Probability and Statistics

Prof. Suzanne Aigrain

Department of Physics
UNIVERSITY OF OXFORD

MPhys Year 2
Michaelmas Term, 2023

Welcome to Year 2 Probability and Statistics

In Year 1 you saw

- Some statistical recipes for analysing data (in *Introduction to Data Analysis*).
  - Where do those come from?
- Some probability distributions (in *CP4: Mathematical Methods*).
  - How do these relate to real scientific experiments?

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.

Outline

Lecture 1: Science, Probability, and Statistics

Lecture 2: Some Common Probability Distributions

Lecture 3: Expectation, Variance and Correlation

Lecture 4: Likelihood and Inference

Lecture 5: Testing Assumptions and Hypotheses

Lecture 6: Fitting more Complex Models

Lecture 7: Building in Prior Knowledge

Lecture 8: Questions & live exercise

What else?

- A problem set is available for you to attempt and go through with your college tutors.
- Recommended reading:
  - *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;
  - For the very keen: *Data analysis recipes: fitting a model to data* by D. Hogg, J. Bovy & D. Lang, 2010 (https://arxiv.org/abs/1008.4686).

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

Acknowledgements

I am grateful to the Prof. Steve Biller (the previous lecturer for this course) and to Prof. Roberto Trotta (who lectured a similar course at Imperial College London some years ago), for sharing source material with me, which I have drawn on freely to prepare the present lecture notes and the course problem set.

Thanks also to the students and tutors who pointed out typos and errors on these notes and the problem set over the last year. I've tried to correct them, but if you find any more let me know!

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: Suzanne · Aigrain @ 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.

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

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.

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.

In both cases, all probabilities are ***real numbers between 0 and 1***.

The Bayesian definition is more general, and is favoured by most research scientists today. 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

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***:

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

Marginalisation rule

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

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

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,

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

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: $\{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n\}$.
(In this course we use symbols with a hat to denote measured values.)

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:

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

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:

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

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

Lecture 3: Expectation, Variance and Correlation

Lecture 4: Likelihood and Inference

Lecture 5: Testing Assumptions and Hypotheses

Lecture 6: Fitting more Complex Models

Lecture 7: Building in Prior Knowledge

Lecture 8: Questions & live exercise

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:

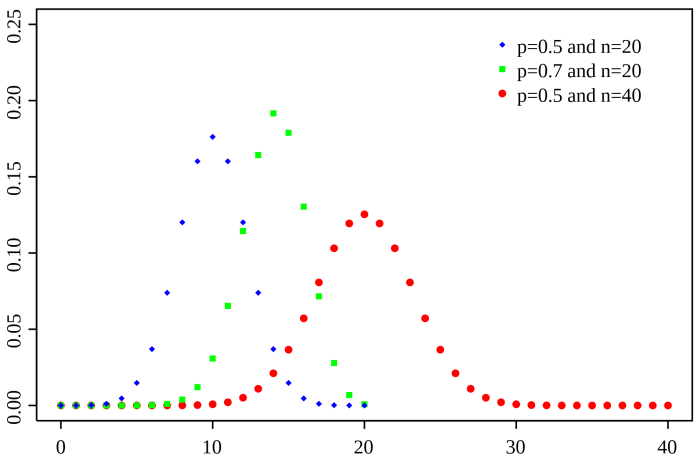
$$P(r|n,p) = \binom{n}{r} p^r \, (1-p)^{n-r}.$$

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



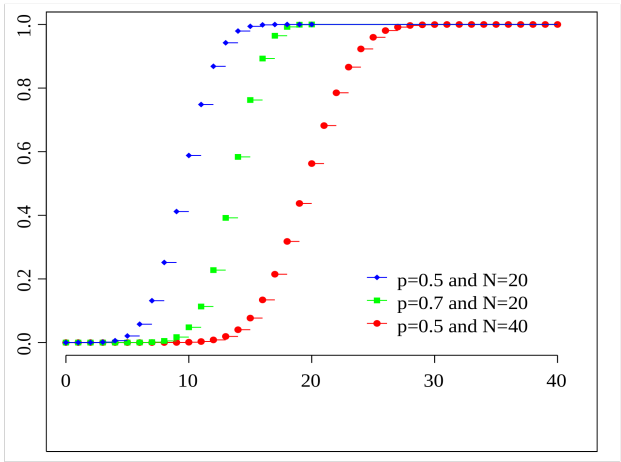


Figure: Pmf and cdf of the binomial distribution for different values of p and n.
Source: Wikipedia.

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?

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

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 cdf 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,
- number of photons or electrons hitting a detector per second
- number of galaxies per square arcminute in the sky

Note that, although n is an integer, λ doesn't have to be.

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

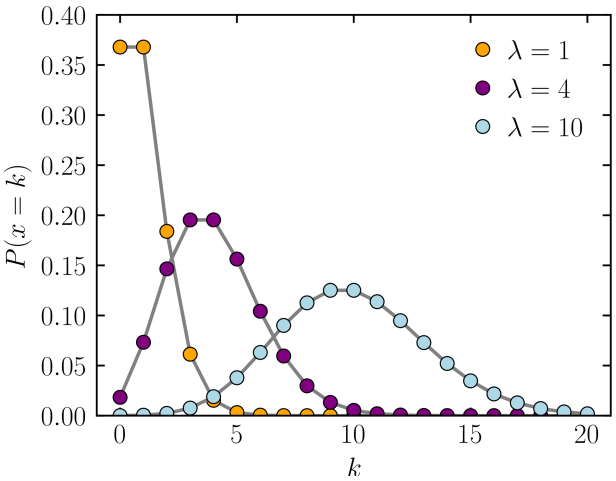
The Poisson cumulative mass function is

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

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



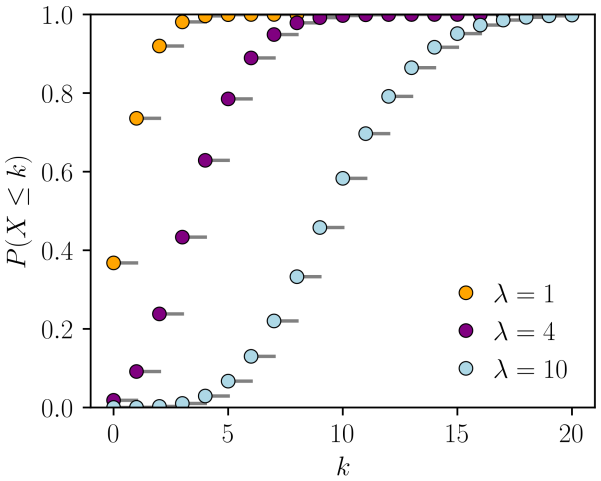


Figure: Pmf and cdf of the Poisson distribution for different values of λ and $t = 1$. Source: Wikipedia.

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?

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

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

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

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

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

Deriving the Poisson distribution

It is also possible to derive the Poisson distribution from the binomial distribution (can you think how?).

Where *not* to use the Poisson distribution

If the rate isn't constant. For example: number of cyclists going down Parks road per hour.

If the fact that one event has occurred makes it more likely that more events will occur. For example: earthquakes.

...

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.

The exponential distribution

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

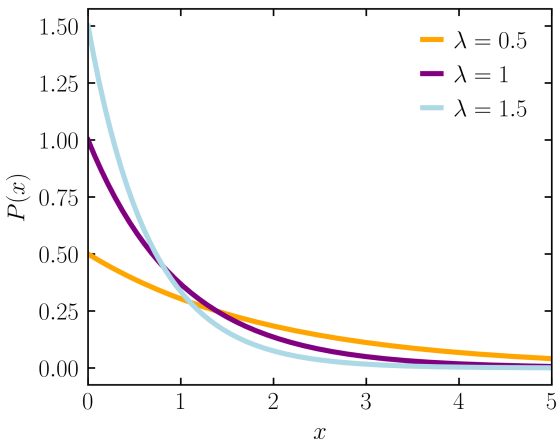
The probability that no event occurs in time interval t is given by the Poisson distribution with $n = 0$: $P(0|\lambda, t) = \exp(-\lambda\, t)$.

The probability 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



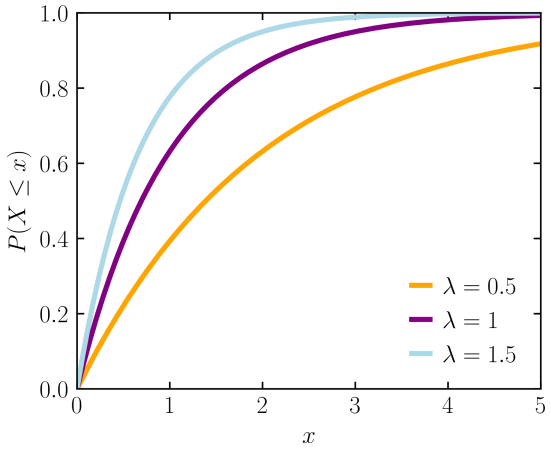


Figure: Pmf and cdf of the exponential distribution for different values of λ.
Source: Wikipedia.

The exponential distribution

Example: You arrive at a bus stop just as the bus is leaving. There are around 5 busses per hour on this route. What is the probability that you will have to wait more than 15 minutes?

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

Plugging in the numbers, $\lambda = 5/60$ buses/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

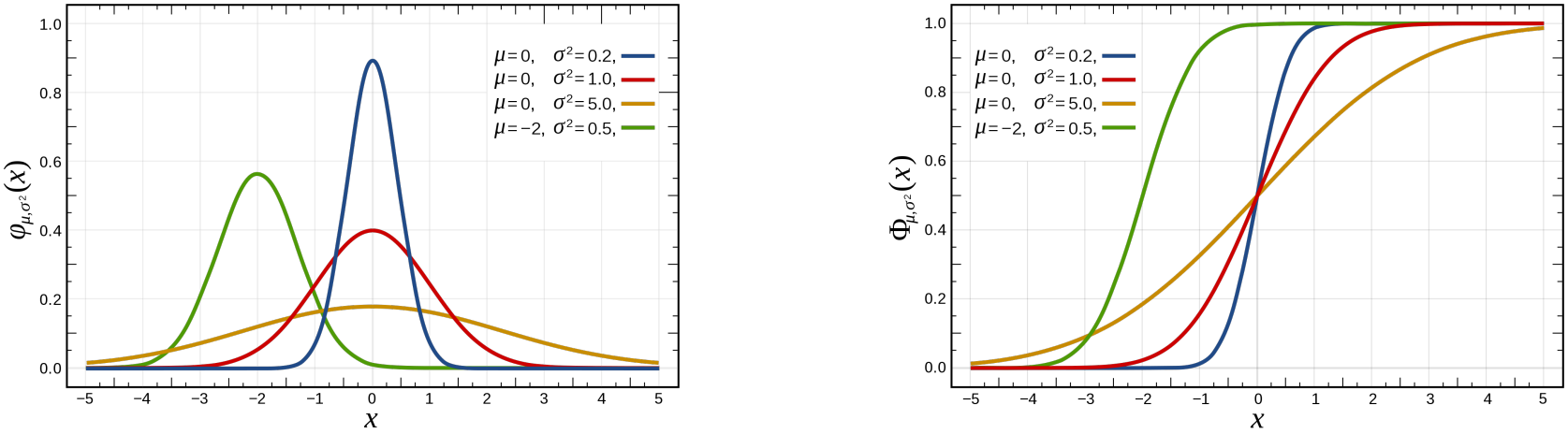


Figure: Pdf and cdf of the Gaussian distribution for different values of μ and σ. Source: Wikipedia.

Deriving the Gaussian

The Gaussian can be derived as the large n limit of the binomial, and the large λt limit of the Poisson distribution.

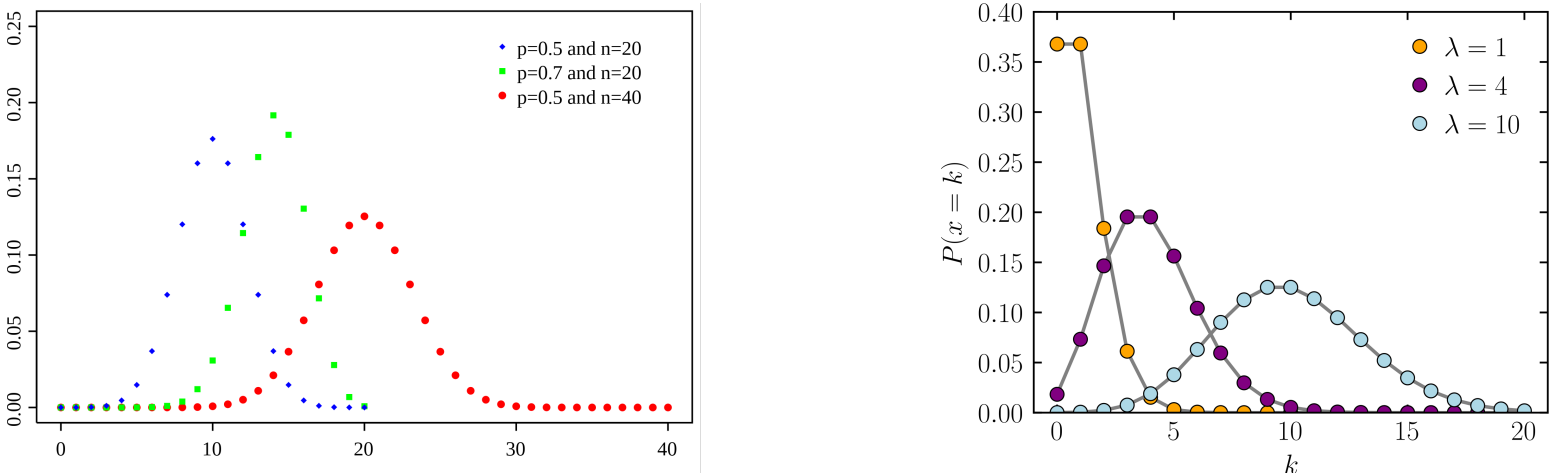


Figure: Binomial (left) and Poisson (right) pmfs.

The approximation is actually very good even for moderate values of the relevant parameters.

The Central Limit Theorem

Let X_1, X_2, ..., X_N be a collection of independent RVs with finite expectation value μ and finite variance σ^2.

For $N \to \infty$, their sum $X \equiv X_2 + X_2 + \ldots + X_N$ is Gaussian distributed with mean $N\mu$ and variance $N\sigma^2$.

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, ...).

Outline

Lecture 1: Science, Probability, and Statistics

Lecture 2: Some Common Probability Distributions

Lecture 3: Expectation, Variance and Correlation

Lecture 4: Likelihood and Inference

Lecture 5: Testing Assumptions and Hypotheses

Lecture 6: Fitting more Complex Models

Lecture 7: Building in Prior Knowledge

Lecture 8: Questions & live exercise

Intuition for the Central Limit Theorem (CLT)

Roll a 6-sided die 3 times and compute the sum S of the outcomes. How many ways are there of getting $S = 3$? $S = 18$? $S = 10$??

All the individual pathways are equally likely, so the final distribution reflects the number of ways of getting a given result.

The Galton board is a physical demonstration that the sum of the outcomes of multiple draws from a binomial process also follows a Gaussian-like distribution.

There are several mathematical proofs of the CLT but they are too involved to go through here.

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

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

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?

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

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

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

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 $t \sim \mathcal{E}(\lambda)$:

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

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

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

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 $f(x)$?

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

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

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

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

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

Sample estimators

Say we have a sample of N measurements $\hat{\mathbf{x}} = \{\hat{x}_i\}, i = 1, \ldots, N\}$ of an RV x. We may not know the what distribution the observations were drawn from.

However, we can define ***estimators*** of the key properties of this distribution, such as its mean and variance, computed from the data.

Sample mean

The ***sample mean*** is defined as:

$$\langle \hat{x} \rangle = \frac{1}{N} \sum_{i=1}^{N} \hat{x}_j$$

Using the rules we derived earlier, we can work out expectation value and variance of the sample mean is

$$\begin{aligned} \mathbb{E}\left[\langle \hat{x} \rangle\right] &= \tfrac{1}{N} \textstyle\sum_{i=1}^{N} \mathbb{E}\left[\hat{x}_i\right] = \mu_x \\ \mathrm{Var}\left[\langle \hat{x} \rangle\right] &= \tfrac{1}{N^2} \textstyle\sum_{i=1}^{N} \mathrm{Var}\left[x\right] = \sigma_x^2/N. \end{aligned}$$

As $\langle \hat{x} \rangle$ involves a sum over independent observations, for large N it follows a Gaussian distribution, regardless of the 'true' distribution of x.

Sample variance

The ***sample variance*** is defined as:

$$\sigma^2(\hat{x}) = \frac{1}{N}\sum_{i=1}^{N}(\hat{x}_i - \langle\hat{x}\rangle)^2$$

One can show (see next slide) that the expectation value of the sample variance is

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

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

Expectation value of the sample variance

$$
\begin{aligned}
\mathbb{E}\left[\sigma^2(\hat{x})\right] &= \mathbb{E}\left[\tfrac{1}{N}\textstyle\sum_i\left(\hat{x}_i - \tfrac{1}{N}\sum_j \hat{x}_j\right)^2\right] \\
&= \tfrac{1}{N}\textstyle\sum_i \mathbb{E}\left[\hat{x}_i^2 - \tfrac{2}{N}\hat{x}_i\sum_j \hat{x}_j + \tfrac{1}{N^2}\sum_j \hat{x}_j \sum_k \hat{x}_k\right] \\
&= \tfrac{1}{N}\textstyle\sum_i\left(\mathbb{E}\left[\hat{x}_i^2\right] - \tfrac{2}{N}\mathbb{E}\left[\hat{x}_i^2\right] - \tfrac{2}{N}\sum_{j\neq i}\mathbb{E}\left[\hat{x}_i\,\hat{x}_j\right]\right) \\
&\quad + \tfrac{1}{N^2}\textstyle\sum_j\left(\mathbb{E}\left[\hat{x}_j^2\right] + \sum_{k\neq j}\mathbb{E}\left[\hat{x}_j\,\hat{x}_k\right]\right) \\
&= \tfrac{N-1}{N}\left[\mathbb{E}\left[x^2\right] - \left(\mathbb{E}\left[x\right]\right)^2\right] \\
&= \tfrac{N-1}{N}\,\sigma_x^2.
\end{aligned}
$$

Unbiased sample variance

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

$$\tilde{\sigma}^2(\hat{x}) \equiv \frac{N}{N-1}\,\sigma^2(\hat{x}) = \frac{1}{N-1}\sum_{i=1}^{N}(\hat{x}_i - \langle\hat{x}\rangle)^2.$$

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

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

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

Sample mean and variance: summary

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

$$\langle \hat{x} \rangle = \frac{1}{N} \sum_{i=1}^{N} \hat{x}_j \quad \text{and} \quad \sigma^2(\hat{x}) = \frac{1}{N} \sum_{i=1}^{N} (\hat{x}_i - \langle \hat{x} \rangle)^2,$$

and the expectation value and variance of the sample mean are

$$\mathbb{E}\left[\langle \hat{x} \rangle\right] = \mathbb{E}\left[x\right] \quad \text{and} \quad \mathrm{Var}\left[\langle \hat{x} \rangle\right] = \frac{\sigma^2(\hat{x})}{N-1}.$$

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.

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.

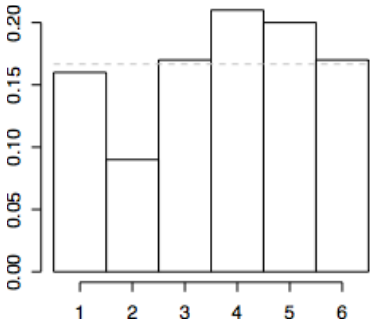


Figure: 100 rolls of a fair, 6-sided dice.
Source: Wikibooks.

For a continuous RV, we can construct a ***histogram***, but it is sensitive to our choice of bins, specially if the sample is small.

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.

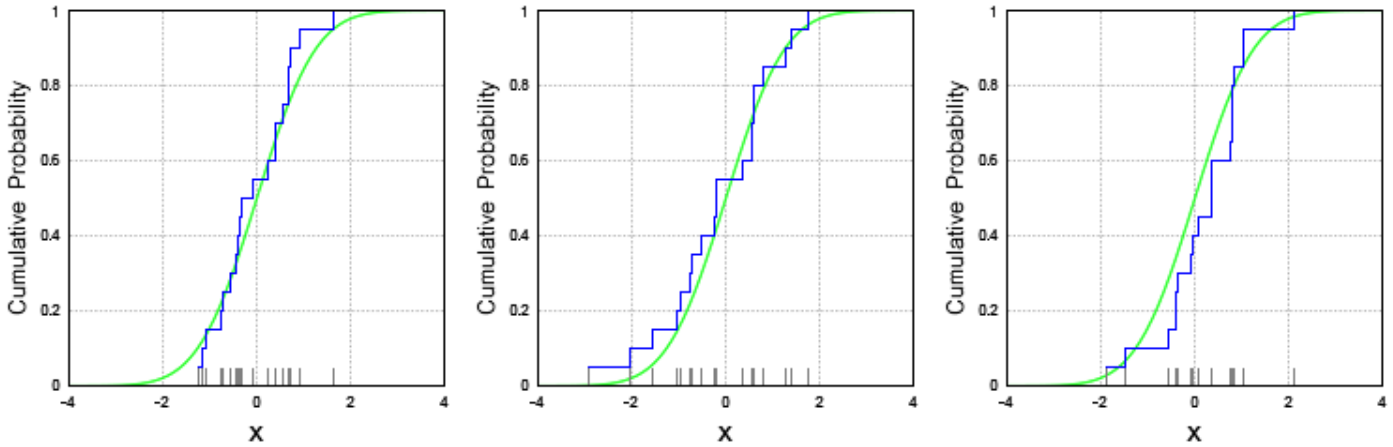


Figure: Three samples from the 'standard' normal distribution ($\mu = \sigma = 1$). In each panel, the green curve is the true cdf, the grey hash marks show the location of 20 observations drawn from that distribution, and the blue step function is the ecdf of that sample. Source: Wikimedia.

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

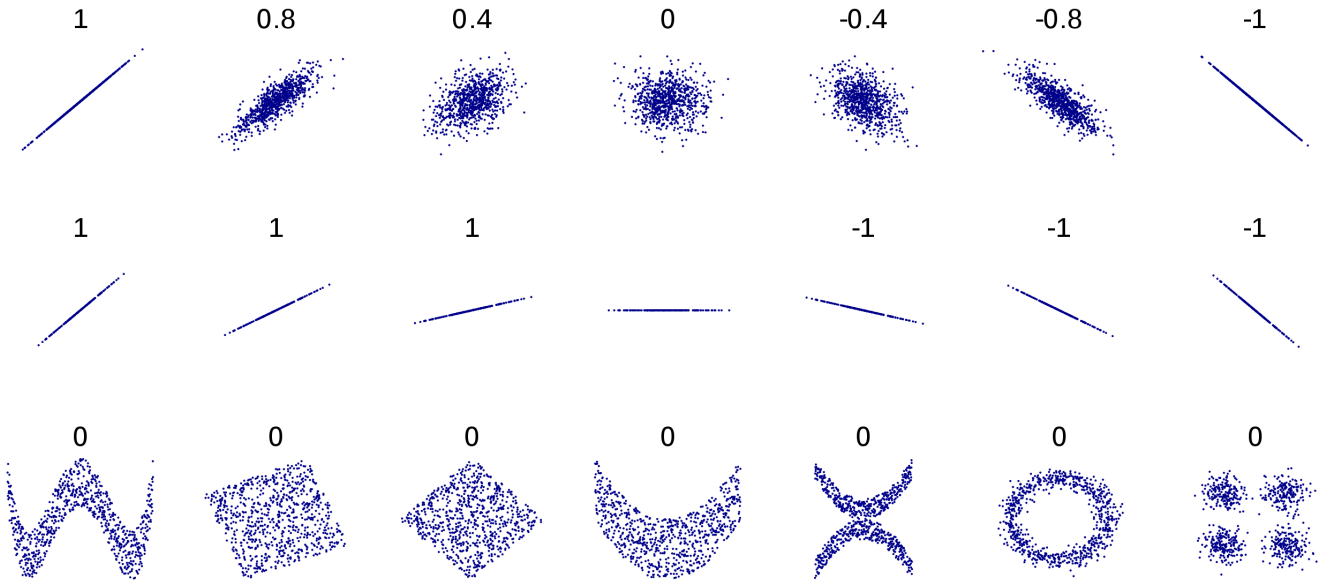


Figure: r_{xy} captures the existence and tightness of a ***linear relationship*** between x and y (but not its slope). Source: Wikipedia..

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***:

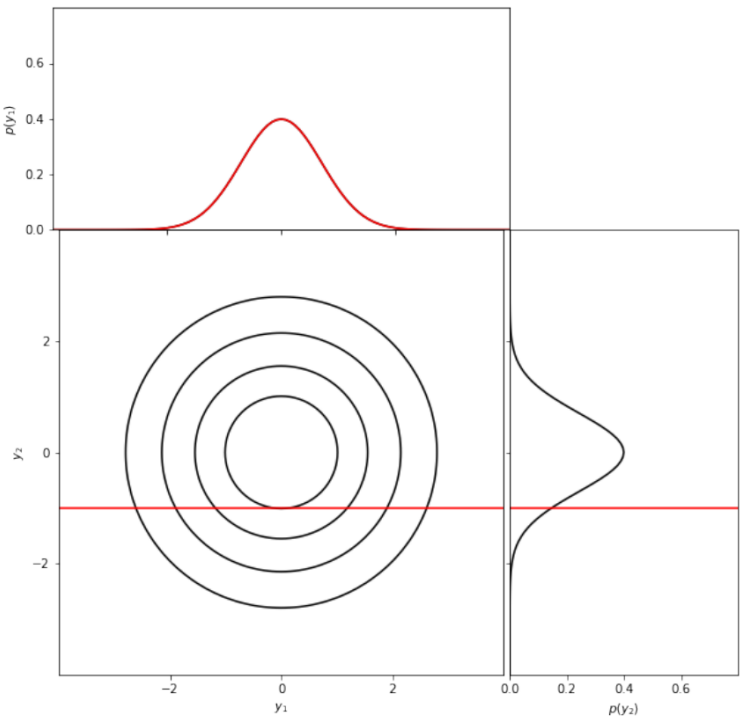
$$p(x, y) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2\pi\sqrt{|\boldsymbol{\Sigma}|}} \exp\left[-\mathbf{r}^{\mathrm{T}}\,\boldsymbol{\Sigma}^{-1}\,\mathbf{r}\right]$$

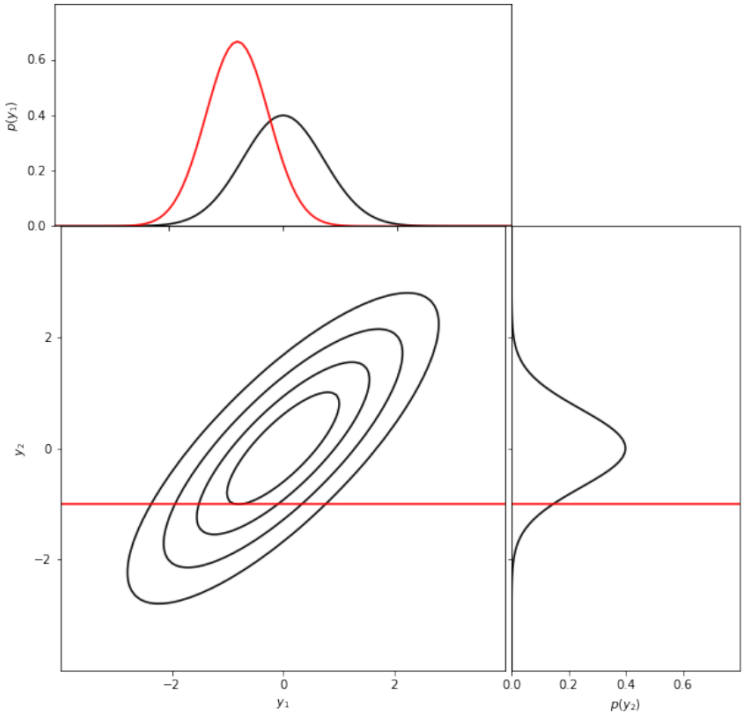
where

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

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

Bivariate Gaussian





Outline

Lecture 1: Science, Probability, and Statistics

Lecture 2: Some Common Probability Distributions

Lecture 3: Expectation, Variance and Correlation

Lecture 4: Likelihood and Inference

Lecture 5: Testing Assumptions and Hypotheses

Lecture 6: Fitting more Complex Models

Lecture 7: Building in Prior Knowledge

Lecture 8: Questions & live exercise

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.

Today we will see how to take external information into account, and work out the full probability distribution over that quantity, subject to some assumptions about the measurement errors.

In later lectures we will see how to relax some of these assumptions and also how to fit more complicated models.

Data, model and parameters

We have an observation $\hat{x}$ (the ***data***) of an RV X, which we believe is drawn from some probability distribution $p(X = \hat{x}|\theta)$ controlled by a ***parameter*** θ. This belief defines the ***model***.

For example:

- If X is the number of heads in 5 consecutive coin tosses and θ is the probability of obtaining heads in one toss, $p(\hat{x}|\theta) = \mathcal{B}(\hat{x}|5, \theta)$.
- If X is the number of buses passing a given stop in $1\,\mathrm{h}$, and θ is the average frequency of buses, $p(\hat{x}|\theta) = \mathcal{P}(\hat{x}|\theta, 1\,\mathrm{h})$.
- If X is a temperature measurement with uncertainty σ and θ is the true temperature $p(\hat{x}|\theta) = \mathcal{N}(\hat{x}|\theta, \sigma^2)$.

Note that there can be multiple observations, multiple parameters, and even multiple models. We will come back to that, but to keep things simple we start with one of each.

The likelihood

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

We re-cast $p(\hat{x}|\theta)$ 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.

Caution!

The likelihood is a probability distribution over the data, ***not*** over θ!

This means we can have

$$\int_{-\infty}^{\infty} \mathcal{L}(\theta)\,\mathrm{d}\theta = \int_{-\infty}^{\infty} p(\hat{x}|\theta)\,\mathrm{d}\theta \neq 1.$$

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

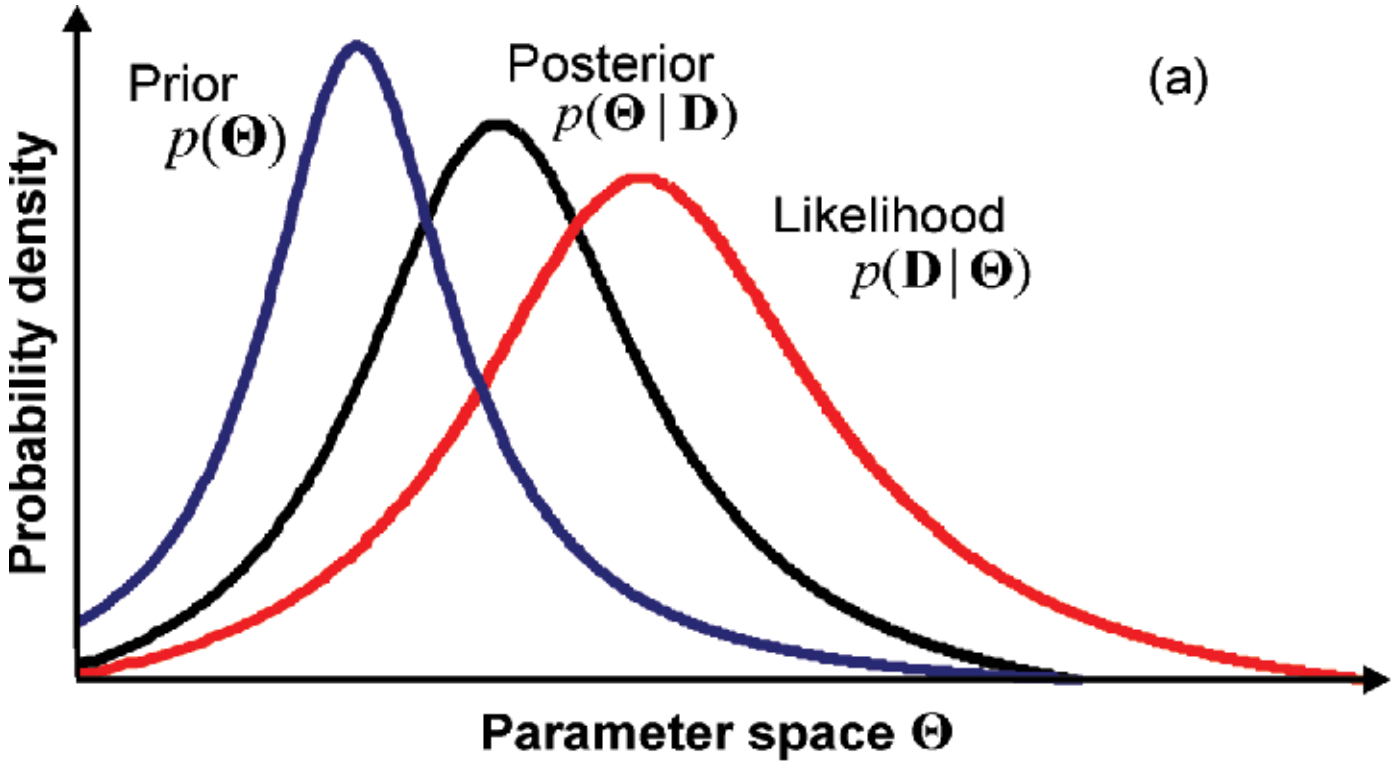


Figure: When we have strong prior beliefs and/or few/imprecise measurements, the prior is narrow compared to the likelihood, and the posterior differs significantly from the likelihood. Source: Xiang & Fackler (2015)

When is it ok to use the likelihood?

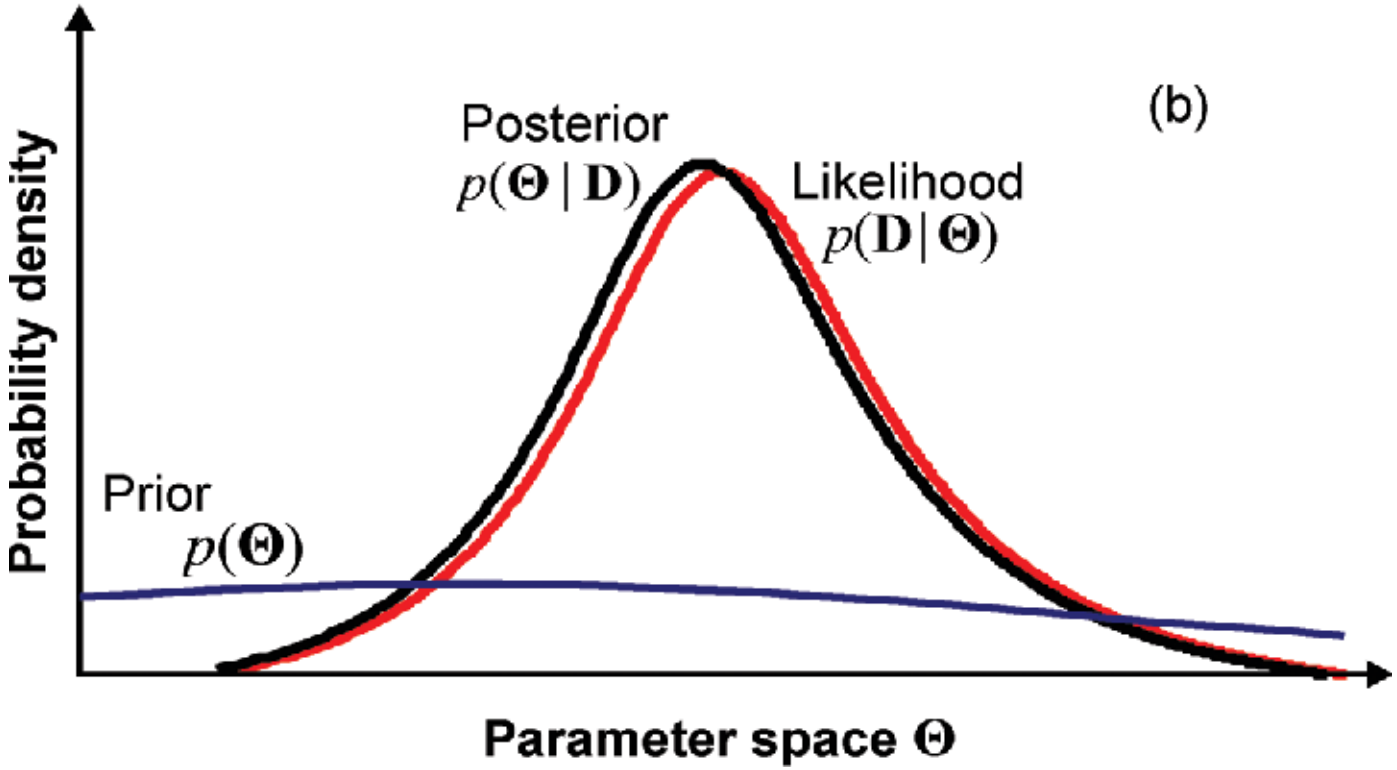


Figure: When we have limited prior knowledge and/or many precise measurements, the prior is broad compared to the likelihood, and the (normalised) likelihood is a good approximation to the posterior. Source: Xiang & Fackler (2015)

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.

Combining observations

We almost never measure anything just once. We generally work with a sample of observations $\hat{\mathbf{x}} = \{\hat{x}_i; i = 1, \ldots, N\}$.

Say we know how to write down a likelihood $p(\hat{x}_i|\theta)$ for one observation. How do we write a likelihood for the whole sample?

For this we need a way to relate the ***joint probability distribution*** $p(\hat{\mathbf{x}})$ to the individual probabilities $p(\hat{x}_i)$.

This is simple enough if the observations are ***independent*** of each other.

Combining independent observations

If the observations in our dataset $\hat{\mathbf{x}} = \{\hat{x}_i, i = 1, \ldots, N\}$ are mutually independent,

$$p(\mathbf{x}) = \prod_{i=1}^{N} p(x_i)$$

Therefore, the likelihood for the dataset as a whole is the product of the individual likelihoods:

$$\mathcal{L}(\theta) = p(\hat{\mathbf{x}}|\theta) = \prod_{i=1}^{N} p(\hat{x}_i|\theta).$$

Taking the logarithm converts the product to a sum:

$$\ln \mathcal{L}(\theta) = \sum_{i=1}^{N} \ln p(\hat{x}_i|\theta).$$

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:

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

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).

Probability distribution for the mean of a Gaussian

The likelihood can be re-written as

$$\mathcal{L}(\mu) = C \times \exp\left[-\frac{(\langle\hat{x}\rangle - \mu)^2}{2\sigma^2/N}\right],$$

where C depends on σ and the $\hat{x}_i$ but not on μ.

If the prior is comparatively flat, we can approximate the posterior as

$$p(\mu|\hat{\mathbf{x}}) = \frac{\mathcal{L}(\mu)\,p(\mu)}{\int \mathcal{L}(\mu)\,p(\mu)\,\mathrm{d}\mu} \approx \frac{\mathcal{L}(\mu)}{\int \mathcal{L}(\mu)\,\mathrm{d}\mu}.$$

The probability distribution for μ is thus a Gaussian with mean $\langle\hat{x}\rangle$ and variance σ^2/N.

Credible and confidence intervals

We have re-derived the mean and variance estimates we already got in lecture 3. However, we now have an explicit probability distribution for μ. This allows us to do more than just evaluate its expectation value and variance.

Integrating the pdf for RV X over a given interval $[a; b]$ gives the probability that X lies in that interval.

Inverting that calculation allows us to compute intervals containing a specific fraction of the total probability mass.

When derived from the full posterior probability distribution for the model parameters, probability intervals are called ***credible intervals***. When derived from the likelihood only, they are called ***confidence intervals***.

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.

Gaussian probability intervals

Recall the Gaussian cdf

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

where $\mathrm{erf}(x)$ is the ***error function***:

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

It gives the probability that a normally distributed variable with mean 0 and variance $1/2$ lies in the interval $[-x; x]$.

Gaussian probability intervals

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

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

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

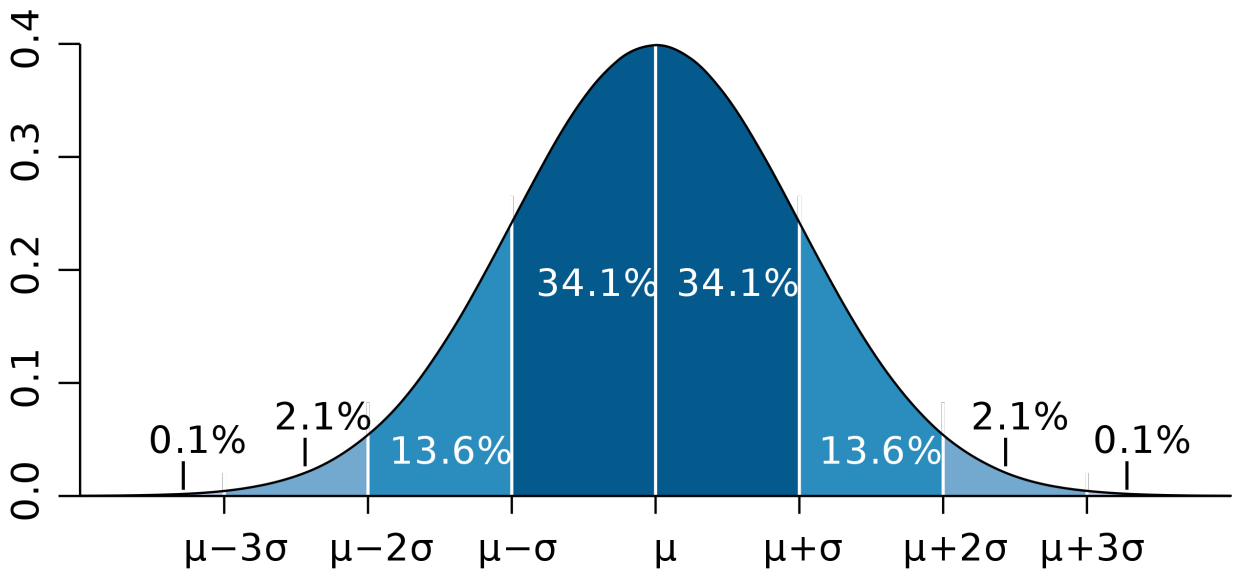


Figure: Credit: Ainali – Own work, CC BY-SA 3.0.

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%.
- ...

Outline

Lecture 1: Science, Probability, and Statistics

Lecture 2: Some Common Probability Distributions

Lecture 3: Expectation, Variance and Correlation

Lecture 4: Likelihood and Inference

Lecture 5: Testing Assumptions and Hypotheses

Lecture 6: Fitting more Complex Models

Lecture 7: Building in Prior Knowledge

Lecture 8: Questions & live exercise

Relaxing our assumptions

In lecture 4 we assumed independent, identically distributed Gaussian noise with known variance and an uninformative prior.

Let us relax some of these assumptions, one at a time.

First consider the assumption of known variance. When we treat σ as 'known' it's usually based on some prior calibration of the experimental apparatus, but such calibrations aren't perfect.

A useful sanity check for the assumed σ is that it should be approximately the same as the (unbiased) sample variance.

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

$$\hat{\epsilon}_i \sim \mathcal{N}(C, \sigma^2)$$

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 $\hat{x}_i = \mu + C + \hat{\epsilon}_i$ where $\hat{\epsilon}_i \sim \mathcal{N}(0, \sigma^2)$.

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

Note the ***systematic error does not reduce with more data***.

What if we don't know σ?

If we don't know σ, we can simply treat it as a parameter and fit for it.

Just as we did for μ, we can maximise $\mathcal{L}$ with respect to σ:

$$\begin{aligned} \ln \mathcal{L}(\sigma, \mu) &= -\tfrac{N}{2} \ln 2\pi - N \ln \sigma - \tfrac{1}{2} \textstyle\sum_{i=1}^{N} \left(\frac{\hat{x}_i - \mu}{\sigma} \right)^2 . \\ \frac{\partial \ln \mathcal{L}}{\partial \sigma} &= -\tfrac{N}{\sigma} - \tfrac{1}{\sigma^3} \textstyle\sum_{i=1}^{N} (\hat{x}_i - \mu)^2 . \end{aligned}$$

Plugging in the MLE for μ, setting the derivative to zero and solving for σ^2 gives:

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N} \sum_{i=1}^{N} (\hat{x}_i - \mu_{\mathrm{ML}})^2 = \frac{1}{N} \sum_{i=1}^{N} (\hat{x}_i - \langle \hat{x} \rangle)^2 = \sigma^2(\hat{x}).$$

MLE estimate of the variance of a Gaussian

The MLE for the variance of a Gaussian is the sample variance.

However, we know that the sample variance is biased. For small N, the unbiased sample variance, $\tilde{\sigma}^2(\hat{x}) = \sigma^2(\hat{x})N/(N-1)$, is a better approximation of the true variance than the MLE estimate.

Chi-squared distribution

To discuss the probability distribution of the sample variance, we first need to introduce the ***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):

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

Chi-squared distribution

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

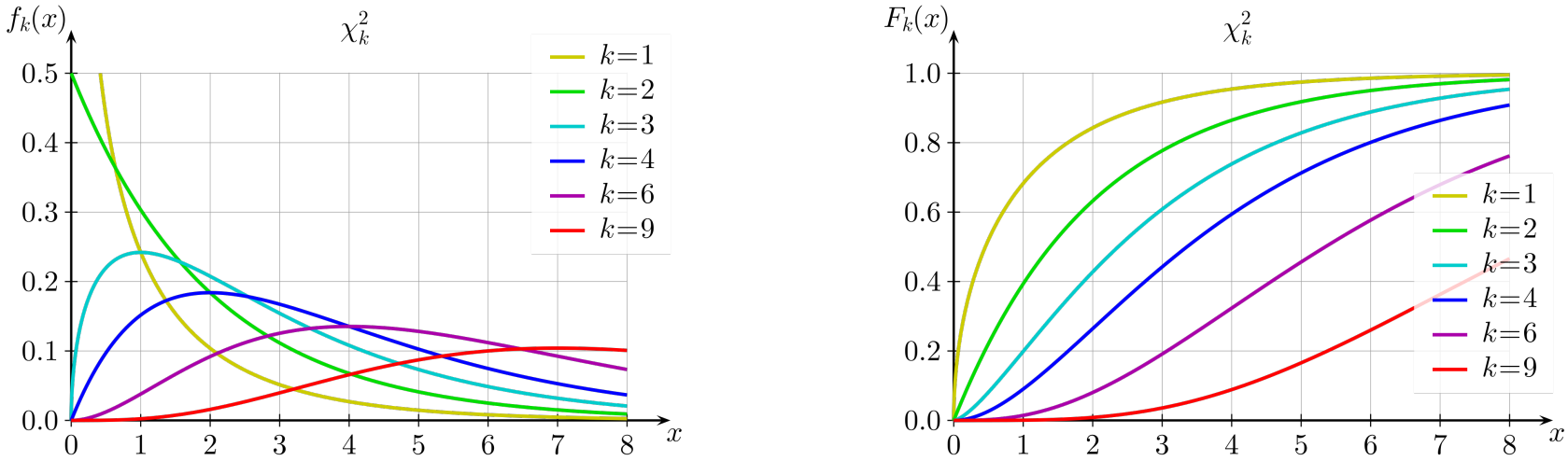


Figure: Pdf and cdf of the chi-squared distribution for different values of k. The expectation value is k and the variance is $2k$. For large k, it approaches a normal distribution. Source: Wikipedia.

Distribution of the sample variance

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

$$\hat{r}_i \equiv \frac{\hat{x}_i - \mu}{\sigma} \sim \mathcal{N}(0, 1),$$

and thus

$$S^2 \equiv \sum_{i=1}^{N} \hat{r}_i^2 = \frac{\sum_{i=1}^{N} (\hat{x}_i - \mu)^2}{\sigma^2} \sim \chi^2_N.$$

Distribution of the sample variance

Recall the definition of the unbiased sample variance then obeys

$$\tilde{\sigma}^2(\hat{x}) \equiv \frac{1}{N-1}\sum_{i=1}^{N}(\hat{x}_i - \langle\hat{x}\rangle)^2 \sim \frac{\sigma^2}{N-1}\,\chi^2_{N-1}.$$

Why $N-1$ d.o.f.'s? One d.o.f. is 'used' by replacing the true mean μ with the sample mean $\langle\hat{x}\rangle$. (A more formal proof involves Cochran's theorem.)

This distribution can be used to perform a sanity check in the case when you ***think*** you know the true variance σ...

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.

What if we have relevant prior information

If the prior is informative, the likelihood isn't a good approximation to the posterior anymore.

We have to evaluate the posterior pdf explicitly and it may not be Gaussian.

Probability intervals for general pdfs

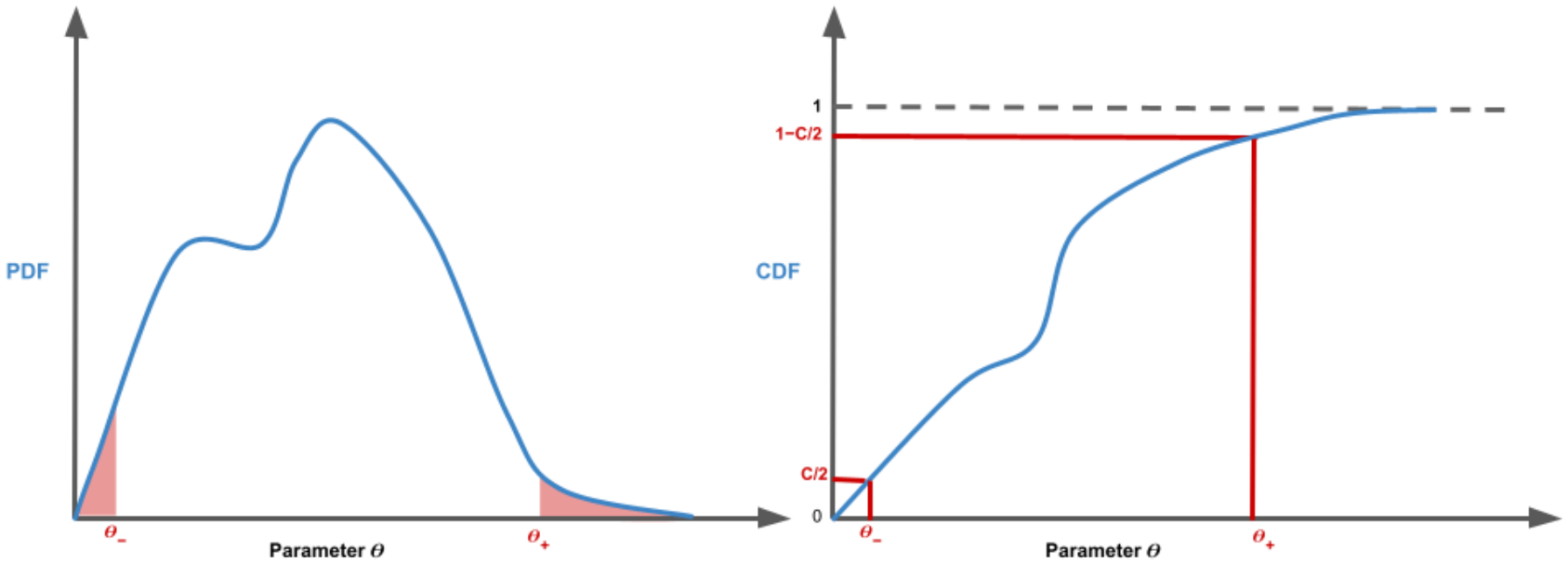


Figure: If the pdf is asymmetric, we can still define a confidence interval. Here, we define a two-sided interval by excluding the 'tails' of the distribution until $1 - C$ of the total area under the pdf is left. We could also have defined a symmetric interval about the median, expectation value (mean) 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

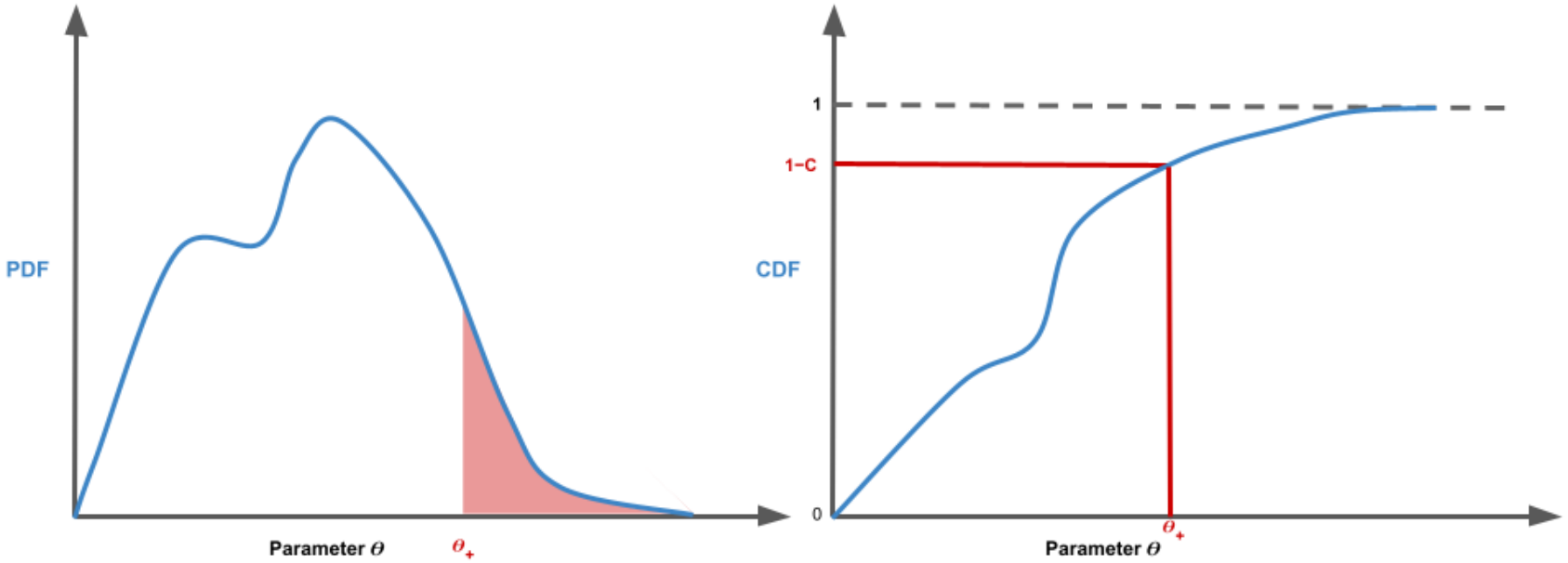


Figure: If we are interested in a lower or upper limit, we can evaluate a one-sided confidence interval (this applies to Gaussians too).

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

The probability of getting a $\geq 3.16\,\sigma$ departure from the mean of a Poisson distribution with rate $\lambda \approx 98.6$ from random statistical fluctuations alone is $p = 8 \times 10^{-4}$.

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 $p = 8 \times 10^{-4}$? Binomial!

$$\begin{aligned} P(r > 1|n,p) &= \sum_{r=1}^{\infty} \mathcal{B}(r|n,p) = 1 - \mathcal{B}(0|n,p) \\ &= 1 - \binom{n}{0} p^0 (1-p)^{n-0} = 1 - (1-p)^n = 0.25! \end{aligned}$$

This is known as the ***Bonferroni correction***.

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

Hypothesis testing: the null hypothesis

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 observation
- work out the pdf of this statistic
- integrate it to obtain a confidence interval

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 $\langle\hat{x}\rangle$.

Thanks to the CLT, we know that the quantity:

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

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

Can I reject the null hypothesis? Case 2: unknown variance

If the true variance is unknown, we can construct a t-***value***

$$t \equiv \frac{\langle \hat{x} \rangle - \mu}{\tilde{\sigma}(\hat{x})/\sqrt{N}}$$

where $\tilde{\sigma}^2(\hat{x})$ is the unbiased sample variance.

$$t = \langle \hat{r} \rangle \times \frac{\sigma}{\tilde{\sigma}(\hat{x})} = \langle \hat{r} \rangle \times \frac{1}{\sqrt{N-1}} \frac{\sigma}{\tilde{\sigma}(\hat{x})} \times \sqrt{N-1}.$$

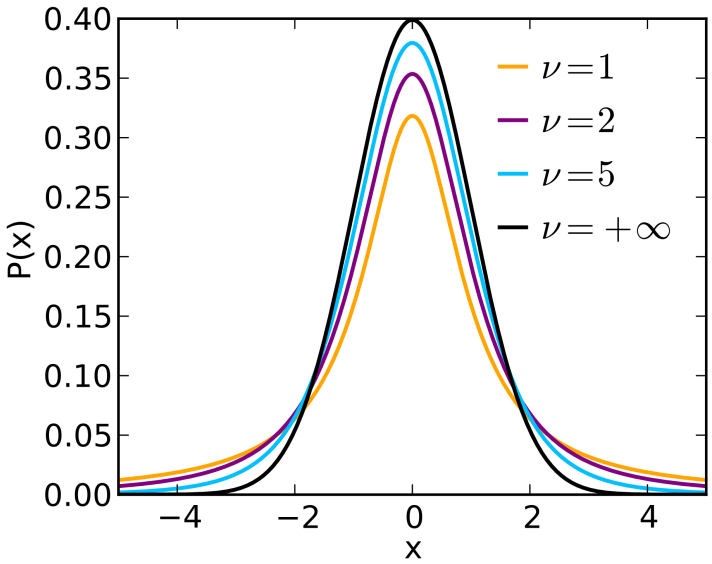
We know that $\langle \hat{r} \rangle$ follows a standard normal distribution, and that $(N-1)\,\tilde{\sigma}^2(\hat{x})/\sigma^2$ follows a chi-squared distribution with $N-1$ d.o.f.s.

The resulting distribution for t is known as a ***Student's*** t ***distribution*** with $N-1$ d.o.f.s:

$$t \sim \mathcal{T}_{N-1}.$$

Student's t distribution

$$p(x,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{2\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$



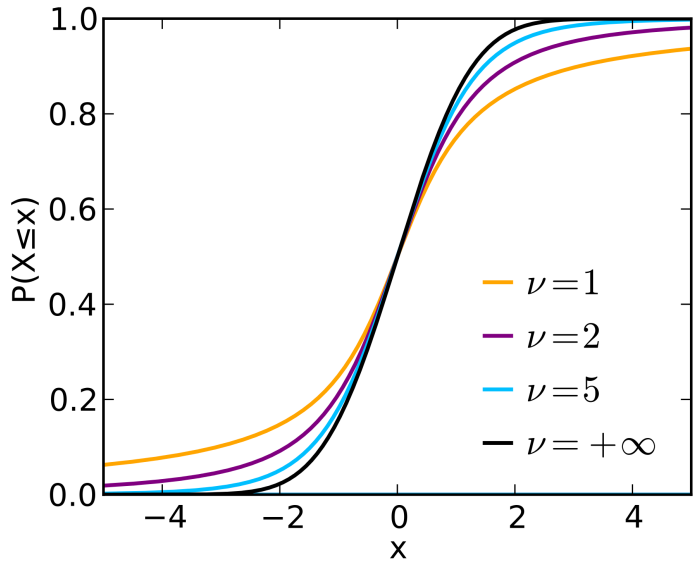


Figure: Pdf and cdf of the Student's t distribution for different values number of degrees of freedom ν. The expectation value is $\mathbb{E}[t] = 0$ (for $\nu > 1$) and the variance $\mathrm{Var}[t] = \nu/(\nu-2)$ (for $\nu > 2$). For large ν, the Student's t distribution approaches the standard normal distribution, but for low ν it has much heavier tails. Source: Wikipedia.

Can I reject the null hypothesis? Case 2: unknown variance

We obtain N measurements $\{\hat{x}_i\}$ of x, with unknown variance, and compute the sample mean $\langle\hat{x}\rangle$ and the unbiased sample variance $\tilde{\sigma}^2(\hat{x})$.

As we've just seen, the t-value:

$$\langle\hat{t}\rangle = \frac{\langle\hat{x}\rangle - \mu}{\tilde{\sigma}(\hat{x})/\sqrt{N}} \sim \mathcal{T}(N-1)$$

To evaluate confidence intervals for this situation we would use a Student's t distribution look-up table and calculator.

For example, $P(|\hat{t}| > 3) = 0.04$ for $N = 5$, 0.004 for $N = 50$ . . . Note $P(|\hat{t}| > 3) \gg P(|\hat{r}| > 3)$!

This procedure is called a one-sample Student's t-test.

Do two RVs share the same mean?

In an image from an astronomical sky survey, I have identified what looks like an overdensity of stars in a particular part of the sky. Have I discovered a new star cluster?

I can divide the image into a grid of equal-area cells, then count the number of stars in each cell.

Is the mean number of stars per cell larger for the cells which overlap with my hypothetical cluster than for the cells which do not?

The number of stars in each cell is drawn from a Poisson distribution, but the sample mean over a group of cells is normally distributed.

Do two RVs share the same mean?

We have two samples, containing N_1 and N_2 cells respectively, with respective means $\langle \hat{x}_1 \rangle$ and $\langle \hat{x}_2 \rangle$ and unbiased sampled variances $\tilde{\sigma}^2(\hat{x}_1)$ and $\tilde{\sigma}^2(\hat{x}_2)$.

We want to check if the two samples share the same true mean, but we do not know if they share the same true variance.

Defining $\Delta = \langle \hat{x}_1 \rangle - \langle \hat{x}_2 \rangle$, we can estimate is variance as

$$\mathrm{Var}\left[\Delta\right] = \mathrm{Var}\left[\langle \hat{x}_1 \rangle\right] + \mathrm{Var}\left[\langle \hat{x}_2 \rangle\right] = \frac{\tilde{\sigma}^2(\hat{x}_1)}{N_1} + \frac{\tilde{\sigma}^2(\hat{x}_2)}{N_2}.$$

Welch's t-test

It can be shown that $\Delta / \sqrt{\mathrm{Var}\left[\Delta\right]}$ approximately follows a Student's t distribution with

$$\frac{\left[\frac{\tilde{\sigma}^2(\hat{x}_1)}{N_1} + \frac{\tilde{\sigma}^2(\hat{x}_2)}{N_2}\right]^2}{\frac{\tilde{\sigma}^4(\hat{x}_1)}{N_1^2(N_1-1)^2} + \frac{\tilde{\sigma}^4(\hat{x}_2)}{N_2^2(N_2-1)^2}}$$

degrees of freedom.

This can be used to evaluate the significance of my overdensity. (NB: This is ***not*** how new star clusters are actually identified anymore!)

Simplifications exist if the variance of the two samples can be assumed to be the same.

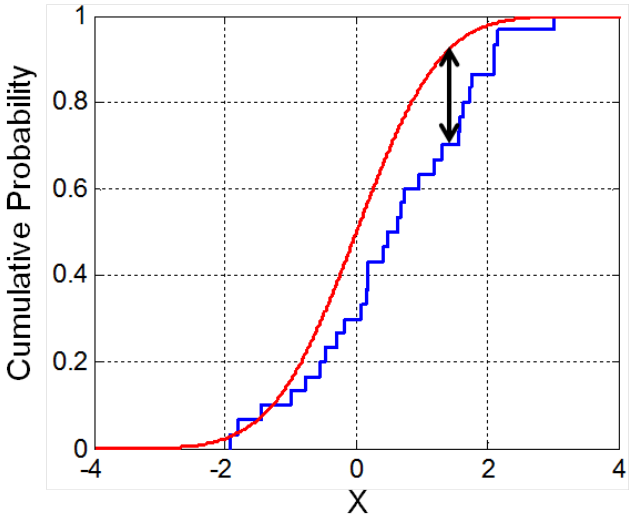
Is this RV drawn from that distribution?

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.

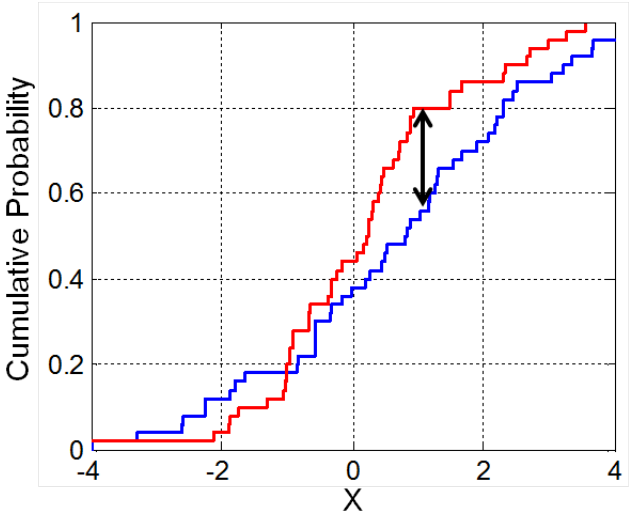
Do two RVs share the same mean?

Consider two datasets $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$ containing n and m observations, respectively. Are they drawn from the same distribution?

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

$$D_{n,m} = \max_{\hat{x}} |F_{1,n}(\hat{x}) - F_{2,m}(\hat{x})|.$$

where $F_{1,n}(\hat{x})$ and $F_{2,m}(\hat{x})$ are the ecdf's of the two samples.



Two-sample K-S test

For large samples, the *null hypothesis* that the two samples are drawn from the same distribution can be rejected with confidence α if

$$D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{n\,m}},$$

where

$$c(\alpha) = \sqrt{\frac{1}{2}\,\ln\left(\frac{\alpha}{2}\right)}.$$

Outline

Lecture 1: Science, Probability, and Statistics

Lecture 2: Some Common Probability Distributions

Lecture 3: Expectation, Variance and Correlation

Lecture 4: Likelihood and Inference

Lecture 5: Testing Assumptions and Hypotheses

Lecture 6: Fitting more Complex Models

Lecture 7: Building in Prior Knowledge

Lecture 8: Questions & live exercise

Fitting more complex models

In lecture 4 we looked at a model of the form

$$\hat{x}_i = \mu + \hat{\epsilon}_i$$

where $\hat{\epsilon}_i$ is i.i.d. Gaussian noise. This model has one free parameter μ.

Let us now consider a slightly more complex situation

$$\hat{y}_i = f(x_i, \boldsymbol{\theta}) + \hat{\epsilon}_i,$$

where the observations are now $\hat{y}_i$, taken at (known) coordinates x_i.

For example, we could have

$$f(x) = a\,x + b \quad \text{or} \quad f(x) = a\,e^{-bx},$$

both with parameters $\boldsymbol{\theta} = \{a, b\}$.

Fitting more complex models

As before, the MLEs for the parameters can be found by maximising the log-likelihood with respect to the parameters. We now have a system of M simultaneous equations with M unknowns (where M is the number of free parameters):

$$\begin{cases} \dfrac{\partial \ln \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_1} = 0 \\ \dfrac{\partial \ln \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_2} = 0 \\ \vdots \\ \dfrac{\partial \ln \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_M} = 0 \end{cases}$$

Maximum likelihood and least squares

The log likelihood in the case of independent, zero-mean Gaussian noise with variance σ_i^2 is:

$$\ln \mathcal{L}(\boldsymbol{\theta}) = -\frac{N}{2} \ln 2\pi - \sum_i \ln \sigma_i - \frac{\chi^2}{2},$$

where

$$\chi^2 \equiv \sum_{i=1}^{N} \frac{[\hat{y}_i - f(x_i, \boldsymbol{\theta})]^2}{\sigma_i^2}.$$

Maximising the likelihood is thus equivalent to minimizing the ***chi-squared*** χ^2. This is known as ***least squares*** fitting (because χ^2 the sum of the squared, normalised residuals).

Distribution of the χ^2

The χ^2 computed with the (unknown) true parameter values . . .

. . . is the sum of the squares of N draws from a standard normal distribution, and thus follows a chi-square distribution with N dofs.

The χ^2 computed with the parameter MLEs . . .

. . . follows a chi-square distribution with $N - M$ dofs, where M is the number of parameters we are fitting for.

Note that M is the number of 'physically meaningful' parameters, but does not include any 'nuisance' parameters.

Using χ^2 to check the goodness of fit

If the model is correct, the probability of Q obtaining a χ^2 equal to or larger than the best-fit value, due to statistical fluctuations in the noise, is

$$Q = 1 - F_{\chi^2}(\chi^2; N - M),$$

where $F_{\chi^2}(x; k)$ is the cdf for the chi-square distribution with k degrees of freedom.

As a rule of thumb,

- $Q > 0.1$ indicates a fairly good fit;
- $0.001 < Q < 0.1$ suggests the observations may be slightly non-normal, or their uncertainties under-estimated.
- $Q < 0.001$ indicates a poor fit, and suggests a different model should be used.

Linear basis models

Linear basis models can be written

$$f(x) = \sum_j w_j \, B_j(x),$$

where the B_j's are pre-determined functions of x (known as ***basis functions*** or, in the machine learning jargon, ***features***), and the w_j's are ***coefficents*** (or ***weights***).

Popular linear basis models include polynomials ($B_j(x) = x^j$), but also Fourier decompositions (sines and cosines), sigmoid functions, ...

For such models, the system of equations to solve to optimize the likelihood is linear, and an analytical solution exists.

For non-linear models, the MLE must be found numerically.

Example: fitting a straight line

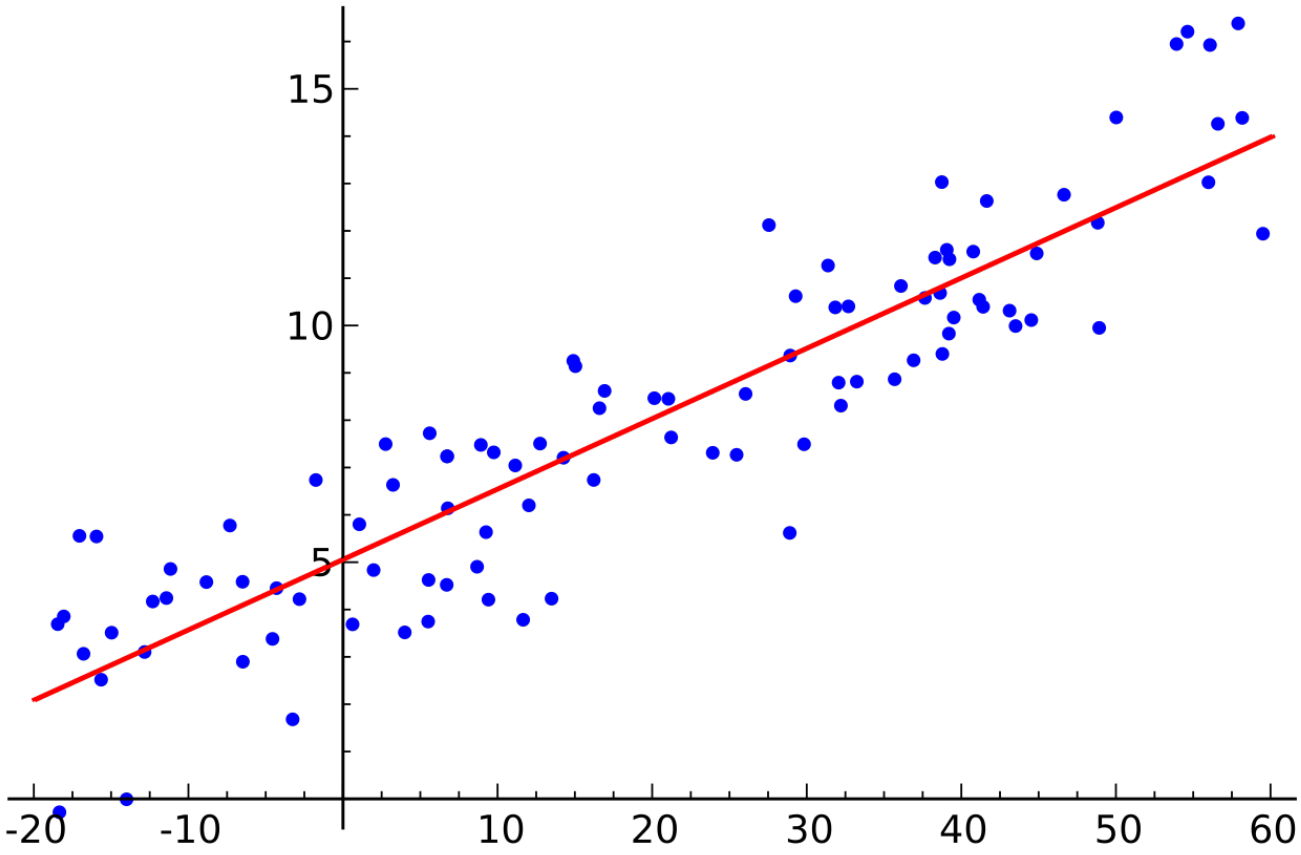


Figure: Simple linear regression example. All observations share the same variance. Source: Wikipedia.

Example: fitting a straight line

We want to fit the function $f(x) = a_1\,x + a_2$.
Start by writing down the chi-squared:

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{N} \frac{(\hat{y}_i - a_1\,x_i - a_2)^2}{\sigma_i^2} \\ &= \sum_{i=1}^{N} \frac{\hat{y}_i^2 + a_1^2\,x_i^2 + a_2^2 - 2(a_1\,x_i\,\hat{y}_i + a_2\,\hat{y}_i - a_1\,a_2\,x_i)}{\sigma_i^2}\end{aligned}$$

Taking derivatives with respect to a_1 and a_2:

$$\begin{aligned}\frac{\partial\chi^2}{\partial a_1} &= 2\sum_{i=1}^{N} \frac{a_1\,x_i^2 + a_2\,x_i - x_i\,\hat{y}_i}{\sigma_i^2} \\ \frac{\partial\chi^2}{\partial a_2} &= 2\sum_{i=1}^{N} \frac{a_1\,x_i + a_2 - \hat{y}_i}{\sigma_i^2}\end{aligned}$$

Fitting a straight line: system of equations

Set the derivatives to zero and solve for the parameter MLEs:

$$a_1 \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} + a_2 \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} = \sum_{i=1}^{N} \frac{x_i\,\hat{y}_i}{\sigma_i^2}$$

$$a_1 \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} + a_2 \sum_{i=1}^{N} \frac{1}{\sigma_i^2} = \sum_{i=1}^{N} \frac{\hat{y}_i}{\sigma_i^2}$$

These are known as the ***normal equations*** for the fit. We can re-write the them in matrix form

$$\begin{bmatrix} \sum_i \frac{x_i^2}{\sigma_i^2} & \sum_i \frac{x_i}{\sigma_i^2} \\ \sum_i \frac{x_i}{\sigma_i^2} & \sum_i \frac{1}{\sigma_i^2} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \sum_i \frac{x_i\,\hat{y}_i}{\sigma_i^2} \\ \sum_i \frac{\hat{y}_i}{\sigma_i^2} \end{bmatrix}$$

Linear basis models

As previously mentioned, polynomials are linear basis models:

$$f(x) = \sum_{j=1}^{M} a_j X_j(x) \quad \text{where} \quad X_j = x^{M-j}.$$

The X_j's are the ***basis functions*** functions of the model, and the a_js the coefficients to be fit for.

The straight line model we considered on the previous slide

corresponds to $M = 2$, with $X_1(x_i) = x_i$ and

$$X_2(x_i) = x_i^0 = 1.$$

Let us re-write the normal equations in a general form that would apply to any linear basis model before solving them.

General linear least squares

The matrix equation we previously derived can be re-written

$$\mathbf{A}^{\mathrm{T}}\mathbf{A}\,\mathbf{a} = \mathbf{A}^{\mathrm{T}}\mathbf{b}$$

where:

$\mathbf{A}$ is an $N \times M$ matrix with elements $A_{ij} = X_j(x_i)/\sigma_i$. $\mathbf{A}$ is known as the ***design matrix***;

$\mathbf{a}$ is an M-element vector of parameters a_j, containing the coefficients of each basis function;

$\mathbf{b}$ is an N-element vector with elements $b_i = \hat{y}_i/\sigma_i$.

Solving linear least squares

To solve for $\mathbf{a}$ we need to invert the $M \times M$ matrix $\mathbf{A}^T\mathbf{A}$:

$$\mathbf{a}_{\mathrm{ML}} = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\,\mathbf{A}^{\mathrm{T}}\mathbf{b}.$$

where

$$[\mathbf{A}^T\mathbf{A}]_{jk} = \sum_{i=1}^{N} \frac{X_j(x_i)\,X_k(x_i)}{\sigma_i^2}$$

and

$$[\mathbf{A}^T\mathbf{b}]_j = \sum_{i=1}^{N} \frac{X_j(x_i)\,\hat{y}_i}{\sigma_i^2}.$$

Linear least squares: solution

We can then write the general solution:

$$a_{j,\mathrm{ML}} = \mathbf{C}\,\mathbf{A}^{\mathrm{T}}\,\mathbf{b} = \sum_{k=1}^{M} C_{jk}\left[\mathbf{A}^{\mathrm{T}}\,\mathbf{b}\right]_{ki} = \sum_{k=1}^{M} C_{jk}\left[\sum_{i=1}^{N}\frac{X_k(x_i)\,\hat{y}_i}{\sigma_i^2}\right].$$

where we have defined

$$\mathbf{C} \equiv (\mathbf{A}^{\mathrm{T}}\,\mathbf{A})^{-1}.$$

Solution for straight line fit

In the straight line case, the solution can be written:

$$\begin{aligned} a_{1,\mathrm{ML}} &= (S\,S_{xy} - S_x\,S_y)\,/\Delta, \\ a_{2,\mathrm{ML}} &= (S_{xx}\,S_y - S_x\,S_y)\,/\Delta, \end{aligned}$$

where

$$S \equiv \sum_{i=1}^{N} \frac{1}{\sigma_i^2}, \qquad S_x \equiv \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2}, \qquad S_{xx} \equiv \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2},$$

$$S_y \equiv \sum_{i=1}^{N} \frac{\hat{y}_i}{\sigma_i^2}, \qquad S_{xy} \equiv \sum_{i=1}^{N} \frac{x_i\,\hat{y}_i}{\sigma_i^2},$$

and

$$\Delta \equiv S\,S_{xx} - (S_x)^2.$$

Linear least squares: parameter uncertainties

The parameter variances are then:

$$\begin{aligned}
\sigma^2(a_{j,\mathrm{ML}}) &= \sum_{i=1}^{N} \sigma_i^2 \left(\frac{\partial a_{j,\mathrm{ML}}}{\partial \hat{y}_i} \right)^2 = \sum_{i=1}^{N} \sigma_i^2 \left(\sum_{k=1}^{M} \frac{C_{jk} X_k(x_i)}{\sigma_i^2} \right)^2 \\
&= \sum_{k=1}^{M} C_{jk} \sum_{l=1}^{M} C_{jl} \sum_{i=1}^{N} \frac{X_k(x_i)\, X_l(x_i)}{\sigma_i^2} \\
&= \sum_{k=1}^{M} C_{jk} (\mathbf{C}\, \mathbf{C}^{-1}) = C_{jj}.
\end{aligned}$$

The diagonal elements of $\mathbf{C}$ are the parameter variances (squared uncertainties).

Parameter uncertainties for straight line fit

In the straight line case, the partial derivatives can be written

$$\begin{aligned} \frac{\partial a_{1,\mathrm{ML}}}{\partial \hat{y}_i} &= \frac{1}{\sigma_i^2}\,\frac{S\,x_i - S_x}{\Delta} \\ \frac{\partial a_{2,\mathrm{ML}}}{\partial \hat{y}_i} &= \frac{1}{\sigma_i^2}\,\frac{S_{xx} - x_i\,S_x}{\Delta} \end{aligned}$$

and thus the parameter variances are

$$\sigma^2(a_{1,\mathrm{ML}}) = \frac{S}{\Delta} \quad \text{and} \quad \sigma^2(a_{2,\mathrm{ML}}) = \frac{S_{xx}}{\Delta}.$$

Linear least squares: parameter covariance matrix

Just as the diagonal elements of $\mathbf{C}$ give the parameter variances, the off-diagonal elements give the covariances between pairs of parameters.

$$\mathrm{Cov}\left[a_{j,\mathrm{ML}}, a_{k,\mathrm{ML}}\right] = C_{jk} \quad \text{where} \quad [\mathbf{C}^{-1}]_{jk} = \left[\sum_{i=1}^{N} \frac{X_j(x_i)\, X_k(x_i)}{\sigma_i^2}\right]$$

In the straight line case, this reduces to:

$$\mathrm{Cov}\left[a_{1,\mathrm{ML}}, a_{2,\mathrm{ML}}\right] = -\frac{S_x}{\Delta}.$$

Linear least squares: parameter distributions

In a linear basis model, the parameter MLEs are linear combinations of the observations. Therefore, if the observations are normally distributed with known variances, the parameters are also normally distributed.

Even if the observations are non-normal, for large N the CLT ensures that the parameter MLEs are approximately normally distributed.

The joint distribution over the parameter MLEs is then 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 $|\mathbf{C}| \geq 0$).

Confidence intervals in > 1-D

At the end of the last lecture we saw that the joint probability distribution over the coefficients of a linear basis model is a multi-variate Gaussian. How can we construct confidence intervals from this?

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

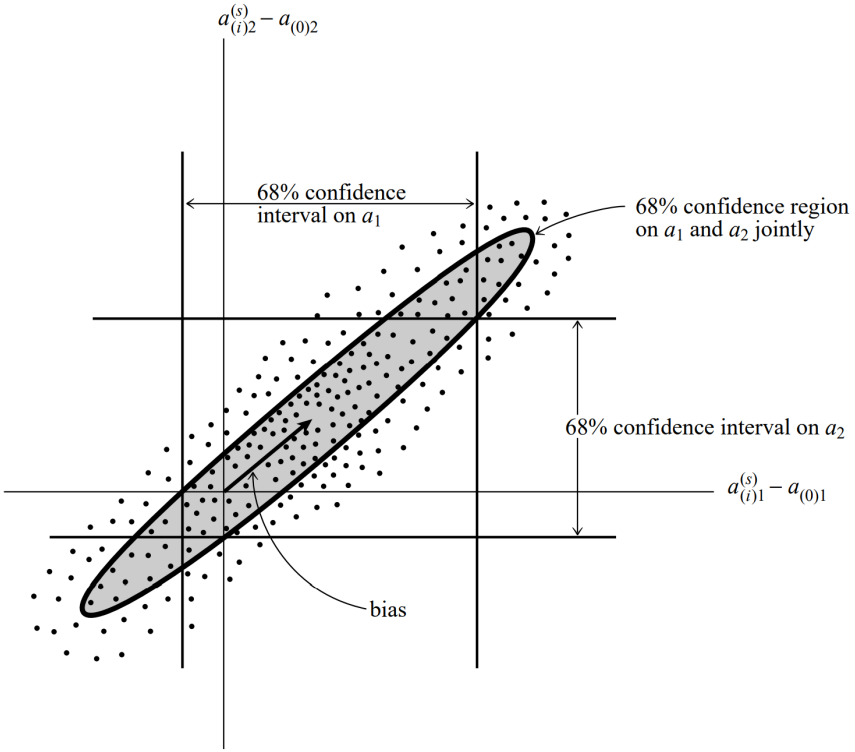


Figure: Confidence intervals in 1 and 2 dimensions. The same fraction of measured points (here 68%) lies (i) between the two vertical lines, (ii) between the two horizontal lines, (iii) within the ellipse. Source: *Numerical Recipes: The Art of Scientific Computing*, Press et al.

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 cdf to get the pdf:

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

Plugging in $R = \sigma$, for example, give s $P(R = \sigma) = 0.393$, 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 $R_C = \sigma\sqrt{-2\ln(1 - C)}$.

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

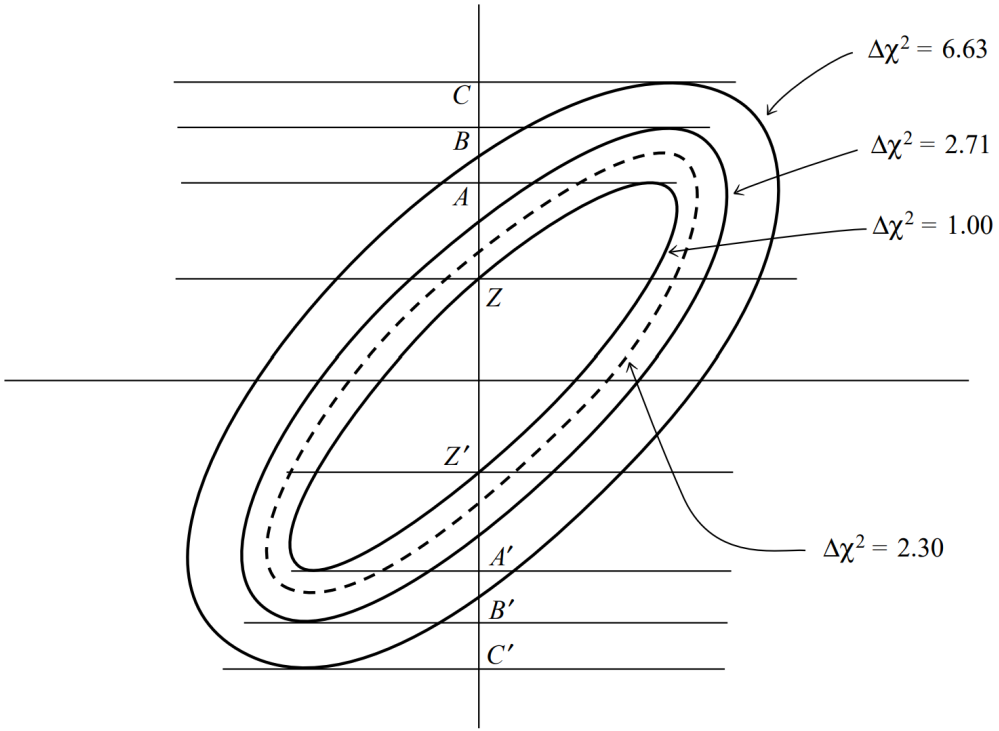


Figure: Confidence region ellipses corresponding to values of χ^2 larger than the fitted minimum. The solid curves, with $\Delta\chi^2 = 1.00, 2.71, 6.63$ project onto one-dimensional intervals AA', BB', CC'. These intervals – not the ellipses themselves – contain 68.3%, 90%, and 99% of normally distributed 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.

Outline

Lecture 1: Science, Probability, and Statistics

Lecture 2: Some Common Probability Distributions

Lecture 3: Expectation, Variance and Correlation

Lecture 4: Likelihood and Inference

Lecture 5: Testing Assumptions and Hypotheses

Lecture 6: Fitting more Complex Models

Lecture 7: Building in Prior Knowledge

Lecture 8: Questions & live exercise

Least squares with unknown observational uncertainties

If we do not know the σ_i, we can estimate a global data uncertainty σ after performing a least-squares fit.

We first find the parameter MLEs by setting all the σ_i to unity.

We can then compute an (unbiased) estimate of the sample variance of the residuals:

$$\sigma^2(\hat{y}_i|f,\boldsymbol{\theta}_{\mathrm{ML}}) = \frac{1}{N-M}\sum_{i=1}^{N}[\hat{y}_i - f(x_i,\boldsymbol{\theta}_{\mathrm{ML}})]^2$$

This precludes using the χ^2 to check the goodness of fit, and confidence intervals must be estimated using a Student's t test.

Non-linear models

If the model is non-linear in the parameters, the system of equations to solve to minimize χ^2 or maximise $\mathcal{L}$ is also nonlinear and in general we can't solve it analytically.

Sometimes we can't even write down the derivatives of the χ^2 with respect to the parameters (or we are too lazy to do so).

In that case we can start with an initial guess for the parameters and try to improve the χ^2 iteratively, using numerical or analytical partial derivatives, until convergence is reached. This is known as ***local optimization*** and there are many algorithms for doing it.

However there can be local minima. To mitigate against those, try different starting points or use a global optimization scheme.

Non-linear least squares: parameter covariance matrix

If the model is non-linear, the joint and marginal distributions over the parameters are not Gaussian.

However, any non-linear function can be approximated locally by a linear one. Doing this near the MLE parameter set is equivalent to making a Gaussian approximation to the likelihood (***Laplace approximation***), or equivalently a quadratic approximation to the χ^2.

The local curvature of the χ^2 is given by the ***Hessian matrix*** $\mathbf{H}$:

$$H_{kl} = 2\sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left[\frac{\partial f(x_i, \boldsymbol{\theta})}{\partial \theta_k} \frac{\partial f(x_i, \boldsymbol{\theta})}{\partial \theta_k} \right]$$

The inverse of the Hessian matrix is the covariance matrix over the parameters $\mathbf{C} = \mathbf{H}^{-1}$. (You can check for yourself that this expression for $\mathbf{C}$ reduces to the one we derived earlier in the linear case.)

Limitations of the Laplace approximation

Sometimes we have a multi-modal likelihood, or strong non-linear correlations between parameters (banana-shaped contours).

Even if the likelihood is unimodal and approximately Gaussian, When high confidence levels are required, we have to go further from the optimum, where the Gaussian approximation deteriorates.

In those cases, using the Hessian matrix to estimate the parameter (co)variances can give rise to misleading confidence intervals or regions.

In such situations, we must vary all the parameters ***jointly*** to map out the shape of the likelihood (or posterior). Doing this on a grid becomes impractical as soon as we have more than 2 or 3 parameters. A popular alternative is Markov Chain Monte Carlo (MCMC).

Observational variance as a free parameter

If the observational uncertainties are unknown, they can be fit for alongside the other parameters when maximising the log likelihood:

$$\begin{aligned} \mathcal{L} &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left(-\frac{\chi^2}{2}\right) \\ \ln\mathcal{L} &= -0.5N\ln 2\pi - N\ln\sigma - 0.5\chi^2 \end{aligned}$$

As a free parameter (σ) now appears outside the χ^2 term, maximising the likelihood is no longer equivalent to minimizing the χ^2.

Even if the fitted function $f(x)$ was linear in the parameters, the model now includes σ as a free parameter, and therefore the model as a whole is non-linear.

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.

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

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

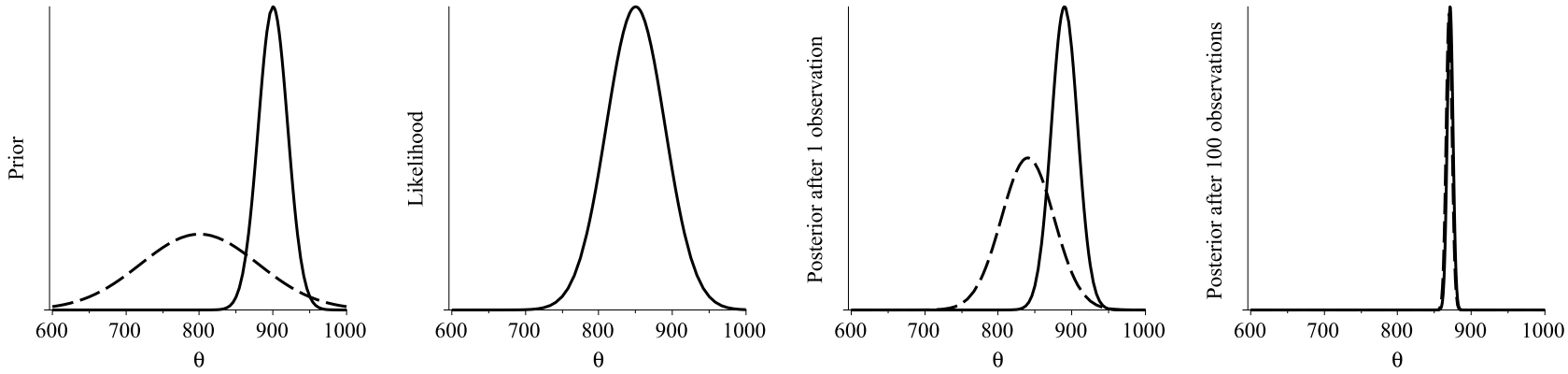


Figure: Two scientists with different prior beliefs $p(\theta)$ about the value of a quantity θ (shown by the two different curves in the first panel) observe one datum with likelihood $\mathcal{L}(\theta)$ (second panel), after which their posteriors $p(\theta|\hat{x})$ (third panel) represent their 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.

Log uniform prior

For a quantity that must always be positive (such as a standard deviation or frequency, for example), a uniform distribution in log space is often appropriate.

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

An advantage of this prior is that it is invariant under a change of coordinates, for example

- standard deviation $\sigma \longrightarrow$ precision $\beta \equiv 1/\sigma$
- frequency $\nu \longrightarrow$ period $T = 1/\nu$

This property makes it a Jeffreys prior.

Uninformative prior

Sometimes we need to write down a prior that embodies the weakest possible assumptions about our parameter.

Claude Shannon sought to define a quantity that measures the amount of new information we gain when measuring some RV X.

He defined this ***self-information*** so as to satisfy the following 3 axioms:

- An event with probability 100% is perfectly unsurprising and yields no information.
- The less probable an event is, the more surprising it is and the more information it yields.
- If two independent events are measured separately, the total amount of information is the sum of the self-informations of the individual events.

Shannon information

Given an RV X with probability mass (or density) function $p_X(x)$, the ***self-information*** of measuring X as outcome x is defined as

$$I_X(x) \equiv -\log_b[p_X(x)].$$

The choice of base b defines the unit of information: $b = 2$ gives a *binary digit* or ***bit***, $b = e$ a ***nat*** and $b = 10$ a ***hartley***.

If two variables X and Y are independent, their ***joint information*** is:

$$\begin{aligned} I_{X,Y}(x,y) &= -\log_b[p_{X,Y}(x,y)] = -\log_b[p_X(x)\,p_Y(y)] \\ &= I_X(x) + I_Y(y). \end{aligned}$$

Information theory underpins not just statistical mechanics and data analysis but a huge range of topics, e.g data compression, quantum computing, linguistics, bioinformatics...

Shannon entropy

A closely related concept is that of ***Shannon entropy***, which quantifies the average level of "information", "surprise", or "uncertainty" inherent to the variable's possible outcomes.

Given a discrete RV X with *sample space* Ω distributed according to $p_X(x)$, the entropy is

$$H(X) \equiv -\sum_{x \in \Omega} p_X(x) \log_b[p_X(x)] = \mathbb{E}\left[I(x)\right]$$

Compare to the definition of ***Gibbs entropy*** in statistical mechanics:

$$S = -k_{\mathrm{B}} \sum_i p_i \ln p_i.$$

Uninformative priors

The least informative prior we can chose for a given parameter is the probability distribution with the maximum entropy, subject to whatever other constraints we have, including the sample space (or support) of the distribution.

For a parameter with a fixed mean and variance, the Gaussian is the maximum entropy distribution.

When we don't know the mean and variance but we know that they exist, we can let the variance of the prior tend to infinity, recovering an (improper) uniform prior.

Bayesian model comparison

You should ***never*** compare two models with different numbers of parameters using the maximum likelihood or χ^2. Adding more parameters ***always*** reduces the χ^2, but it doesn't mean the model is better.

To compare different models, or hypotheses, we need to evaluate the marginal likelihood, or evidence, for each of them

$$\mathcal{Z}_M = p(\hat{x}|M) = \int_\Omega p(\hat{x}|\theta)\, p(\theta)\, \mathrm{d}\theta.$$

This enables us to compute the Bayes factor $\mathcal{B}_{21} = \mathcal{Z}_2/\mathcal{Z}_1$ which represents the ratio of the likelihoods for models 1 and 2 integrated over the full parameter space of each model.

Evaluating and interpreting the Bayes factor

To evaluate $\mathcal{Z}$, in general we must ***sample*** $p(\hat{x}|\theta)\,p(\theta)$ at numerous locations in the multi-dimensional parameter space and then numerically integrate it. This can be done using e.g. nested sampling but it is computationally expensive.

If $\mathcal{B}_{21} > 1$, then model 2 is preferred over model 1.

However, $\mathcal{B}_{21}$ can be extremely sensitive to noise fluctuations and the details of the numerical evaluation procedure, so we set the bar high when using it for hypothesis testing.

Typically, $\mathcal{B}_{21} \leq 10^3$ is considered barely worth mentioning, and $\mathcal{B}_{21} \geq 10^{10}$ is required for one model to be strongly supported over another.

Bayesian information criterion

A 'quick and dirty' estimate of $\ln \mathcal{Z}$ (up to a constant) can be derived by using Laplace's method to integrate the posterior.

This leads to the ***Bayesian Information Criterion*** (BIC):

$$\mathrm{BIC} = M \ln N - 2 \ln \mathcal{L}_{\max}$$

where M is the number of parameters, N is the number of observations, and $\mathcal{L}_{\max}$ is the maximum value of the likelihood.

The first term in the BIC penalises more complex models with more parameters: it acts as Occam's razor.

If the observational uncertainties are known, the BIC reduces to $\mathrm{BIC} = M \ln N + \chi^2_{\min}$.

Bayesian information criterion: assumptions and limitations

The BIC derivation assumes that

- The likelihood can be approximated with a Gaussian near its peak
- The prior is approximately linear near the peak of the likelihood
- $N \gg M$

Using the BIC when these assumptions are not valid (especially the last one) is dangerous.

Support for the "correct" model increases with more data

Say we are comparing models 1 & 2, which have M_1 and M_2 parameters, respectively, with $M_2 > M_1$. Let's suppose that model 2 is correct and model 1 is not.

Model 2 is preferred over model 1 if

$$\ln \mathcal{B}_{21} = (\Delta \ln \mathcal{Z})_{21} = \ln \mathcal{Z}_2 - \ln \mathcal{Z}_1 > 0.$$

Using the BIC approximation, model 2 is preferred over model 1 if

$\Delta\mathrm{BIC}_{12} = \mathrm{BIC}_1 - \mathrm{BIC}_2 > 0$, where we have used the fact that $\mathrm{BIC} \approx \mathrm{constant} - \ln \mathcal{Z}$.

In the least-squares case

$\mathrm{BIC} = M \ln N + \chi^2_{\mathrm{min}}$, so $\Delta\mathrm{BIC}_{12} = (M_1 - M_2) \ln N + (\chi^2_{1,\mathrm{min}} - \chi^2_{2,\mathrm{min}})$.

Support for the "correct" model increases with more data

If model 2 is correct, $\chi^2_{2,\min}$ follows a chi-square distribution with $N - M_2$ dofs, and $\mathbb{E}\left[\chi^2_{2,\min}\right] = N - M_2$.

If model 1 is incorrect, it should give a poorer fit, even accounting for the difference in the number of parameters $\mathbb{E}\left[\chi^2_{1,\min}\right] > N - M_1$.

Let's say that $\mathbb{E}\left[\chi^2_{1,\min}\right] = \alpha(N - M_1)$ where $\alpha > 1$.

Therefore:

$$\begin{aligned}\mathbb{E}\left[\Delta\mathrm{BIC}_{12}\right] &= (M_1 - M_2)\ln N + \mathbb{E}\left[\chi^2_{1,\min}\right] - \mathbb{E}\left[\chi^2_{2,\min}\right] \\ &= (M_1 - M_2)\ln N + (\alpha - 1)N + M_2 - \alpha M_1\end{aligned}$$

As $M_2 > M_1$, the first term scales as $-\ln N$, but second scales as $+N$, so $\mathbb{E}\left[\Delta\mathrm{BIC}_{12}\right] \propto N$ for large N.

Outline

Lecture 1: Science, Probability, and Statistics

Lecture 2: Some Common Probability Distributions

Lecture 3: Expectation, Variance and Correlation

Lecture 4: Likelihood and Inference

Lecture 5: Testing Assumptions and Hypotheses

Lecture 6: Fitting more Complex Models

Lecture 7: Building in Prior Knowledge

Lecture 8: Questions & live exercise

Live exercise: fitting a straight line to data

1. Go to colab.research.google.com and open a new notebook.
2. Simulate a dataset containing $N = 20$ data points with independent identically distributed Gaussian noise plus a linear trend.
3. Fit the data with a straight line. How good is the fit?
4. Compare the MLE values of the parameters (slope and intercept) to the true values.
5. Evaluate a 95% confidence interval for each of the parameters.
6. ***[Harder!]*** Plot a 95% joint confidence region for the parameters.
7. Fit the data with a second order polynomial. Is the fit better? Which model is preferred?

The end

Thank You!