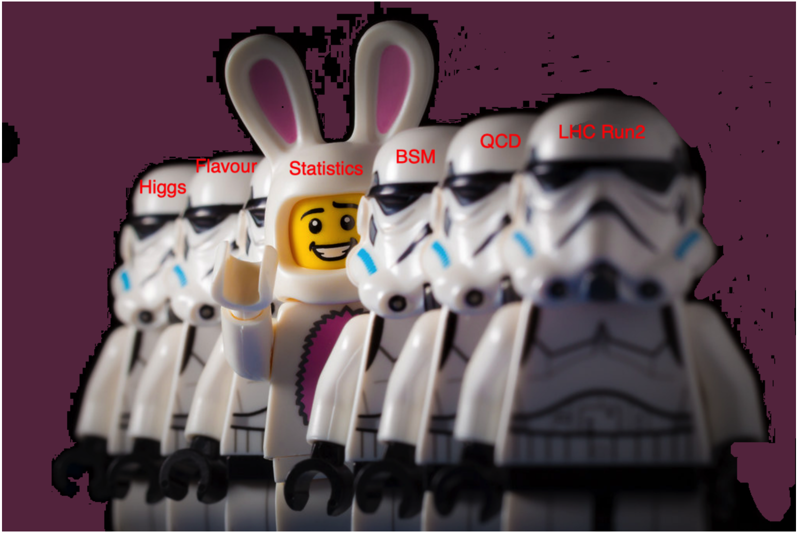
Statistics for Particle Physics
Lecture 1: From Poissons to p-values

Roger Barlow
Huddersfield University

ISPAD-2022

14th March 2022

Why Statistics among these high-power lectures?



Why Statistics among these high-power lectures?

What is probability P_A of A?

Four possible answers:

- P_A is number obeying certain mathematical rules.
- P_A is a real property of A that determines how often A happens
- For N trials in which A occurs N_A times, P_A is the limit of the frequency N_A/N for large N
- P_A is my subjective belief that A will happen, measurable by seeing what odds I will accept in a bet.

The frequentist and subjective (or Bayesian) uses are both useful and need a closer look...

Mathematical

Kolmogorov Axioms:

A. N. Kolmogorov

For all $A \subset S$
$P_A \geq 0$
$P_S = 1$
$P_{(A \cup B)} = P_A + P_B$ if $A \cap B = \phi$ and $A, B \subset S$

From these simple axioms a complete and complicated structure can be erected. E.g. show $P_{\overline{A}} = 1 - P_A$, and show $P_A \leq 1$....

But!!!

This says *nothing* about what P_A actually means.

Kolmogorov had frequentist probability in mind, but these axioms apply to any definition.

Classical

or Real probability

Evolved during the 18th-19th century

Developed (Pascal, Laplace and others) to serve the gambling industry.

Two sides to a coin - probability $\frac{1}{2}$ for each face

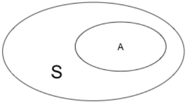
Likewise 52 cards in a pack, 6 sides to a dice...

Answers questions like 'What is the probability of rolling more than 10 with 2 dice?'

Problem: cannot be applied to continuous variables. Symmetry gives different answers working with θ or $sin\theta$ or $cos\theta$. Bertan's paradoxes.

Frequentist

The usual definition taught in schools and undergrad classes



$P_A = \lim_{N\to\infty} \frac{N_A}{N}$
N is the total number of events in the ensemble (or collective)

The probability of a coin landing heads up is $\frac{1}{2}$ because if you toss a coin 1000 times, one side will come down $\sim$ 500 times.

The lifetime of a muon is $2.2\mu s$ because if you take 1000 muons and wait $2.2\mu s$, then $\sim$ 368 will remain.
The probability that a DM candidate will be found in your detector is *[insert value]* because of 1,000,000 (simulated) DM candidates *[insert value × 1,000,000]* passed the selection cuts

Important

P_A is not just a property of A, but a joint property of A and the ensemble.

Problems (?) for Frequentist Probability. 1/2

More than one ensemble

German life insurance companies pay out on 0.4% of 40 year old male clients. Your friend Hans is 40 today. What is the probability that he will survive to see his 41st birthday?
99.6% is an answer (if he's insured)
But he is also a non-smoker and non-drinker - so maybe 99.8%?
He drives a Harley-Davidson - maybe 99.0%?
All these numbers are acceptable

What is the probability that a K^+ will be recognised by your PID?
Simulating lots of K^+ mesons you can count to get $P = N_{acc}/N_{tot}$
These can be divided by kaon energy, kaon angle, event complexity... and will give different probabilities ... All correct.

Problems (?) for Frequentist Probability. 2/2

There may be no Ensemble

What is the probability that it will rain tomorrow?
There is only one tomorrow. It will either rain or not. P_{rain} is either 0 or 1 and we won't know which until tomorrow gets here
Suppose the forecast predicts rain, and records show the forecast is correct 80% of the time.

Bad Statement

There is an 80% probability of rain tomorrow

Good Statement

The statement 'It will rain tomorrow' has an 80% chance of being true

We say 'It will rain tomorrow' with 80% confidence.

More relevant example

What is the probability that there is a SUSY particle below 2 TeV?

Bayes' theorem

Bayes' Theorem applies (and is useful) in any probability model

Conditional Probability: $P(A|B)$: probability for A, given that B is true.
Example: $P(\spadesuit A) = \frac{1}{52}$ and $P(\spadesuit A|Black) = \frac{1}{26}$

Theorem

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

Proof.

The probability that A and B are both true can be written in two ways

$$P(A|B) \times P(B) = P(A\&B) = P(B|A) \times P(A)$$

Throw away middle term and divide by $P(B)$

□

Bayes' theorem

Examples

Example

$$P(\spadesuit A|Black) = \frac{P(Black|\spadesuit A)}{P(Black)} P(\spadesuit A) = \frac{1}{\frac{1}{2}} \times \frac{1}{52} = \frac{1}{26}$$

Example

Example: In a beam which is 90% π, 10% K, kaons have 95% probability of giving no Cherenkov signal; pions have 5% probability of giving none. What is the probability that a particle that gave no signal is a K?

$$P(K|no\ signal) = \frac{P(no\ signal|K)}{P(no\ signal)} \times P(K) = \frac{0.95}{0.95 \times 0.1 + 0.05 \times 0.9} \times 0.1 = 0.68$$

This uses the (often handy) breakdown:

$$P(B) = P(B|A) \times P(A) + P(B|\overline{A}) \times (1 - P(A))$$

Bayesian Probability

Probability P_A expresses your belief in A.
1 means certainty, 0 means total disbelief

Intermediate values can be calibrated by asking whether you would prefer to bet on A, or on a white ball being drawn from an urn containing a mix of white and black balls.

This avoids the limitations of frequentist probability - coins, dice, kaons, rain tomorrow, existence of SUSY can all have probabilities.

Bayesian Probability and Bayes Theorem

Re-write Bayes' theorem as

$$P(Theory|Data) = \frac{P(Data|Theory)}{P(Data)} \times P(Theory)$$

$$Posterior \propto Likelihood \times Prior$$

Works sensibly

Data predicted by theory boosts belief - moderated by probability it could happen anyway

Can be chained.

Posterior from first experiment can be prior for second experiment. And so on. (Order doesn't matter)

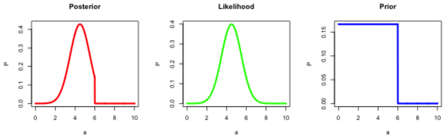
From Prior Probability to Prior Distribution

Suppose theory contains parameter a: (mass, coupling, decay rate...)

Prior probability distribution $P_0(a)$

$\int_{a_1}^{a_2} P_0(a)\,da$ is your prior belief that a lies between a_1 and a_2

$\int_{-\infty}^{\infty} P_0(a)\,da = 1$ (or: your prior belief that the theory is correct)
Simple number $P(data|theory) \rightarrow$ Likelihood function $L(x|a)$
Bayes' Theorem given data x the posterior is : $P_1(a) \propto L(x|a)P_0(a)$
Example: measure 4.5 ± 1.0 but you know it is less than 6



If range of a infinite, $P_0(a)$ may be vanishingly small ('improper prior'). Not a problem. Just normalise $P_1(a)$

Shortcomings of Bayesian Probability

Subjective Probability

Your $P_0(a)$ and my $P_0(a)$ may be different. How can we compare results?

What is the right prior?

Is the wrong question.

'Principle of ignorance' - take $P(a)$ constant (uniform distribution). But then not constant in a^2 or $\sqrt{a}$ or $\ln a$, which are equally valid parameters.

Jefffreys' Objective Priors

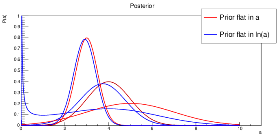
Choose a flat prior in a transformed variable a' for which the Fisher information, $-\left\langle \frac{\partial^2 L(x;a)}{\partial a^2} \right\rangle$ is flat. Not universally adopted for various reasons.

With lots of data, $P_1(a)$ decouples from $P_0(a)$. But not with little data..

Right thing to do: try several forms of prior and examine spread of results ('robustness under choice of prior')

Just an example

Measure $a = 4.0 \pm 1.0$.
Likelihood is Gaussian (coming up!)



Taking a prior uniform in a gives a posterior with a mean of 4.0 and a standard deviation of 1.0 (red curve).
Prior uniform in $\ln a$ shifts the posterior (blue curve). Some difference.
For $a = 3.0 \pm 0.5$ the posteriors are pretty similar
for $a = 5.0 \pm 2.0$ they are really different.
Different priors lead to different posteriors - maybe significantly different.

Exercise

Try this for yourself with various values

Frequentist or Bayesian?

Both useful concepts. Use both. But don't get them confused.

Probability Distributions and pdfs

Integer Values

Numbers of positive tracks, numbers of identified muons, numbers of events..
Generically call this r. Probabilities $P(r)$

Real-number Values

Energies, angles, invariant masses...
Generically call this x. Probability Density Functions $P(x)$.
$P(x)$ has dimensions of $[x]^{-1}$. $\int_{x_1}^{x_2} P(x) dx$ or $P(x)\, dx$ are probabilities
Sometimes also use cumulative $C(x) = \int_{-\infty}^{x} P(x')\, dx'$

Pdfs and Likelihoods

If the pdf has a parameter, $P(x, a)$ it is mathematically the same as the likelihood $L(x|a)$ but handled as a function of x rather than a function of a.
Pdfs are normalised $\int dx$. Likelihoods $\int da$ are not.

Mean, Standard deviation, and expectation values

From $P(r)$ or $P(x)$ can form the Expectation Value

$$< f >= \sum_r f(r)P(r) \qquad \text{or} \qquad < f >= \int f(x)P(x)\,dx$$

Sometimes written $E(f)$
In particular the mean $\mu =< r >= \sum_r rP(r)$ or $< x >= \int xP(x)\,dx$
and higher moments $\mu_k =< r^k >= \sum_r r^k P(r)$ or $< x^k >= \int x^k P(x)\,dx$
and central moments
$\mu'_k =< (r-\mu)^k >= \sum_r (r-\mu)^k P(r)$ or $< (x-\mu)^k >= \int (x-\mu)^k P(x)\,dx$

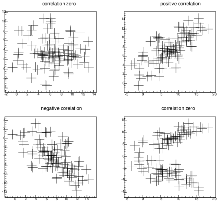
The Variance and Standard Deviation

$\mu'_2 = V = \sum_r (r-\mu)^2 P(r) =< r^2 > - < r >^2$
or $\int (x-\mu)^2 P(x)\,dx =< x^2 > - < x >^2$
The standard deviation is the square root of the variance $\sigma = \sqrt{V}$

Statisticians usually use variance. Physicists usually use standard deviation
Skew is $< (x- < x >)^3 > /\sigma^3$ and Kurtosis is $< (x- < x >)^4 > /\sigma^4 - 3$

Covariance and Correlation

2-dimensional data (x, y)
Form $< x >, < y >, \sigma_x$ etc
Also other quantities



Covariance

$$Cov(x, y) =< (x - \mu_x)(y - \mu_y) >=< xy > - < x >< y >$$

Correlation

$$\rho = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

ρ lies between 1 (complete correlation) and -1 (complete anticorrelation).
$\rho = 0$ if x and y are independent.

Covariance and Correlation (continued)

Many Dimensions $(x_1, x_2, x_3 \ldots x_n)$

Covariance matrix $\mathbf{V}_{ij} =< x_i x_j > - < x_i >< x_j >$

Correlation matrix $\rho_{ij} = \frac{\mathbf{V}_{ij}}{\sigma_i \sigma_j}$

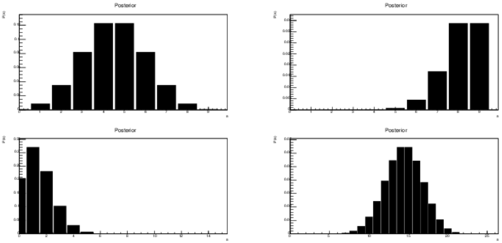
Diagonal of $\mathbf{V}$ is σ_i^2

Diagonal of ρ is 1.

The Binomial Distribution

Binomial: Number of successes in N trials, each with probability p of success

$$P(r;p,N) = \frac{N!}{r!(N-r)!}p^r q^{N-r} \qquad (q \equiv 1-p)$$



Binomial distributions for
(1) $N = 10, p = 0.6$
(2) $N = 10, p = 0.9$
(3) $N = 15, p = 0.1$
(4) $N = 25, p = 0.6$

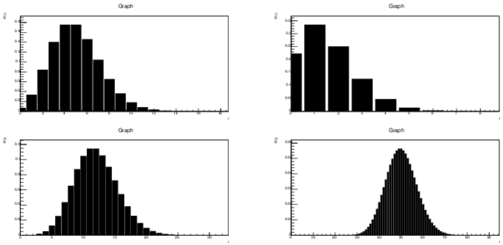
Mean $\mu = Np$, Variance $V = Npq$, Standard Deviation $\sigma = \sqrt{Npq}$

The Poisson Distribution

Number of events occurring at random rate λ

$$P(r;\lambda) = e^{-\lambda}\frac{\lambda^r}{r!}$$

Limit of binomial as $N \to \infty$, $p \to 0$ with $np = \lambda = constant$



Poisson distributions for
(1) $\lambda = 5$
(2) $\lambda = 1.5$
(3) $\lambda = 12$
(4) $\lambda = 50$

Mean $\mu = \lambda$, Variance $V = \lambda$, Standard Deviation $\sigma = \sqrt{\lambda} = \sqrt{\mu}$

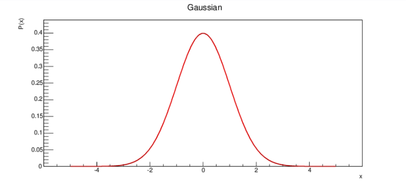
Meet this a lot as it applies to event counts - on their own or in histogram bins

The Gaussian

The Formula

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

The Curve



Only 1 Gaussian curve, as μ and σ are just location and scale parameters

Properties

Mean is μ and standard deviation σ. Skew and kurtosis are 0.

The Central Limit Theorem

Why the Gaussian is so important

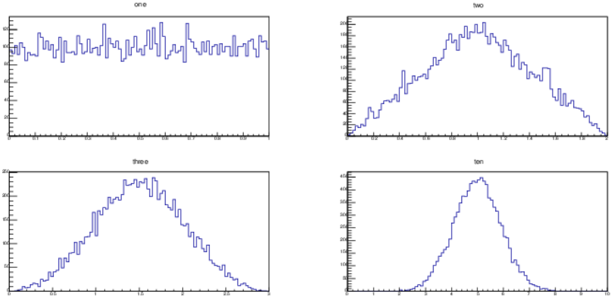
If the variable X is the sum of N variables $x_1, x_2 \dots x_N$ then

1. Means add: $< X >=< x_1 > + < x_2 > + \cdots < x_N >$
2. Variances add: $V_X = V_1 + V_2 + \dots V_N$
3. If the variables x_i are independent and identically distributed (i.i.d.) then $P(X)$ tends to a Gaussian for large N

(1) is obvious
(2) is pretty obvious, and means that standard deviations add in quadrature, and that the standard deviation of an average falls like $\frac{1}{\sqrt{N}}$
(3) applies whatever the form of the original $p(x)$

Demonstration

Take a uniform distribution from 0 to 1. It is flat. Add two such numbers and the distribution is triangular, between 0 and 2.



With 3 numbers, it gets curved. With 10 numbers it looks pretty Gaussian

Gaussian or Normal?

Statisticians call it the ‘Normal’ distribution. Physicists don’t. But be prepared.

Even if the distributions are not identical, the CLT tends to apply, unless one (or two) dominates.

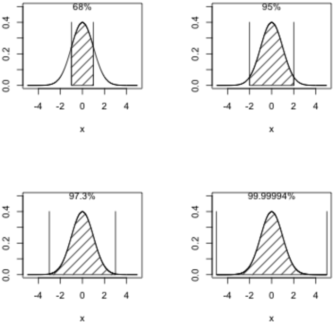
Most ‘errors’ fit this, being compounded of many different sources.

The Central Limit Theorem is amazingly powerful. Non-Gaussian distributions are nothing to be scared of.

Some important facts about the Gaussian...

Deviations of many sigma are unlikely

68% of samples lie within 1 sigma
95% lie within two sigma
99.7% lie within 3 sigma
and so on



Quantify by p-value: probability of a deviation this large (or larger)
32%, 5%, 0.3%
(Need to decide whether deviations in one direction only or both)
Small p-values means your original assumption (called H_0) is implausible.
More about this in Lecture 3.

N sigma language

One tailed

p	N
0.158	1
0.022	2
1.35×10^{-3}	3
3.17×10^{-5}	4
2.87×10^{-7}	5

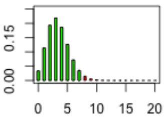
Two tailed

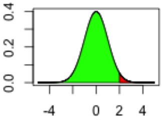
p	N
0.317	1
0.0455	2
0.00270	3
6.33×10^{-5}	4
5.73×10^{-7}	5

N-sigmas are easier to handle than small p-values so they are often used
(Functions to do this are available in ROOT or Python or R or whatever...)

Example

You observe 8 events, with an expected background of 3.4.
Poisson probability of 8 or more events is 2.3%, equivalent to 1.99 sigma





Final thoughts

Frequentist and Bayesian probability are both useful concepts. Be prepared to use both. But don't get them confused.

You will meet the Poisson often, the Binomial occasionally, and the Gaussian all the time.

The Central Limit Theorem is powerful and beautiful and, above all, useful

The p-value is the probability of getting a result this weird

p-values are often expressed in terms of equivalent sigmas, even though there is no actual Gaussian involved