Probability and Statistics



MPhys Year 2

Michaelmas Term, 2026

Welcome to Year 2 Probability and Statistics

- " The result of this experiment was inconclusive, so we had to use statistics"
- " There are 3 kinds of lies: Lies, dammned lies and statistics "

Bad stats competition

Send me your examples of real-life ‘bad stats‘

- Bad visualisation - results/inference not wrong but presented to lead a different conclusion. e.g Narrow y-axis range to magnify differences that are not significant.
- Bad inference - incorrect inference of what the data is saying. Monetary prize offered for this category.
- Sources should come from the press/journals/blogs etc with a few lines (or more) on what is wrong. Send to me via email.
- Deadline: Friday wk 3
- Discussion in lecture 8

Why should you engage in this lecture course?

- Characterizing data, and drawing inferences from it is an integral part to being a scientist.
- Is science a "fact", or is it a "degree of uncertainty"?
- Bad statistics is really bad. You want to be able to tell the difference

This course aims to prepare you for the *Practical Course* by giving you a (slightly) deeper introduction to probability theory and its applications to data analysis.

More generally, the concepts we will cover are fundamental to how we do science.

What this course aims to teach

- The syllabus is narrow
- Depending what you learnt at school you may have covered a fair bit of this already
- Hope there will be new items for all of you.
- We'll try and be a little more rigourous with the mathematics
- It is not the aim to teach you all you will need to know about statistics
- It is the aim to give you the tools to ask the right questions.
- It is the aim to give the basics so that when the time comes you can understand textbooks and papers and determine if that is the information you are looking for.

Where is this course assessed / useful

- Explicitly on the Part A syllabus. [Disclaimer: I'm not an examiner so its not my call.]
- Many Part B exams use graphs and implicitly expect base knowledge presented here.
- Saves some time during the practical course
- May help you in your project(s) as part of your course.
- Helps you look smarter to your project or internship supervisor.
- After you leave here, many of you will work with data and inference.

Studies and real life

The answers to statistics questions (e.g those on the problem sheet) have well defined answers.

Most practical course experiments have been tried and tested over the years and the right answer to the way to do the associated data analysis is reasonably defined.

How you approach data analysis on something new is not. The questions to ask are

- Is what I'm doing sensible?
- What possible biases could I end up with?
- Is this the best way?
- What can I anticipate potentially going wrong, what statistical tests will I use, and when will these be deployed?

Outline

Lecture 1: Science, Probability, and Statistics

Lecture 2: Some Common Probability Distributions, Expectation and Variance

Lecture 3: Correlation, Central Limit Theorem, Uncertainty propagation, Estimators

Lecture 4: Estimators, Data Modeling, Maximum Likelihood

Lecture 5: Confidence intervals, Least squares and Hypothesis testing

Lecture 6: Hypothesis testing, Confidence intervals, Systematic uncertainties and checks

Lecture 7: Combinations, Bayesian Statistics

Lecture notes

Two versions available on Canvas for Lectures 1-7

- Ones on display during the lecture which will have white boxes for me to fill in. You could fill these in as we go if you want.
- Ones that have the white boxes already filled in. If you don't want to fill in during the lecture you can use these to annotate the existing notes

Ultimately these are maths lectures - you will need to go back over them and make sure you can get from one line to the next.
Lecture 8 will go through the bad stats examples - slides will be made available after the lecture.

What else?

- A problem set is available for you to attempt and go through with your college tutors.
- Recommended reading:
  - Statistical Data Analysis by G. Cowan, Oxford University Press
  - *Data Analysis: A Bayesian Tutorial* by D. Sivia & J. Skilling, Oxford University Press, 2006 (2^{nd} edition);
  - *Practical statistics for Astronomers* by J. W. Wall & C. R. Jenkins, Cambridge University Press, 2003;
  - *Statistics A guide to the use of statistical methods in the physical sciences* by R. J Barlow, John Wiley & Sons

  Note that all four of the above go significantly beyond the syllabus of this course.

Acknowledgements

I am grateful to the Prof. Suzanne Aigrain
There are a series of python notebooks written by Prof. Aigrain, available on the canvas page.
The ordering of these is different to the lectures here, but I do recommend working through them at some point to see how the concepts covered in lectures can be used for specific situations.

Questions?

- Raise your hand during or at the end of the lecture
- You can come and speak to me at the end of the lecture
- Email me: sneha · malde @ physics · ox · ac · uk

Deductive versus inductive logic

In mathematics and in games of chance, we use ***deductive logic***: start from some axiom or cause and deduce the consequences.

In real world problems, we observe the consequences (experimental data) and seek to infer the causes (the laws of Physics). For this we need a different kind of logic, called ***inductive logic*** – or plausible reasoning.

What do we do as scientists?

A highly simplified workflow for a scientist might be:

- Start with a ***theory*** that explains what we know so far about some aspect of the real world.
- Design an ***experiment*** to test that theory (make a ***prediction***)
- Carry out the experiment to get some ***data***
- Compare the data to the prediction: are they ***consistent***?
  - If yes, belief in our theory is reinforced, and we can start another test
  - If not, we need to modify our theory.
- Sometimes theory doesn't make a prediction, you just try and measure something. Then you compare your measurement to other measurements.

What do we do as scientists?

Even this idealised workflow raises important questions:

- How do we account for errors in the data?
- What do we mean by consistent?
- What about all the other possible theories we didn't consider?

Statistics are numbers (constants) we construct from the data that enable us to quantify the comparison between data and theories.

To define statistics that are useful, we need probability theory.

Definitions: Probability : Frequentist

Two definitions that are very different conceptually:

Frequentist definition: The probability $P(X)$ of some proposition X represents the ***frequency*** with which this proposition is realised in some (often hypothetical) large number of trials, or repetitions, of a given experiment.

Dice rolling, picking coloured balls out of sacks, picking cards.
Probability of decay before time t of a radioactive isotope.
Probability of finding a molecule with speed higher than x if parameters of the gas are known.
The Frequentist definition is well known to you.
It is viewed as "objective".

Definitions: Probability : Bayesian

Bayesian definition: The probability $P(X)$ of some proposition X represents the experimenter's ***degree of belief*** in the validity of that proposition given all available information.

Probability that someone in your college wins the prize. Probability that it rains tomorrow. Not repeatable events. It is viewed as "subjective". This is also well known to you.
In both cases, all probabilities are ***real numbers between 0 and 1***.

The Bayesian definition is more general. However, for many situations you'll encounter as an undergraduate the former is sufficient.

When should I use which?

Both definitions are ok (and agree) when

- Our experiment is repeatable and self-contained
- Our measurements are precise and accurate enough

The Bayesian definition works irrespective of these limitations.

Frequentist statistics are useful when

- We lack a generative model
- We can't afford expensive calculations

We will look at the two methods in closer detail later. Bayes' theorem is true for both - don't conflate that with Baysian statistics.

Definitions: Sample space

Let A, B, C denote possible outcomes of an experiment. The ***sample space*** or ***state space*** of the problem Ω is the set of all possible outcomes.

For a coin toss or a draw from a pack of cards, Ω is a discrete list.

For a measurement of a continuous scalar variable, Ω might be the set of real numbers, $\mathbb{R}$ (we will come back to this).

Definitions: Joint and conditional probability

The ***joint probability*** of A and B, written $P(A, B)$ is the probability that both A and B are true.

The ***conditional probability*** of B given A, $P(B|A)$, is the probability that B is true if we already know that A is true.

Conditional probabilities are closely related to the notions of ***correlation*** and ***independence***.

Basic rules: The sum rule

The ***sum rule***:

$$P(A) + P(\overline{A}) = 1,$$

where $\overline{A}$ denotes "not A".

If the sample space is $\Omega = \{A_1, A_2, \ldots, A_n\}$,

$$P(A_1) + P(A_2) + \ldots + P(A_n) = \sum_{i=1}^{n} P(A_i) = 1,$$

Basic rules: The product rule

The ***product rule***:

$$P(A, B) = P(A|B)\, P(B)$$

If A and B are ***independent*** of each other, knowing B doesn't change what we know about A.

$$P(A|B) = P(A).$$

Therefore, A independent of $B \iff$

$$P(A, B) = P(A)\, P(B).$$

Bayes theorem

Joint probabilities are commutative:

$$P(B, A) = P(A, B)$$

Plug in the product rule and rearrange to get ***Bayes' theorem***:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Lets try an example

Marginalisation rule

Combining the product and the sum rules gives the ***marginalisation rule***:

$$\begin{aligned} P(A) &= P(A, B_1) + P(A, B_2) + \ldots + P(A, B_n) \\ &= \sum_{i=1}^{n} P(A, B_i) \\ &= \sum_{i=1}^{n} P(A|B_i)\, P(B_i) \end{aligned}$$

Example

Definitions: Random variable

A ***random variable*** (RV) is a function mapping the sample space Ω of possible outcomes of a random process to the space of real numbers $\mathbb{R}$.

For example, when tossing a coin once, the RV X can be defined as:

$$X = \begin{cases} 0 & \text{if coin lands tails up,} \\ 1 & \text{if coin lands heads up.} \end{cases}$$

When throwing a die once ... $X = \{1, 2, 3, 4, 5, 6\}$.

When drawing a ball from an urn containing red and white balls ...

$$X = \begin{cases} 0 & \text{if ball is white,} \\ 1 & \text{if ball is red} \end{cases}$$

Discrete and continuous RVs

The previous examples were all ***discrete*** RVs (only a specific set of outcomes is possible). An RV can also be ***continuous***.

For example, if X is an RV describing a temperature measurement in degrees Kelvin, X will be a positive real number.

Definitions: Probability mass function

Every RV has an associated ***probability distribution***.

For a discrete RV, it is called the ***probability mass function*** (pmf), and gives the probability of each outcome.

The quantity $P(X = x_i) = P_i$ represents the probability that the RV X takes the value x_i.

Hereafter we will use the shorthand notation $P(x_i)$ to mean $P(X = x_i)$.

For example, if X represents the outcome of throwing a fair, 6-sided dice once ... $P(x_i) = 1/6 \; \forall \; x_i \in \{1, 2, 3, 4, 5, 6\}$ and 0 otherwise.

Definitions: Probability density function

Every RV has an associated ***probability distribution***.

For a continuous RV, it is called the ***probability density function*** (pdf), denoted by $p(X)$.

The quantity $p(x)\mathrm{d}x$ represents the probability that the RV X takes a value between x and $x + \mathrm{d}x$.

The choice of probability distribution to associate to a given RV is dictated by the nature of the random process that gives rise to it.

Definitions: Cumulative probability distribution

The ***cumulative distribution function*** (cdf) gives the probability $P(X \leq x_i)$ that the RV X takes a value less than or equal to x_i.

For a discrete RV,

$$C(x_i) = \sum_{j=1}^{i} P(x_j).$$

In the dice example discussed above ... $C(x_i) = x_i/6$

For a continuous RV defined over the real line,

$$C(x) = \int_{-\infty}^{x} p(y)\,\mathrm{d}y.$$

How do probability distributions relate to the real world?

When we make a measurement, nature selects a value from the sample space, with a probability given by the associated pmf or pdf.

Frequentist interpretation: if we repeated this measurement an infinite number of times, the relative frequency of each outcome would be the same as the pmf or pdf.

The outcome of a series of measurements is called a ***sample***. It is a series of real numbers: $\{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n\}$.

We model RVs but our data are constants

Almost all experiments involve randomness:

- some physical processes are truly random, e.g. radioactive decay;
- some are deterministic on microscopic scales, but effectively random, e.g. a coin toss;
- any observation of a continuous variable (random or not) is always affected by random ***measurement errors***.

Therefore, models used to interpret data almost always involve RVs.

Once we have measured something, though, we can't change it. ***Data are constants.***

Statistics, which are computed from the data, are also constants.

The uniform distribution

We encountered this already in the case of tossing a coin, or rolling a dice, once.

For n equally probable outcomes, the ***uniform distribution*** $\mathcal{U}(n)$ has pmf:

$$P(x_i) = \frac{1}{n} \text{ for } i = 1, \ldots, n$$

and cdf:

$$C(x_i) \equiv P(x \leq x_i) = \frac{i}{n} \text{ for } i = 1, \ldots, n.$$

The uniform distribution

For a continuous RV defined over the interval $[a, b]$, the ***uniform distribution*** $\mathcal{U}(a, b)$ has pdf:

$$p(x)\,\mathrm{d}x = \frac{\mathrm{d}x}{b-a}$$

and cdf:

$$C(x) = \int_a^x p(y)\mathrm{d}y = \frac{x-a}{b-a}$$

The uniform distribution

Note that the uniform probability distribution is defined only for RVs with finite sample spaces.

This excludes (e.g) discrete integer RVs, or continuous real RVs, without upper or lower bounds.

For these cases it's possible to define something similar to a uniform pmf or pdf, but un-normalised.

Outline

Lecture 1: Science, Probability, and Statistics

Lecture 2: Some Common Probability Distributions, Expectation and Variance

Lecture 3: Correlation, Central Limit Theorem, Uncertainty propagation, Estimators

Lecture 4: Estimators, Data Modeling, Maximum Likelihood

Lecture 5: Confidence intervals, Least squares and Hypothesis testing

Lecture 6: Hypothesis testing, Confidence intervals, Systematic uncertainties and checks

Lecture 7: Combinations, Bayesian Statistics

Bad stats competition

Send me your examples of real-life ‘bad stats‘

- Bad visualisation - results/inference not wrong but presented to lead a different conclusion. e.g Narrow y-axis range to magnify differences that are not significant.
- Bad inference - incorrect inference of what the data is saying. Monetary prize offered for this category.
- Sources should come from the press/journals/blogs etc with a few lines (or more) on what is wrong. Send to me via email.
- Deadline: Friday wk 3
- Discussion in lecture 8

Outline

Yesterday:

- Basic probability and Bayes' theorem
- Random variable
- Uniform probability distribution

Today:

- Other common probability distributions
- Central Limit Theorem

The binomial distribution

The ***binomial distribution*** $\mathcal{B}(n, p)$ describes the probability of obtaining r "successes" in a sequence of n independent trials, each of which has probability p of success.

Here, "success" can be defined as one specific outcome in a binary process (such as tossing or drawing one ball from an urn containing two different colours).

The binomial distribution

To work this out consider:

- The probability of obtaining r "successes" in a row is p^r.
- The probability of obtaining $n - r$ "failures" in a row is

$$(1-p)^{(n-r)}.$$

- The number of ways of re-ordering the successes and failures is

$$\binom{n}{r} \equiv C_r^n = \frac{n!}{(n-r)!\,r!}.$$

The binomial distribution

Putting all this together, the binomial distribution has pmf:

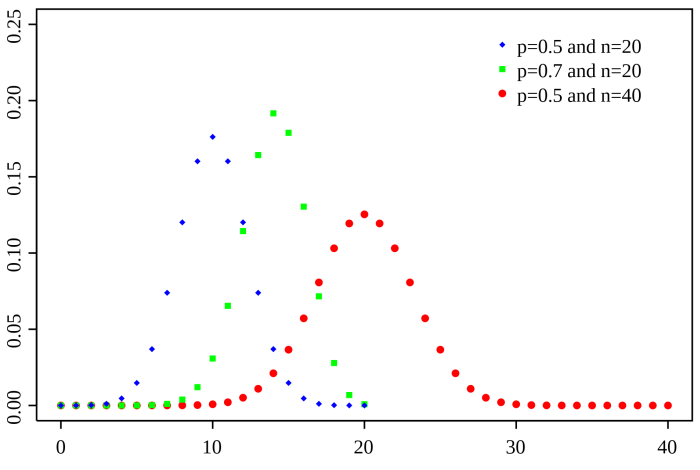
$$P(r|n,p) = \binom{n}{r} p^r \, (1-p)^{n-r}.$$

The cmf is

$$P(r \leq k|n,p) = I_{1-p}(n-k, k+1),$$

where $I_x(a,b)$ is the *regularized incomplete beta function*.

The binomial distribution



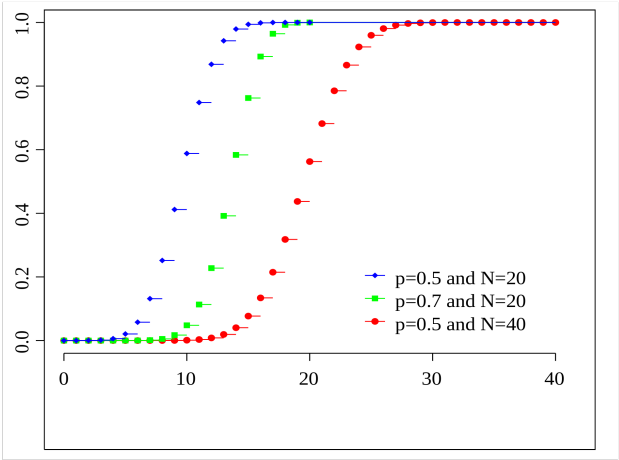


Figure: Pmf and cmf of the binomial distribution for different values of p and n. Source: Wikipedia.

Binomial distribution examples

A biased coin has probability of landing heads $p = 0.7$. You toss it $n = 10$ times. What is the probability of getting $r = 6$ heads?

$$\begin{aligned} P(r = 6|n = 10, p = 0.7) &= \tfrac{10!}{6!\,4!}\, 0.7^6\, 0.3^4 \\ &= 210 \times 0.112 \times 0.008 = 0.2 \end{aligned}$$

I toss a coin 20 times and obtain 15 heads. Is my coin biased?

This is an inference question, which we don't know how answer yet. However, we can use the cmf to work out the probability of getting at least 15 heads from 20 tosses of a fair coin:

$$\begin{aligned} P(r \geq 15|20, 0.5) &= 1 - P(r \leq 14|20, 0.5) = 1 - I_{0.5}(6, 15) \\ &= 1 - 0.979 = 0.021 \end{aligned}$$

The Poisson distribution

The ***Poisson distribution*** $\mathcal{P}(\lambda)$ describes the probability of obtaining a number n of events in a process where events occur with a fixed average rate λ and independently of each other.

Examples:

- number of planes landing at Heathrow per hour (during the working day),
- number of photons or electrons hitting a detector per second (assuming constant source)
- number of galaxies per square arcminute in the sky

Note that, although n is an integer, λ doesn't have to be.
Watch out for things that are not independent (e.g earthquakes)

Deriving the Poisson distribution

Consider a very small time interval δt, so small that the number of events occurring during it can only be 0 or 1.

The probability of 1 event happening in time interval δt is

$$P_1(\delta t) = \lambda\, \delta t.$$

(Throughout this derivation, we denote $P_n(t)$ the probability of n events occurring in time interval t.)

The probability of 0 events in the same time interval δt is then simply

$$P_0(\delta t) = 1 - \lambda\, \delta t.$$

Deriving the Poisson distribution

Now we consider the probability of n events happening in time $t + \delta t$.

There are two ways to get this

- n events in the time interval $[0, t]$ and 0 in $[t, t + \delta t]$, or
- $n - 1$ events in the time interval $[0, t]$ and 1 in $[t, t + \delta t]$.

Putting these together gives

$$\begin{aligned} P_n(t+\delta t) &= P_n(t)\,P_0(\delta t) &+ P_{n-1}(t)\,P_1(\delta t) \\ &= P_n(t)\,(1-\lambda\,\delta t) &+ P_{n-1}(t)\,\lambda\,\delta t \end{aligned}$$

Deriving the Poisson distribution

Rearranging:

$$\frac{P_n(t+\delta t) - P_n(t)}{\delta t} = \lambda[P_{n-1}(t) - P_n(t)].$$

Taking the limit $\delta t \to 0$:

$$\frac{\mathrm{d}P_n(t)}{\mathrm{d}t} = \lambda[P_{n-1}(t) - P_n(t)].$$

You can check for yourself that this differential equation is satisfied by the expression

$$P_n(t) = \frac{(\lambda t)^n}{n!}\, e^{-\lambda\, t}.$$

The Poisson distribution

For problems involving the time domain, the probability of n events happening in a time t is given by the Poisson distribution:

$$P(n|\lambda, t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}.$$

Note that this is a discrete pmf in the number of events n, and not a continuous pdf in t. To calculate the probability of getting n events per unit time interval simply set $t = 1$ in the above equation.

For problems involving the spatial domain, simply replace t by the relevant spatial measures (length l, area A, solid angle Ω, ...)

The Poisson distribution

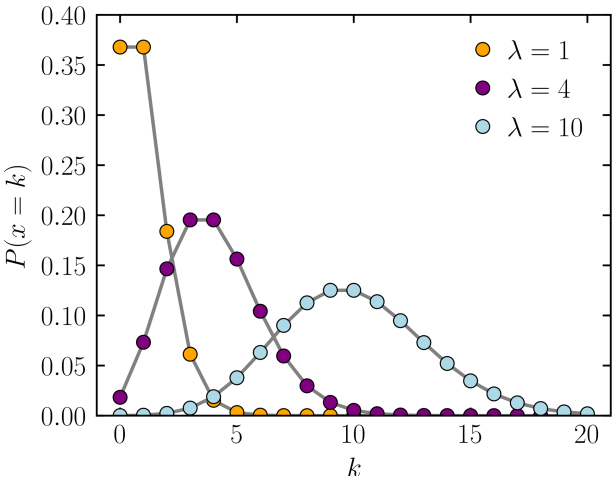
The Poisson cumulative mass function is

$$P(n \leq k|\Lambda) = e^{-\Lambda} \sum_{i=0}^{k} \frac{\Lambda^i}{i!} = \frac{\Gamma(k+1, \Lambda)}{k!},$$

where, Λ is the average number of events in time t: $\Lambda \equiv \lambda\, t$, and $\Gamma(s, x)$ is the *upper incomplete Gamma function*:

$$\Gamma(s, x) = \int_x^\infty y^{s-1}\, e^{-y}\, \mathrm{d}y.$$

The Poisson distribution



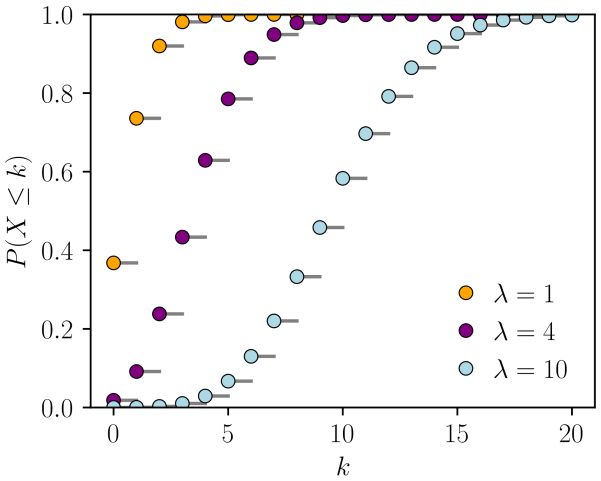


Figure: Pmf and cmf of the Poisson distribution for different values of λ and $t = 1$. Source: Wikipedia.

Poisson distribution examples

A particle detector measures protons which are emitted with an average rate $\lambda = 4.5/\mathrm{s}$. What is the probability of measuring 6 protons in 2 seconds?

$$P(n = 6|\lambda = 4.5, t = 2) = \frac{(4.5 \times 2)^6}{6!} e^{-4.5\times 2} \approx 0.091.$$

In a post office, people arrive at the counter at an average rate of 4 customers per minute. What is the probability of 6 people arriving in a minute?

$$P(n = 6|\lambda = 4, t = 1) = \frac{(4 \times 1)^6}{6!} e^{-4\times 1} \approx 0.104.$$

The exponential distribution

The exponential distribution $\mathcal{E}$ describes the interval between consecutive events in a Poisson process.

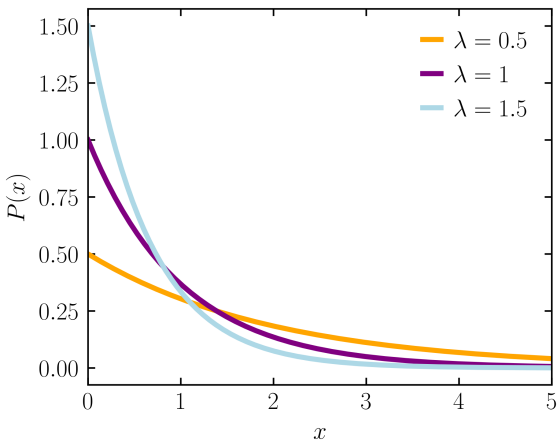
The probability that no event occurs in time interval t is given by the Poisson distribution with $n = 0$: $P(0|\lambda, t) = \exp(-\lambda\, t)$.

The probability that one event occurs during the interval $[t, t + \delta t]$ is given by $P(1|\lambda, \delta t) = \lambda\, \delta t$.

The probability that the first event occurs after time t is thus

$$\begin{aligned} p(t)\,\mathrm{d}t &= \lim_{\delta t \to 0}\{P(0|\lambda, t)\, P(1|\lambda, \delta t)\} \\ &= \lim_{\delta t \to 0}\{\exp(-\lambda\, t)\, \lambda\, \delta t\} \\ &= \lambda\, \exp(-\lambda\, t)\, \mathrm{d}t. \end{aligned}$$

The exponential distribution



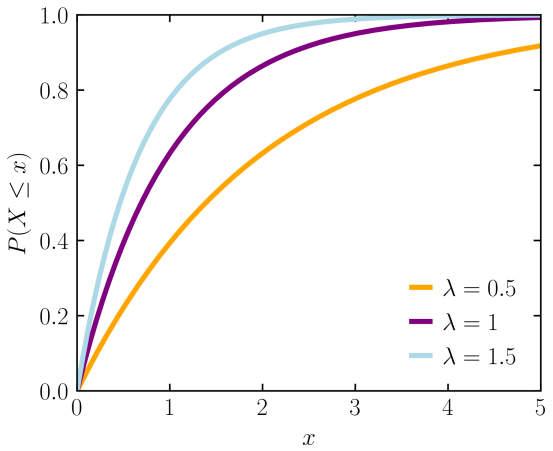


Figure: Pdf and cdf of the exponential distribution for different values of λ.
Source: Wikipedia.

The exponential distribution

Example: Your geiger counter fires. There are around 5 counts per hour with the sample you have. What is the probability that you will have to wait more than 15 minutes for the next signal?

$$P(t > t_0|\lambda) = \int_{t_0}^{\infty} p(t)\,\mathrm{d}t = \int_{t_0}^{\infty} \lambda\, e^{-\lambda\, t}\mathrm{d}t = \left[-e^{-\lambda\, t}\right]_{t_0}^{\infty} = e^{-\lambda\, t_0}$$

Plugging in the numbers, $\lambda = 5/60$ signal/min, $t_0 = 15$ min, $P(t > t_0|\lambda) = e^{-5/4} \approx 0.29$.

The exponential distribution

Example continued: You have waited for 10 minutes already. What is the probability you will have to wait for more than 15 minutes now?

We need to evaluate $P(t > t_0 + t_1 \,|\, t > t_0)$, where $t_0 = 10$ min and $t_1 = 15$ min. Using the definition of conditional probability:

$$P(t > t_0 + t_1 \,|\, t > t_0) = \frac{P(t > t_0 + t_1,\, t > t_0)}{P(t > t_0)} = \frac{P(t > t_0 + t_1)}{P(t > t_0)},$$

where we have used the fact that $t > t_0 + t_1$ necessarily implies $t > t_0$. We know that $P(t > \tau) = e^{-\lambda\,\tau}$. Therefore:

$$P(t > t_0 + t_1 \,|\, t > t_0) = \frac{e^{-\lambda\,(t_0+t_1)}}{e^{-\lambda\,t_0}} = e^{-\lambda\,t_1} = P(t > t_1)$$

The exponential distribution has no memory!

The Gaussian or normal distribution

The ***Gaussian or normal distribution*** $\mathcal{N}(\mu, \sigma^2)$ is defined as:

$$p(x|\mu,\sigma)\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]\,\mathrm{d}x.$$

The corresponding cdf is

$$P(x|\mu,\sigma) = \frac{1}{2}\left[1+\mathrm{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right],$$

where erf is the *error function*;

$$\mathrm{erf}(z) \equiv \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}\,\mathrm{d}t.$$

The Gaussian or normal distribution

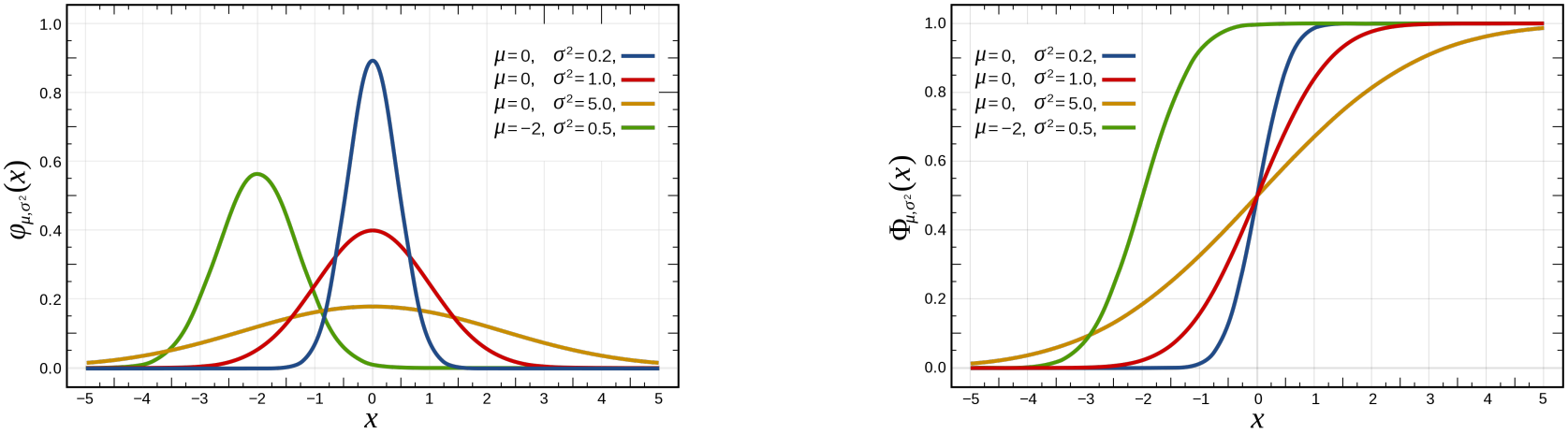


Figure: Pdf and cdf of the Gaussian distribution for different values of μ and σ.
Source: Wikipedia.

Parameters

The distributions we are considering depend on ***parameters*** (such as n and p for the binomial, λ and t for the Poisson), which control the shape of the distribution.

If we know the value of the parameters, we can compute the probability of an observation (as done it the examples above). This is the subject of ***probability theory***, which concerns itself with the theoretical properties of the distributions.

The inverse problem of making inferences about the parameters from the observed samples (i.e., learning about the parameters from the observations made) is the subject of ***statistical inference***, and will be addressed later.

Expectation value

The *expectation value* or *mean* of a discrete RV X is defined as

$$\mathbb{E}[X] \equiv \sum_{i=1}^{n} x_i \, P(x_i)$$

where the sum is taken over the entire sample space.

It quantifies the *location* of the distribution.

This can differ from the *most probable value* or *mode*, which is the value of x_i at which the pmf peaks.

In this course we will often use the notation $\mathbb{E}[X] \equiv \mu_X$.

Variance

The ***variance*** or *dispersion* of a discrete RV X quantifies the *spread* of the distribution. It is defined as

$$\begin{aligned}
\mathrm{Var}\left[X\right] &\equiv \mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^2\right] \\
&= \mathbb{E}\left[X^2 - 2X\left(\mathbb{E}\left[X\right]\right) + \left(\mathbb{E}\left[X\right]\right)^2\right] \\
&= \mathbb{E}\left[X^2\right] - 2\left(\mathbb{E}\left[X\right]\right)\left(\mathbb{E}\left[X\right]\right) + \left(\mathbb{E}\left[X\right]\right)^2, \\
&= \mathbb{E}\left[X^2\right] - \left(\mathbb{E}\left[X\right]\right)^2,
\end{aligned}$$

where in the last step we used the fact that the expectation value is a linear operation.

The square root of the variance is the ***standard deviation*** and is usually denoted by the symbol σ, so that $\mathrm{Var}\left[X\right] = \sigma_X^2$.

Expectation value for the uniform distribution

If $X \sim \mathcal{U}(n)$:

$$\mathbb{E}[X] = \frac{n+1}{2}.$$

Example: What is the expectation value of the outcome for tossing a fair, 6-sided die?

$$\mathbb{E}[X] = \frac{7}{2} = 3.5.$$

The expectation value does not necessarily belong to the sample space.

Expectation value and variance for the binomial distribution

If $X \sim \mathcal{B}(n, p)$:

$$\begin{aligned} \mathbb{E}[X] &= n\,p, \\ \mathrm{Var}[X] &= n\,p\,(1-p). \end{aligned}$$

Example: A fair coin is tossed n times. What is the expectation value of the number of heads, H? What is the variance?

$$\mathbb{E}[H] = 0.5n \text{ and } \mathrm{Var}[H] = 0.25n.$$

Expectation value and variance for the Poisson distribution

If $X \sim \mathcal{P}(\lambda, t)$:

$$\begin{aligned} \mathbb{E}[X] &= \lambda t, \\ \mathrm{Var}[X] &= \lambda t. \end{aligned}$$

Example: An X-ray telescope receives $\lambda = 0.2$ photons per second from a distant source. What are the expectation value and variance for the number of photons detected in a 200 second observation?

$$\mathbb{E}[n] = \mathrm{Var}[n] = 0.2 \times 200 = 40.$$

All other sources of error aside, what is the ***Signal-to-Noise Ratio (SNR)*** for this observation?

$$\mathrm{SNR} = \mathbb{E}[n] / \sqrt{\mathrm{Var}[n]} = \sqrt{\mathbb{E}[n]} = \sqrt{0.2 \times 200} \approx 6.3.$$

Expectation value and variance for continuous distributions

The expectation value of the continuous RV x with pdf $p(x)\,\mathrm{d}x$ is defined as

$$\mathbb{E}\left[x\right] \equiv \int x\, p(x)\,\mathrm{d}x.$$

Its variance or dispersion is defined as

$$\begin{aligned}
\mathrm{Var}\left[x\right] &\equiv \mathbb{E}\left[\left(x - \mathbb{E}\left[x\right]\right)^2\right] \\
&= \mathbb{E}\left[x^2\right] - \left(\mathbb{E}\left[x\right]\right)^2 \\
&= \int x^2\, p(x)\,\mathrm{d}x - \left[\int x\, p(x)\,\mathrm{d}x\right]^2 .
\end{aligned}$$

Expectation value and variance for continuous distributions

Exponential distribution: if $p(t) = \lambda e^{-\lambda t}$:

$$\mathbb{E}[t] = \frac{1}{\lambda} \qquad \text{and} \qquad \mathrm{Var}[t] = \frac{1}{\lambda^2}.$$

Gaussian distribution: if $x \sim \mathcal{N}(\mu, \sigma^2)$:

$$\mathbb{E}[x] = \mu \qquad \text{and} \qquad \mathrm{Var}[x] = \sigma^2.$$

Outline

Lecture 1: Science, Probability, and Statistics

Lecture 2: Some Common Probability Distributions, Expectation and Variance

Lecture 3: Correlation, Central Limit Theorem, Uncertainty propagation, Estimators

Lecture 4: Estimators, Data Modeling, Maximum Likelihood

Lecture 5: Confidence intervals, Least squares and Hypothesis testing

Lecture 6: Hypothesis testing, Confidence intervals, Systematic uncertainties and checks

Lecture 7: Combinations, Bayesian Statistics

Bad stats competition

Send me your examples of real-life ‘bad stats‘

- Bad visualisation - results/inference not wrong but presented to lead a different conclusion. e.g Narrow y-axis range to magnify differences that are not significant.
- Bad inference - incorrect inference of what the data is saying. Monetary prize offered for this category.
- Sources should come from the press/journals/blogs etc with a few lines (or more) on what is wrong. Send to me via email.
- Deadline: Friday wk 3
- Discussion in lecture 8

Expectation value and variance for continuous distributions

The expectation value of the continuous RV x with pdf $p(x)\,\mathrm{d}x$ is defined as

$$\mathbb{E}\left[x\right] \equiv \int x\, p(x)\,\mathrm{d}x.$$

Its variance or dispersion is defined as

$$\begin{aligned}
\mathrm{Var}\left[x\right] &\equiv \mathbb{E}\left[\left(x - \mathbb{E}\left[x\right]\right)^2\right] \\
&= \mathbb{E}\left[x^2\right] - \left(\mathbb{E}\left[x\right]\right)^2 \\
&= \int x^2\, p(x)\,\mathrm{d}x - \left[\int x\, p(x)\,\mathrm{d}x\right]^2 .
\end{aligned}$$

Expectation value and variance for continuous distributions

Exponential distribution: if $p(t) = \lambda e^{-\lambda t}$:

$$\mathbb{E}[t] = \frac{1}{\lambda} \qquad \text{and} \qquad \mathrm{Var}[t] = \frac{1}{\lambda^2}.$$

Gaussian distribution: if $x \sim \mathcal{N}(\mu, \sigma^2)$:

$$\mathbb{E}[x] = \mu \qquad \text{and} \qquad \mathrm{Var}[x] = \sigma^2.$$

The Central Limit Theorem

If you take the sum X of N independent variables, x_i, where $i = 1, 2, 3, \ldots, N$, each taken from a distribution of mean μ_i and variance V_i or σ_i^2, the distribution for X:

1. Has an expectation value

$$\langle X \rangle = \sum_i \mu_i$$

2. Has variance

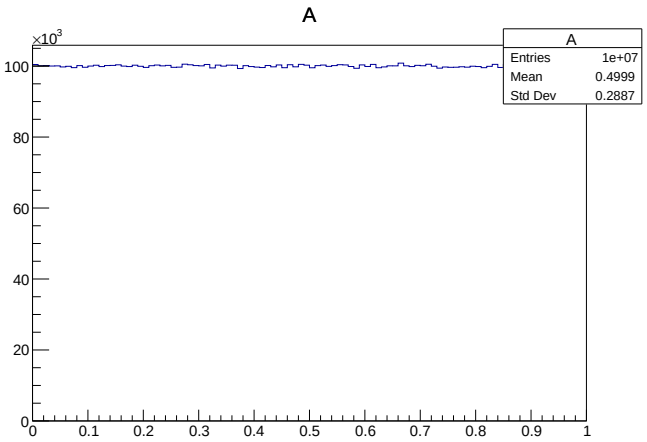
$$V(X) = \sum_i V_i = \sum_i \sigma_i^2$$

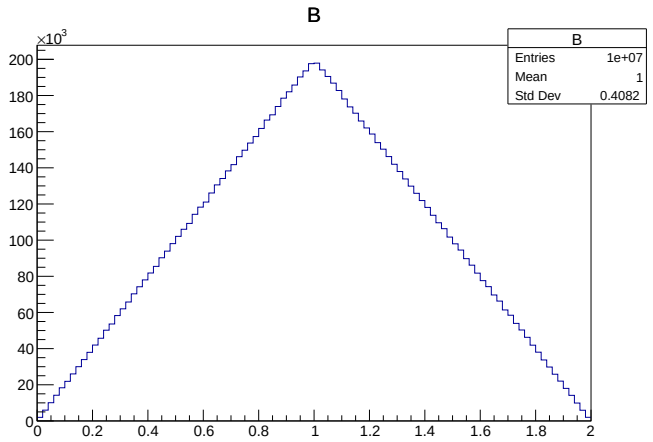
3. Becomes Gaussian as $N \to \infty$

It doesn't matter what distributions the individual variables X_i were drawn from.

CLT: Example

Distribution A is a uniform one [0,1]
Distribution B takes 2 values from A at random and sums them





$$\mu_A = 0.5$$

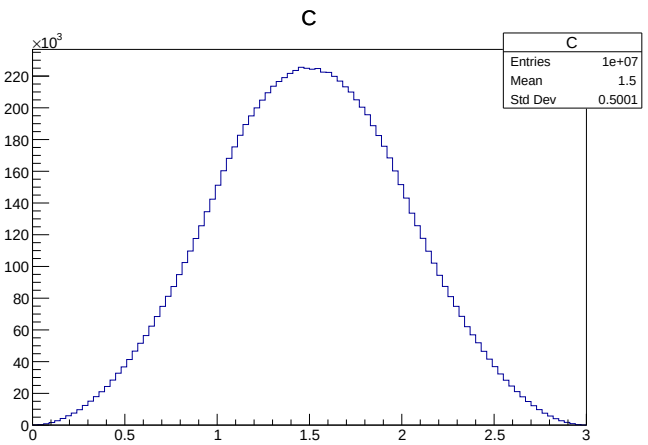
$$V_A = 0.083$$

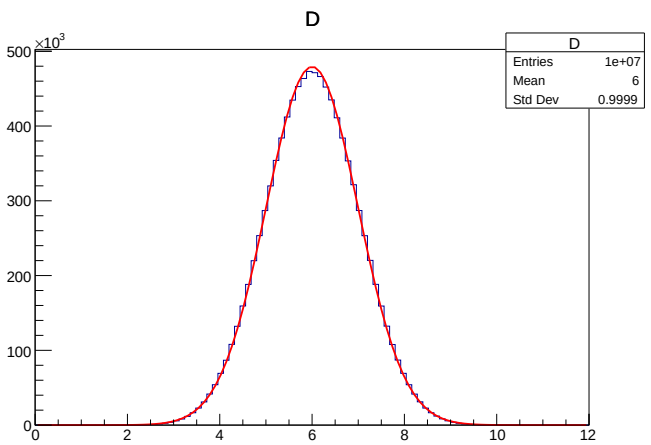
$$\mu_B = 1.0$$

$$V_B = 2 * 0.083 = 0.166$$

CLT: Example

Distribution C takes 3 values from A at random and sums them
Distribution D takes 12 values from A at random and sums them





$$\mu_C = 1.5$$

$$V_C = 3 \times 0.083 = 0.249$$

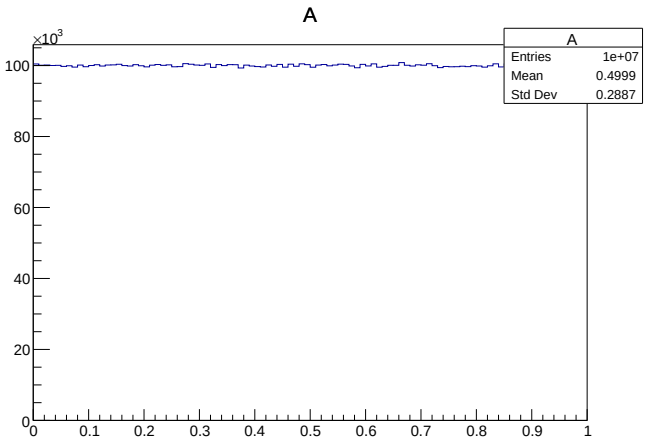
$$\mu_D = 6.0$$

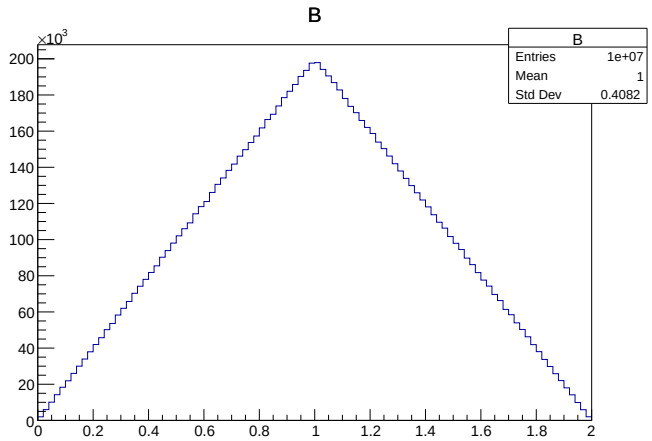
$$V_D = 12 \times 0.083 = 1.00$$

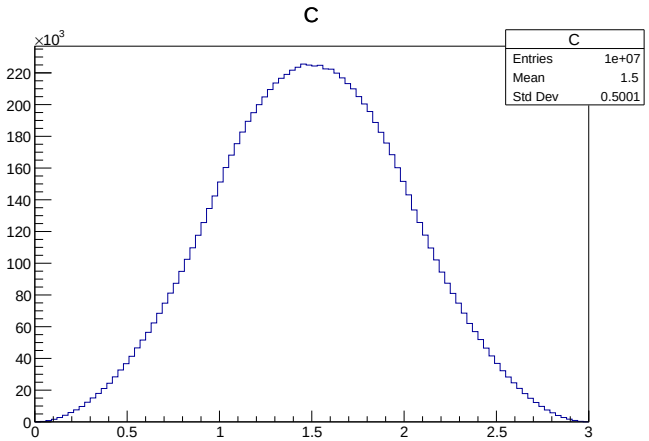
Note

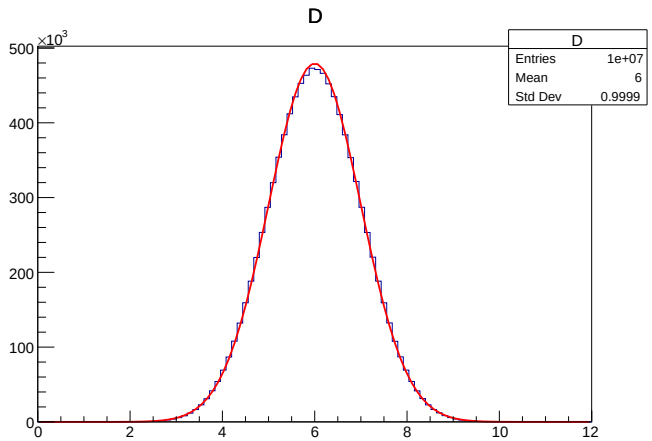
that distribution D is almost Gaussian at $N = 12$

CLT: Example









A: Uniform number distribution, B: Two numbers from the distribution added together, C: Three numbers added together, D: Twelve numbers added together

The Central Limit Theorem

If you take the sum X of N independent variables, x_i, where $i = 1, 2, 3, \ldots, N$, each taken from a distribution of mean μ_i and variance V_i or σ_i^2, the distribution for X:

1. Has an expectation value
$$\langle X \rangle = \sum_i \mu_i$$
2. Has variance
$$V(X) = \sum_i V_i = \sum_i \sigma_i^2$$
3. Distribution of X becomes Gaussian as $N \to \infty$

It doesn't matter what distributions the individual variables X_i were drawn from.

This makes the Gaussian extremely general. In particular, it is generally appropriate to describe ***measurement errors*** on physical quantities (lengths, masses, temperatures, . . .).

Propagation of uncertainty

Say we have a RV x with known expectation value μ_x and variance σ_x^2.

Now consider a function $f(x)$. Like x, it is also a random variable. If x changes, then so does $f(x)$.

What can we say about the expectation value and variance of $f(x)$?

e.g I know the error on x (σ_x), but what is the error on $f(x)$ when $f(x) = ax^2 + bx + c$

Propagation of uncertainty

Taylor expand $f(x)$ about μ_x, keeping only 1^{st} order terms:

$$f(x) \approx f(\mu_x) + (x - \mu_x) \left.\frac{\partial f}{\partial x}\right|_{x=\mu_x}$$

The expectation value of $f(x)$ is:

$$\mu_f \equiv \mathbb{E}\left[f(x)\right] \approx \mathbb{E}\left[f(\mu_x)\right] + \mathbb{E}\left[x - \mu_x\right] \left.\frac{\partial f}{\partial x}\right|_{x=\mu_x} = f(\mu_x)$$

as $\mathbb{E}\left[x - \mu_x\right] = 0$. So far, so trivial.

Propagation of uncertainty

The variance of $f(x)$ is

$$\begin{aligned}
\sigma_f^2 \equiv \mathrm{Var}\left[f(x)\right] &= \mathbb{E}\left[(f(x) - \mathbb{E}\left[f(x)\right])^2\right] \\
&= \mathbb{E}\left[(f(x) - f(\mu_x))^2\right] \\
&= \mathbb{E}\left[\left((x-\mu_x)\ \left.\frac{\partial f}{\partial x}\right|_{x=\mu_x}\right)^2\right] \\
&= \mathbb{E}\left[(x-\mu_x)^2\right]\left(\left.\frac{\partial f}{\partial x}\right|_{x=\mu_x}\right)^2 \\
&= \sigma_x^2\left(\left.\frac{\partial f}{\partial x}\right|_{x=\mu_x}\right)^2 .
\end{aligned}$$

This applies ***irrespective of the probability distribution for*** x, so long as μ_x and σ_x are defined.

Propagation of uncertainty

The derivation can be generalised to functions of multiple variables $f(x, y, \ldots)$, provided that the RVs $x, y, \ldots$ are independent:

$$\sigma_f^2 = \sigma_x^2 \left(\left. \frac{\partial f}{\partial x} \right|_{x=\mu_x} \right)^2 + \sigma_y^2 \left(\left. \frac{\partial f}{\partial y} \right|_{y=\mu_y} \right)^2 + \ldots .$$

For example

$$\text{If } f = ax + b, \quad \sigma_f = |a|\,\sigma_x.$$

$$\text{If } f = x + y, \quad \sigma_f = \sqrt{\sigma_x^2 + \sigma_y^2}$$

$$\text{If } f = x^n, \quad \sigma_f = |n|\,\mu_x^{n-1}\,\sigma_x$$

$$\text{If } f = 1/x, \quad \sigma_f = \sigma_x/\mu_x^2$$

Are two RVs correlated?

We have two RVs x and y, both of which are measured experimentally (and hence have measurement uncertainties), and we want to find out if they are ***independent*** of each other. If not, we say that they are ***correlated***.

The ***covariance*** between two RVs x and y is defined as:

$$\mathrm{Cov}\left[x, y\right] \equiv \mathbb{E}\left[\left(x - \mathbb{E}\left[x\right]\right)\left(y - \mathbb{E}\left[y\right]\right)\right].$$

Note that $\mathrm{Cov}\left[x, x\right] = \mathrm{Var}\left[x\right]$.

Covariance and correlation

The ***Pearson correlation coefficient*** is defined as:

$$\begin{aligned} \rho_{xy} &\equiv \mathrm{Cov}\left[x, y\right] / (\sigma_X \, \sigma_Y) \\ &= \frac{\mathbb{E}[xy] - \mathbb{E}[x]\,\mathbb{E}[y]}{\sqrt{\mathbb{E}[x^2] - (\mathbb{E}[x])^2} \cdot \sqrt{\mathbb{E}[y^2] - (\mathbb{E}[y])^2}} \end{aligned}$$

This separates the co-dependence from the individual variances.

Pearson's sample correlation coefficient

Given observations $\{\hat{x}_i\}$ and $\{\hat{y}_i\}$, the ***sample Pearson correlation coefficient*** is defined as:

$$r_{xy} \equiv \frac{\sum_i \left(\hat{x}_i - \langle\hat{x}\rangle\right)\left(\hat{y}_i - \langle\hat{y}\rangle\right)}{\sqrt{\sum_i \left(\hat{x}_i - \langle\hat{x}\rangle\right)^2}\sqrt{\sum_i \left(\hat{y}_i - \langle\hat{y}\rangle\right)^2}}$$

where $\langle\hat{x}\rangle$ and $\langle\hat{y}\rangle$ are the sample means of $\{\hat{x}\}$ and $\{\hat{y}\}$, respectively.

Pearson's sample correlation coefficient: examples

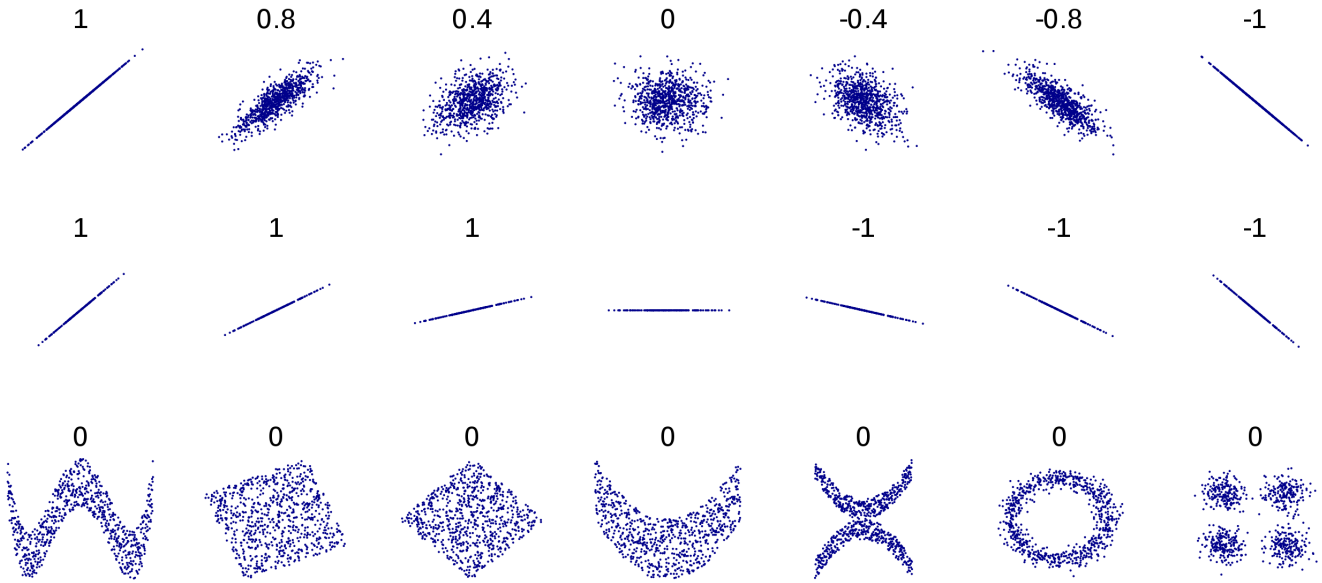


Figure: r_{xy} captures the existence and tightness of a ***linear relationship*** between x and y (but not its slope). Source: Wikipedia..

Computing r_{xy} on the ranked variables $R(x)$ and $R(y)$ gives the ***Spearman's rank-order correlation coefficient***. It is better at capturing non-linear (but still monotonic) dependence.

Correlated Gaussian RVs

If x and y are both normally distributed and have Pearson correlation coefficient ρ, i.e.

$$x \sim \mathcal{N}(\mu_x, \sigma_x^2),\ y \sim \mathcal{N}(\mu_x, \sigma_x^2) \text{ and } \mathrm{Cov}\,[x, y] = \rho\,\sigma_x\,\sigma_y,$$

then the joint probability distribution of x and y is a ***bivariate Gaussian distribution***:

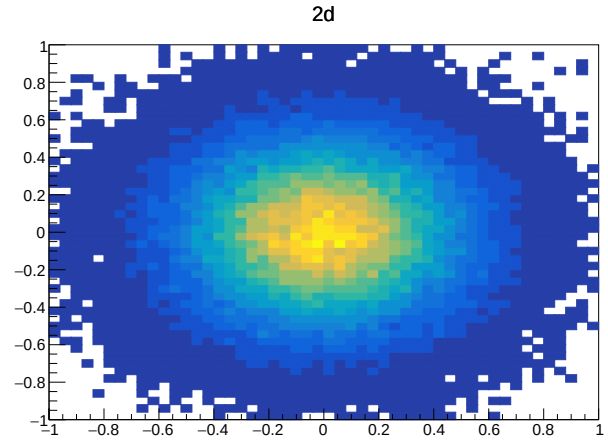
$$p(x, y) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2\pi\sqrt{|\boldsymbol{\Sigma}|}} \exp\left[-\mathbf{r}^{\mathrm{T}}\,\boldsymbol{\Sigma}^{-1}\,\mathbf{r}\right]$$

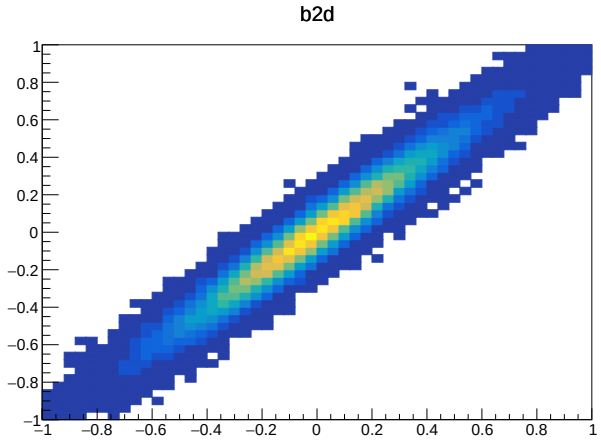
where

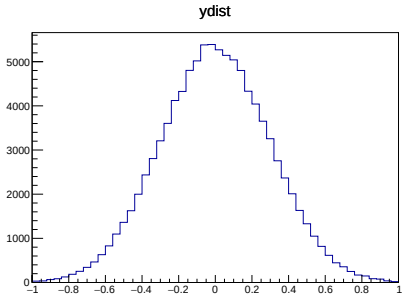
$$\boldsymbol{\mu} \equiv \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \quad \boldsymbol{\Sigma} \equiv \begin{bmatrix} \sigma_x^2 & C \\ C & \sigma_y^2 \end{bmatrix}, \quad \mathbf{r} \equiv \begin{bmatrix} x - \mu_x \\ y - \mu_y \end{bmatrix},$$

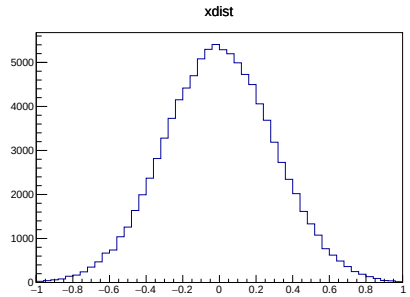
and $|\boldsymbol{\Sigma}|$ is the determinant of the ***covariance matrix*** $\boldsymbol{\Sigma}$.

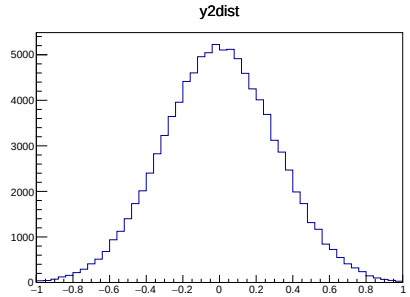
Examples of correlations











Projections:

Left	y distribution of top left distribution
Middle	x distribution of both
Right	y distribution of top right distribution

Error propagation with correlated variables

Assume now that $f = f(x_1, x_2, x_3 ... x_n)$. The elements of $\boldsymbol{x}$ can have correlations. The means of the x_i are known, μ_i, as is the covariance matrix V_{ij}.
What is the variance σ_f^2 ?

$$\sigma_f^2 \approx \sum_{i,j=1}^{n} \left[\frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} \right]_{x=\mu} V_{ij}$$

Sample estimators

There is a random variable x. We don't (to maintain generality) know what distribution it is, but know it has a true mean, μ_x, and variance, σ_x.

We have a sample of measurements, x_i where $i = 1, \ldots N$. These x_i are drawn from the parent distribution x.

How do we define good estimators of the key properties of x using the data x_i.

What is a good estimator?

A good estimator is *consistent*, *unbiased*, *efficient*.

I want to know the true count rate (per min) of a source in the practical lab. The demonstrator tells me that she thinks its about 15. I take a 100 (n) separate measurements, where each interval of counting is 1 minute. To estimate the rate i can do the following

1. Choose the most common count rate
2. Order the count rates in increasing order and take the middle one
3. Add up the count rates and divide by N
4. Add up the first 10 count rates, divide by 10, ignore the other data
5. Add them all up and divide by (N-1) (because stats textbooks seemingly insert factors of $N-1$)
6. Multiply all the count rates and take the Nth root.
7. Ignore the data and just write 15 in my logbook
8. Add every other measurement and divide by $N/2$
9. Add the smallest and largest and divide by 2

All these are an estimator but which are good, and which is the best? [Credit to R. Barlow]

Sample estimators: Consistent

An estimator is consistent if it tends to the true value as the number of data values tends to infinity:

1. ~~Choose the most common count rate~~
2. ~~Order the count rates in increasing order and take the middle one~~
3. Add up the count rates and divide by N
4. ~~Add up the first 10 count rates, divide by 10, ignore the other data~~
5. Add them all up and divide by (N-1) (because stats textbooks seemingly insert factors of $(N-1)$)
6. ~~Multiply all the count rates and take the Nth root.~~
7. ~~Ignore the data and just write 15 in my logbook~~
8. Add every other measurement and divide by $N/2$
9. ~~Add the smallest and largest and divide by 2~~

Sample estimators: Unbiased

An estimator is unbiased, if its expectation is equal to the true value.

$$
\begin{aligned}
\bar{x} &= \frac{x_1+x_2+\ldots+x_N}{N} \\
\langle \bar{x} \rangle &= \left\langle \frac{x_1+x_2+\ldots+x_N}{N} \right\rangle \\
\langle \bar{x} \rangle &= \frac{\langle x_1 \rangle+\langle x_2 \rangle+\ldots+\langle x_N \rangle}{N} \\
\langle \bar{x} \rangle &= \frac{N\mu_x}{N} = \mu_x
\end{aligned}
$$

Sample estimators: Unbiased

1. ~~Choose the most common count rate~~
2. ~~Order the count rates in increasing order and take the middle one~~
3. Add up the count rates and divide by N
4. ~~Add up the first 10 count rates, divide by 10, ignore the other data~~
5. ~~Add them all up and divide by (N-1) (because stats textbooks seemingly insert factors of $(N-1)$)~~
6. ~~Multiply all the count rates and take the Nth root.~~
7. ~~Ignore the data and just write 15 in my logbook~~
8. Add every other measurement and divide by $N/2$
9. ~~Add the smallest and largest and divide by 2~~

Sample estimators: Efficient

An estimator is efficient if its variance is small.

$$\text{Let } X = \sum_{i=1}^{N} x_i; \text{ Then } \bar{x} = X/N$$
$$\text{From C.L.T Var}\left[X\right] = \sum_{i=1}^{N} \sigma_x^2$$
$$\text{Var}\left[N\bar{x}\right] = \text{Var}\left[X\right] = N\sigma_x^2$$
$$\text{Var}\left[\bar{x}\right] = \frac{N\sigma_x^2}{N^2} = \frac{\sigma_x^2}{N}$$

Hence the estimator that is most efficient is the one that uses the most data.

To reduce the uncertainty on your mean by a factor 10, you need 100 times more data

Maths in full

Sample estimators: Efficient

1. ~~Choose the most common count rate~~
2. ~~Order the count rates in increasing order and take the middle one~~
3. Add up the count rates and divide by N
4. ~~Add up the first 10 count rates, divide by 10, ignore the other data~~
5. ~~Add them all up and divide by (N-1) (because stats textbooks seemingly insert factors of $(N-1)$)~~
6. ~~Multiply all the count rates and take the Nth root.~~
7. ~~Ignore the data and just write 15 in my logbook~~
8. ~~Add every other measurement and divide by $N/2$~~
9. ~~Add the smallest and largest and divide by 2~~

And hence we can pick our best estimator for the mean.

Therefore $\bar{x}$ is a good estimator for μ_x.

Sample variance: Want to estimate σ_x^2

We see that the sample mean, $\bar{x}$, is a good estimator for the population mean, μ_x.

Does it follow that the sample variance is a good estimator for the true variance?

We could try computing s^2, the ***sample variance***, and see if is a good estimator where :

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

We'll find out tomorrow.

Outline

Lecture 1: Science, Probability, and Statistics

Lecture 2: Some Common Probability Distributions, Expectation and Variance

Lecture 3: Correlation, Central Limit Theorem, Uncertainty propagation, Estimators

Lecture 4: Estimators, Data Modeling, Maximum Likelihood

Lecture 5: Confidence intervals, Least squares and Hypothesis testing

Lecture 6: Hypothesis testing, Confidence intervals, Systematic uncertainties and checks

Lecture 7: Combinations, Bayesian Statistics

Last lecture

We started looking at estimators - using the data to estimate true population parameters

A good estimator is consistent, unbiased, efficient

- Consistent - estimator tends to the true value as the number of data points tends to infinity.
- Unbiased - the expectation value of the estimator is equal to the true value.
- An estimator is efficient if the variance is small.

We found that the sample mean is a good estimator for the true mean.

Sample variance: Want to estimate σ_x^2

We could try computing s^2, the ***sample variance***, and see if is a good estimator where :

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

Expectation value of the sample variance

$$\begin{aligned} \sigma_x^2 &= \mathbb{E}\left[x_i^2\right] - \left[\mathbb{E}\left[x_i\right]\right]^2 \implies \mathbb{E}\left[x_i^2\right] = \sigma_x^2 + \mu_x^2 \\ \mathrm{Var}\left[\bar{x}\right] &= \mathbb{E}\left[\bar{x}^2\right] - \left[\mathbb{E}\left[\bar{x}\right]\right]^2 \implies \mathbb{E}\left[\bar{x}^2\right] = \frac{\sigma_x^2}{N} + \mu_x^2 \end{aligned}$$

$$\begin{aligned} \mathbb{E}\left[s^2(x)\right] &= \mathbb{E}\left[\tfrac{1}{N}\textstyle\sum_i (x_i - \bar{x})^2\right] = \tfrac{1}{N}\mathbb{E}\left[\textstyle\sum_i (x_i^2 - 2x_i\bar{x} + \bar{x}^2\right] \\ &= \tfrac{1}{N}\left[\left(\textstyle\sum_i \mathbb{E}\left[x_i^2\right]\right) - 2N\mathbb{E}\left[\bar{x}^2\right] + \mathbb{E}\left[\textstyle\sum_i \bar{x}^2\right]\right] \\ &= \tfrac{1}{N}[N(\sigma_x^2 + \mu_x^2) - N(\tfrac{\sigma_x^2}{N} + \mu_x^2)] \\ &= \tfrac{N-1}{N}\,\sigma_x^2. \end{aligned}$$

Sample variance: Want to estimate σ_x^2

We could try computing s^2, the ***sample variance***, and see if is a good estimator where :

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

We determine that the expectation value of the sample variance is

$$\mathbb{E}\left[s^2(x)\right] = \frac{N-1}{N}\,\sigma_x^2 \neq \sigma_x^2$$

This means that the sample variance systematically under-estimates the population variance, particularly for small N.
This estimator is biased.

Unbiased sample variance

We can define an ***unbiased sample variance***:

$$\tilde{s}^2(x) \equiv \frac{N}{N-1}\, s^2(x) = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2.$$

such that $\mathbb{E}\left[\tilde{s}^2(x)\right] = \sigma_x^2$. This is known as the Bessel bias correction.

Plugging this into the expression for the variance of the sample mean:

$$\mathrm{Var}\left[\bar{x}\right] = \frac{\sigma_x^2}{N} \approx \frac{\tilde{s}^2(x)}{N} = \frac{s^2(x)}{N-1}.$$

"Random $N-1$ factors" are not random. They are there to correct the bias of some estimators.

Sample mean and variance: summary

For a sample of N measurements $\{x_i, i = 1, \ldots, N\}$ of an RV x, the sample mean and variance are:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{and} \quad s^2(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2,$$

and the expectation value and variance of the sample mean are

$$\mathbb{E}[\bar{x}] = \mathbb{E}[x] \quad \text{and} \quad \mathrm{Var}[\bar{x}] = \frac{s^2(x)}{N-1}.$$

This result relies on the following assumptions:

- Every observation is independent of every other observation;
- Every observation is drawn from the same probability distribution;
- This probability distribution has a finite variance.

We have ***not*** assumed anything about the shape of the distribution of x.

What if we want to know other things other than the mean or the variance?

Empirical distribution functions

Can we use the data to learn about the shape of the distribution it is drawn from? We can plot an ***empirical distribution function***. This simply shows the frequency of different outcomes in the sample space.

For a discrete RV, this is a simple ***bar chart***, normalised so the height of the bars adds up to unity.

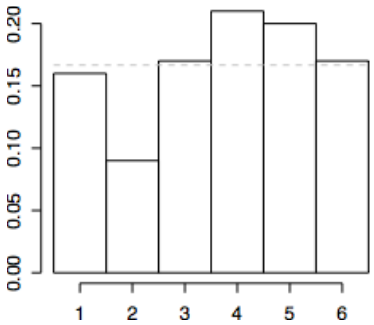
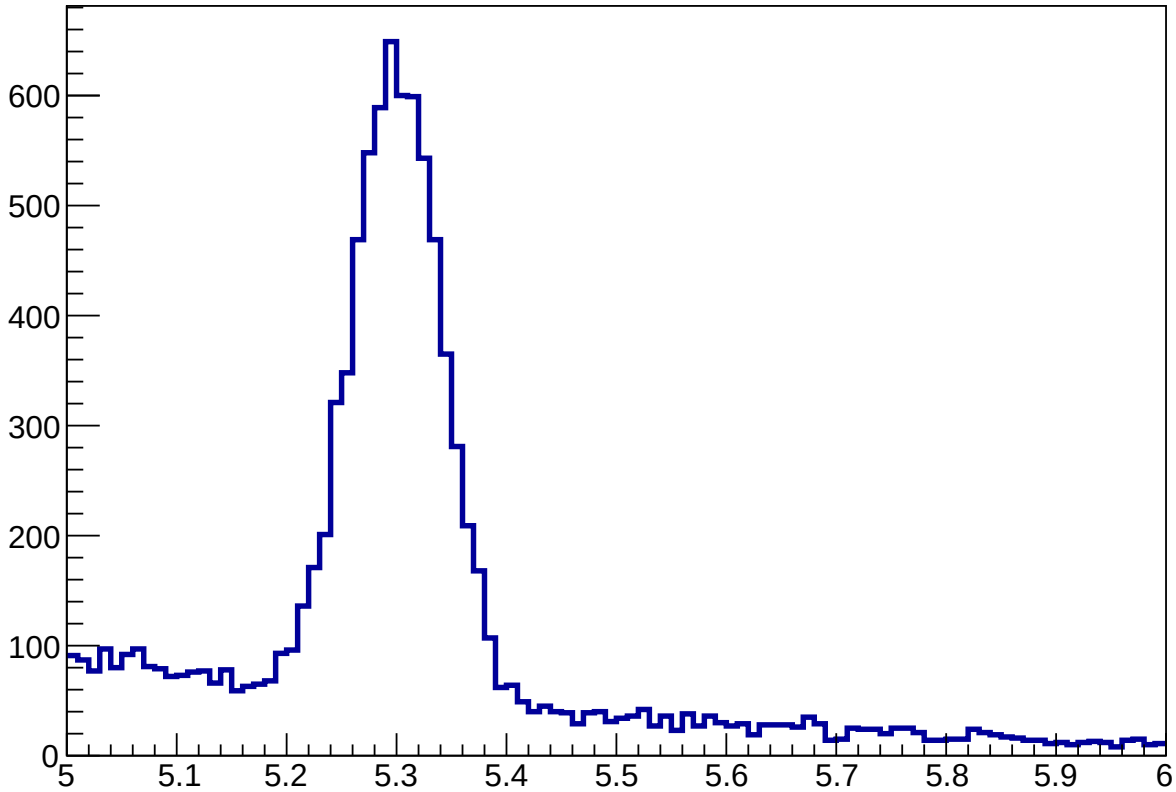


Figure: 100 rolls of a fair, 6-sided dice.
Source: Wikibooks.

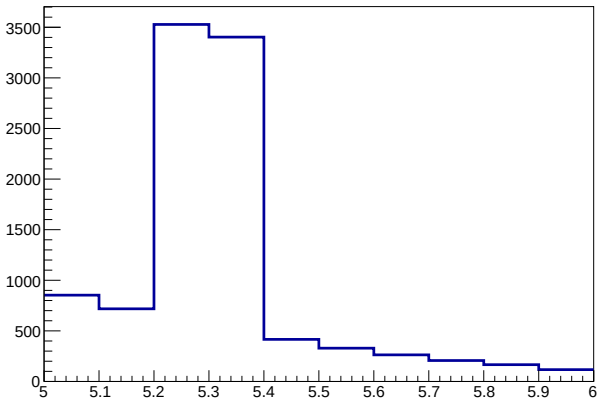
Continuous Random Variables

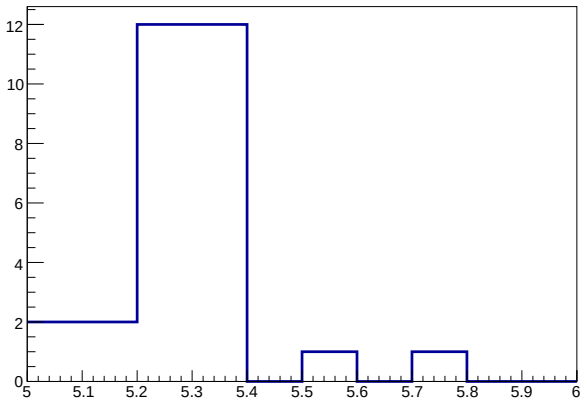
For a continuous Random Variable we can draw a histogram



Continuous Random Variables

For a continuous Random Variable we can draw a histogram - but this can be sensitive to the binning.





Left figure is identical data to previous slide, Right figure is 30 events drawn at random from the previous slide.

Empirical distribution functions

To avoid binning, we can plot the ***empirical cumulative distribution function*** (ecdf). This is a step function which rises by $1/n$ at each of the n data points.

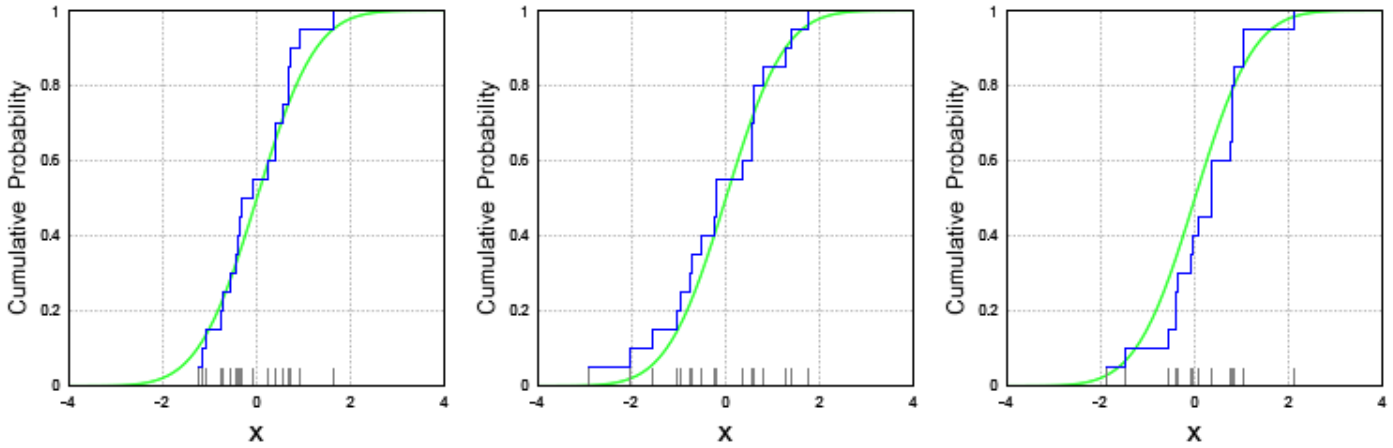
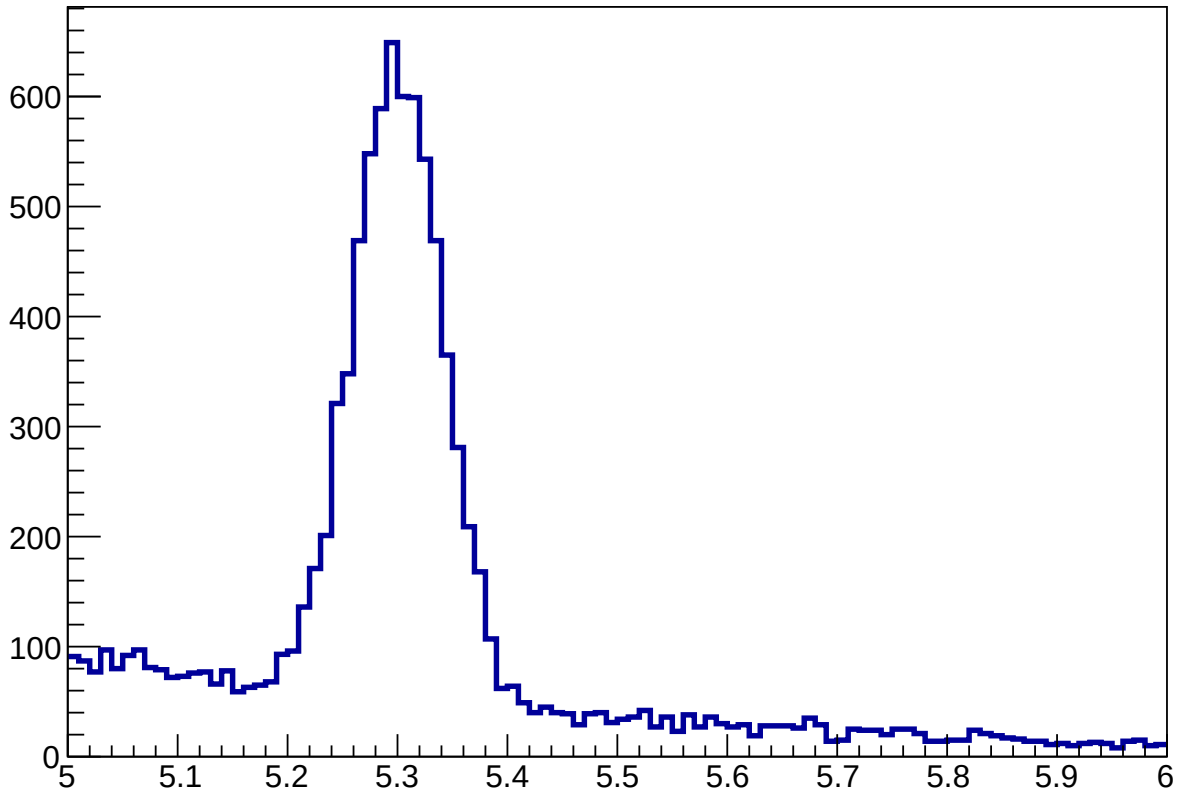


Figure: Three samples from the 'standard' normal distribution ($\mu = \sigma = 1$). In each panel, the green curve is the true cdf, the grey hash marks show the location of 20 observations drawn from that distribution, and the blue step function is the ecdf of that sample. Source: Wikimedia.

Modeling data

In the previous lecture, we saw how, using repeat observations of a quantity, we can estimate the mean and variance of the quantity being observed, in the absence of any other information.



The mean and variance from this data is rather useless.

The experiment

- There is a signal (the peak) and some background.
- I understand my detector so I know that the signal peak can be modeled by a Gaussian. I also know I can assume that the signal events all have the same variance
- Experience tells me the background follows a decaying exponential shape
- Hence the data could be described by 4 parameters, the mean and width of the signal peak, the slope of the bkg, and the fraction of events that are signal.
- $P(m, m+dm_1) = [f \times G(m|\mu, \sigma) + (1-f) \times \mathrm{Exp}(-m|\lambda)]dm$
- I only care about the mean, but i have to determine the other ones along the way.

Maximum Likelihood

- Common method of estimation beyond the mean and variance of the sample
- Consider the probability of observing the first data point in the interval $[m_1, m_1 + dm]$, the second between $[m_2, m_2 + dm]$, the third ....
- The joint probability of observing this data (assuming observations in the data are independent) is then

$$\mathcal{L}(m_1, m_2 ... m_N | a) = \prod_i^N P(m_i | a)$$

Where the m_i are your data points, and the a are the parameters of the PDF. $\mathcal{L}$ is called the likelihood

The likehihood

The data are fixed. However we can vary the parameters a and see how the likelihood changes.
Note we can take the logarithm and work with the sum rather than a product.

$$\ln \mathcal{L}(a) = \sum_{i=1}^{N} \ln P(m_i|a).$$

Maximum Likelihood Estimators

To find the estimate of the parameters a we find the values of a that maximise the probability. For more than one parameter this gives a system of simultaneous equations:

$$\begin{cases} \dfrac{\partial \ln \mathcal{L}(\boldsymbol{a})}{\partial a_1} = 0 \\ \dfrac{\partial \ln \mathcal{L}(\boldsymbol{a})}{\partial a_2} = 0 \\ \vdots \\ \dfrac{\partial \ln \mathcal{L}(\boldsymbol{a})}{\partial a_M} = 0 \end{cases}$$

Typically you would use a computer to do this, do not write your own maximisation routine, there are plenty out there.
In fact they tend to be minimisation routines, so you need to feed the algorithm $-\ln \mathcal{L}$.
However we can learn some properties by thinking of simple cases that are solvable analytically.

Application: repeat observations with Gaussian noise

Say we have measured some physical quantity μ multiple times and want to combine our measurements to estimate μ_{ML}.

Each observation $\hat{x}_i$ is affected by some random measurement error or noise $\hat{\epsilon}_i$:

$$\hat{x}_i = \mu + \hat{\epsilon}_i \qquad \Longleftrightarrow \qquad \hat{\epsilon}_i = \hat{x}_i - \mu.$$

We often assume that the noise is:

- ***Gaussian***: $p(\hat{\epsilon}_i) = \mathcal{N}(0, \sigma_i^2)$;
- ***independent***: $p(\hat{\epsilon}) = \prod_i p(\hat{\epsilon}_i)$;
- ***identically distributed***: $\sigma_i = \sigma \, \forall \, i$;
- Additionally, to start with, we will assume σ is known.

Application: repeat observations with Gaussian noise

Under these assumptions, the likelihood for $\hat{x}_i$ is

$$\mathcal{L}_i(\mu) = p(\hat{x}_i|\mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(\hat{x}_i - \mu)^2}{2\sigma^2}\right].$$

The likelihood for the dataset as a whole is

$$\mathcal{L}(\mu) = \prod_{i=1}^{N} \mathcal{L}_i(\mu) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left[-\sum_{i=1}^{N} \frac{(\hat{x}_i - \mu)^2}{2\sigma^2}\right].$$

Taking the logarithm:

$$\ln \mathcal{L}(\mu) = -\frac{N}{2} \ln 2\pi - N \ln \sigma - \frac{1}{2} \sum_{i=1}^{N} \left(\frac{\hat{x}_i - \mu}{\sigma}\right)^2.$$

MLE for the mean of a Gaussian

To find the MLE, maximise $\ln \mathcal{L}$ with respect to μ:

$$\frac{\partial \ln \mathcal{L}}{\partial \mu} = \frac{1}{\sigma^2} \left(N\mu - \sum_{i=1}^{N} \hat{x}_i \right) = 0$$

As $\sigma^2 > 0$, this is satisfied if:

$$\mu_{\mathrm{ML}} = \frac{1}{N} \sum_{i=1}^{N} \hat{x}_i = \langle \hat{x} \rangle$$

The MLE for the mean of a Gaussian is the sample mean (and hence is equal to its expectation value).

Now assume we didn't know σ^2

To find the MLE, simultaneously maximize:

$$\frac{\partial \ln \mathcal{L}}{\partial \mu} = 0 \implies \sum_{i=1}^{N} (\hat{x}_i - \mu) = 0$$

$$\frac{\partial \ln \mathcal{L}}{\partial \sigma} = 0 \implies \frac{-N}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{N} (\hat{x}_i - \mu)^2 = 0$$

$$\mu_{ML} = \bar{x}$$

$$\sigma^2_{ML} = \frac{1}{N} \sum_{i=1}^{N} (\hat{x}_i - \bar{x})^2$$

But we know this to be biased! (more on this in a few minutes)

What if the uncertainties vary?

We now consider the case where the measurement uncertainties are known, but they change from one observation to the next.

Replace σ by σ_i in the likelihood:

$$\ln \mathcal{L}(\mu) = -\frac{N}{2} \ln 2\pi - \sum_i \ln \sigma_i - \frac{1}{2} \sum_{i=1}^{N} \left(\frac{\hat{x}_i - \mu}{\sigma_i} \right)^2 .$$

Maximising this with respect to μ gives

$$\mu_{\mathrm{ML}} = \frac{\sum_i w_i \hat{x}_i}{\sum_i w_i}, \text{ where } w_i = \frac{1}{\sigma_i^2}.$$

This is the ***weighted sample mean***, with ***inverse variance*** weights.

For $\sigma_i = \sigma$, this reduces to the unweighted sample mean, as expected.

What if the uncertainties vary?

One can show that the likelihood is still a Gaussian function of μ, with mean μ_{ML} and variance

$$\sigma^2(\mu) = \left[\sum_{i=1}^{N} \frac{1}{\sigma_i^2}\right]^{-1}.$$

For $\sigma_i = \sigma$, this reduces to $\sigma^2(\mu) = \sigma^2/N$, as expected.

One can also define a weighted (unbiased) sample variance

$$\tilde{\sigma}^2(\hat{x}) = \frac{\sum_i w_i(\hat{x}_i - \langle\hat{x}\rangle)^2}{\sum_i w_i - \frac{\sum_i w_i^2}{\sum_i w_i}}$$

where the second term in the denominator is the equivalent of the Bessel bias correction.

Uncertainty of the estimator

We use the data to determine μ_{ML} and σ_{ML} by maximising the likelihood.

What is the uncertainty on those estimates?

More formal proofs exist - here we take a more conceptual approach. The range of likely values for the parameter a is determine by the width of $\mathcal{L}$ near the maximum.

Uncertainty of the estimator

$$l = \ln \mathcal{L}$$

Take a taylor expansion around the maximum. a_0 is the ML estimator

$$l = l_{max} + l'(a - a_0) + \frac{1}{2} l''(a - a_0)^2$$

By recasting l'' and noting that at maximum l'=0 :

$$l = l_{max} - \frac{1}{2} \left(\frac{a - a_0}{c} \right)^2$$

$$\mathcal{L} = A \exp \left[-\frac{1}{2} \left(\frac{a - a_0}{c} \right)^2 \right]$$

Uncertainty of the estimator

With this we can set two equivalent definitions of the uncertainty of the parameter

- $[-\partial^2 l/\partial a^2]^{-\frac{1}{2}}$
- The change in a required to reduce l from its maximum value by 0.5. Note that this is 0.5 absolute, not a relative factor

Pros of MLE

MLE are consistent but as we saw, they are in general biased. This bias reduces with N. So why do we accept this? It is the trade-off for other good qualities:

- Each event is used, no issues with bins and histograms
- Invariance of ML estimators with parameter transformation: We look for a peak in $\mathcal{L}$, given by a this is also the peak in a^2 etc If we determine decay rate λ_0 or the decay time τ_0 then

$$\lambda_0 = \frac{1}{\tau_0}$$

Cons of MLE

- Small biases, particularly at low N
- Parameters at the physical limit don't behave well (the derivative is not defined)
- Multi-parameter likelihood can have local maxima which have to be thoughtfully removed.
- There is no "goodness-of-fit" metric

Outline

Lecture 1: Science, Probability, and Statistics

Lecture 2: Some Common Probability Distributions, Expectation and Variance

Lecture 3: Correlation, Central Limit Theorem, Uncertainty propagation, Estimators

Lecture 4: Estimators, Data Modeling, Maximum Likelihood

Lecture 5: Confidence intervals, Least squares and Hypothesis testing

Lecture 6: Hypothesis testing, Confidence intervals, Systematic uncertainties and checks

Lecture 7: Combinations, Bayesian Statistics

Last Lecture

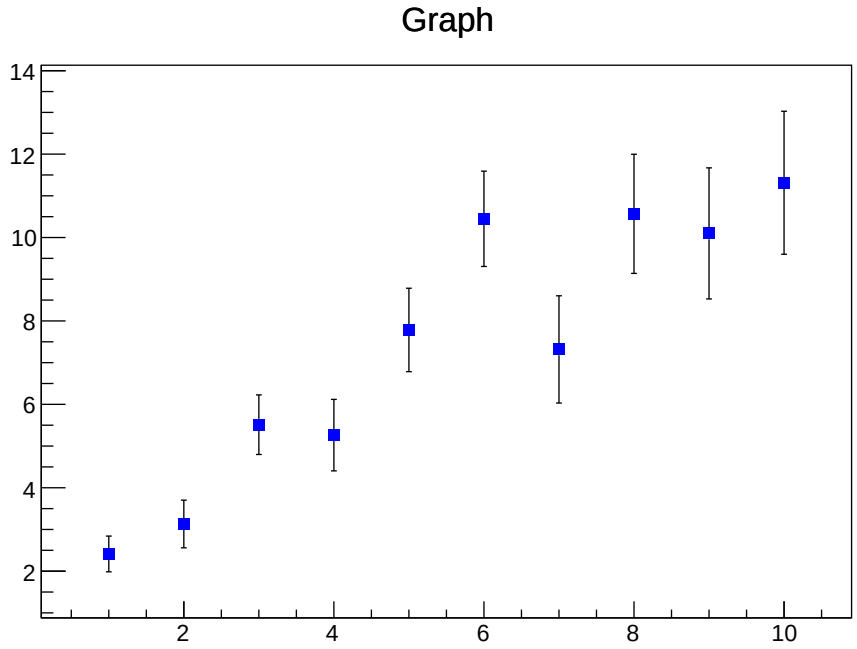
Looked at the maximum likelihood method for extracting a value for a parameter from the data.

We saw that the probability distribution for μ is thus a Gaussian with mean $\bar{x}$ and variance σ^2/N.

Another option for extracting parameters is the least-squares technique

Maximum likelihood and least squares

Data : (x_i, y_i); x_i known exactly, y_i have some uncertainty σ_i.
Theory: $y = f(x|a)$, where a are some parameters we want to estimate
y is the true quantity, the measurements y_i have been smeared from this by σ_i.
Invoking the CLT, the distribution of y_i around their true value will be Gaussian.



Maximum likelihood and least squares

Consider the probability for observing a particular y_i

$$P(y_i|a) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(\frac{-(y_i - f(x_i|a)^2}{2\sigma_i^2}\right)$$

The log-likelihood is then:

$$\ln \mathcal{L} = -\frac{1}{2}\sum_i \left(\frac{y_i - f(x_i|a)}{\sigma_i}\right)^2 - \sum_i \sqrt{2\pi}\ln\sigma_i$$

To maximise, you need to minimise the least squares, Z

$$Z = \sum_i \left(\frac{(y_i - f(x_i|a)}{\sigma_i}\right)^2$$

If y_i are not independent

Don't read too much into the last slide. The connection shows the relation, but you always knew that minimising the least squares was a sensible thing to do.

If y_i are not fully independent - imagine that part of the uncertainty is the same for all (in the same direction) and some other component is randomly distributed then:

$$Z = \sum_{i,j=1}^{N} (y_i - f(x_i|a))(V^{-1})_{ij}(y_j - f(x_j|a))$$

where V_{ij} are elements of the covariance matrix

Without proof (available in text books) the error on the parameter is where the value of Z changes by 1.
Recall that for the ML, the error on the parameter is where the value of $\ln\mathcal{L}$ changes by 0.5

Confidence intervals

From last lecture: The probability distribution for μ is thus a Gaussian with mean $\bar{x}$ and variance σ^2/N

Integrating the pdf for RV X over a given interval $[a; b]$ gives the probability Y$\%$ for that confidence interval.

In the frequentist world, this means that if you ran the experiment lots of times and got lots of different intervals then Y$\%$ of those intervals would contain the true value X.

That's rather convoluted, in the Bayesian interpretation it means the probability of X lying in that interval is Y%.

Inverting that calculation allows us to compute intervals associated to a specific value Y.

Gaussian probability intervals

Recall the Gaussian cdf

$$P(t;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{t} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx,$$

where $P(t;\mu,\sigma)$ is the probability that a measurement x is less than t.

$$\mathrm{erf}(x) \equiv \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\,\mathrm{d}t.$$

Is a useful substitution to help with calculations. It gives the probability that normally distributed variable $\mathcal{N}(0, \frac{1}{2})$ lies in the interval $[-x, x]$.

Gaussian probability intervals

If $x \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\begin{aligned} p(\mu - \kappa\sigma < x < \mu + \kappa\sigma) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mu-\kappa\sigma}^{\mu+\kappa\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \mathrm{d}y \\ &= \frac{2}{\sqrt{\pi}} \int_0^{\kappa/\sqrt{2}} \exp(-z^2)\, \mathrm{d}z \\ &= \mathrm{erf}(\kappa/\sqrt{2}). \end{aligned}$$

Because the error function cannot be evaluated analytically, it is usually evaluated numerically and the results saved in a ***look-up table*** for reference.

Gaussian probability intervals

κ "number of sigma"	$P(\lvert x-\mu\rvert < \kappa\,\sigma)$ probability content	 usually called
1	0.683	1σ
2	0.954	2σ
3	0.997	3σ
4	0.99993	4σ
5	0.99999943	5σ
1.64	0.90	90% confidence
1.96	0.95	95% confidence
2.54	0.99	99% confidence
3.29	0.999	99.9% confidence

Gaussian probability intervals

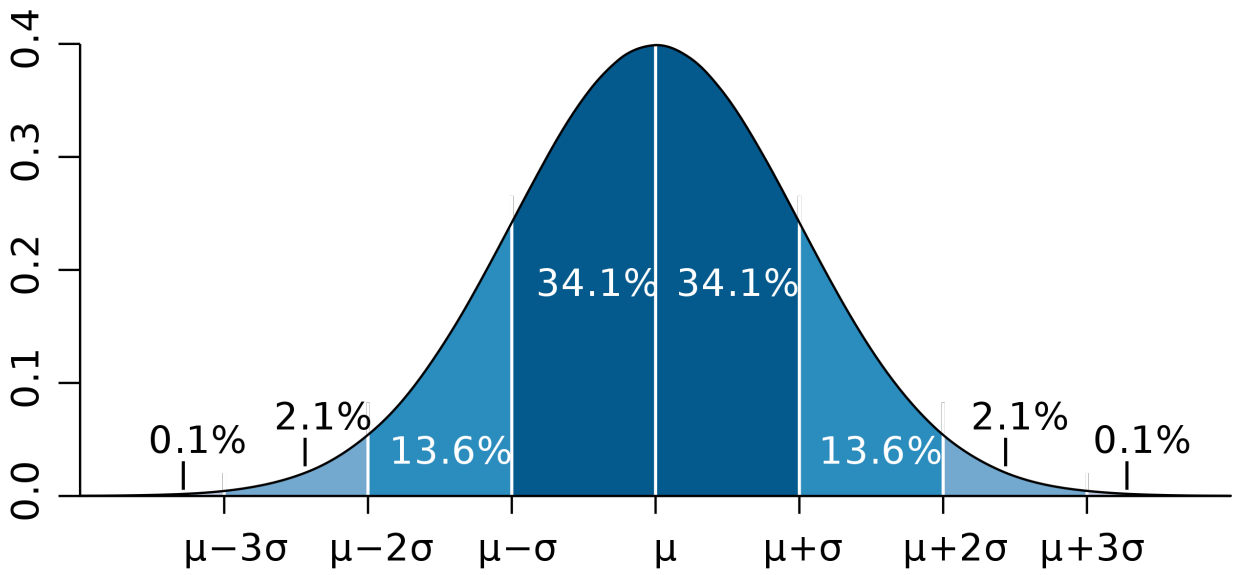


Figure: Credit: Ainali – Own work, CC BY-SA 3.0.

A common mistake is to state that the full width at half maximum of the Gaussian is 2σ. Actually it is $2\sqrt{2\ln 2}\sigma \approx 2.35\sigma$.

What do we mean by result = value ± uncertainty?

Most of the time, when we report a value for a physical quantity as $x = \mu \pm \sigma$, we implicitly mean not only that x has expectation value μ and variance σ^2, but also that its probability distribution is Gaussian.

We are therefore implying that

- the probability that x lies in the interval $[\mu - \sigma; \mu + \sigma]$ is 68.3%.
- the probability that x lies in the interval $[\mu - 2\sigma; \mu + 2\sigma]$ is 95.4%.
- the probability that x lies in the interval $[\mu - 3\sigma; \mu + 3\sigma]$ is 99.7%.
- ...

Well this is a common interpretation, but it is Bayesian, keep the frequentist one in mind if you are going down that route.

Probability intervals

There are many ways of defining probability intervals:

- symmetric intervals about the expectation value (or the most probable value, or any other value)
- starting from both ends of the range and moving 'up' the pdf to exclude 'tails'
- starting from one end only to get upper or lower limit.

Probability intervals for general pdfs

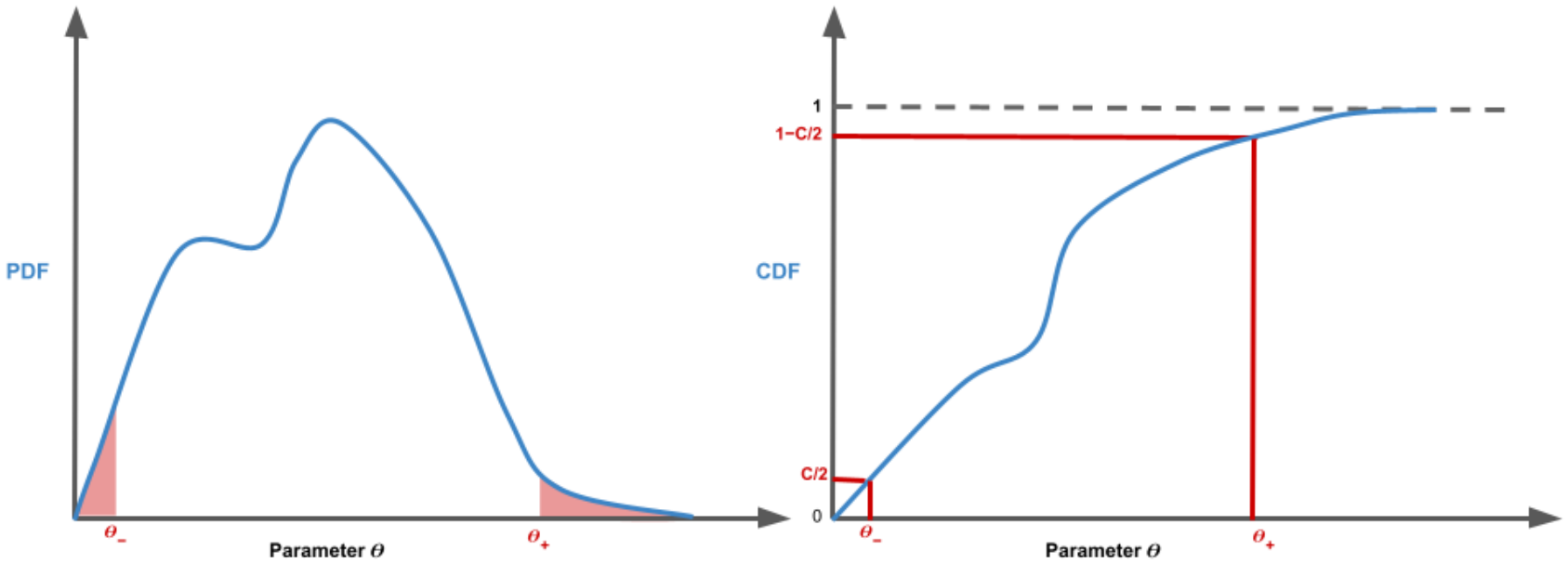


Figure: If the pdf is asymmetric, we can still define a confidence interval. Here, we define a two-sided interval by excluding the 'tails' of the distribution until $1 - C$ of the total area under the pdf is left. We could also have defined a symmetric interval about the median, expectation value (mean) or most probable value. For a Gaussian these would all be the same, but in non-Gaussian cases one must specify what kind of interval one is reporting.

One-sided confidence interval

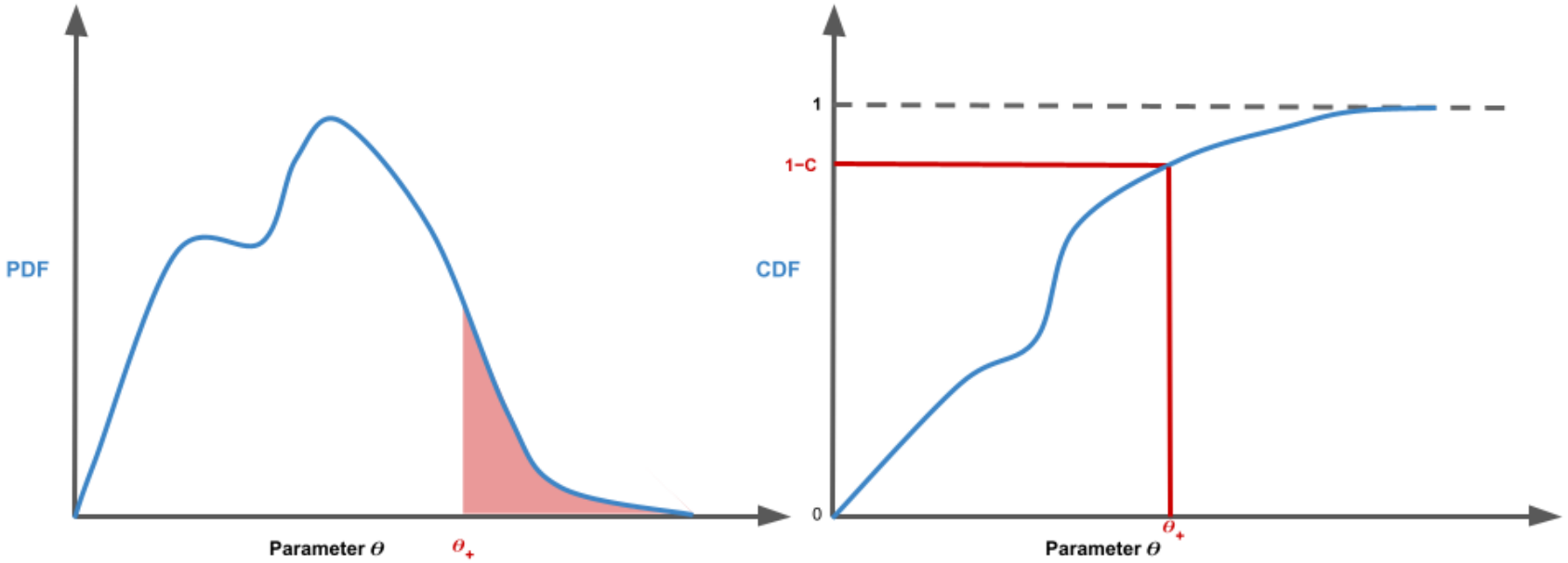


Figure: If we are interested in a lower or upper limit, we can evaluate a one-sided confidence interval (this applies to Gaussians too).

- We can determine parameters from data using either ML or χ^2
- From each we can determine an uncertainty on the parameter and relate that to confidence intervals.
- Being able to determine a parameter doesn't tell you if that determination is correct or not.
- Let's now turn to hypothesis testing

Linear or exponential ?

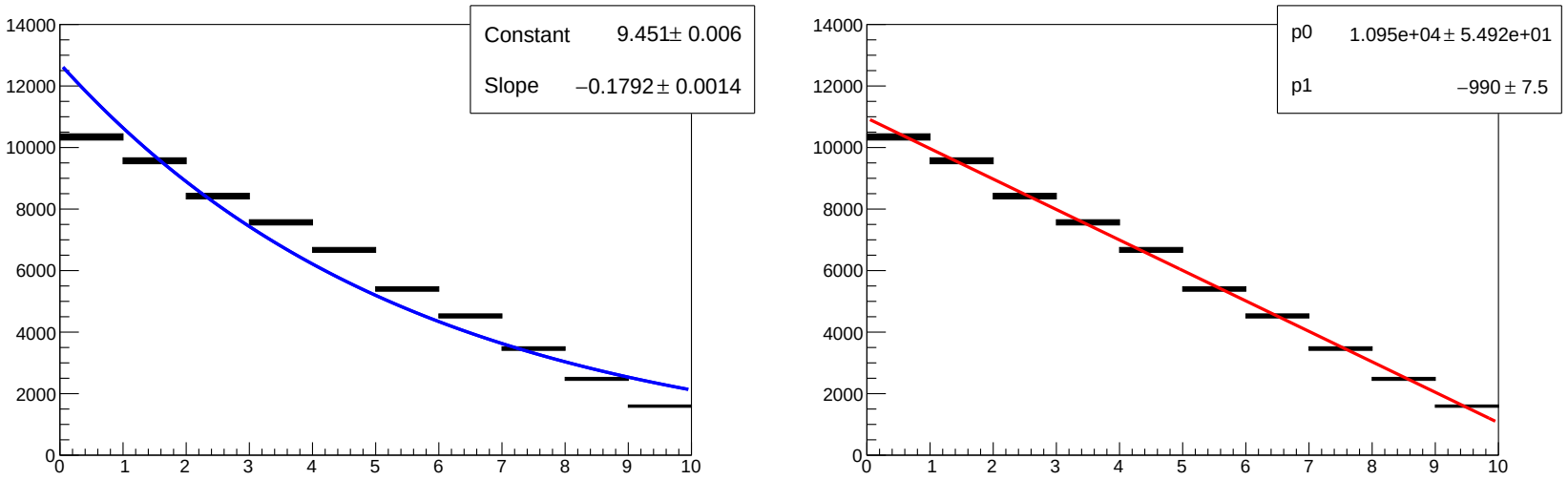


Figure: The same data fit by an (left) exponential (right) linear pdf

The same data can be fit by either function. The comparison of the fitted function and the data tell me clearly that the linear function is giving a better fit.

Linear or exponential ?

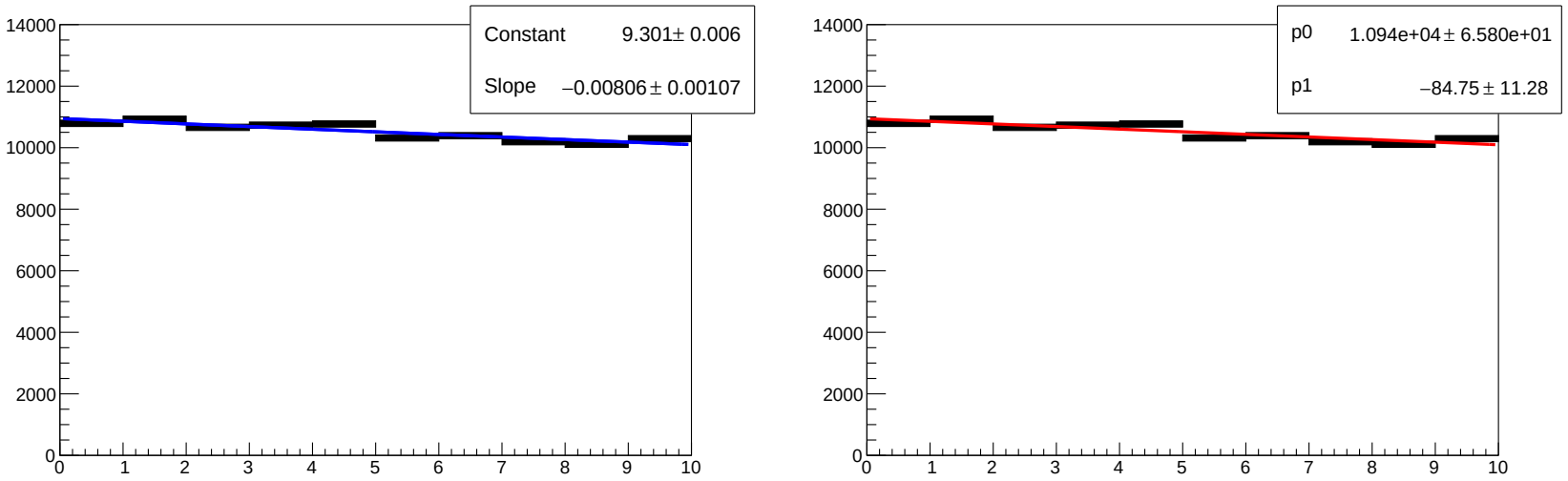


Figure: The same data fit by an (left) exponential (right) linear pdf

This time can't tell the difference. Both the functions are equally good (how do i know this)?

Linear or exponential ?

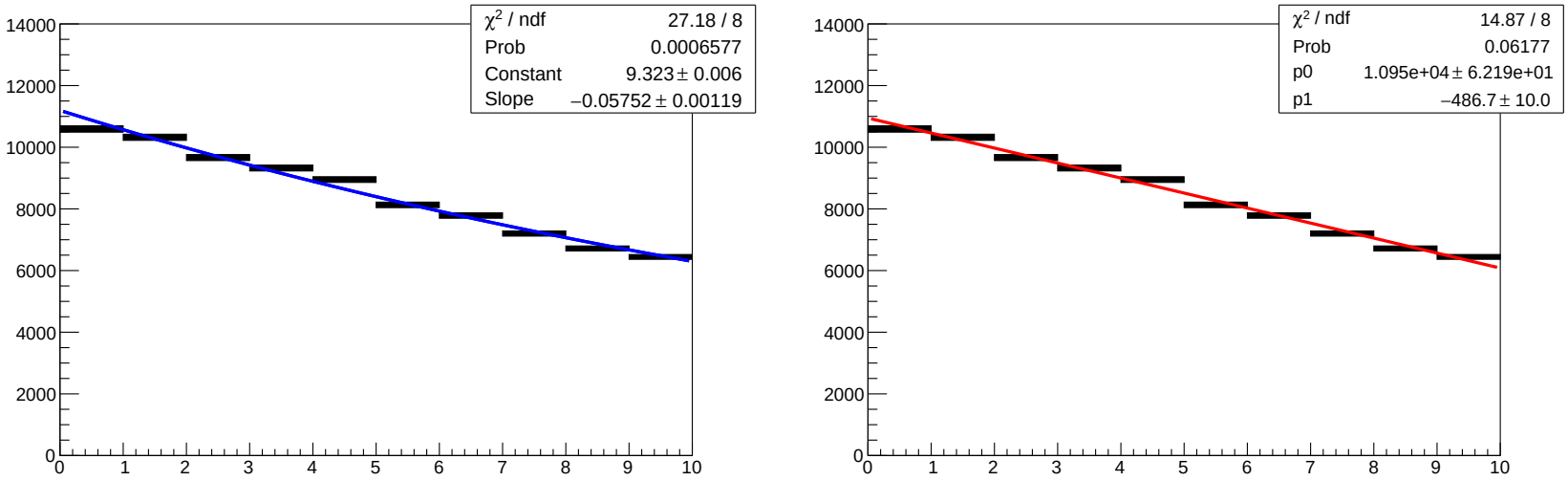


Figure: The same data fit by an (left) exponential (right) linear pdf

This time I can tell a difference but it seems small? Can quantify by the value of the χ^2.

Chi-squared distribution

If X_1, X_2, ..., X_k are k independent RVs with a standard normal distribution:

$$X_i \sim \mathcal{N}(0, 1),$$

then the sum of their squares

$$Z \equiv \sum_{i=1}^{k} X_i^2$$

follows a chi-squared distribution with k ***degrees of freedom*** (d.o.f.'s):

$$Z \sim \chi_k^2.$$

Example: $X_i = \frac{\mathrm{Data_i - Fit_i}}{\sigma_i}$

Chi-squared distribution

$$p(x,k) = \frac{x^{\frac{k}{2}-1}\, e^{-\frac{x}{2}}}{2^{\frac{k}{2}}\, \Gamma\left(\frac{k}{2}\right)} \text{ if } x > 0, \text{ and } 0 \text{ otherwise.}$$

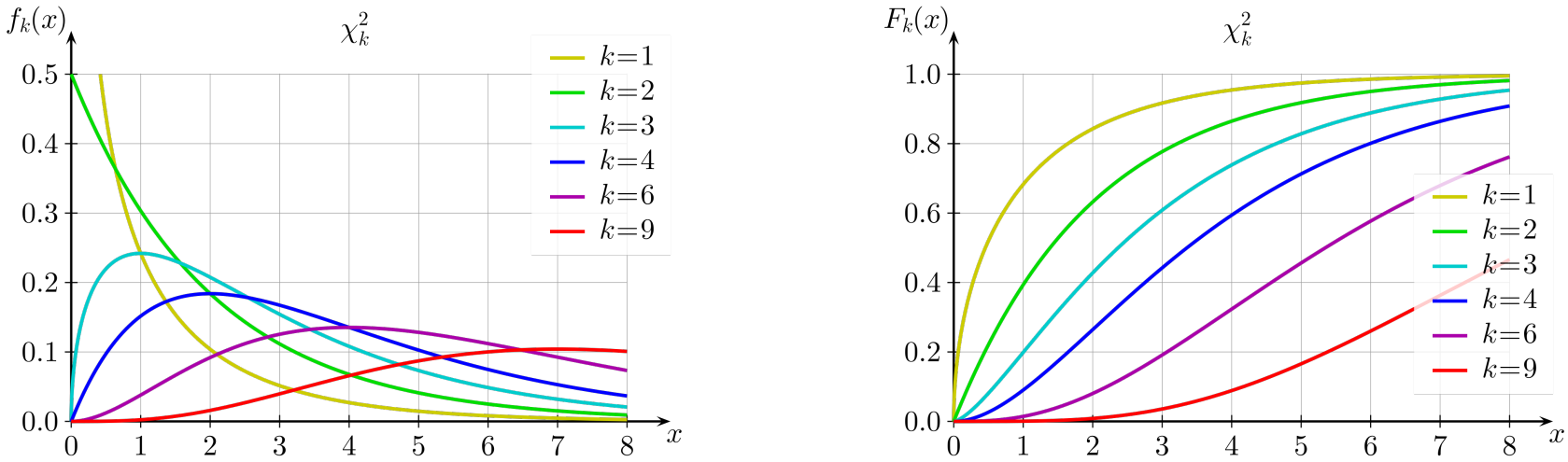


Figure: Pdf and cdf of the chi-squared distribution for different values of k. For large k, it approaches a normal distribution. Source: Wikipedia.

$$\mathbb{E}\left[p(x,k)\right] = k \text{ and } \mathrm{Var}\left[p(x,k)\right] = 2k$$

χ^2 fitting further details

If the following three conditions are met

- The y_i are distributed with known variance (or known covariance if not independent)
- The hypothesis $f(x_i|a)$ is linear in the parameters a
- The functional form of the hypothesis is correct

then the minimum value of Z is distributed according to a χ^2 distribution with k degrees of freedom

where [k = N - b] where N is the number of data points and b the number of parameters that vary while minimising Z

Hypothesis testing

- Hence we can use the χ^2 distribution to test our "goodness of fit" or the hypothesis that the data are described by a specific function.
- Since $\mathbb{E}\left[\chi^2(x,k)\right] = k$ then is it usual to consider $Z_{\min}/k$, commonly written as χ^2/ndof. Expect the observed value to be near 1 if all the conditions are met.
- Large values of $Z_{\min}$ are unlikely and so our hypothesis is more likely to be wrong.
- Use the χ^2 distribution to determine the probability that you would have got a larger value of Z by integrating it from Z to infinity
- Look up tables exist.

Linear or exponential ?

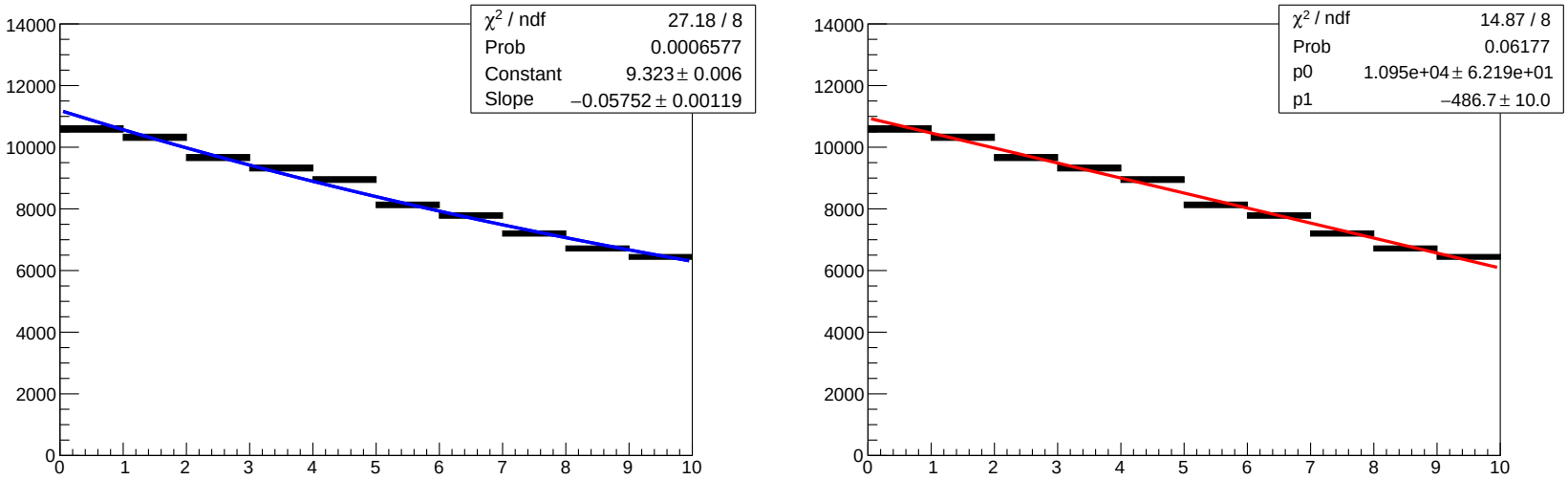


Figure: The same data fit by an (left) exponential (right) linear pdf

	χ^2/ndof	Probability (%)
Expo	3.40	0.07
Linear	1.86	6.18

Linear or Cubic ?

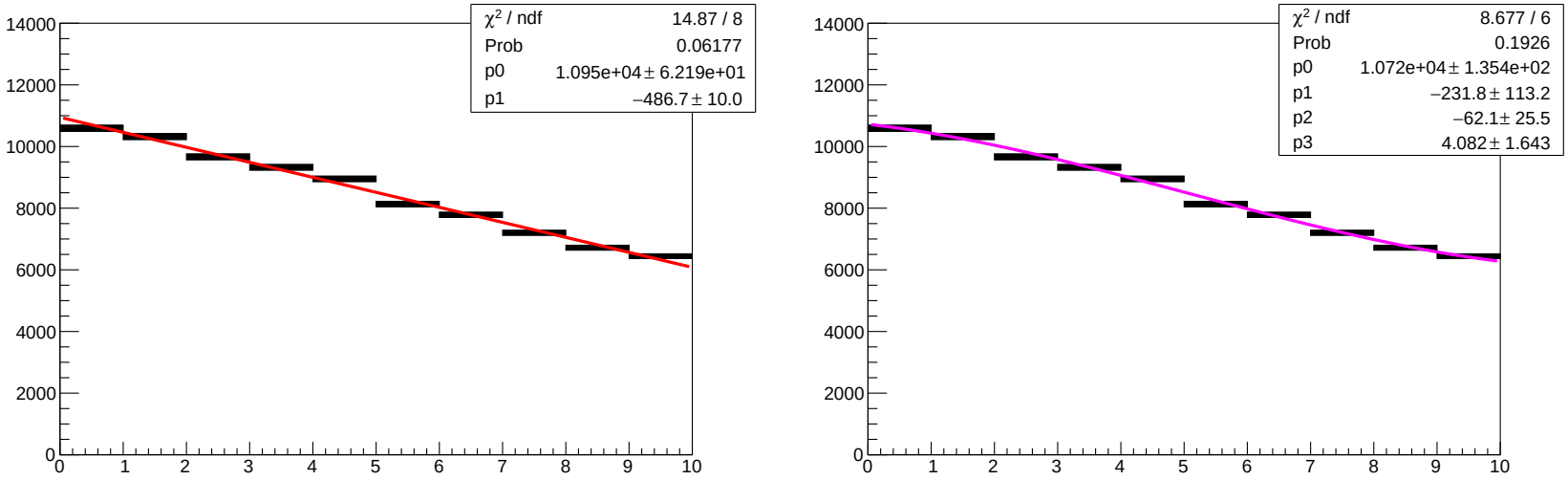


Figure: The same data fit by an (left) linear (right) cubic pdf

	χ^2/ndof	Probability (%)
Linear	1.86	6.18
Cubic	1.45	19.26

Cubic fit looks better visually, and by numerical metrics

Hypothesis testing

- $\mathcal{H}_0 =$ The null hypothesis
- Set a probability limit p_0
- If $\mathrm{P}(\mathcal{H}_0) \leq p_0$ we reject the null hypothesis

Which is it?

	χ^2/ndof	Probability (%)
Expo	3.40	0.07
Linear	1.86	6.18
Cubic	1.45	19.26

- If we set our probability cutoff at 1% then we reject the exponential
- If we set our probability cutoff at 10% then we also reject the linear
- Infact the cubic is a case of overfitting,
- The data were generated with a linear function
- The lower fit probability is merely chance

Where to set the probability limit

- It is our choice on where to set the limit of rejecting the hypothesis.
- You should set that before you compute whether or not the condition is met.
- Say we choose to reject the hypothesis is $(p < 0.05)$ (i.e the chance of observing $Z_{\min}$ larger was less than 5%).
- Then in 5% of cases, we will reject $\mathcal{H}_0$ when it was correct.
- We know this, we have decided to live with that.
- But if our errors are wrong then this can happen more (or less) than 5% of the time. We are less ok with being wrong more than 5% of the time so this is now a problem.
- Could reduce the p value of acceptance - make it 1%.... but

Where to set the probability limit

- What if we fail to reject H when it was in fact wrong?
- We happen to get a value of $Z_{\min}$ smaller than expected
- If we set p smaller then the chance of this increases.
- The probability is not easy to estimate. It depends how wrong the hypothesis is (how wrong the functional form is)

What we choose is a trade off between the two. The importance of knowing your variances is clear....

Hypothesis testing

So far have considered a χ^2 distribution, but we can be a lot more general about hypothesis testing. Lets look at a few examples.
Say we have a standard theory that predicts a value μ for a quantity x.

We call the proposition $x = \mu$ the ***null hypothesis***.

We then take a series of measurements of that quantity. Are our results consistent (or not) with the hypothesis?

To answer this we must:

- define a summary statistic that quantifies the departure from the null hypothesis in our sample, given the precision of the experiment and the number of observations
- work out the pdf of this statistic
- integrate it to obtain a confidence interval

Hypothesis testing: single-point deviation

An X-ray telescope monitors a particular star for a whole year, during which it receives a total of 36000 X-ray photons from the star.

On one particular day, $\hat{x}_i = 130$ photons were detected. Has the star undergone an outburst?

$$\langle \hat{x} \rangle = \frac{36000}{365} = 98.6, \qquad \lambda \approx \langle \hat{x} \rangle, \qquad \sigma_\lambda = \sqrt{\lambda} \approx \sqrt{\langle \hat{x} \rangle}$$

as photon counts obey a Poisson distribution.
Mean is large could approximate to $\mathcal{N}(98.6, 98.6)$.

Try both ways

Hypothesis testing: single-point deviation

Therefore, the anomalously high rate observed on that particular day corresponds to a

$$\frac{\hat{x}_i - \lambda}{\sigma_\lambda} \approx \frac{130 - 98.6}{\sqrt{98.6}} \approx 3.16\,\sigma$$

departure from the mean.
Poisson : $P(X \geq 129) = 0.00199 \approx 0.2\%$
Normal : $P(X \geq 129) = 0.00078 \approx 0.1\%$

Is it a fluctuation

The probability of getting a $\geq 3.16\,\sigma$ departure from the mean $\lambda \approx 98.6$ from random statistical fluctuations alone is $p = 8 \times 10^{-4}$.
Full Poisson calculation (more compute power) gives slightly higher at 0.2%

We can set our $p-$value at 1%. Is the correct conclusion then to reject the notion that this is just a fluctuation?

Outline

Lecture 1: Science, Probability, and Statistics

Lecture 2: Some Common Probability Distributions, Expectation and Variance

Lecture 3: Correlation, Central Limit Theorem, Uncertainty propagation, Estimators

Lecture 4: Estimators, Data Modeling, Maximum Likelihood

Lecture 5: Confidence intervals, Least squares and Hypothesis testing

Lecture 6: Hypothesis testing, Confidence intervals, Systematic uncertainties and checks

Lecture 7: Combinations, Bayesian Statistics

Hypothesis testing: single-point deviation

An X-ray telescope monitors a particular star for a whole year, during which it receives a total of 36000 X-ray photons from the star.

On one particular day, $\hat{x}_i = 130$ photons were detected. Has the star undergone an outburst?

$$\langle \hat{x} \rangle = \frac{36000}{365} = 98.6, \qquad \lambda \approx \langle \hat{x} \rangle, \qquad \sigma_\lambda = \sqrt{\lambda} \approx \sqrt{\langle \hat{x} \rangle}$$

as photon counts obey a Poisson distribution.

Therefore, the anomalously high rate observed on that particular day corresponds to a

$$\frac{\hat{x}_i - \lambda}{\sigma_\lambda} \approx \frac{130 - 98.6}{\sqrt{98.6}} \approx 3.16\,\sigma$$

departure from the mean.

Trial factors and the look-elsewhere effect

Full Poisson calculation gives p=0.2% and Gaussian approximation gives p=0.1%.

We can set our $p-$value at 1%. Is the correct conclusion then to reject the notion that this is just a fluctuation?

Trial factors and the look-elsewhere effect

The probability of getting 130 counts departure from the mean of a Poisson distribution with rate $\lambda \approx 98.6$ from random statistical fluctuations alone is $p = 0.002$.

But the telescope observed the star every day for a year. On each day there was a probability p of observing such an outburst by chance.

What is the probability of getting at least one success in $n = 365$ trials where each trial has a probability of success $p = 0.002$? Binomial!

$$\begin{aligned} P(r > 1|n, p) &= \sum_{r=1}^{\infty} \mathcal{B}(r|n, p) = 1 - \mathcal{B}(0|n, p) \\ &= 1 - \binom{n}{0} p^0 (1-p)^{n-0} = 1 - (1-p)^n = 0.51 \end{aligned}$$

This is known as the ***Bonferroni correction*** or the "look-elsewhere" effect.

It's not usually this simple! Are the tests truly independent?

Can I reject the null hypothesis? Case 1: known variance

We obtain N measurements $\{\hat{x}_i\}$ of x, each with known variance σ^2, and compute the sample mean $\bar{x}$.

Thanks to the CLT, we know that the quantity:

$$\langle \hat{r} \rangle = \frac{\bar{x} - \mu}{\sigma/\sqrt{N}} \sim \mathcal{N}(0, 1)$$

To evaluate confidence intervals for this situation we would use a standard normal distribution look-up table or calculator.

For example, $P(|\langle \hat{r} \rangle| > 3) = 0.003$ (regardless of N). If we observe $|\langle \hat{r} \rangle| > 3$, the null hypothesis is rejected at $> 99.7\%$ confidence.

NB: in some cases it's more appropriate to use a one-sided confidence interval.

Probability of exceeding limits with a Gaussian pdf

κ	$P(\lvert x-\mu\rvert > \kappa\,\sigma)$	$P(x > \mu+\kappa\,\sigma)$
1	0.3173	0.1587
2	0.0455	0.0228
3	0.0027	0.0014
4	0.00006	0.00003

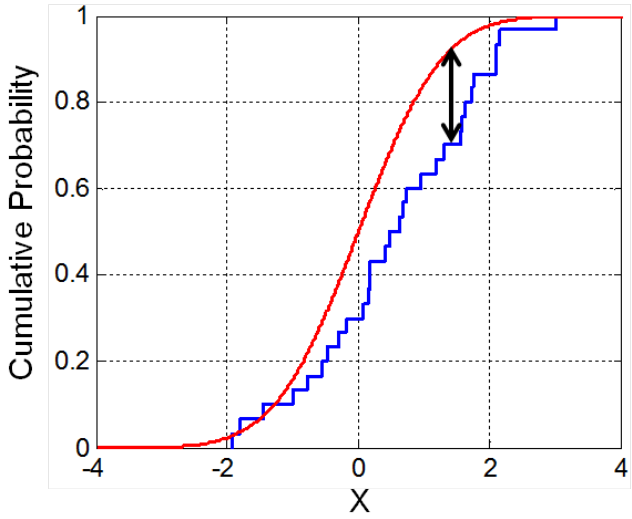
Is this distribution of RV drawn from that parent distribution?

Consider a dataset $\hat{\mathbf{x}}$ containing n observations of an RV x. Does x have a given probability distribution with cdf $F(x)$?

The one-sample ***Kolmogorov-Smirnov (K-S) statistic*** is

$$\hat{D}_n = \max_{\hat{x}} |F_n(\hat{x}) - F(x)|.$$

where $F_n(\hat{x})$ is the empirical cumulative distribution function (ecdf) of the observations.



One-sample K-S test

Alexey Kolmogorov showed that, under the null hypothesis, the probability distribution for $\sqrt{n}D_n$ asymptotically approaches the so-called Kolmogorov distribution.

(The approximation is poor for small N, but a small-sample correction exists.)

Nikolai Smirnov then used this distribution to compute critical values K such that $p(D_n \geq K) = 1 - \alpha$, for different values of n and α.

If $\hat{D}_n \geq K$, the null hypothesis is rejected with confidence α.

Examples K-S test

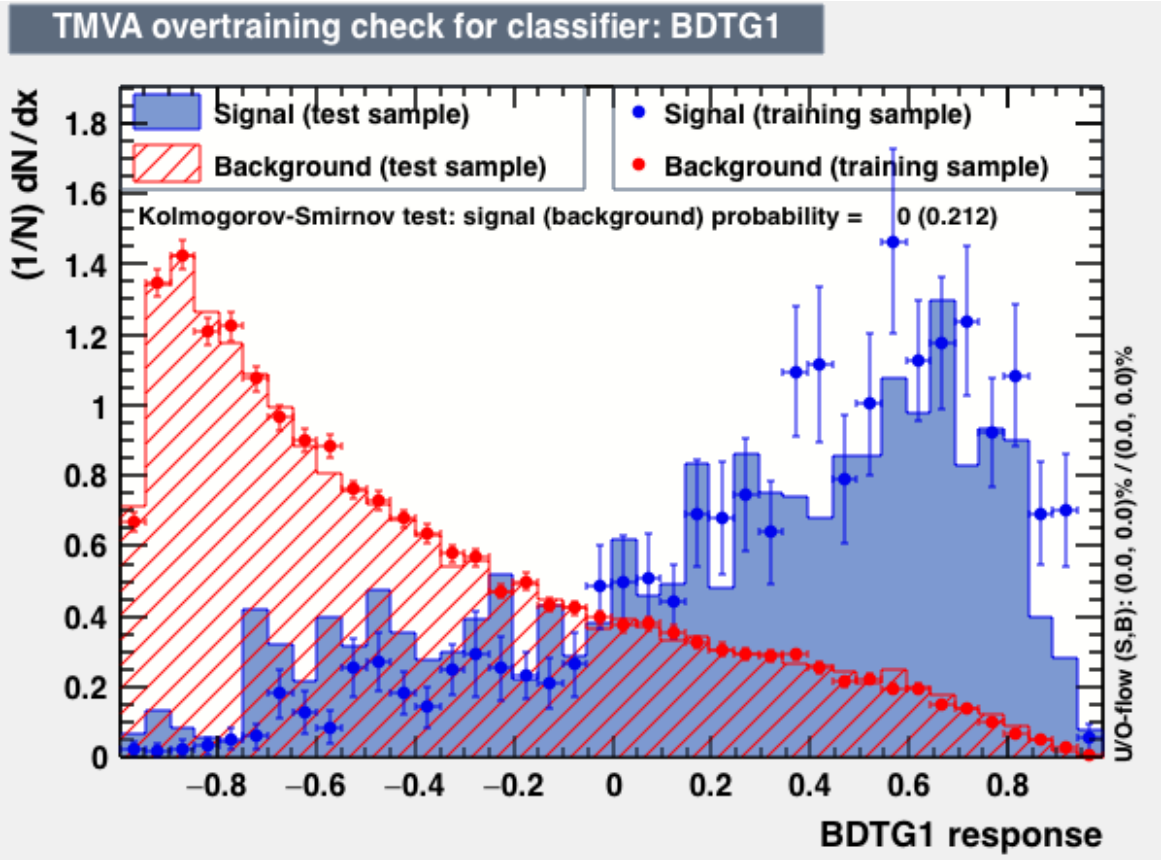


Figure:
Credit:https://root-forum.cern.ch/t/change-in-tmva-bdtg-results/29268

Return to confidence intervals

Confidence intervals in > 1-D

When we determine a single parameter the baseline assumption is that its probability distribution is Gaussian to set the 1-D confidence intervals. When fitting multiple parameters the joint distribution over the MLEs becomes a multi-variate Gaussian:

$$\mathbf{a}_{\mathrm{ML}} \sim \mathcal{N}(\mathbf{a}, \mathbf{C})$$

The covariance matrix $\mathbf{C}$ is always ***symmetric*** and ***positive semi-definite*** (meaning that $|\mathbf{C}| \geq 0$).

If we only care about one of the parameters, we can ***marginalise*** the joint distribution over all the other parameters (that is, project the distribution onto just one dimension), and use the resulting ***marginal distribution*** to estimate confidence intervals.

Alternatively, we can define a ***confidence region*** in 2 or more dimensions, that contains a certain fraction of the probability mass.

Confidence intervals vs confidence region

When reporting confidence regions, we choose both the confidence level and the shape of the region.

A natural choice of shape is one that is bounded by a constant value of the joint probability density function.

If the likelihood is (approximately) Gaussian, these regions are ellipses (in 2-D) or ellipsoids (in 3 and more dimensions) centred on the MLE values.

1-D confidence intervals vs 2-D confidence regions

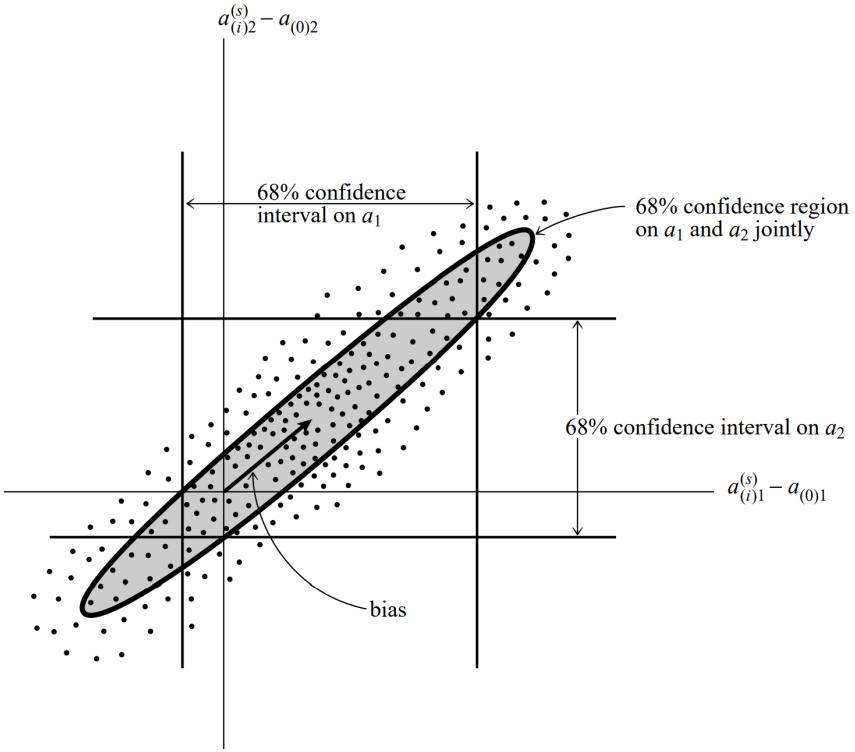


Figure: Confidence intervals in 1 and 2 dimensions. The same fraction of measured points (here 68%) lies (i) between the two vertical lines, (ii) between the two horizontal lines, (iii) within the ellipse. Source: *Numerical Recipes: The Art of Scientific Computing*, Press et al.

1-D confidence intervals vs 2-D confidence regions

Multi-dimensional confidence regions span wider ranges than 1-D confidence intervals with the same confidence level.

This is true even if the parameters are uncorrelated. Consider two parameters a_1 and a_2 with joint distribution:

$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix} \right) .$$

In polar coordinates

$$p(r) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{r^2}{2\sigma^2}\right) \quad \text{where} \quad \begin{aligned} r &= \sqrt{a_1^2 + a_2^2} \\ \theta &= \arctan(a_2/a_1) \end{aligned} .$$

1-D confidence intervals vs 2-D confidence regions

Integrate the pdf to get the cdf:

$$\begin{aligned} P(R) &= \int_{\theta=0}^{2\pi} \int_{r=0}^{R} r\, p(r)\, \mathrm{d}r\, \mathrm{d}\theta \\ &= \int_{r=0}^{R} \frac{r}{\sigma^2} \exp\left(-\frac{r^2}{2\sigma^2}\right) \mathrm{d}r \\ &= 1 - \exp\left(-\frac{R^2}{2\sigma^2}\right) \end{aligned}$$

Plugging in $R = \sigma$, for example, gives $P(R = \sigma) = 0.393$, not 0.683 as would be the case in 1-D.

To get a given confidence level α, set $P(R) = \alpha$ and solve for R which gives $R_\alpha = \sigma\sqrt{-2\ln(1-\alpha)}$.

1-D confidence intervals vs 2-D confidence regions

This also applies to the general case

$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_1 \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \rho \\ \rho & \sigma^2 \end{bmatrix} \right).$$

becausewe can always find a 2×2 matrix $\mathbf{A}$, such that the transformed parameters $\mathbf{a}' = \mathbf{A}\,\mathbf{a}$ have covariance $\mathbf{C}' = \sigma^2\mathbf{I}$.

Confidence intervals and regions from $\Delta\chi^2$

When using least-squares fitting, symmetric confidence intervals or regions about the MLEs correspond to lines or surfaces of constant $\chi^2 = \chi^2_{\rm min} + \Delta\chi^2$.

The values of $\Delta\chi^2$ corresponding to a given confidence level are taken from the cdf of a chi-square distribution with D degrees of freedom, where D is the number of dimensions of the confidence region (i.e. the number of parameters being considered simultaneoulsy). NB: this is different from the number of dofs used to check the goodness of fit!

For $D = 1$, $\Delta\chi^2 = 1, 2.71.6.63$ corresponds to 68.3% (1-σ), 90% and 99% confidence levels, respectively. For $D = 2$, the 68.3% confidence level corresponds to $\Delta\chi^2 = 2.3$.

1-D confidence intervals vs 2-D confidence regions

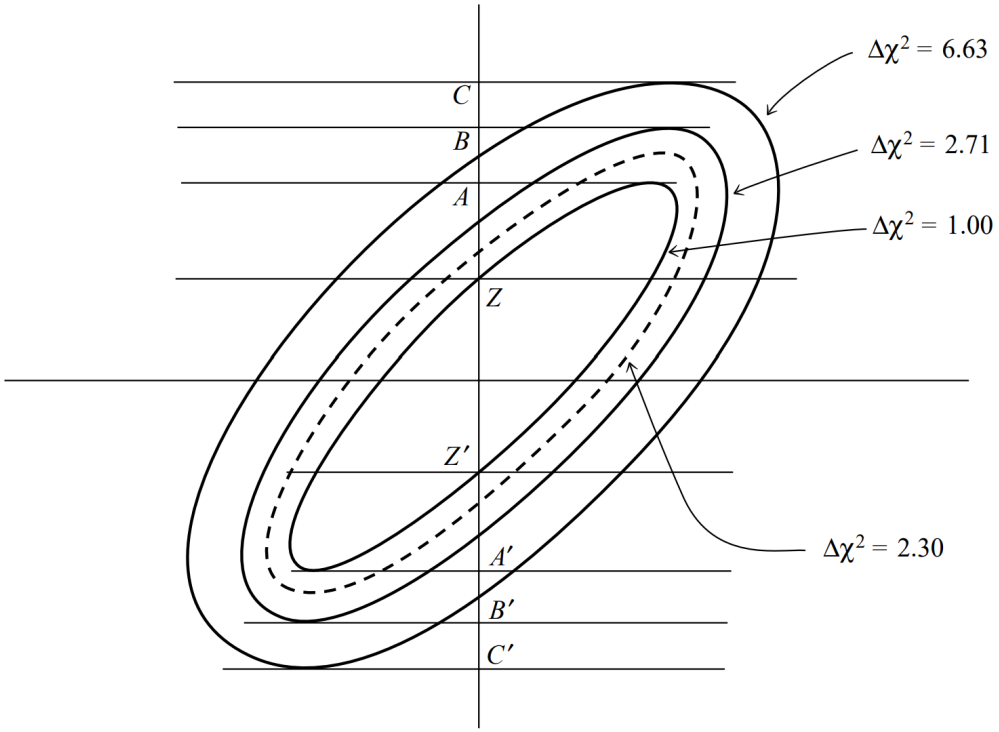


Figure: Confidence region ellipses corresponding to values of χ^2 larger than the fitted minimum. The solid curves, with $\Delta\chi^2 = 1.00, 2.71, 6.63$ project onto one-dimensional intervals AA', BB', CC'. These intervals – not the ellipses themselves – contain 68.3%, 90%, and 99% of normally distributed data. The ellipse that contains 68.3% of normally distributed data is shown dashed, and has $\Delta\chi^2 = 2.30$. Source: *Numerical Recipes: The Art of Scientific Computing*, Press et al.

Systematic Uncertainties

- So far we've mainly considered statistical errors - ones where the estimate of the standard deviation of the estimator reduces the more measurements you make
- These derive from true random processes - either fundamental from the physics process of interest, or through a series of things in the equipment.
- There are systematic uncertainties too which can be a bit more mysterious
- In the practical lab you can hand-wavily talk about these, or slap some large ones on to the dodgy result to make it ok
- But when we set uncertainties confidence intervals are implied - hence this has meaning,
  - Underestimating errors - more likely to reject hypothesis
  - "Being conservative" makes your confidence intervals devoid of meaning. Undoes any good work you did do.

Systematic uncertainties

- Mistakes
- True systematic uncertainties
- Cross checks

Systematic mistakes

Or as they are more commonly known ***mistakes***
Consider this dataset:

$$3.45, 3.49, 3.42, 3.44, 3.46, 3.46, 3.94, 3.45, 3.47$$

There is an outlier

$$3.45, 3.49, 3.42, 3.44, 3.46, 3.46, 3.94, 3.45, 3.47$$

The study of statistics tells us this is an outlier - but not what to do with it. That's up to us. Looks like a transcription error

Ideally you understand why the mistake happened and fix it. Mistakes do happen, no problem with that, we have tools to find them so use them.

What if the noise doesn't have zero mean?

Some forms of experimental error (e.g. thermal noise in electrical circuits, background noise in astronomical images) are always positive.

More generally, many physical processes lead to ***systematic*** noise

$$\epsilon \sim \mathcal{N}(C, \sigma_C^2)$$

where $C \neq 0$. Usually C is known (from the physics of the system or from careful calibration of the experiment), albeit with some uncertainty σ_C.

We can write $x_i = \mu + C + \hat{\epsilon}_i + \hat{\epsilon}_i^C$ where $\hat{\epsilon}_i \sim \mathcal{N}(0, \sigma^2)$ and $\hat{\epsilon}_i^C \sim \mathcal{N}(0, \sigma_C^2)$.

Then $\mu_{\mathrm{ML}} = \bar{x} - C$ and $\sigma^2(\mu_{\mathrm{ML}}) = \frac{\sigma^2}{N} + \sigma_C^2$.

What if the noise doesn't have zero mean?

Then $\mu_{\mathrm{ML}} = \langle \hat{x} \rangle - C$ and $\sigma^2(\mu_{\mathrm{ML}}) = \frac{\sigma^2}{N} + \sigma_C^2$.

- The mean is corrected by the bias.
- Always correct for ***known*** biases.
- An additional error is gained from the uncertainty in that bias
- The ***systematic error does not reduce with more data***.
- Systematic error analysis should come with the design of the experiment, not at the end
- No point taking days of data if your total uncertainty is rapidly dominated by σ_c
- In principle σ_c can be reduced through other experiments, better equipment, etc

Are the measurements consistent?

Consider two measurements of a parameter a from two different experiments:

$$a_1 \pm \sigma_{a_1}; a_2 \pm \sigma_{a_2}$$

Are they consistent? Let's assume σ_{a_1} and σ_{a_2} are uncorrelated. We have

$$\Delta = |a_1 - a_2|; \sigma_\Delta = \sqrt{\sigma_{a_1}^2 + \sigma_{a_2}^2}$$

Then can compute

$$\frac{\Delta}{\sigma_\Delta}$$

If the number is near 1 we accept that they are consistent and if larger than 4 they are not, and if inbetween its a choice on "how many σ discrepancy we allow", noting that if checking 20 pairs we ***expect*** one pair to be different by at least 2σ,

Checking results

It is common to try and check your analysis. One way is to split the data (often by a characteristic that shouldn't matter) and check the subset.

Do an experiment to measure something. Leave it running for a week. Heating is on in the lab on weekdays. Shouldn't be a problem your equipment shouldn't behave any differently with that type of temperature change.

$$a_{MF} \pm \sigma_{a_{MF}}; a_T \pm \sigma_{a_T}$$

MF = Monday to Friday, T = total (all days) Very clearly $\sigma_{a_{MF}}$ and σ_{a_T} are correlated, so what is σ_Δ?

Checking results

$$\sigma_\Delta^2 = \sigma_{a_T}^2 + \sigma_{a_{MF}}^2 - 2\rho\sigma_{a_{MF}}\sigma_{a_T}$$

$$\rho = \frac{\sigma_{a_T}}{\sigma_{a_{MF}}}$$

$$\sigma_\Delta^2 = \sigma_{a_{MF}}^2 - \sigma_{a_T}^2$$

and so the significance of the shift in results becomes:

$$\frac{\Delta}{\sqrt{\sigma_{a_{MF}}^2 - \sigma_{a_T}^2}}$$

Checking results

The example is rather simple. A harder problem : You fit some parameters from a binned distribution of your RV. You can do this fit with a different number of bins. You test to see that you are not sensitive to your bin choice by using an alternate fit with different bins.

You expect there to be some wobble in the results, but how big is an indication of a problem?

$$\sigma_\Delta^2 = \sigma_{a_1}^2 + \sigma_{a_2}^2 - 2\rho\sigma_{a_1}\sigma_{a_2}$$

What is the value of ρ?

Problems like this are often best solved by creating multiple sets of simulated data. Then you can check to see what Δ is expected to be between two methods, and whether your observed Δ is within that distribution or in the tails.

Checks not systematic uncertainties

- In an experiment you can make lots of choices, e.g threshold value, cutoff values, measuring instrument, bin size, fit function etc. etc
- you should check that your measurement isn't biased by these choices
- DO NOT add the small shifts you see in these checks to the systematic uncertainty.
- This penalizes diligence. Alternatively lets you set the size of the uncertainty as large as you like, you just need to keep finding checks until you hit the target.
- If your check doesn't pass, find the underlying cause and take that into account applying a systematic uncertainty if necessary

Outline

Lecture 1: Science, Probability, and Statistics

Lecture 2: Some Common Probability Distributions, Expectation and Variance

Lecture 3: Correlation, Central Limit Theorem, Uncertainty propagation, Estimators

Lecture 4: Estimators, Data Modeling, Maximum Likelihood

Lecture 5: Confidence intervals, Least squares and Hypothesis testing

Lecture 6: Hypothesis testing, Confidence intervals, Systematic uncertainties and checks

Lecture 7: Combinations, Bayesian Statistics

Systematic uncertainties and covariance matrices

Imagine, we fixed the mistakes, corrected the biases, performed the checks, and evaluated the remaining systematic errors.

Lets think about two measurements, x_1 and x_2 that have common systematic error σ_S (e.g due to the common use of a material expansion parameter that has an uncertainty in its value)
The variance of x_1 is

$$\mathrm{Var}\left[x_1\right] = \mathbb{E}\left[x_1^2\right] - \mathbb{E}\left[x_1\right]^2$$

Let x_1 be comprised of a part x_1^R with random error σ_1 and x_1^S with systematic error σ_S

$$\mathrm{Var}\left[x_1\right] = \mathbb{E}\left[(x_1^R + x_1^S)^2\right] - \mathbb{E}\left[x_1^R + x_1^S\right]^2$$

Building covariance matrices

$$\mathrm{Var}\left[x_1\right] = \mathbb{E}\left[(x_1^R + x_1^S)^2\right] - \mathbb{E}\left[x_1^R + x_1^S\right]^2$$

x_1^R and x_1^S are independent of each other so

$$\mathrm{Var}\left[x_1\right] = \sigma_1^2 + \sigma_s^2$$

and similarly

$$\mathrm{Var}\left[x_2\right] = \sigma_2^2 + \sigma_s^2$$

then

$$\begin{aligned}
\mathrm{cov}(\mathrm{x}_1, \mathrm{x}_2) &= \mathbb{E}\left[x_1 x_2\right] - \overline{x_1} \times \overline{x_2} \\
&= \mathbb{E}\left[(x_1^R + x_1^S)(x_2^R + x_2^S)\right] - \overline{x_1^R + x_1^S} \times \overline{x_2^R + x_2^S} \\
&= \mathbb{E}\left[x_1^S x_2^S\right] - \overline{x_1^S} \times \overline{x_2^S} \\
&= \sigma_s^2
\end{aligned}$$

Covariance Matrix

Hence the covariance matrix becomes

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 + \sigma_s^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_2^2 + \sigma_s^2 \end{pmatrix}$$

Can generalise this to more measurements and even those where some systematic uncertainties appear in some and not others.

Combining measurements

Already looked at the way to do this with fully independent measurements:

$$\mu = \frac{\sum_i w_i \, \hat{x}_i}{\sum_i w_i}, \text{ where } w_i = \frac{1}{\sigma_i^2}.$$

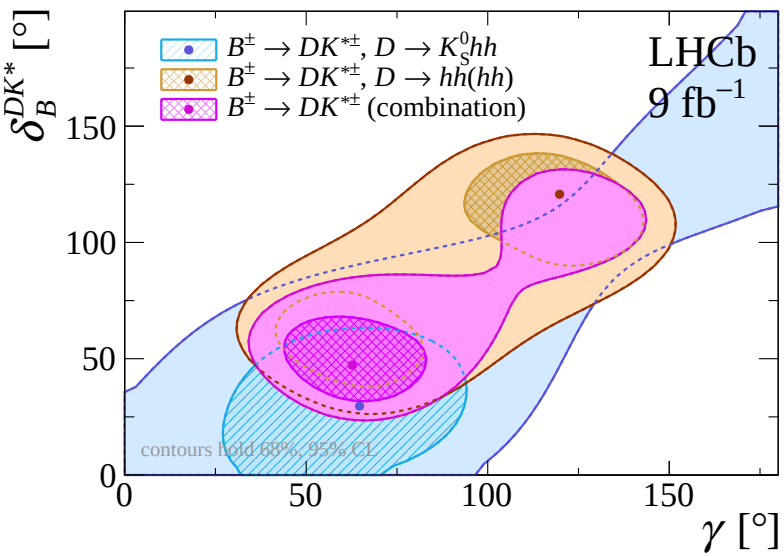
Impact of correlations is to change the weight function which becomes

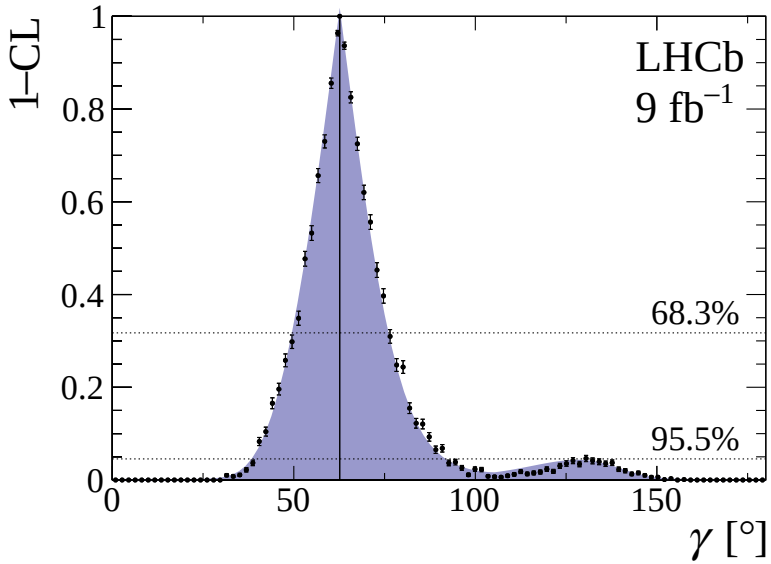
$$w_i = \frac{\sum_{j=1}^{N} (\mathbf{V})_{ij}^{-1}}{\sum_{k,l=1}^{N} (\mathbf{V})_{kl}^{-1}} \text{ which satisfies } \sum_{i=1}^{N} w_i = 1$$

$$\sigma_\mu^2 = \sum_{i,j=1}^{N} w_i w_j \mathbf{V}_{ij}$$

which reduces to the formula we had earlier in the case of no correlation.
http://cds.cern.ch/record/183996/files/OUNP-88-05.pdf
The formula is given so that you know that such things are possible.

Example of CL and combinations





Human Bias - bad or good?

The human can be the biggest cause of the bias - when results conform to what you believe you don't do too much and when they don't agree you go looking for the mistake: That's human nature. You need to be aware of that and the statistical methods you choose to gain inference. On the other hand is you gut feeling always wrong? Surely prior knowledge and understanding have a place in inference?

Frequentist vs Bayesian

But prior knowledge is an important thing. Do we look only at the data infront of us and remove all prior knowledge? Or should we incorporate that knowledge into our inference?

If we do the latter then how do we incorporate prior knowledge robustly and without a vauge "belief" concept.

Some History

- What does it mean to determine the mass of Saturn from orbital data?
- What is the meaning of the uncertianty you might assign to that?

Frequentist vs Bayesian

- Frequentist: Defines probability as the limit of an event's frequency in the limit of infinite trials.
- Frequentist statistics assigns probabilities to data.
- Bayesian: Defines probability as a statement of belief or plausibility.
- Bayesian statistics assigns probabilities to hypotheses.

The likelihood

In general we want to use the data to learn more about the parameter(s). This is called ***inference***.

We have $p(\hat{x}|\theta)$ as a function of θ, and call it the ***likelihood***:

$$\mathcal{L}(\theta) \equiv p(\hat{x}|\theta)$$

The likelihood is equal to the probability of observing the data ***given*** the model and parameter(s).

It plays a fundamental role in data analysis. Writing down a likelihood is usually the first thing you do when modelling data.

Bayes theorem for inference

The likelihood is the probability of the data given the model and its parameters. But we *have* the data and want to *infer* the parameters. We need $p(\theta|\hat{x})$, not $p(\hat{x}|\theta)$.

Bayes' theorem allows us to get from one to the other:

$$p(A|B) = \frac{p(B|A)\,p(A)}{p(B)}$$

Applying Bayes' theorem to the likelihood

$$p(\theta|\hat{x}) = \frac{p(\hat{x}|\theta)\,p(\theta)}{p(\hat{x})}.$$

The components of Bayes' theorem

$$\text{posterior pdf} = \frac{\text{likelihood} \times \text{prior pdf}}{\text{marginal likelihood}}$$

The ***likelihood*** $p(\hat{x}|\theta)$ describes how well the model explains the data.

The ***prior*** $p(\theta)$ encodes our prior knowledge, or beliefs, about θ before we measured $\hat{x}$. Unlike the likelihood, it *is* a probability distribution over θ.

The ***posterior*** $p(\theta|\hat{x})$ is also a probability distribution over θ, but ***updated*** to take the data into account.

The ***marginal likelihood*** $p(\hat{x}) = \int p(\hat{x}|\theta)\, p(\theta)\, \mathrm{d}\theta$ is essentially a normalisation constant. It is often ignored in parameter inference, as it doesn't depend on θ. (But it is important in model comparison.)

Bayesian inference

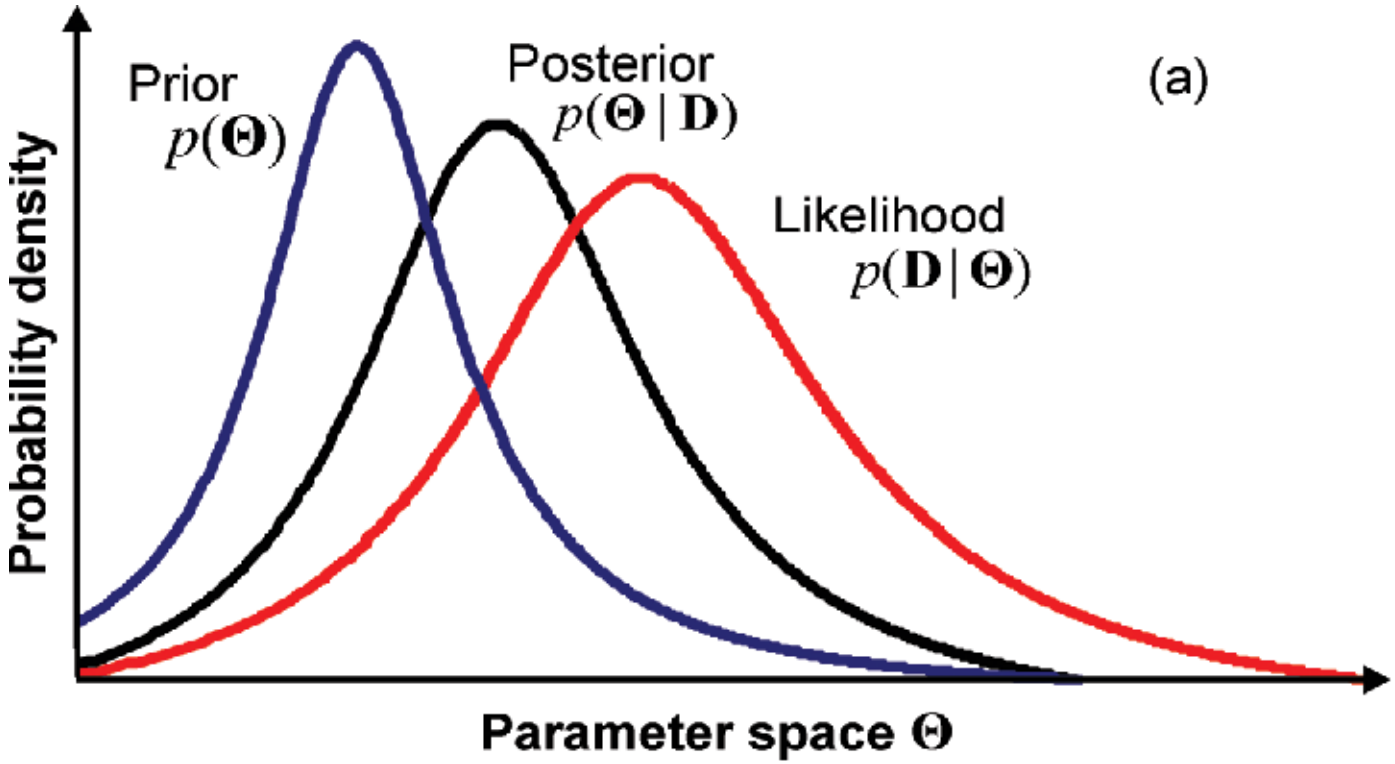


Figure: When we have strong prior beliefs and/or few/imprecise measurements, the prior is narrow compared to the likelihood, and the posterior differs significantly from the likelihood. Source: Xiang & Fackler (2015)

When is it ok to use the likelihood?

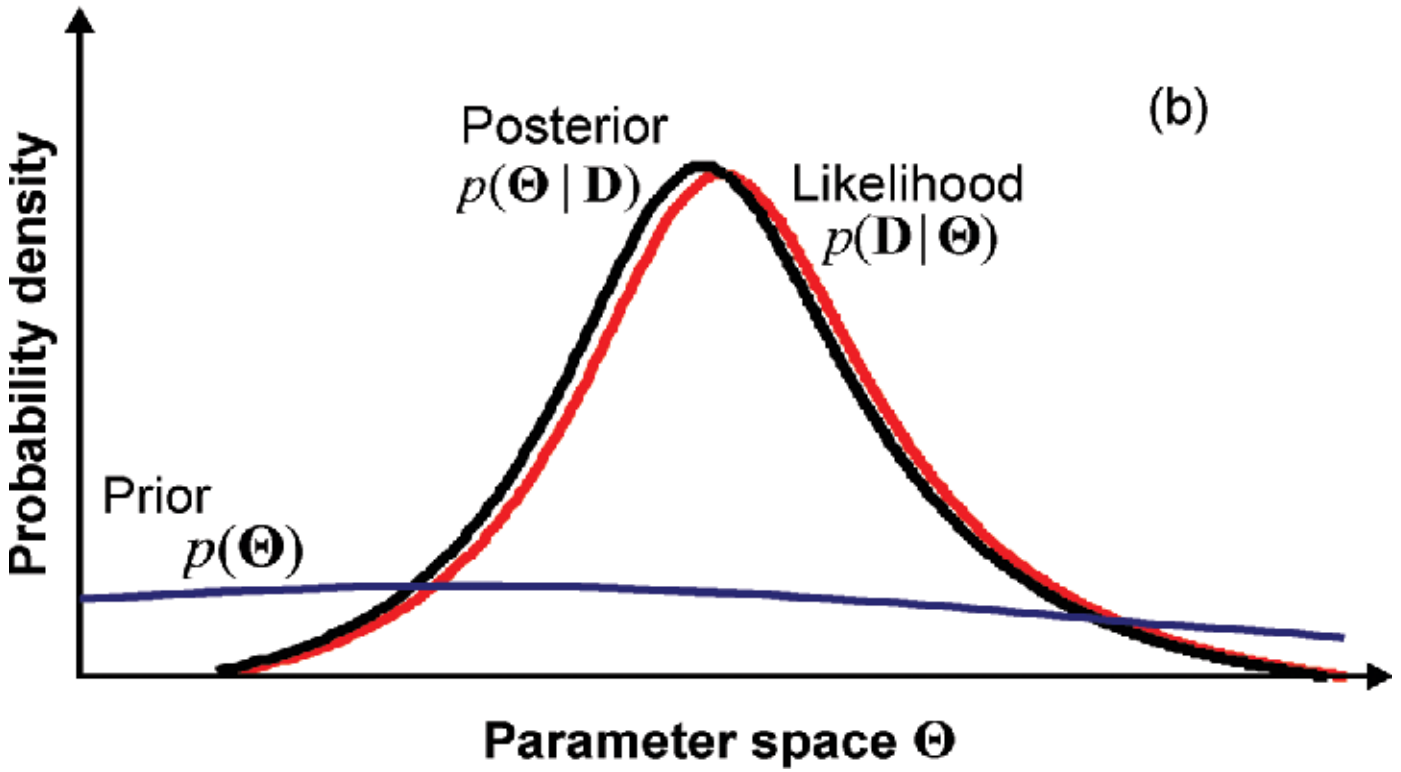


Figure: When we have limited prior knowledge and/or many precise measurements, the prior is broad compared to the likelihood, and the (normalised) likelihood is a good approximation to the posterior. Source: Xiang & Fackler (2015)

Fitting for the parameter(s)

The most probable value of θ *given all available information* is the one that maximises the posterior:

$$\theta_{\mathrm{MAP}} = \max_{\theta} p(\theta|\hat{x}),$$

also known as the ***Maximum A Posteriori (MAP)*** estimate of θ.

The most probable value of θ *given the data only* is the one that maximises the likelihood:

$$\theta_{\mathrm{ML}} = \max_{\theta} p(\hat{x}|\theta),$$

also known as the ***Maximum Likelihood Estimate (MLE)*** of θ.

When the likelihood is a good approximation of the posterior, the MLE is close to the MAP estimate – and is often much easier to evaluate.

Incorporating prior information when modeling data

Recall Bayes' theorem for inference

$$p(\theta|\hat{x}) = \frac{p(\hat{x}|\theta)\,p(\theta)}{p(\hat{x})} = \frac{\mathcal{L}(\theta)\,p(\theta)}{\mathcal{Z}}$$

When we have relevant prior information we ***should always*** incorporate it by folding it into $p(\theta)$. Not doing so can lead to drastically wrong conclusions.

When working with the likelihood alone, we are ***implicitly*** assuming a flat (uniform) prior ranging from $-\infty$ to $+\infty$ (so it can't be normalised, we say it is ***improper***.)

Gaussian priors and likelihood

Sometimes we can draw on previous measurements (even if they are from a different experiment) to define our prior. Often this will mean using a Gaussian prior of the form $p(\theta) = \mathcal{N}(\mu_{\rm pre}, \sigma^2_{\rm pre})$. If the likelihood is also Gaussian,
$\mathcal{L}(\theta) = \mathcal{N}(\mu_{\rm meas}, \sigma^2_{\rm meas})$, one can show that the posterior will be Gaussian too: $p(\theta|\hat{x}) = \mathcal{N}(\theta_{\rm post}, \sigma^2_{\rm post})$, with mean and variance:

$$\mu_{\rm post} = \frac{\mu_{\rm pre}\,\sigma^2_{\rm meas} + \mu_{\rm meas}\,\sigma^2_{\rm pre}}{\sigma^2_{\rm pre} + \sigma^2_{\rm meas}} \quad \text{and} \quad \sigma^2_{\rm post} = \frac{\sigma^2_{\rm pre}\,\sigma^2_{\rm meas}}{\sigma^2_{\rm pre} + \sigma^2_{\rm meas}}.$$

When the prior and likelihood belong to the same family (or related families) of distributions, and the resulting posterior can be expressed analytically, we say that they are conjugate with each other.

Gaussian priors inside least-squares

A Gaussian prior can be thought of as an additional data point consisting of an observation of the parameter with value μ_0 and uncertainty σ_0.

This can be incorporated directly in the χ^2 framework as an additional term.

$$\chi^2 = \left(\sum_{i=1}^{N} \frac{[\hat{y}_i - f(x_i, \theta)]^2}{\sigma_i^2} \right) + \frac{(\theta - \mu_{\rm pre})^2}{\sigma_{\rm pre}^2}.$$

Bayesian inference is sequential

Bayesian inference works by ***updating*** our state of knowledge about a parameter (or hypothesis) as new data flow in. The posterior from a previous cycle of observations becomes the prior for the next one.

Given enough successive experiments, the final posterior will become independent of the initial prior.

Converging views in Bayesian inference

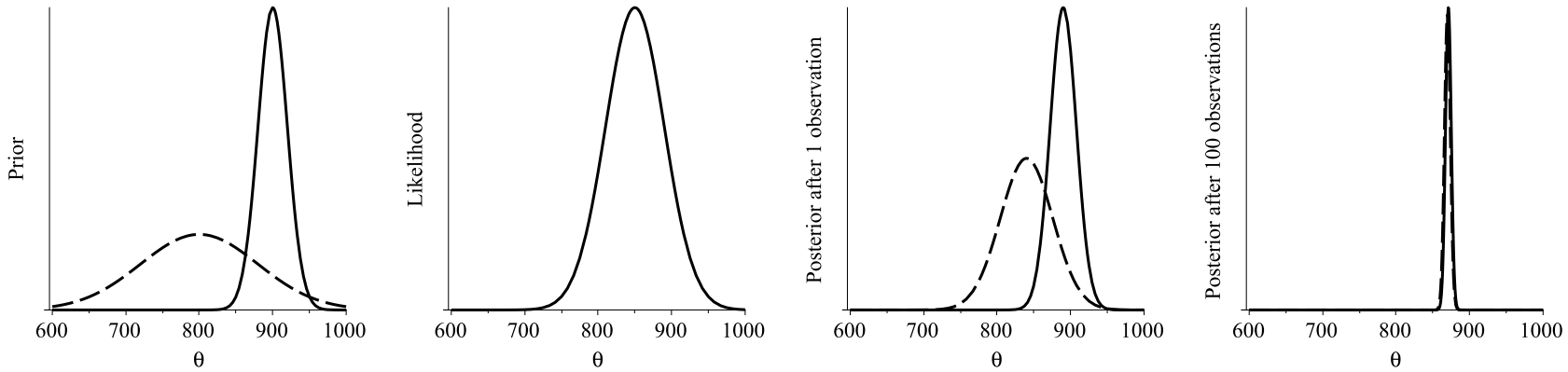


Figure: Two scientists with different prior beliefs $p(\theta)$ about the value of a quantity θ (shown by the two different curves in the first panel) observe one datum with likelihood $\mathcal{L}(\theta)$ (second panel), after which their posteriors $p(\theta|\hat{x})$ (third panel) represent their updated states of knowledge about the parameter. This posterior then becomes the prior for the next observation. After observing 100 data points, the two posteriors have become essentially indistinguishable (fourth panel). Source: R. Trotta.

Uniform prior

Sometimes we have restrictions on the sample space of our parameter. For example, if θ represents a temperature in degrees Kelvin, or a variance, it must always be positive. We can chose a prior that is only defined over $\mathbb{R}^+$.

A common choice of prior when only the range of allowed values is known is a uniform distribution

$$p(\theta)\,\mathrm{d}\theta = \frac{\mathrm{d}\theta}{\theta_{\max} - \theta_{\min}}.$$

It's important to realise that this is not a trivial choice. The limited range can significantly affect the resulting posterior.

Comparison

	Frequentist	Bayesian
Crux	Probability of the events observed	Probability of the theory
Interpretation	Focused on observed effects	In the context of prior beliefs and their update
Sample Size	Large sample required for low bias	Reasonable with small data samples if prior is informative
Computation	Simpler	Posterior estimation can be extremely CPU intensive
Uncertainties	point estimates with uncertainties and confidence intervals	Posterior distributions model the parameter uncertainty and credible intervals

Your Choice

- You can choose to be one or the other
- Both are legitimate forms of analysis
- They answer different questions
- With the exception of one or two places what we've covered so far is relevant to both.
- In the "mathematical problem solving" you will be more familiar with the Frequentist
- In life you will be familiar with the Bayesian
- What you must not do is confuse and mix them up.
- You can choose the method that is most useful for the problem at hand.

Cartoon

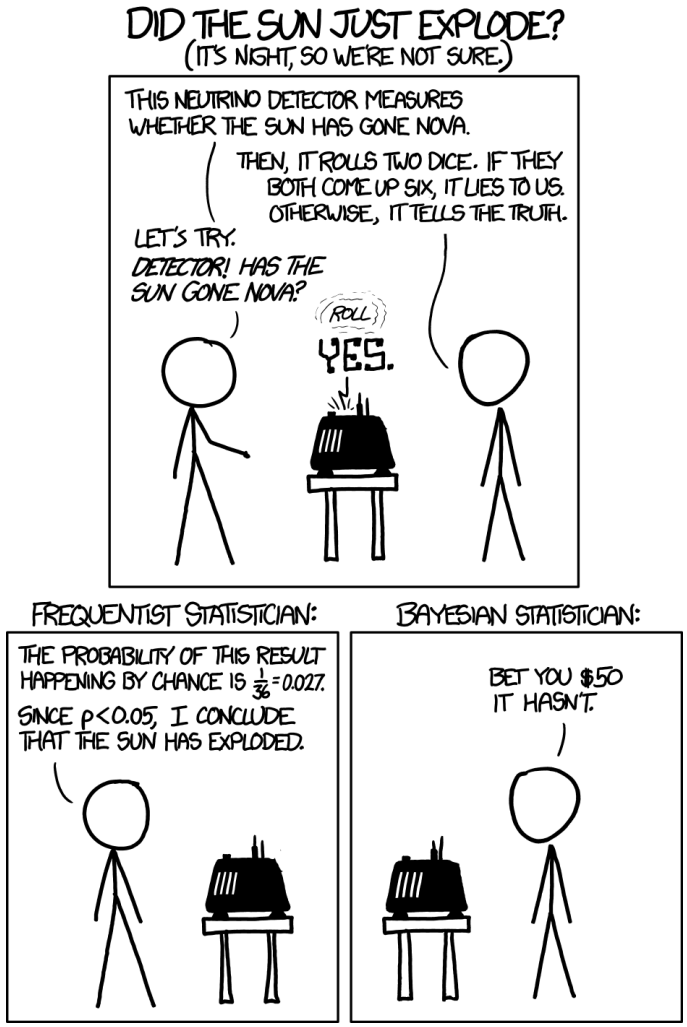


Figure: http://www.xkcd.com

Some Data Collection of my own....