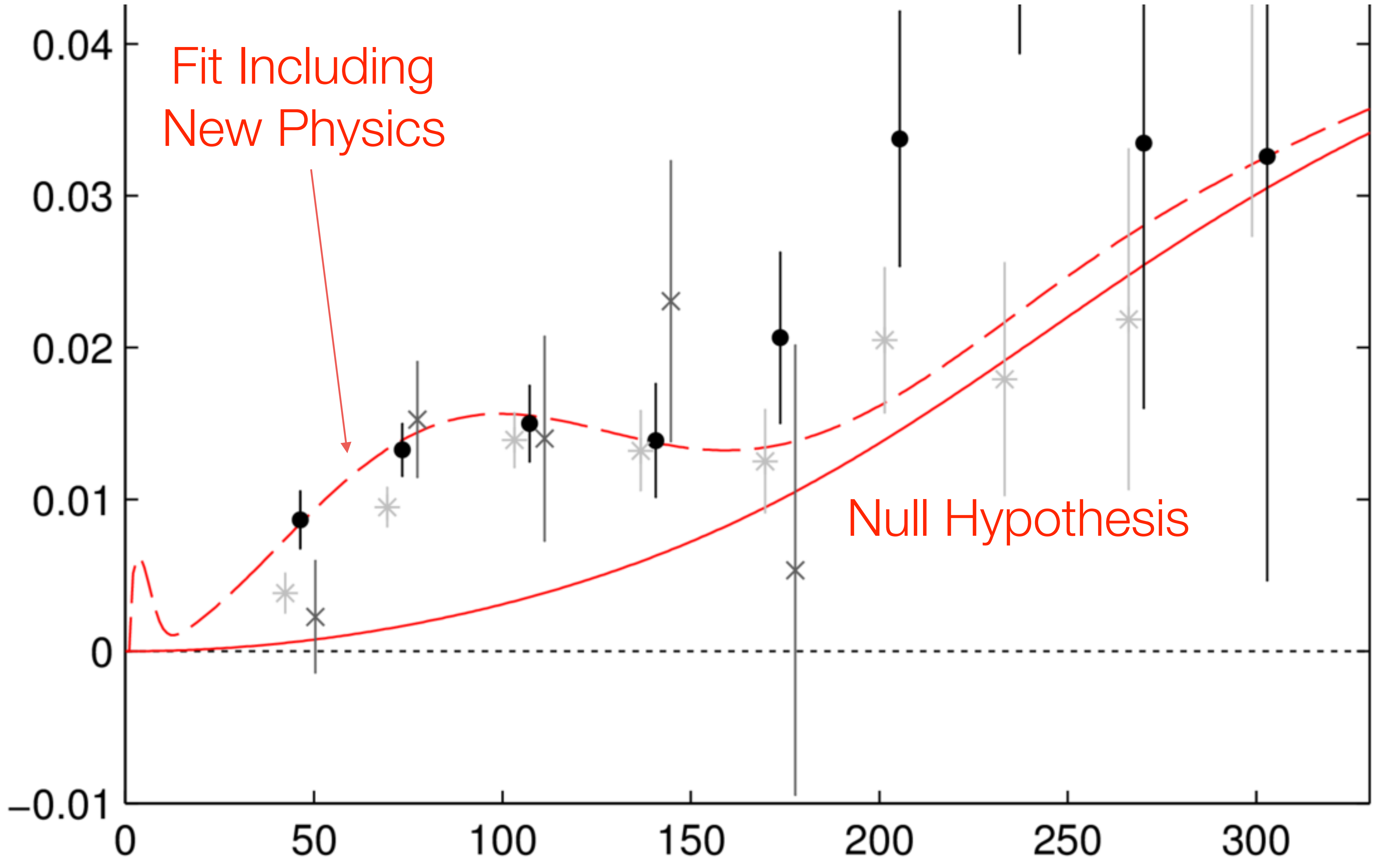
$$\chi^2_{reduced} \gg 1$$

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



Chi-By-Eye

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected + \sigma^2_{expected}}$$



Gaussian/Poisson Uncertainty is Everywhere

- Thanks to basic statistics, and Siméon Poisson, an estimate of the uncertainty on data points is generically sqrt(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
- When in doubt, take the square root of something

Exercise 1

- Read in data from "FranksNumbers.txt"
  - 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 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

http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2019/AdvancedMethodsAppliedStatistics2019.html

Exercise 1 pt.2

- Using the eq. y=x*0.48 + 3.02, calculate the Pearson's χ^2 for each data set
    - Write your own method
    - Bonus: use a class or external package to get value
- Using the same eq. calculate a χ^2 where the uncertainty on each data point is ±1.22
- From the two χ^2, what is a better reflection of the uncertainty?
    - ±1.22 or sqrt(events)?

Some chi-squared Remarks

- A chi-squared distribution is based on gaussian ‘errors’, so beware when errors/uncertainties are not gaussian
  - 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 has a pretty good
  - 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 most variables will produce a gaussian distribution at 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 is online
- Read “Not Normal: the uncertainties of scientific measurements”, there will be a discussion next class

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)