STATISTICAL ANALYSIS & INFERENCE

(in a nutshell)

Vinh Q Tran

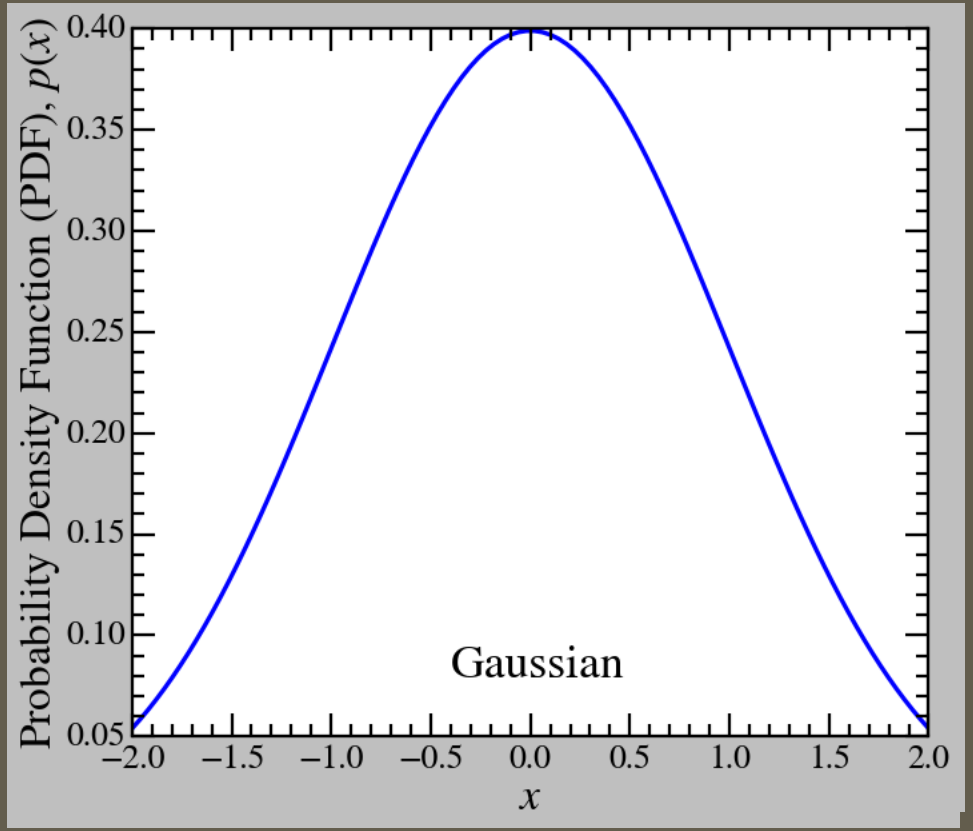
Massachusetts Institute of Technology (MIT)

Distributions

Basis of Probability

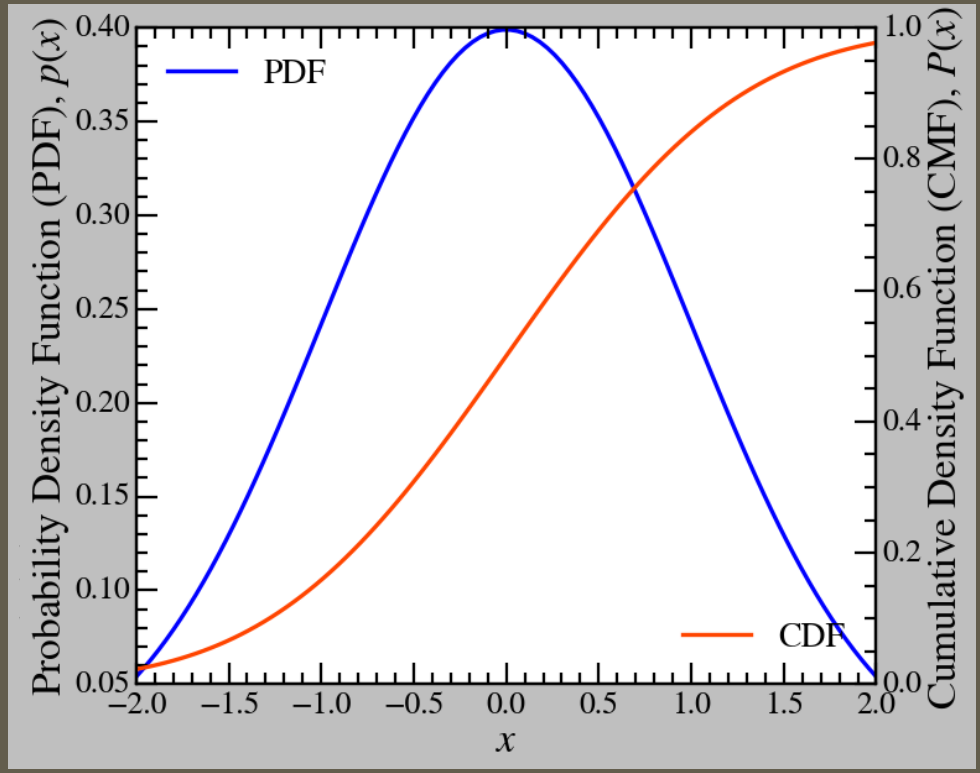
What is a Distribution?

- Random variable (x): outcome (or possible outcome) of something, e.g. measurement (length of string, mass of particle,..), number on dice, …
- Discrete v.s. continuous random variable.
- Distribution: describe how the possible values of the random variable spread/cluster, i.e. how they "distributed", their probability.
- Distributions can be calculated from samples of random variables ($\{x\}$), i.e. values of measurements.



PDF (PMF), CDF (CMF)

- PDF (PMF): probability distribution (mass) function, $p(x)$ describes how the possible values of the random variable, x, are distributed (main description of the distribution).
- CDF (CMF): cumulative distribution (mass) function, $P(x)$ describes the probability of the random variable having values less or equal than x. $P(x_{\max}) = 1$ for CMF, $P(\infty) = 1$ for CDF.



Mean, Variance, Standard Deviation, Median, and Mean Absolute Deviation

- Mean:

  Sample: $\bar{x} = \frac{\sum \mathrm{x}}{\mathrm{N}}$

  Distribution: $\langle \mathrm{x} \rangle = \sum \mathrm{p(x)} \,.\, \mathrm{x} = \int \mathrm{p(x)} \,.\, \mathrm{x} \,.\, \mathrm{dx}$

- Variance:

  Sample: $s^2 = \frac{\sum (\mathrm{x} - \bar{x})^2}{\mathrm{N} - 1}$

  Distribution: $\sigma^2 = \left\langle \left(x - \langle x \rangle \right)^2 \right\rangle$

- Standard Deviation (std)

  Sample: $s = \sqrt{s^2}$

  Distribution: $\sigma = \sqrt{\sigma^2}$

- Median: $\mathrm{med}(x)$, the "middle" value of the sample or the value x at which $P(x) = 0.5$.
- Median Absolute Deviation (mad): $\mathrm{med}\left(\left| x - \mathrm{med}(x) \right| \right)$.
- Med and mad show the spread of the distribution (or sample) with lesser effect from outliers when compared to mean and std.

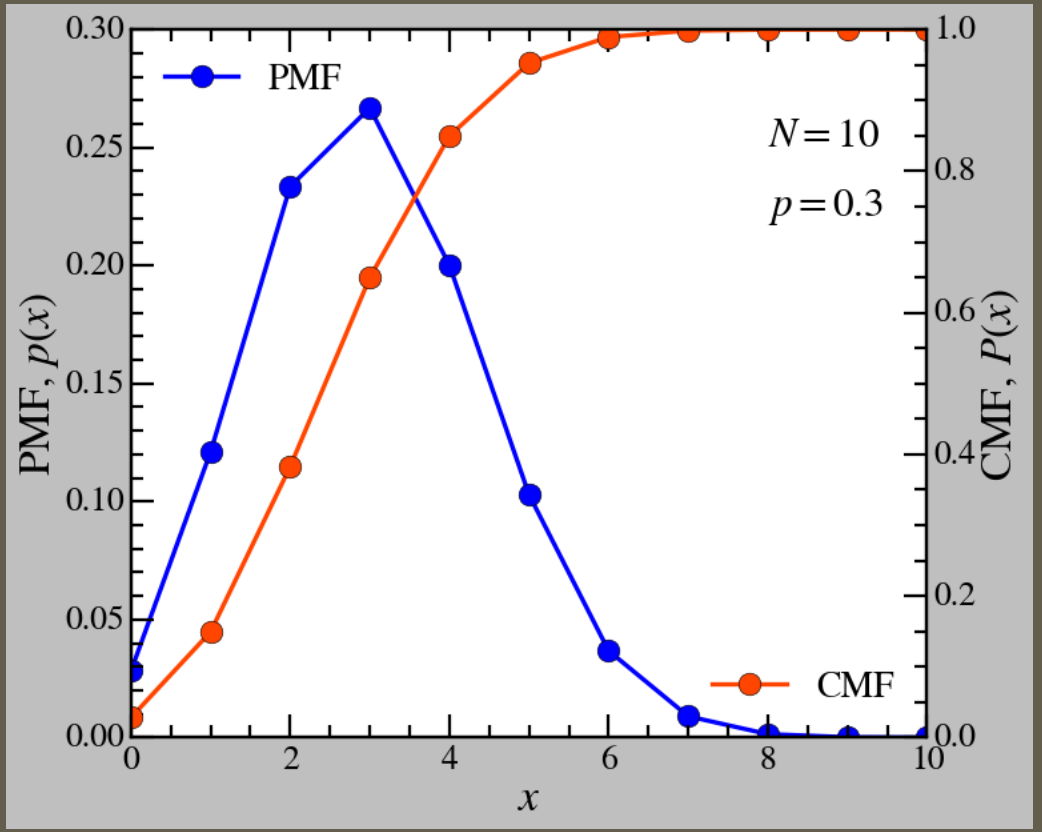
Common Distributions

Binomial

Having N numbers, each has the probability of p of being A and $1-p$ of being B, what is the probability that n numbers out of N is A?

$$p(n) = \binom{N}{n} p^n (1-p)^{N-n} = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

- $\langle n \rangle = Np$
- $\sigma^2 = N(1-p)p$



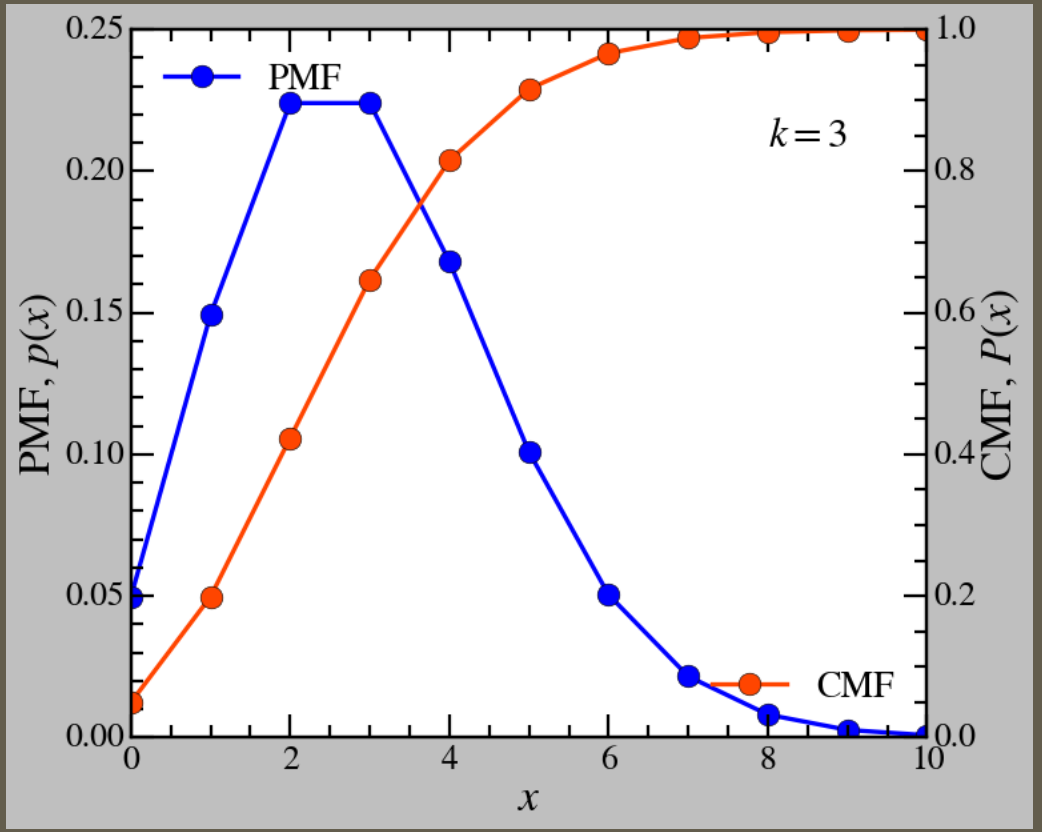
Poisson

Given the constant mean rate of probability that an event occurring λ and the time interval t, what is the probability that n events happen?

$$p(n) = \frac{(\lambda t)^n e^{-n}}{n!}$$

- $\langle n \rangle = \lambda t$
- $\sigma^2 = \lambda t$

(!) The binomial distribution become the Poisson distribution when $n \to \infty,\ p \to 0$.



Gaussian

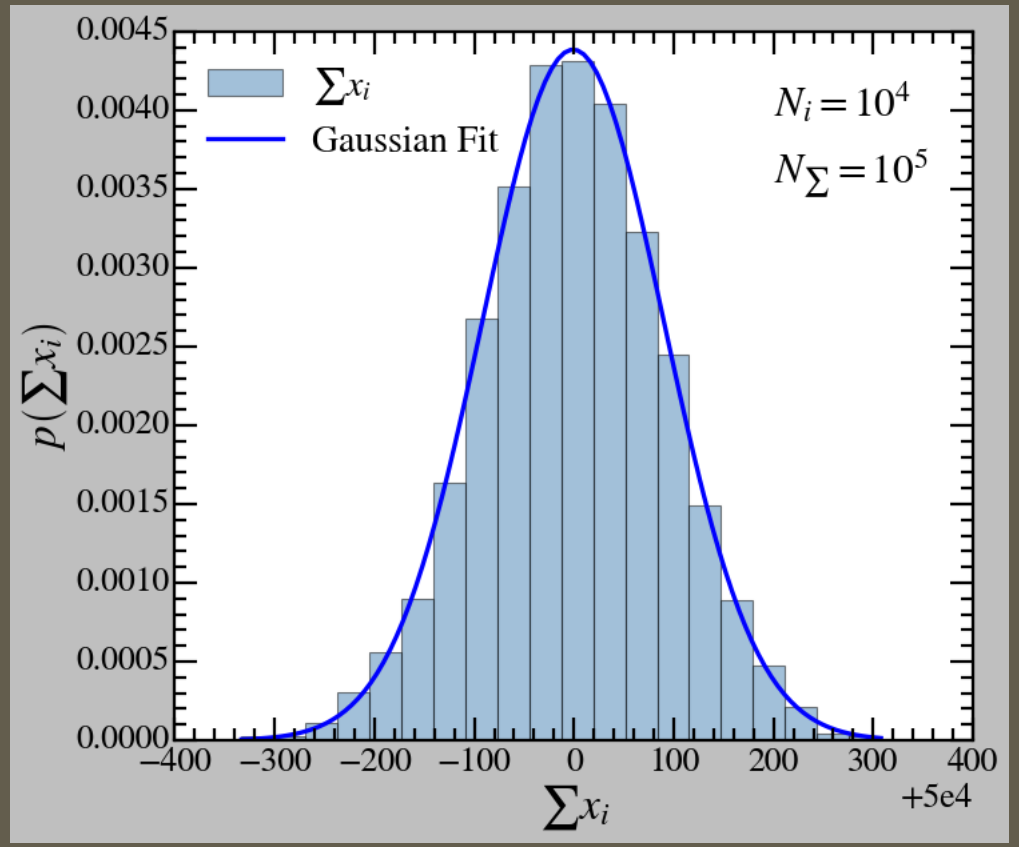
The "normal" distribution

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- $\langle x \rangle = \mu$
- $\sigma^2 = \sigma^2$

(!) The gaussian distribution becomes the Poisson distribution when $\mu = \sigma^2$ and $\lambda \to \infty$.

(!) Central limit theorem: the distribution of the sum of large enough independent identical random variables is Gaussian (under certain condition*).



$$p(x_i) \equiv U(0,1)$$

Sampling

What is Sampling?

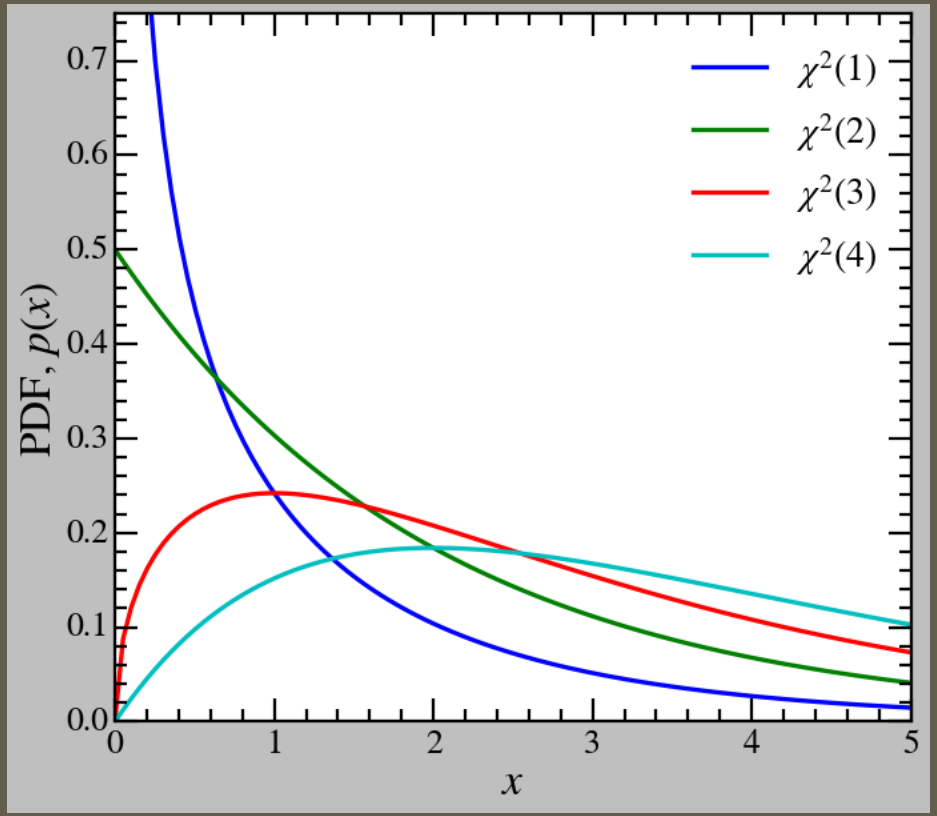
- Given a distribution $p(x)$, return a set of number $\{x_i\}$.
- The distribution of x_i must follow $p(x)$.
- Built-in sampling functions (scipy.stats), most commonly used are the uniform $U(0,1)$ and normal $N(\mu, \sigma)$ sampling, serving as basis for the sampling of other distributions.
- Inversion sampling:
  - sample y_i from $U(0,1)$
  - $x_i = CDF^{-1}(y_i)$.
- Rejection sampling: *ask me later.

Chi Square Distribution

$\chi^2(x, k)$

- $Y = \sum_{i=1}^{k} X_i^2$
- $X_i = \dfrac{x_i - \mu_i}{\sigma_i}$
- $p(x_i) \equiv N(x_i,\ \mu_i, \sigma_i)$

- $p(Y) = \chi^2(Y, k)$
- $\langle Y \rangle = k$
- $\sigma^2 = 2k$



Statistical Uncertainty

Sources of Uncertainty

- Limited resolution
- Uncertainty from instruments
- Counting error
- Statistical spread of sample
- Systematic errors and errors from physical phenomena (often calibratable)
- …

Measuring Uncertainty & Error

- Resolution (angular resolution of telescope)
- Test on instrument (measure noise from sensor)
- Poisson uncertainty in counting error
- Standard deviation of sample as error
- Monte Carlo methods
- Physics models and simulations for calibration
- Error propagation
- …

Error propagation

- Assume measured variables are gaussian distributed.
- Error of $y = f(x_1, x_2, \ldots, x_n)$, assuming x_i are not correlate

$$\Delta y = \sqrt{\left(\frac{dy}{dx_1}\right)^2 \Delta x_1^2 + \left(\frac{dy}{dx_2}\right)^2 \Delta x_2^2 + \ \ldots + \left(\frac{dy}{dx_n}\right)^2 \Delta x_n^2}$$

Likelihood Statistics

Likelihood and Inference

Likelihood, Prior, and Posterior

- Conditional probability: given a condition B, the probability of A is $p(A \mid B)$.
- Likelihood: The probability of the observed measurement given a prior $p(X \mid \theta)$.
- Prior: An assumption about the underlaying mechanism/physics of the system θ.
- Posterior (probability): Given the observed measurement, what is the probability of the prior being correct $p(\theta \mid X)$.
- Example: Flip a "fair" coin 10 times, 3 heads,

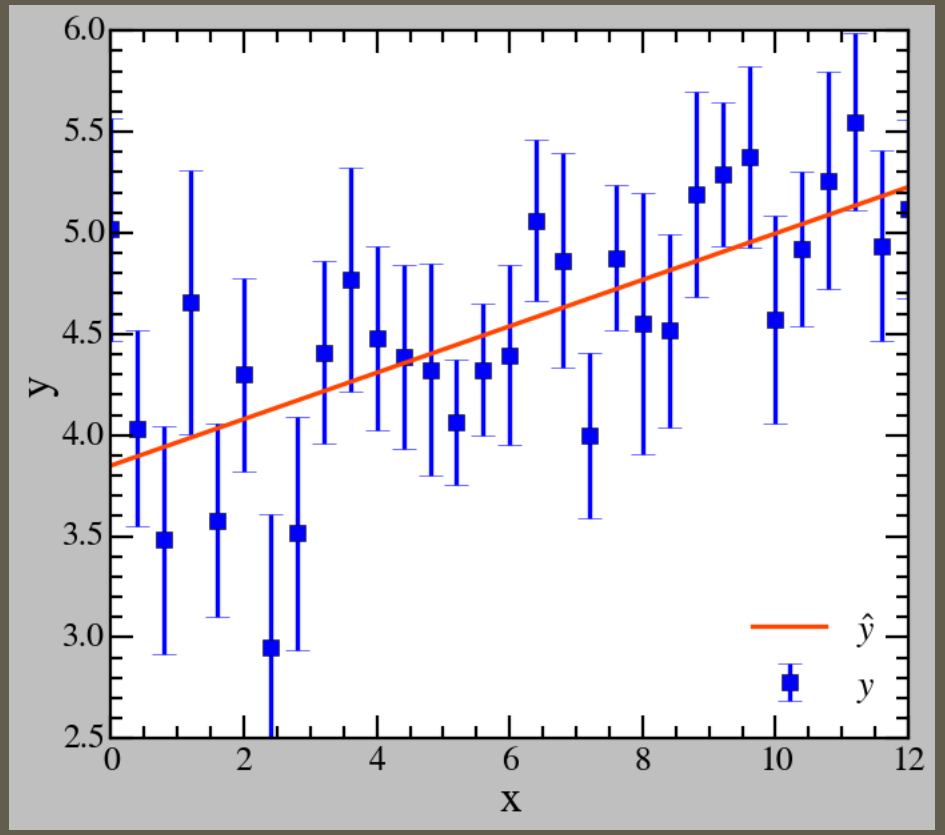
$$p(\theta \mid X) = \frac{p(X \mid \theta)p(\theta)}{p(X)}$$

- $p(\theta)$: the probability of the prior θ.
- $p(X)$: can be understand as the probability of the measurement, but serve more as a normalizing factor $\int \mathrm{d}\theta p(\theta \mid X) = 1.$
- $p(X) = \int \mathrm{d}\theta p(X \mid \theta)p(\theta).$

Maximum Likelihood Fitting

Least-square Fitting

- Residual:
  - $\chi^2(\theta) = \sum_{i=1}^{N} \left(y_i - \hat{y}_i(\theta) \right)^2$
  - $\hat{y}_i(\theta) = f(x_i, \theta)$.
- Least-square fitting criterion: minimize the residual to find the "best fit" parameters/prior θ.
- With errors
  - $\chi^2(\theta) = \sum_{i=1}^{N} \frac{\left(y_i - \hat{y}_i(\theta) \right)^2}{\sigma_i^2}$.

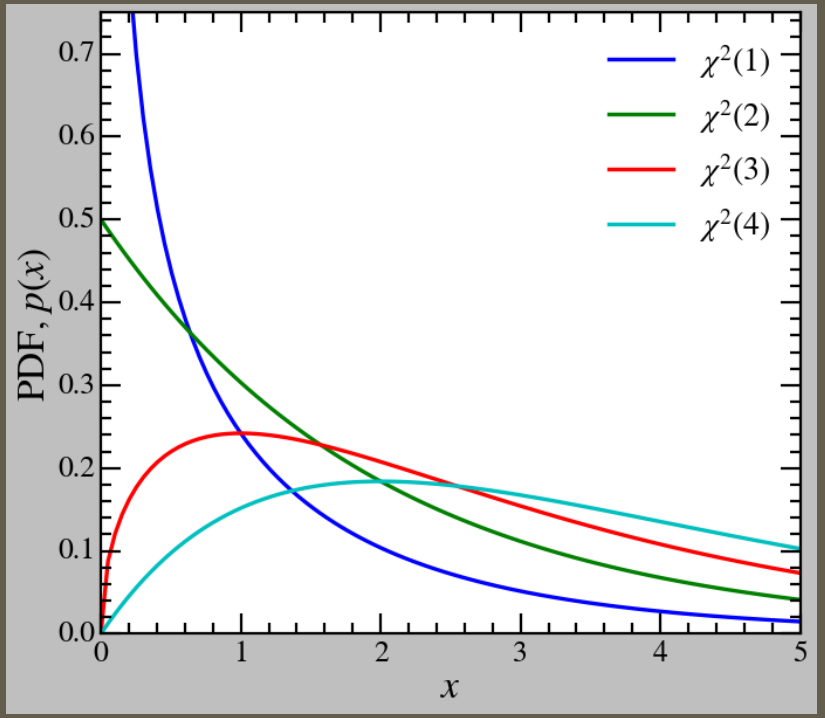


Prior: $\hat{y} = f(x, \theta) = ax + b$

$\theta = (a, b)$

Why Is the Residual Called χ^2?

- Residual with errors
  - $$\chi^2(\theta) = \sum_{i=1}^{N} \frac{\left(y_i - \hat{y}_i(\theta)\right)^2}{\sigma_i^2}.$$
- Assuming noise is gaussian
  - $$Y = \sum_{i=1}^{k} \left(\frac{x_i - \mu_i}{\sigma_i}\right)^2$$
  - $p(x_i) \equiv N(x_i,\ \mu_i, \sigma_i)$
  - $p(Y) = \chi^2(Y, k).$
- The residual follow a Chi square distribution



Least-square ≡ Maximum Likelihood

- Given the prior $f(x,\theta)$ and the the measurement $\{(x_i, y_i)\}$, what is the likelihood, assuming measurements have gaussian nature?

$$LH = \prod_{i=1}^{N} N(y_i, \hat{y}_i(\theta), \sigma_i)$$

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{\left(y_i - \hat{y}_i(\theta)\right)^2}{2\sigma_i^2}}$$

$$LLH = \ln(LH)$$

$$= -\frac{1}{2}\sum_{i=1}^{N} \frac{\left(y_i - \hat{y}_i(\theta)\right)^2}{\sigma_i^2} + C$$

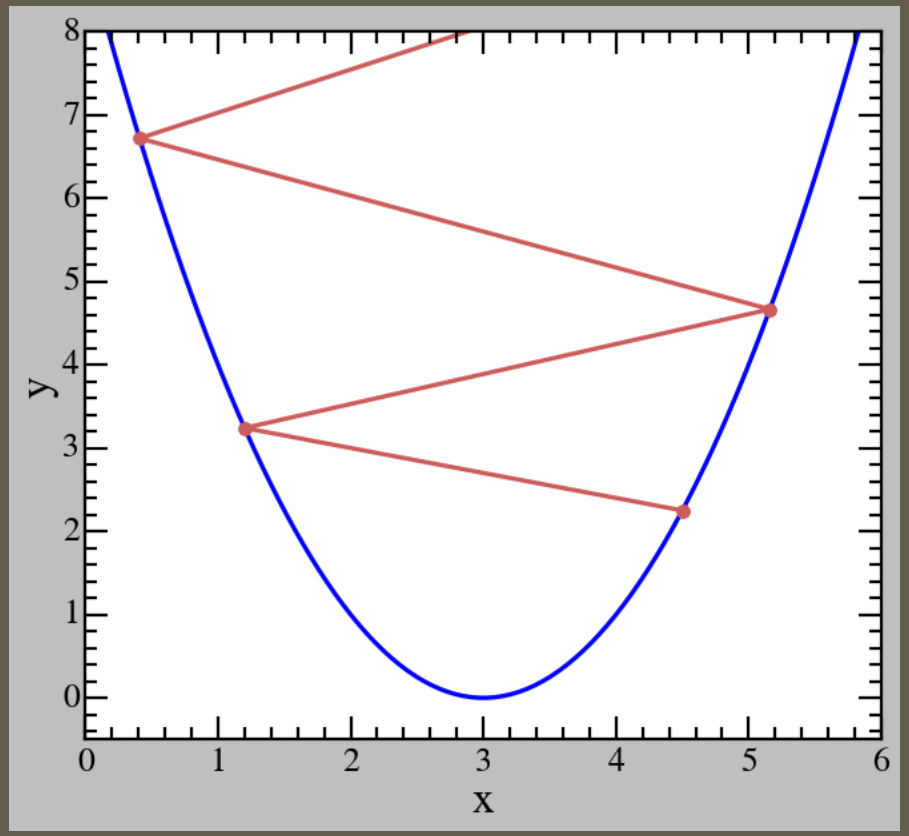
$$= -\frac{1}{2}\chi^2(\theta) + C$$

- Minimizing the residual/least-square error is equivalent to maximizing the likelihood.

Gradient Descend

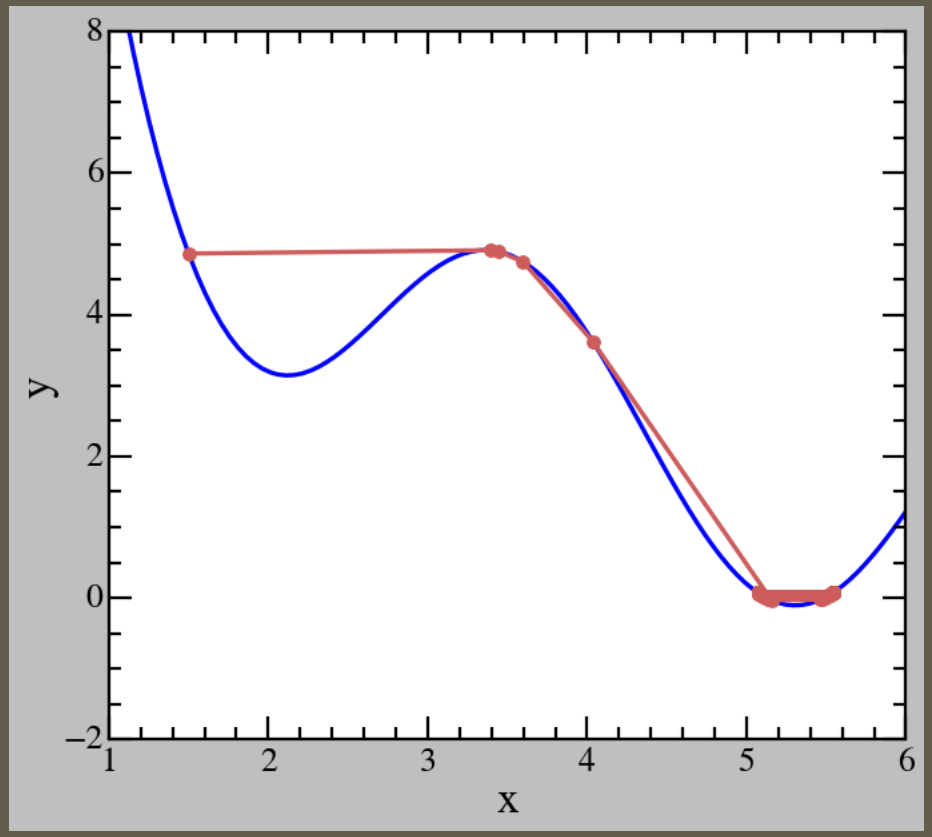
Physics Motivation

- Problem: Given a function $y = f(x)$, find $x_{\min}$ that minimize y.
- A commonly used method: gradient descend.
- Gradient descend: Take inspiration from physics, treat $y = f(x)$ similar to the gravitational potential $U(x)$, let the point naturally "fall down" to the lowest point.
- Iterative step: $x_{k+1} = x_k - \alpha f'(x_k)$.
- α: the learning rate.



Local Minimum

- Choosing a bad initial guess for the gradient descend can result in a local minimum.
- A potential solution is increasing the learning rate α.
- Another is to use more complicated algorithms, e.g. stochastic gradient descend,..



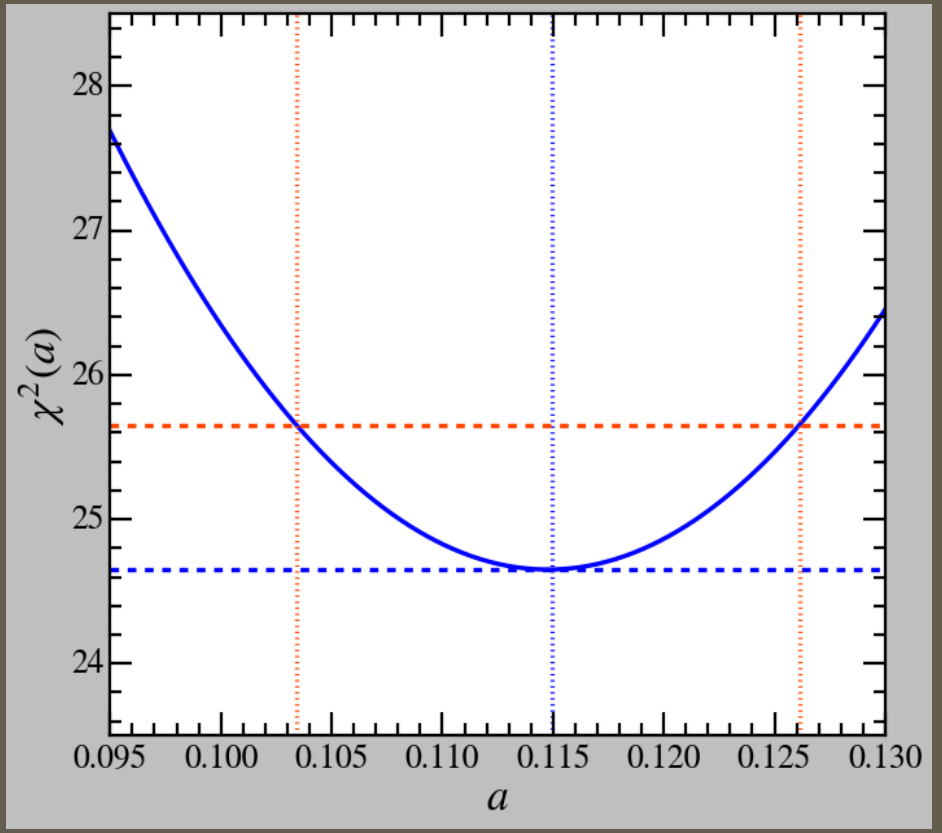
Uncertainty in Fitted Parameters

Uncertainty in Fitted Parameters

- Reminder: The residual follow a χ^2 distribution with $\mathrm{N}^* = N - \mathrm{DOF}(\theta)$ DOFs (N is the number of data point, $\mathrm{DOF}(\theta)$ is the number of parameters of the prior.
- Fixing all except 1 fitting parameter θ_1 would reduce the chi square distribution to 1 DOF,

$$\chi^2(\theta) = \chi^2(\theta_1, 1) + C = \frac{(\theta_1 - \theta_{1,0})^2}{\sigma_{\theta_1}^2} + C.$$

- Taking $\theta_{1,0}$ to be the parameter with the highest likelihood, varying θ_1 by σ_{θ_1} would vary $\chi^2(\theta)$ by 1 and vice versa.

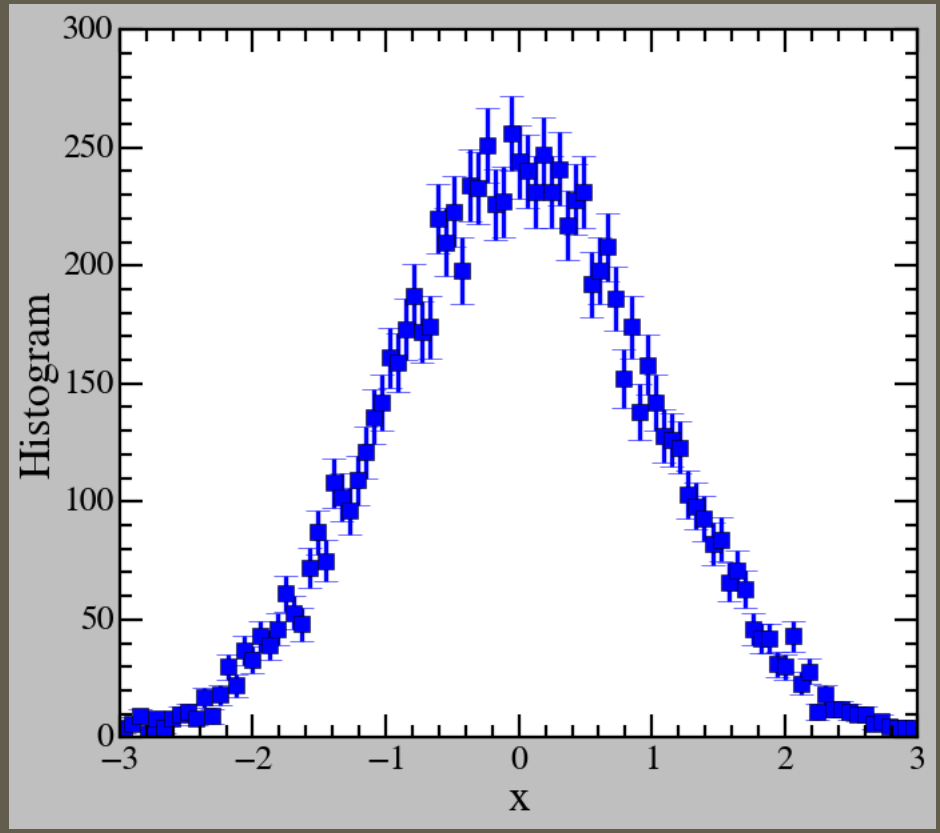


$$a = 0.115 \pm 0.011$$

Cubic Spline Method

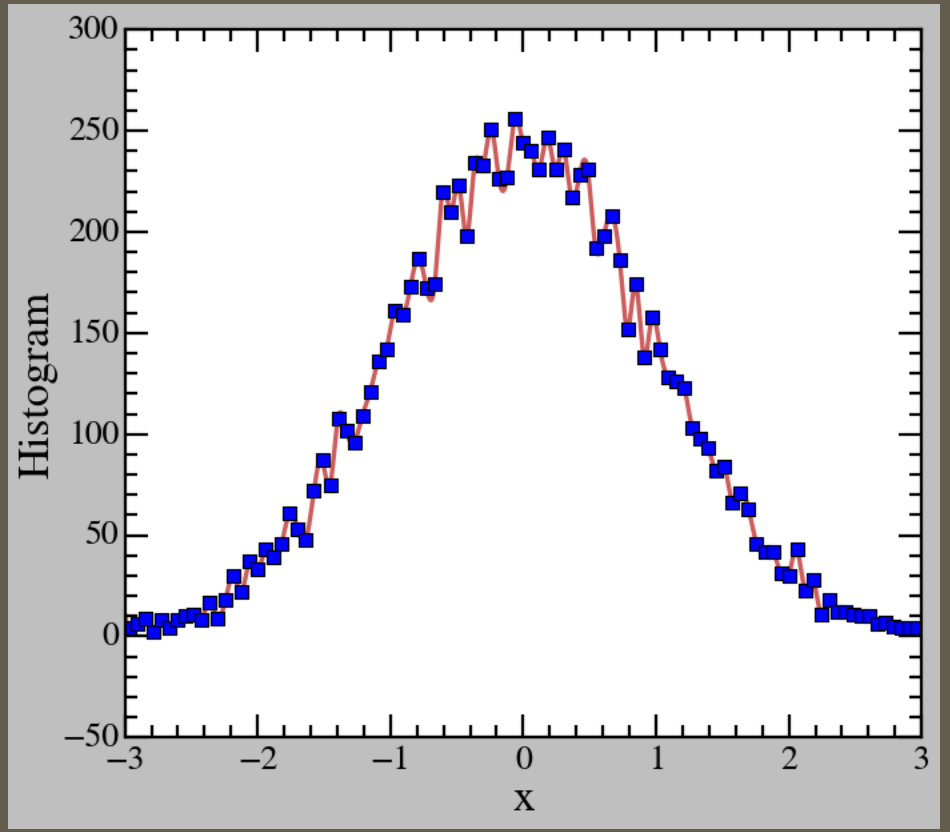
Histogram

- Histogram: a type of plot showing how many points x_i of a sample $\{x_i\}$ are within each of a set of range $\left\{\left[x_{low,i}\,;x_{high,i}\right]\right\}$ called bins.
- Histogram describe the distribution of a sample. A normalized histogram is basically a probability density function $p(x)$.
- Example: sampling 10000 numbers from a normal/gaussian distribution.
- Error bars are from (Poisson) counting errors.



From Scatter Points to Function

- Problem: what if we want to sample points from the distribution described by a histogram via the inversion sampling method, i.e. require a function form or interpolation of the CDF.
- (Cubic) spline fit: every 5 points, fit a third order polynomial and ensuring continuity (i.e. requiring the function and first order derivative to be continuous).
- Does not work very well when uncertainties are high or point are too dense, in which case, the fit function similar to a linear interpolation. Also does not work well with extrapolate, unless



Benefits and (more) Limits

- Derivative and integral are easier to perform
- Spline fitting and interpolation reduce computing cost in many complicated functions, especially those that require integration (talk to me, my main research employ a lot of these techniques).
- Spline fit always include original data points, so they may be overfitting.
- Again, edge cases and extrapolation require special care.
- Employing spline fit is more art than science!

Hypothesis Testing

Basis of Hypothesis Testing

Bayes' Theorem

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)}$$

- $p(\theta)$: the probability of the prior θ
- $p(X|\theta)$: likelihood
- $p(\theta|X)$: posterior probability
- $p(X)$:

can be understand as the probability of the measurement, but serve more as a normalizing factor

$$\int d\theta p(\theta|X) = 1;\ p(X) = \int d\theta p(X|\theta)p(\theta)$$

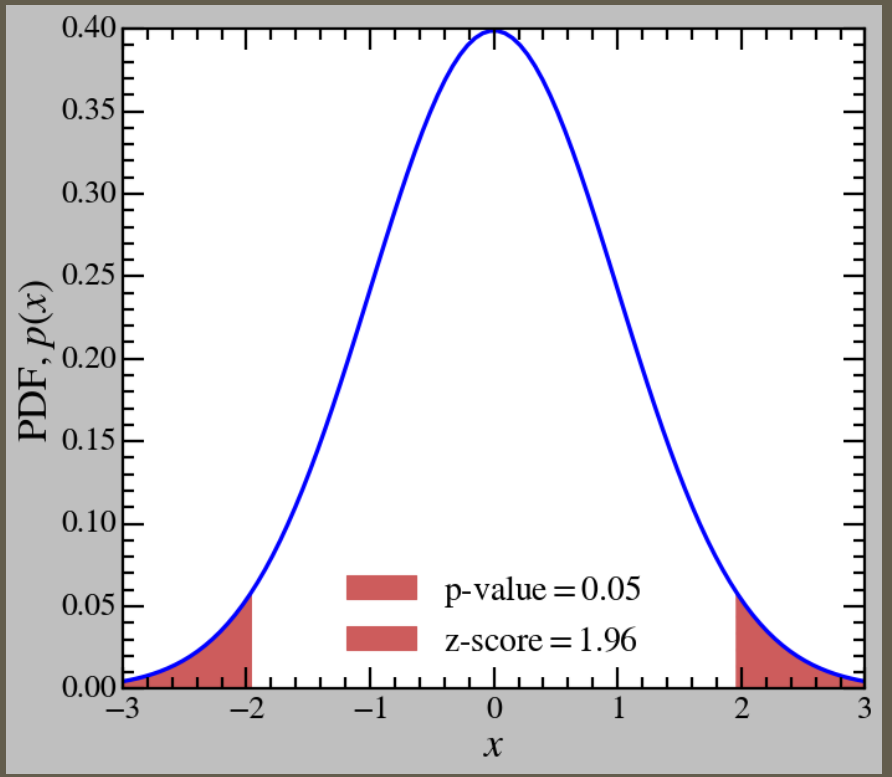
- Example 1: 2 coins, 1 fair 1 bias, choose 1, flip 10 times, 3 heads 7 tails. What is the probability of the coin being fair?
- Example 2: Monty Hall problem. 3 doors, 2 empty 1 with prize. You choose door 1. Door 2 is shown to be empty. Do you switch to door 3 or not?

p-value & Null Hypothesis

- Null hypothesis (H_0): the prior that there is nothing "special", e.g. a constant function $y = b$, only background signal/noise..
- Alternative hypothesis (H_1): the prior that there is something "interesting", e.g. signal from objects, new physics..
- p-value: the cumulative likelihood that the null hypothesis can explain the observed data/measurement, i.e. the probability that fluctuation from the uncertainty in the null hypothesis/prior can result in the observed data/measurement.
- A general standard is if p-value $< \alpha = 0.05$, the null hypothesis is most likely false.

Visualizing p-value

- p-value is often taken as the CDF of the distribution that governs the outcome of the measurement, i.e. the chi square distribution. This is the single-tail p-value. However, beware of high p-value in the chi square distribution!
- The double-tail p-value takes into account both ends of the distribution, i.e. all regions with the probability distribution function less or equal than that of the observed likelihood. This is more trust worthy in most case, exceptions include the chi square distribution.
- p-value can be translate to z-score, which describe, in the context that the distribution in question is mapped to a gaussian/normal distribution, how many standard deviations



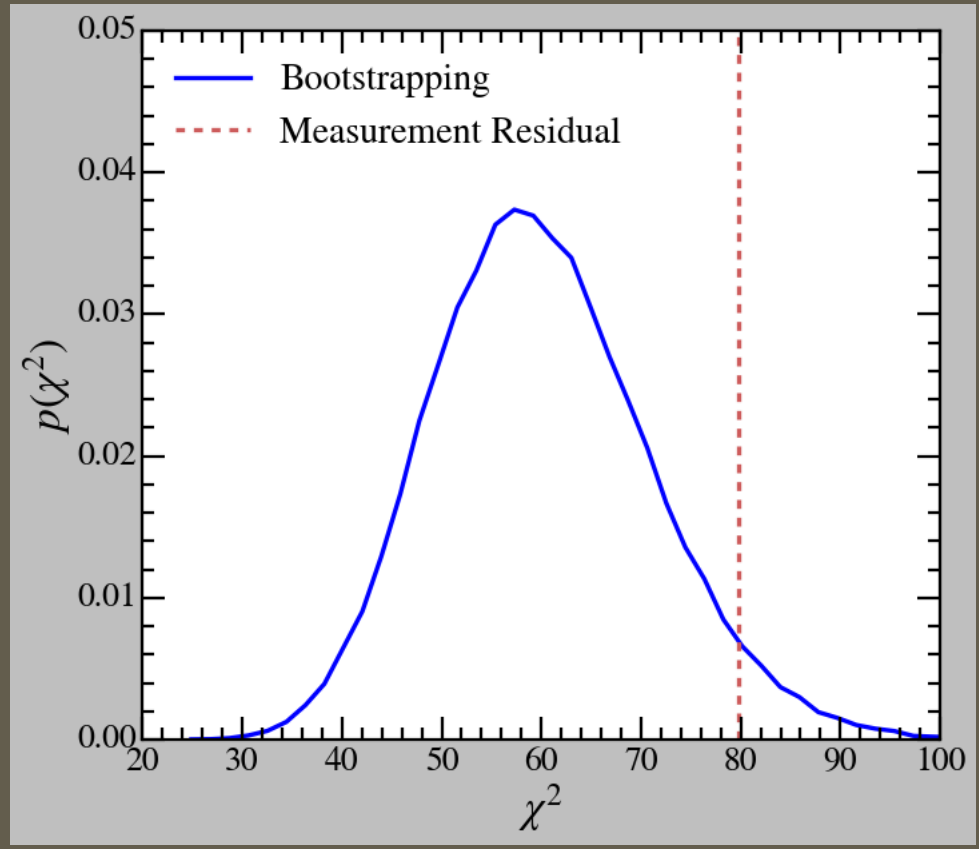
Bootstrapping

In the Case of Complicated Null Hypothesis

- Example: Absorption line with a black-body background

$$F(\lambda, T) = \frac{2hc^2}{\lambda^5} \frac{1}{e^{\frac{hc/\lambda}{k_B T}} - 1}$$

- Null hypothesis: only black-body background. Let's try fitting the background.
- Bootstrapping: a Monte Carlo method, from the null hypothesis, try to sample/simulate the measurement a large number of time, and inspect the distribution of the resulting configurations.
- In this case, let's look at the distribution of the residual χ^2. Different hypothesis tests have different

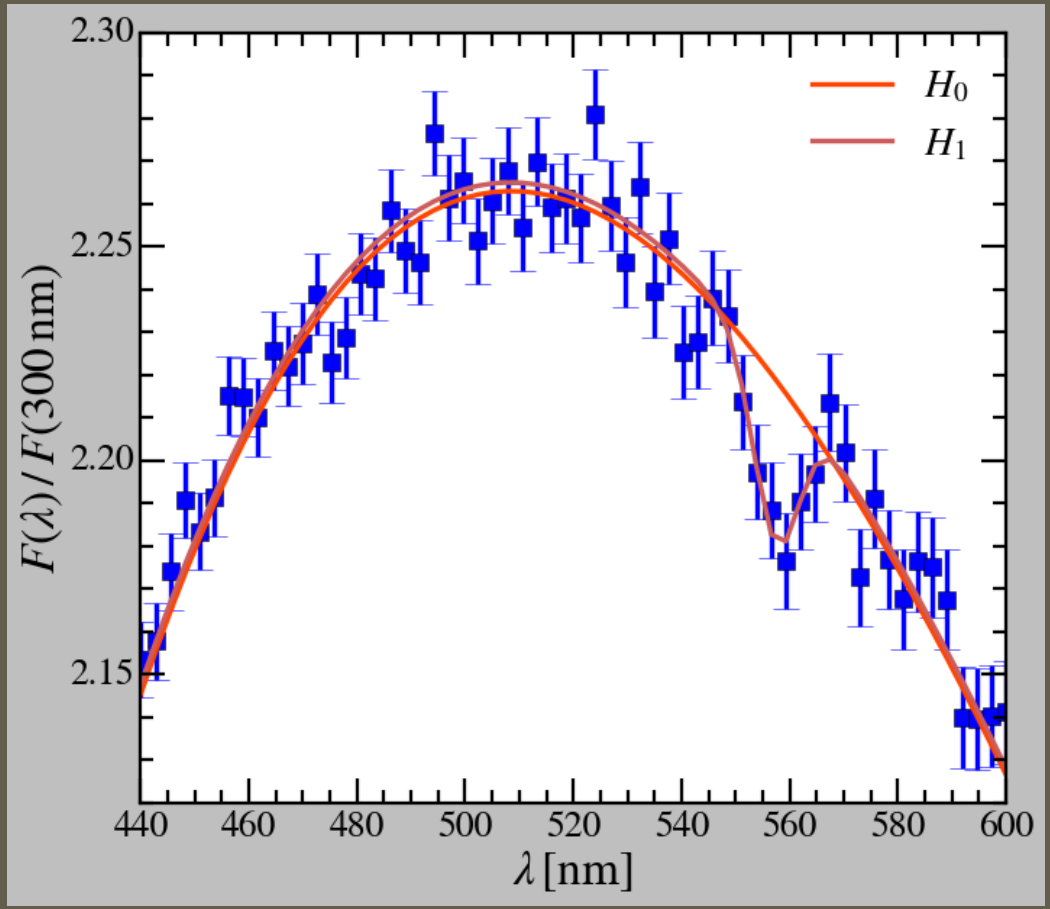


Likelihood-ratio Testing

Null Hypothesis v.s. Alternative Hypothesis

- Example: Absorption line with a black-body background
- Null hypothesis H_0: only black-body background.
- Alternative hypothesis H_1: black-body background with an absorption feature, modeled with a gaussian function form: $I(\lambda) = I_{BB}(\lambda) - I_{ap} \cdot N\left(\lambda, \lambda_{ap}, \sigma_{ap}\right)$.
- Reminder: $\chi^2(\theta) = -2\ln\left(p(X|\theta)\right) + C$

$$\frac{p(X|H_0)}{} = \exp\left(\chi^2(H_0) - \chi^2(H_1) \right)$$



Null Hypothesis v.s. Alternative Hypothesis

- From Wilks' theorem:

$$-2\ln(\Lambda) = -2\ln\left(\frac{p(X|H_0)}{p(X|H_1)}\right) = \chi^2(H_0) - \chi^2(H_1)$$

follows a chi square distribution with $\Delta N = N_1 - N_0$ DOFs (N_1 and N_0 are the number of free parameters in H_1 and H_0, respectively) under the null hypothesis.

- $-2\ln(\Lambda)$ represents the probability that under the null hypothesis, the observation/measurement can happen.
- p-value can be quoted!