Statistics for experimental particle physicists

Andrew Fowlie

29th March 2015

Contents

I will give two lectures

1. Motivation
2. Reading
3. Probability
4. Distributions
5. Point estimation
6. Exercises
7. Confidence intervals
8. Hypothesis testing
9. The look-elsewhere effect
10. Exercises

I will put these lecture notes and useful material on my web-page: ▸ http://hep.kbfi.ee/index.php/Members/AndrewFowlie

Lecture 1: Overview of probability and statistics

Motivation I

Your motivation, not really mine

- We are physicists — why are we discussing statistics?
- Abstract for the Higgs discovery (Aad et al. 2012b):

> *This observation, which has a significance of 5.9 standard deviations, corresponding to a background fluctuation probability of 1.7×10^{-9}, is compatible with the production and decay of the Standard Model Higgs boson*

- If we want to infer information from an experiment, we must use statistics. Typically, we want to reject explanations for observations, *e.g.*, the SM without a Higgs boson is rejected

Motivation II

Your motivation, not really mine

- Statistics is critical to LHC experiments, because we have counting experiments.
- If we expected 100 events from backgrounds and saw 110, what does it mean? Is there a signal?
- That isn't the only source of uncertainties, though
- Historical note: statistical methodologies applied in mid-nineties at LEP and Tevatron
- The discovery of the W-boson in 1983 did not mention *e.g.*, 5σ or any significance (Arnison et al. 1983) (low backgrounds)

In summary, you should learn statistics because you want to be able to justify conclusions from an experiment, *e.g.*, we've discovered a Higgs boson

Motivation

My motivation, probably not yours

- Hopefully, as well as showing you how to calculate things, I can hint that statistics is more than a tool in science
- Scientific methods, choosing between theories *etc*, are linked to statistics
- That is, statistics isn't a tool in science; science is, in a way, built on statistics

This sounds quite grand; now we must get back down to Earth with some introductory material

Reading

I like these books:

- James, F. (2006). *Statistical methods in experimental physics* (Second). World Scientific — frequentist, calculations
- Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. (G. L. Bretthorst, Ed.). Cambridge University Press — Bayesian, interpretations
- Lyons, L. [L.]. (1989). *Statistics for nuclear and particle physicists*. Cambridge University Press — frequentist, calculations

Unfortunately, since mid-nineties, particle physics used arcane test statistics, *e.g.*, CL_s, that aren't in text books

Probability

What is it?

We will later distinguish between probability and statistics, but we will need ideas about both (Hájek 2012):

1. Bayesian: degree of belief in a proposition? Chelsea will probably win the European cup. Probability stems from our ignorance.
2. Quasi-logical: if the sun is shining, it's probably hot
3. Physical: the frequency or propensity with which a physical system exhibits a particular behavior

Physical probability

Does physical probability exist?

- My opinion: most probability is Bayesian, *e.g.*, a coin toss is "random" because we are ignorant of its initial conditions
- Repeated trials tell us something about how those initial conditions vary between trials (Jaynes 2003)
- Physical probabilities in quantum mechanics and QFT

Probability

Which one do we need?

For the purposes of calculating things for experimental physics, we need physical probability. In particular we need frequentist probability:

> *Probability is the frequency with which an outcome occurs in repeated trials:*
>
> $$P = \lim_{n \to \infty} \frac{k}{n} \tag{1}$$

- Almost all statistics in collider physics is frequentist
- "Let the data speak for themselves" (Fisher)

Obvious objections to Frequentism

- If trials are identical, won't outcomes be identical? Not if physical probability
- Can we make identical trials? No but can control relevant factors
- Can we realize infinite limit? No but we make enough trials
- What about events that can't be repeated? They don't have probabilities

Frequentism

Something to remember

Keep in mind that if we are doing frequentist statistics:

- Probability, and many of our statements, concern hypothetical experiments that we haven't performed
- Foreshadowing our later calculations, we will often say, *e.g.*, if we repeated this experiment many times, we'd get a result like this one less than 5% of the time
- We're finding properties of our experiment; not properties of any underlying physical theory. Some people call this "in the data"

I will highlight this later

Axioms

Kolmogorov's axioms and some basics

- If events X_i are exclusive, probabilities must satisfy

$$\begin{aligned} P(\Omega) &= 1 \\ P(X_i) &\geq 0 \\ P(X_i \text{ or } X_j) &= P(X_i) + P(X_j) \end{aligned} \tag{2}$$

- You can find justification for Kolmogorov's axioms with "Dutch-book arguments"
- If you don't follow Kolmogorov, you'll lose money
- Do we have to define probability with recourse to money? Isn't it a fundamental branch of mathematics?

Important formulas

- If events are not exclusive, it follows that

$$P(X_i \text{ or } X_j) = P(X_i) + P(X_j) - P(X_i \text{ and } X_j) \tag{3}$$

  Think of a Venn diagram

- We also want conditional probability:

$$P(X_i \text{ given } X_j) \equiv P(X_i|X_j) \equiv \frac{P(X_i \text{ and } X_j)}{P(X_j)} \tag{4}$$

  This formula is also known as Bayes' theorem

- As mathematics, it's trivial, but if it comes along with Bayesian interpretations of probability, it's contentious

Distributions

- A random event is an event with more than one possible outcome
- Possible outcome is a random variable
- Probability of various outcomes given by probability distribution
- If I roll 100 coins, discrete random variable could be the number of heads (binomial distribution)
- Another discrete random variable is the number of fatalities by horse kicking in the Prussian army (Poisson distribution)

Distributions I

Continuous random variables

- We mentioned discrete random variables; what about continuous?
- We need a probability density function (PDF):

$$p(x) = \lim_{\Delta x \to 0} \frac{P(x - \frac{1}{2}\Delta x < X < x + \frac{1}{2}\Delta x)}{\Delta x} \tag{5}$$

- I denote densities with lower-case p and probabilities with upper-case P
- Example: errors in an experiment (Gaussian). I'll return to this
- This is density with unit $[1/x]$
- $P(a < X < b) = \int_a^b p(x)\,\mathrm{d}x$

Distributions II

Continuous random variables

- It has a cumulative density function (CDF)

$$F(x) = \int_{-\infty}^{x} p(x')\,\mathrm{d}x' \tag{6}$$

- Non-trivial changes of variables with a Jacobian, $\mathcal{J}$

$$p(x)\,\mathrm{d}x = p(y = f(x))\,\mathrm{d}y \Rightarrow p(y) = \sum \frac{p(x)}{|\mathcal{J}|} \tag{7}$$

- Density could describe several continuous random variables, in which case

$$p(x) = \int p(x, y)\,\mathrm{d}y \tag{8}$$

Distributions III

Continuous random variables

- This is called marginalization. It is useful for eliminating "nuisance" parameters, such as systematics
- But, as you'll see, there is a another more common way

Properties of distributions

Expectation

- $E(g) = \int_\Omega g(x)p(x)\,\mathrm{d}x$
- Linear: $E(ag(x) + bh(x)) = aE(g(x)) + bE(h(x))$
- $\mu = E(x)$ is called the mean. The mean isn't always well-defined (Cauchy distribution), but it usually is
- $\sigma^2 = E((x-\mu)^2) = E(x^2) - E(x)^2$ is called the variance, σ is called the standard deviation
- From this definition, no connection between σ and probability content
- Mean gives location, whereas standard deviation gives spread

Properties I

Sample mean

- Suppose we make N random samples from a distribution. We may calculate the sample mean (continuous random variable):

$$\bar{x} = \frac{1}{N} \sum x_i \tag{9}$$

- Expectation of the mean: $E(\bar{x}) = \frac{1}{N} \sum E(x_i) = \frac{1}{N} N\mu = \mu$

Properties II

Sample mean

- Variance of the mean:

$$
\begin{aligned}
E(\bar{x}^2) - E(\bar{x})^2 &= \frac{1}{N^2} E(\sum x_i \sum x_i) - \frac{1}{N^2} E(\sum x_i)^2 \\
&= \frac{1}{N^2} \left[\sum E(x_i^2) - \sum E(x_i)^2 \right] \\
&+ \frac{1}{N^2} \left[\sum E(x_i x_j) - \sum E(x_i)(E(x_j) \right] \\
&= \frac{1}{N} \left[E(x_i^2) - E(x_i)^2 \right] = \frac{1}{N} \sigma^2
\end{aligned}
\tag{10}
$$

- Because samples are independent, so the covariance vanished
- Demonstrates well-known result: standard deviation of sample mean is $\sigma / \sqrt{N}$

Properties

Sample variance

Sample variance (continuous random variable):

$$s^2 = \frac{1}{N} \sum (x_i - \bar{x})^2 \tag{11}$$

- We've already calculated that $E(s^2) = \sigma^2/N$ — check this if it is not obvious
- What about variance of sample variance $E(s^2)^2 - E((s^2)^2)$?
- This is something like the error on the error. Calculation is technical, but result is that $\propto 1/N$, so the error on the error is $\propto 1/\sqrt{N}$
- This is the origin of a rule of thumb for quoting errors in experiments

Quoting results

My rules

- Quote the sample standard deviation and sample mean to the same number of decimal places, for example

$$\bar{x} \pm s = 1.0 \pm 0.1 \tag{12}$$

- How many significant figures for the error? If you have $\gtrsim 100$ samples, two significant figures, $\gtrsim 10^4$ three significant figures, *etc*
- This originates from fact that error on error $\propto 1/\sqrt{N}$

Examples I

Binomial distribution

- Discrete distribution for repeated trials with two possible outcomes
- Probability of k outcomes with probability p in n trials:

$$P_k = \binom{n}{k} p^k (1-p)^{n-k} \tag{13}$$

Examples II

Binomial distribution

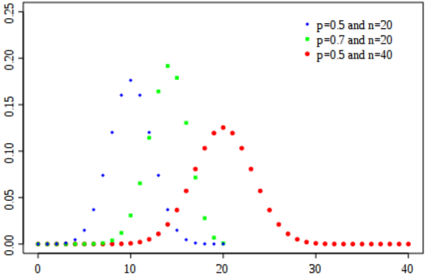


Figure : Binomial distribution for various n and p (wiki)

- For example, tossing a coin
- Expectation $\langle k \rangle = np$

Examples III

Binomial distribution

- Variance $\sigma^2 = np(1-p)$

Examples I

Poisson distribution

- Discrete distribution present in every counting experiment
- Know expected number of "events", λ, in a particular interval, T
- Probability distribution of seeing k events in that interval? Begin with binomial

$$P_k = \binom{n}{k} p^k (1-p)^{n-k} \tag{14}$$

Examples II

Poisson distribution

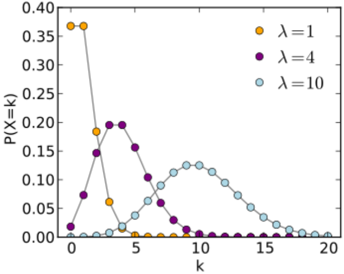


Figure : Poisson distribution for various λ (wiki)

Examples III

Poisson distribution

- Discretize time-interval, such $n = T/\Delta t$ that $p \approx \lambda \Delta t / T$, such that $\langle k \rangle = np = \lambda$:

$$P_k = \frac{(T/\Delta t)!}{k!(T/\Delta t - k)!}(\lambda \Delta t/T)^k (1 - \lambda \Delta t/T)^{T/\Delta t}(1 - \lambda \Delta t/T)^{-k} \tag{15}$$

- Let $n = T/\Delta t \to \infty$ (note that $(1 - a/x)^x \to e^{-a}$ as $x \to \infty$):

$$\begin{aligned} P_k &\to \frac{(T/\Delta t)!}{k!(T/\Delta t - k)!}(\lambda \Delta t/T)^k e^{-\lambda}(1 - \lambda \Delta t/T)^{-k} \\ &= \frac{\lambda^k}{k!}e^{-\lambda} \times \frac{(T/\Delta t)!}{(T/\Delta t - k)!}(\Delta t/T)^k \\ &\to \frac{\lambda^k}{k!}e^{-\lambda} \end{aligned} \tag{16}$$

Examples IV

Poisson distribution

- Expectation $\langle k \rangle = \lambda$
- Variance $\sigma^2 = \lambda$

Normal distribution I

Normal/Gaussian/Bell curve distribution

- Continuous distribution, defined for all reals
- Common notation $\mathcal{N}(\mu, \sigma^2)$ — normal distribution with expectation (and mode and median) μ and variance σ^2

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \tag{17}$$

Normal distribution II

Normal/Gaussian/Bell curve distribution

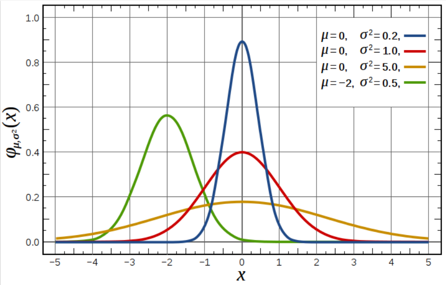
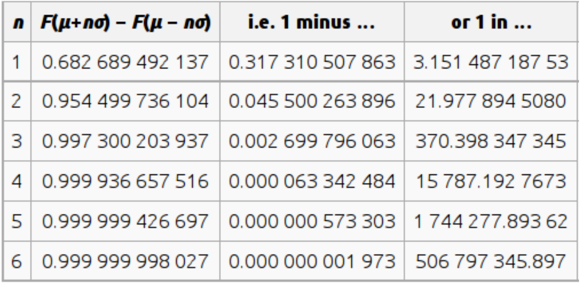


Figure : Normal distribution for various μ and σ (wiki)

- Numbers of standard deviation corresponds to cumulative probabilities (careful about one- or two-tail)

Normal distribution III

Normal/Gaussian/Bell curve distribution



n	$F(\mu+n\sigma) - F(\mu - n\sigma)$	i.e. 1 minus ...	or 1 in ...
1	0.682 689 492 137	0.317 310 507 863	3.151 487 187 53
2	0.954 499 736 104	0.045 500 263 896	21.977 894 5080
3	0.997 300 203 937	0.002 699 796 063	370.398 347 345
4	0.999 936 657 516	0.000 063 342 484	15 787.192 7673
5	0.999 999 426 697	0.000 000 573 303	1 744 277.893 62
6	0.999 999 998 027	0.000 000 001 973	506 797 345.897

Figure : Probability under normal distribution (wiki)

Normal distribution I

Why is this distribution ubiquitous?

- The normal distribution is “natural” — other distributions gravitate towards it!
- If n is great, binomial distribution approximately $\mathcal{N}(np, np(1-p))$
- If $\lambda \gtrsim 10^3$, Poisson distribution approximately $\mathcal{N}(\lambda, \lambda)$
- Central limit theorem: average of large number n of random variables is approximately $\mathcal{N}(\mu, \sigma^2/n)$
- Errors are assumed to be Gaussian, not because we know their distribution, but because this is conservative
- Gaussian is “maximum entropy” distribution — most conservative possible for errors, if we know mean and variance

χ^2-distribution I

If x_i are normally distributed random variables, then

- $\lambda = \sum_{i=1}^{n}(x_i - \mu)^2/\sigma^2$ is χ^2-distributed

$$p(\lambda; n) = \frac{1}{2}\left(\frac{\lambda}{2}\right)^{n/2-1} e^{-X/2}/\Gamma(n/2) \tag{18}$$

χ^2-distribution II

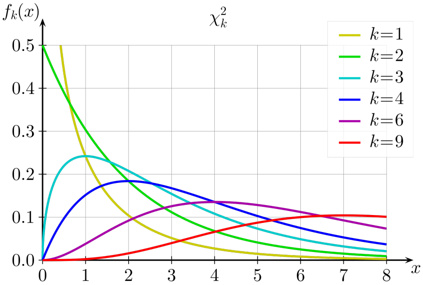


Figure : χ^2-distribution for various degrees of freedom (wiki)

- Expectation $\langle\lambda\rangle = n$
- Variance $\sigma^2 = 2n$

χ^2-distribution III

- n is called the number of degrees of freedom
- This is a common distribution because χ^2 is a common test-statistic and because of Wilk's theorem (forthcoming)

Likelihood function

Suppose we perform an experiment and obtain data D

- The probability of obtaining that data, assuming a particular theory H and various parameters in that theory $\vec{\lambda}$, given by likelihood:
$$\mathcal{L} = p(D|H, \vec{\lambda}) \tag{19}$$
- As a function of D, this is a sampling distribution
- As a function of $\vec{\lambda}$, this is not a distribution
- We (almost always) draw conclusions from likelihood functions — estimate a parameter, confidence interval or a test a hypothesis
- We often, however, first construct a test-statistic from the likelihood

Point estimator

- A physicist might say “measurement”, but in a statistical language, this is an “estimate”
- A point estimate is an estimate of the value of parameter from a finite number of experiments
- Of course, a point estimate should be “as close as possible” to the true value (but we don’t know the true value)
- An estimator is a function or method for making an estimate from experimental data
- An estimator is a random variable (with a distribution)

Estimators I

Desirable properties

Estimator $\hat{\lambda}$ for a parameter λ should:

- Consistent — $\hat{\lambda} \to \lambda$ as more and more data collected
- Unbiased — $E(\hat{\lambda}) = \lambda$, expected to give the correct answer
- Invariant — $\hat{f}(\lambda) = f(\hat{\lambda})$
- Maximum information — no other number could summarize parameter with more information
- Minimum variance — if small variance, more certain that estimate is close to true value

Estimators II

Desirable properties

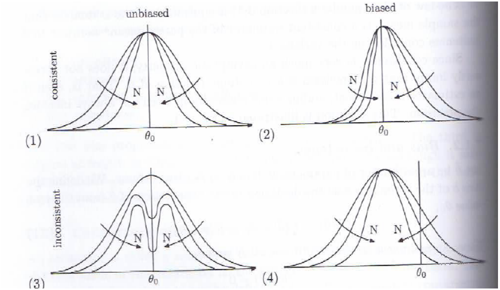


Figure : Example of bias and consistency (James 2006). Distribution of estimator. N is number of observations (arrow is increasing N) and θ_0 is true value

Estimators III

Desirable properties

Estimator should also be practical:

- Simple to present and explain
- Minimal computer/physicist's time
- Robustness — minimal sensitivity to assumptions about *e.g.*, likelihood function

Point estimator I

Maximum likelihood Estimator (MLE)

A common estimator in collider physics is the MLE

- $\hat{\lambda}$ such that
$$\mathcal{L} = p(D|H, \lambda \ldots) \tag{20}$$
is maximized
- Note that maximizing $\mathcal{L}$ is equivalent to minimizing $-\ln \mathcal{L}$ (logarithm monotonic)
- If there are other parameters (*e.g.*, nuisance parameters related to systematics or model par meters), we maximize the likelihood with respect to them all
- *i.e.*, we find global maximum
- This is sometimes called profiling

Point estimator II

Maximum likelihood Estimator (MLE)

- This is asymptotically consistent, unbiased and minimum variance (this is a little technical, see (James 2006))
- Practical — easy to present and understand
- Calculated by minimizing $-\ln \mathcal{L}$, usually numerically
- MLE is used to determine, *e.g.*, Higgs mass at LHC

Point estimator III

Maximum likelihood Estimator (MLE)

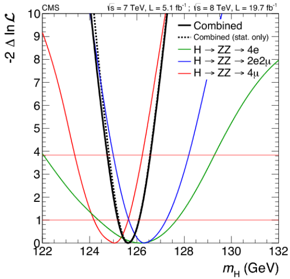


Figure : MLE estimates of the Higgs mass (Chatrchyan et al. 2014)

Other estimators I

There is a lot of theory about estimators. Here are a few remarks:

- In collider physics, MLE is very common
- There are only a few technical cases in which it is a poor choice
- Remember that maximizing probability of data given parameter $\neq$ maximizing probability of parameter given theory
- Alternatives, from Bayesian side, (posterior) mean, mode or median
- But in Bayesian perspective, result is distribution — estimate is a summary
- From frequentist side, *e.g.*, least squares

Estimate parameters with MLE, but keep in mind it isn't the only possibility

Exercises 1

The starred exercises are trickier or probably require a computer. Focus on the unstarred exercises, but think about how you might tackle the starred exercises. Write down expressions even if you cannot numerically evaluate them without a computer.

Please complete all exercises by next week.

Expectations and general probability — seeing numbers:

1. Expected number of dice rolls if I stop rolling once I roll a six?

*2. Expected number of dice rolls if I stop rolling once I roll all numbers on dice?

Exercises II

Conditional probability — a foolish prisoner's dilemma:

3. Guard tells 3 prisoners (Alan, Bob and Colin) he has randomly chosen 2 to release tomorrow. Alan pesters the guard for more information. The guard refuses to tell Alan whether he will be released, but tells him that Bob will be released. Alan is upset; he reasons that his chance to be released was $2/3$, but has fallen to $1/2$ because he and Colin have an equal chance of release. Correct Alan's reasoning.

Trial by binomial distribution:

*4. Consider two juries (variant of Newton-Pepys problem):

Exercises III

- Seven votes from 10 are required to convict. Each juror decides to vote guilty with probability $p = 0.7$.
- Seventy votes from 100 are required to convict. Each juror decides to vote guilty with probability $p = 0.7$.

Which jury is most likely to reach a guilty verdict?

Poisson paradox:

5. Poisson should be probability for rare, independent events. If probability of one event is p, should the probability for 2 events be p^2?
6. Show that it is in fact $\approx p^2/2$.
7. Justify the factor of $1/2$.

Exercises IV

Is there a typo?

*8. I make on average 2 typos per slide. What is the probability that there is at least one typo on this slide?

Switching between distributions:

9. If x is distributed uniformly on $[0,1]$, find $y = f(x)$ such that $y \sim \mathsf{N}(\mu, \sigma)$.
10. What is the distribution of $y = F(x)$, where F the CDF corresponding to the distribution of x?

The sample mean:

Exercises V

*11. Is the sample mean consistent?

*12. Is the sample mean biased?

13. Is the sample mean invariant?

Sample variance:

14. Show that if x_i are uncorrelated,

$$V(\sum a_i x_i) = \sum a_i^2 V(x_i) \tag{21}$$

where $V(Z) = E(Z)^2 - E(Z^2)$, *i.e.*, the variance, and in this case $Z = \sum a_i x_i$

Discovering the significance of 5σ discovery significance

Exercises VI

15. We will discuss hypothesis testing in our next lecture. Physicists like 5σ significance — a confidence equal to the probability under a normal distribution at $x > 5\sigma$ (one-tail). Look up this probability.
16. Is it small enough to convince you?

Maximum likelihood distributions:

17. Our experiment measures a random variable, x, that we believe has an exponential distribution

$$p(x) = \frac{1}{\tau} e^{-t/\tau}. \tag{22}$$

We make n measurements of t, $t_1, \cdots, t_n$. Write down the likelihood function.

Exercises VII

18. Find the MLE for the parameter τ.

Lecture 2: Confidence intervals and hypothesis testing for experimental particle physicists

Confidence intervals I

or interval estimation

- We discussed point estimation and the MLE
- But what if we perform an experiment, and want to present an interval

$$x_l \leq x \leq x_u \tag{23}$$

  that we believe contains the true parameter?
- This is very common — for example, we might want an interval or a limit (one-sided interval) for the photon mass or Higgs mass
- We might desire:
  - Parameterization invariance
  - Pragmatic — easy to calculate and explain

Confidence intervals II

or interval estimation

- In frequentist statistics, we define such intervals by their coverage:

> *If experiment repeated many times, the interval estimated at β confidence will contain the true value of a parameter in a fraction β of experiments*

- This is an important point: interval estimates are constructed in the data — they are properties of hypothetical pseudo-experiments
- $\beta = 68\%$ and 95% common and referred to as 1σ and 2σ (as this is probability under normal distribution)
- The ends of a confidence interval are random variables

Confidence intervals III

or interval estimation

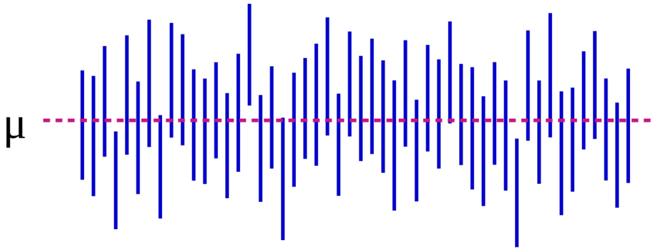


Figure : The bars represent interval estimates in independent experiments and μ is the true value of the parameter. In 95% cases, the interval (the bar) contains the true value μ (wiki)

Example: Neyman construction I

- Let's estimate a parameter θ from data t' at confidence β
- Let's begin by constructing a statement in the data:

$$\beta = \int_{t_l}^{t_u} p(t|\theta)\,\mathrm{d}t \tag{24}$$

- Immediate problem: multiple ways to pick t_l and t_u — they are not unique
- We must pick an ordering rule (this specifies the order in which area under PDF should be included)

Example: Neyman construction II

- Let's pick symmetric ordering rule:

$$(1-\beta)/2 = \int_{-\infty}^{t_l} p(t|\theta)\,\mathrm{d}t = \int_{t_u}^{\infty} p(t|\theta)\,\mathrm{d}t \tag{25}$$

- Solve for interval as function of θ: $t_l = t_l(\theta)$ and $t_u = t_u(\theta)$
- We must invert this solution to obtain an confidence interval for θ
- We invert by solving $t' = t_l(\theta_l)$ and $t' = t_u(\theta_u)$ to obtain an interval (θ_l, θ_u)
- Because interval (t_l, t_u) has correct coverage for data, (θ_l, θ_u) has correct coverage for true parameter

Example: Neyman construction III

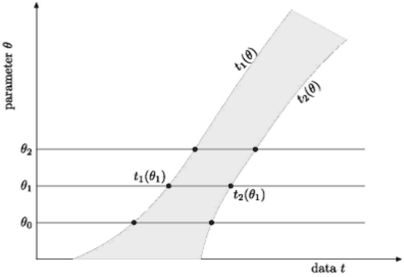


Figure : Constructing a Neyman interval via a confidence belt (James 2006). The inversion from data to theory is found by reading the interval vertically

Flip-flopping

- Flip-flopping: we remarked that we had to pick an ordering rule
- But sometimes, we don't know in advance whether we want to report an upper limit or an interval
- For example, if we're measuring something we think is zero (*e.g.*, photon mass), if we get something close to zero, we want to report an upper limit
- But if we get something much bigger than zero, we might want to report an interval
- This is called/leads to the problem of flip-flopping (incorrect coverage)

Nonphysical values in the interval estimate

- What if our interval includes nonphysical values?
- For example, what if we end up with a negative interval for photon mass?
- We could forbid it, but then we'd have an empty region?
- This is clearly bad news for Neyman construction

Unified approach

AKA Feldman-Cousins

- Addresses the problems with Neyman's construction concerning an ordering rule/nonphysical values
- Feldman-Cousins:

> *Ordering rule is that you should include μ within the physical region with the largest values of*
>
> $$R(\mu) = \frac{p(d|\mu)}{p(d|\hat{\mu})} \qquad (26)$$
>
> *until the required confidence is reached*

- This isn't a priori one or two-tailed — it could be either!

Profile likelihood I

- Similar to Feldman-Cousins construction
- Calculate log-likelihood ratio test statistic:

$$\lambda(\mu) = -2\ln\frac{p(d|\mu,\hat{\hat{\nu}})}{p(d|\hat{\mu},\hat{\nu})} \tag{27}$$

- We have profiled a nuisance parameter ν
- This is a random variable
- By Wilk's theorem, it is often approximately χ^2-distributed

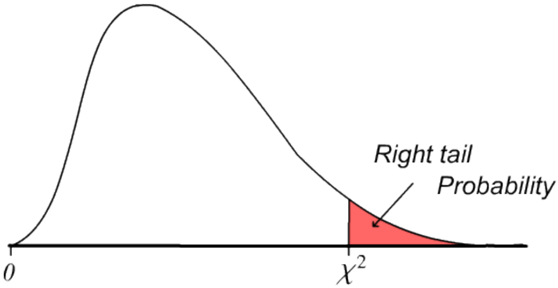
$$\lambda(\mu) \sim \chi^2_1 \tag{28}$$

- Exclude $\lambda(\mu)$ from interval if

$$P(\lambda > \lambda(\mu)) < 1 - \beta \tag{29}$$

Profile likelihood II

- This clearly has desired coverage
- This is a very common method in high-energy physics because it's easy to include nuisance parameters



Profile likelihood III

Figure : If area in tail beyond $\lambda(\mu)$ less than critical value, don't include μ in interval

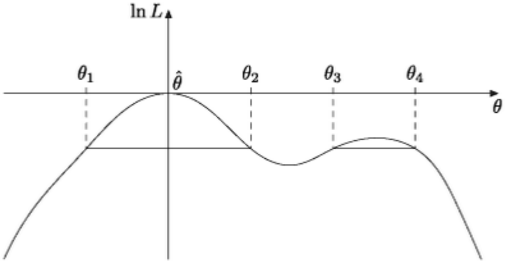


Figure : The "intervals" found from profile likelihood might not be contiguous

CL$_s$ upper limits I

- Particle physicists invented their own method for presenting upper limits in counting experiments
- Suppose we expect $s(\mu) + b$ events in signal hypothesis and b in background hypothesis, with unknown parameter μ
- We observe o events
- Calculate CL$_s$ statistic

$$\mathrm{CL}_s(\mu) = \frac{\mathrm{CL}_{s+b}}{\mathrm{CL}_b} = \frac{P(\lambda > \lambda(\mu))|s(\mu) + b)}{P(\lambda > \lambda(0)|b)} \tag{30}$$

- If CL$_s(\mu) < 1 - \beta$, don't include μ in interval
- This over-covers — includes true value more often than it should (conservative)

CL$_s$ upper limits II

- It designed to eliminate cases in which downward fluctuations in background cause signals to be excluded
- For example, if you expect $b = 10$ but observe $o = 1$. You can exclude even very small signals with profile likelihood. Many people don't like this, so use CL$_s$
- CL$_s$ is on shaky theoretical foundations, but it is very commonly used

Hypothesis testing and goodness of fit

- We've used experimental data to estimate parameters and intervals
- But what about testing hypothesis? How do we reject theories with data?
- How do we find whether a particular theory is favored by data?
- With this sort of methodology, we're always trying to reject models (rather than confirm them)

What is a hypothesis

- A hypothesis specifies a PDF for the outcome of an experiment
- A simple hypothesis is a hypothesis with no adjustable parameters
- A composite hypothesis involves free parameters — this is an (infinite) ensemble of simple hypotheses

Hypothesis test

In frequentist statistics,

> *Hypothesis is rejected at β confidence, if, were the experiment repeated and the hypothesis was true, we'd obtain such "extreme" data in only a fraction $1 - \beta$ of the experiments*

- This is not a statement about the probability of the theory
- Statement about probability of obtaining such "extreme" data
- The extremity of the data is formulated with a test statistic

Possible errors

With this definition, we are susceptible to errors:

- Error of first kind (Type-1 error): Reject hypothesis when it is true
- Error of second kind (Type-2 error): Accept hypothesis when it is false

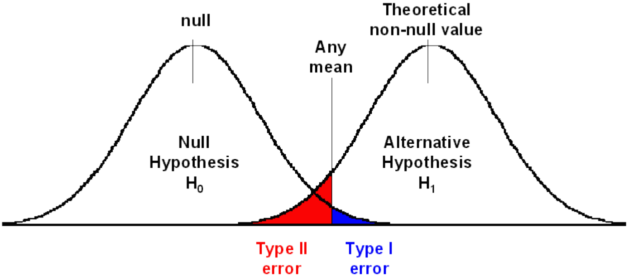


Figure : Example of type-1 and type-2 errors in hypothesis testing

Test-statistic and p-values I

- As mentioned, the extremity of the data is formulated with a test statistic
- A test statistic is a function of experimental data
- It is a random variable

Test-statistic and p-values II

- We may extend our previous definition of hypothesis testing:

The probability of obtaining a test-statistic larger than that obtained is called the p-value:

$$p\text{-value} = P(\lambda > \lambda(\text{observed})|H_0) \qquad (31)$$

If p-value $< 1 - \beta$, reject H_0 at β confidence

Test-statistic and p-values III

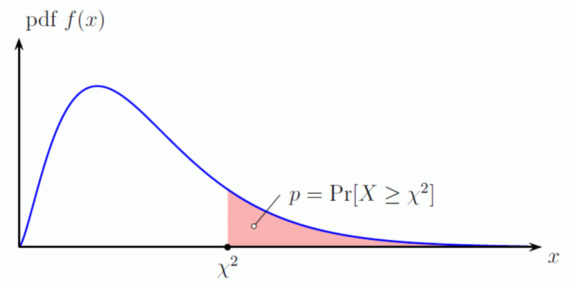


Figure : p-value is the probability in the right-hand tail of test-statistic distribution. If that probability is less than a threshold, reject the hypothesis

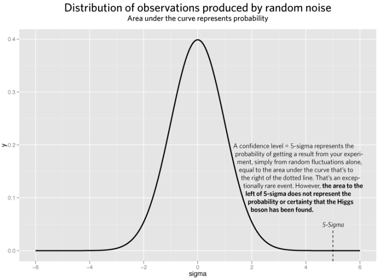
Significance and p-values I

- For historic reasons, it is common in high-energy physics to convert p-values into so-called significances
- That this, the number of deviations that make the same probability in the right-hand tail of a Gaussian

$$p\text{-value} = \int_{-\infty}^{Z} \text{Gauss}(x; \mu = 0, \sigma^2 = 1) \, \mathrm{d}x \tag{32}$$

- High-energy physicists like 5σ significance — which corresponds to a very small p-value
- This minimizes type-1 errors

Significance and p-values II



Distribution of observations produced by random noise

Area under the curve represents probability

A confidence level = 5-sigma represents the probability of getting a result from your experiment, simply from random fluctuations alone, equal to the area under the curve that's to the right of the dotted line. That's an exceptionally rare event. However, **the area to the left of 5-sigma does not represent the probability or certainty that the Higgs boson has been found.**

Figure : Illustration of meaning of 5σ significance for Higgs discovery

Choice of test-statistic I

We've introduced the idea of test-statistic, but we haven't picked one

- It should be consistent — in the limit of large data, it should distinguish correctly between theories
- It shouldn't be biased — *e.g.*, the probability that we accept a hypothesis if it's false should be less than the probability that we accept it if it's true
- There is a trade-off between type-1 and type-2 errors. Decreasing type-1 errors increases type-2 errors
- What's the "best" test-statistic that for a given confidence level, minimizes type-2 errors?

Choice of test-statistic II

- It can be proved that the Neyman-Pearson test-statistic is optimal:

$$\lambda = \frac{p(d|H_0)}{p(d|H_1)} \tag{33}$$

  if we are comparing simple hypotheses H_0 and H_1

Composite hypothesis I

- What about composite hypotheses? Previous result doesn't apply?
- Although it isn't proven to be optimal, we pick a likelihood ratio, and profile the models' free parameters

$$\lambda = \frac{p(d|\hat{x}, H_0)}{p(d|\hat{y}, H_1)} \tag{34}$$

- This is also pragmatic
- If the models are nested (such as a model with Higgs that reduces to a model without a Higgs if the coupling is switched off), we can use Wilk's theorem to find the distribution of $\ln\lambda$
- Wilk's theorem $-2\ln\lambda \sim \chi^2$-distribution
- Distribution of test-statistic also often found by Monte-Carlo

Back to the Beginning. . .

> *This observation, which has a significance of 5.9 standard deviations, corresponding to a background fluctuation probability of 1.7×10^{-9}, is compatible with the production and decay of the Standard Model Higgs boson*

- This statement should now be more clear
- Probability of obtaining such a large test-statistic (such extreme data) under null hypothesis is 1.7×10^{-9}
- With a one-tail Gaussian convention, the significance of that is 5.9σ

Look-elsewhere effect (LEE) I

- If we modify our choice/calculation of test statistic after seeing the data, we alter the distribution of the test statistic
- We are often biased — try to maximize the significance of experimental anomalies. We might:
  - Pick data that maximizes the significance. LEE in the data
  - Pick a test statistic (*e.g.*, pick a particular m_h) that maximizes the significance. LEE in the theory
  - As a consequence, we find a small p-value
- These are local significance's. Nothing wrong with this, as long as you are honest

Look-elsewhere effect (LEE) II

- The global significance reflects the whole procedure — including the fact that you looked at the most significant experiment or test statistic.
- The p-value is always the probability of observing such an extreme test statistic in the null hypothesis
- With local p-values, you are not including all information about the calculation of the test statistic

Look-elsewhere effect (LEE) I

In the data

- I sample one hundred numbers from a black box that I'm told is a standard normal distribution
- I look through my data find one number in 99% tail — I report this, and reject the idea that black box is a standard normal distribution at 99% confidence
- This is madness! I expected to find such a discrepant result in 100 samples. That was a local significance for that particular sample
- I didn't include all information about how my data was chosen

Look-elsewhere effect (LEE) II

In the data

- I should report the global p-value (by *e.g.*, calculating a χ^2 and comparing it with a χ^2-distribution with 100 degrees of freedom)

Look-elsewhere effect (LEE) I

In the theory

- My test statistic is

$$\lambda(m_h) = \ln \frac{\mathcal{L}(D|m_h, \mu = 1)}{\mathcal{L}(D|\mu = 0)} \tag{35}$$

- I haven't specified the Higgs mass
- After seeing the data, I look at $m'_h \equiv \hat{m}_h$, which is such that $\lambda(\hat{m}_h)$ is a maximum. I then forget about how m'_h was chosen
- I report the p-value associated with $\lambda(m'_h)$ (without including how m'_h was chosen)

Look-elsewhere effect (LEE) II

In the theory

- This is a local p-value — it didn't include all information about how test-statistic was chosen
- In hypothetical pseudo-experiments, we would have looked at different $\lambda(m_h)$
- The test statistic was in effect:

$$\hat{\lambda} = \max_{m_h} \lambda(m_h) = \lambda(\hat{m}_h) \tag{36}$$

- Although $\hat{\lambda} = \lambda(\hat{m}_h) = \lambda(m'_h)$, they don't have the same distribution in hypothetical pseudo-experiments and thus don't correspond to the same p-value

Look-elsewhere effect (LEE) III

In the theory

- $\hat{m}_h$ would vary in pseudo-experiments; m_h' wouldn't
- $\hat{\lambda}$ and its distribution give the global p-value, and reflect all information about how test-statistic was chosen

Exercises 1

The starred exercises are trickier or probably require a computer. Focus on the unstarred exercises, but think about how you might tackle the starred exercises. Write down expressions even if you cannot numerically evaluate them without a computer.

Please complete all exercises by next week.

Confident about confidence intervals

1. John hears that the Higgs mass is between 124 GeV and 127 GeV with 95% confidence. "Wow! They know with 95% probability that the mass of the Higgs is between those numbers!", he exclaims. Correct his mistake.

Exercises II

2. Which interval is wider? a 68% interval? or a 95% interval?
3. Are confidence intervals parameterization invariant? *i.e.*, if $y = f(x)$ is $f_u = f(x_u)$?

Tedious example. The lifetime of light-bulbs is known to be approximately normally distributed with $\sigma = 0.1$ years but unknown mean, μ. A random sample of four bulbs last 1.14, 0.963, 0.958 and 1.12 years.

4. Write down the expression for the log-likelihood ratio test-statistic
5. Calculate the MLE for the mean life-time of a bulb

Exercises III

*6. Compute the 68% confidence interval on the mean using the profile likelihood method. You are encouraged to use a computer

Error checking (all fictional)

7. The probability distribution of blood alcohol level reported by a police breathalyzer is $\mathcal{N}(x + 0.1, 0.1^2)$, where x is the amount of alcohol consumed. The police stop Fred in a random check for drink driving. Their threshold for arrest is 0.3. He hasn't been drinking ($x = 0$). What is the probability that he is arrested?
8. The police stop Graham. He has been drinking, $x = 2$. What is the probability that he isn't arrested?

Exercises IV

9. Explain the previous answers in terms of type-1 and type-2 errors in hypothesis testing, include a sketch of the areas under the PDFs.

Quick hypothesis test

10. A manufacturer claims that the mass of flour in a bag is normally distributed with $\mu = 100$ g, $\sigma = 10$ g. I randomly buy a bag a flour and find that it contains only 50 g.
11. What's the probability of obtaining such extreme data, if the manufacturer's claims are true?
12. Should we reject the claims at 3σ?

Exercises V

Higgs testing

13. Check that 1.7×10^{-9} corresponds to 5.9σ, as claimed
14. Read the statistical procedure in Aad, G. et al. (2012a). Combined search for the Standard Model Higgs boson in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector. *Phys.Rev. D86*, 032003. doi:10.1103/PhysRevD.86.032003. arXiv: 1207.0319 [hep-ex]. Write a brief (a few sentences) summary.

Why 5σ?

Exercises VI

15. Read Lyons, L. [Louis]. (2013). Discovering the Significance of 5 sigma. arXiv: 1310.1284 [physics.data-an]. Write a brief (a few sentences) summary.

References I

Aad, G. et al. (2012a). Combined search for the Standard Model Higgs boson in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector. *Phys.Rev. D86*, 032003. doi:10.1103/PhysRevD.86.032003. arXiv: 1207.0319 [hep-ex]

Aad, G. et al. (2012b). Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys.Lett. B716*, 1–29. doi:10.1016/j.physletb.2012.08.020. arXiv: 1207.7214 [hep-ex]

Arnison, G. et al. (1983). Experimental Observation of Isolated Large Transverse Energy Electrons with Associated Missing Energy at $\sqrt{s} = 540$ GeV. *Phys.Lett. B122*, 103–116. doi:10.1016/0370-2693(83)91177-2

References II

Chatrchyan, S. et al. (2014). Measurement of the properties of a Higgs boson in the four-lepton final state. *Phys.Rev. D89*(9), 092007. doi:10.1103/PhysRevD.89.092007. arXiv: 1312.5353 [hep-ex]

Hájek, A. (2012). Interpretations of probability. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Winter 2012).

James, F. (2006). *Statistical methods in experimental physics* (Second). World Scientific.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. (G. L. Bretthorst, Ed.). Cambridge University Press.

Lyons, L. [L.]. (1989). *Statistics for nuclear and particle physicists*. Cambridge University Press.

Lyons, L. [Louis]. (2013). Discovering the Significance of 5 sigma. arXiv: 1310.1284 [physics.data-an]