Statistics & (a bit of) Machine Learning

W. Verkerke
Statistics & Machine Learning

- A really vast topic on which you could spend an entire week lecturing
- Have only 4 hours - so will make some selection of topics here
- Will mostly focus on statistics methods and model building – with a modest excursion into machine learning
- General idea of the course
  - start with simple models – focus on fundamental concepts for those (p-values, bayes vs frequentist)
  - then gradually make models more complex and look at how statistical procedures deal with these (nuisance parameters, systematic uncertainties), but also look on the practical side for physicists – do you understand what happens and how can you debug and validate your complex fits
  - Excursion in multivariate methods and machine learning, when discussing multi-dimensional models (nice connection via NP lemma)
What do we want to know?

- Physics questions we have…
  - Does the (SM) Higgs boson exist?
  - What is its production cross-section?
  - What is its boson mass?

- Statistical tests construct probabilistic statements: $p(\text{theo}|\text{data})$, or $p(\text{data}|\text{theo})$
  - Hypothesis testing (discovery)
  - (Confidence) intervals
  - Measurements & uncertainties

- Result: *Decision* based on tests
  
  "As a layman I would now say: I think we have it"
How do we do this?

• Statistics: if you know distribution \( f(x|\mu, \theta) \) for your observable(s) \( x \) in terms of your parameter of interest \( \mu \) (and other parameters \( \theta \)) then in principle solvable problem
  – In other words if \( f(x|\mu, \theta) \) is known then problem is ‘simple’ (‘just’ follow prescription of statistical procedures)

• Particle physics: connection of theory (SM or its extension) with your parameter \( \mu \) is highly non-trivially connected to your observables \( x \).

• Relation between \( x \) and \( \mu \) can in almost all cases not be analytically formulated, but distribution \( f(x|\mu, \theta) \) can be sampled through (chain of simulation packages)

Root of many of the complexities of HEP data analysis

• Simulation-based knowledge of \( f(x|\mu, \theta) \) often approximate, often in ways that cannot be exactly quantified

Often the really thorny problems (‘theory systematics’) – root of problem is not statistical in nature, but must somehow be accounted for in statistical procedures
Particle physics data analysis – chain of simulation steps

Simulation of ‘soft physics’ physics process

Simulation of high-energy physics process

Simulation of ATLAS detector

LHC data

Analysis Event selection

Reconstruction of ATLAS detector

Proton Structure Function

\[ P(m_{4l}|\text{SM}[m_H]) \]

Observed \( m_{4l} \)

\[ \text{prob(data|SM)} \]
Statistical analysis with HEP data – a multi-step approach

To make HEP data analysis tractable with only samples of $f(x|\mu,\theta)$ known (instead of function itself), inference of $\mu$ from $f(x|\mu,\theta)$ often performed as a multi-step process

1. ‘Event selection’
   - Classification of events according to approximate signal purity using only MC samples of $f(x|\mu,\theta)$

2. ‘reduced sample $D(x')$’ where only a subset of observables is present, or original observables have been replaced with a ‘summary observable’ (e.g. an MVA-based signal purity score)

3. ‘Statistical analysis’
   - Analysis of the reduced data sample for which a probability model $f'(x'|\mu,\theta)$ can be formulated

Wouter Verkerke, NIKHEF
Roadmap of this course

- Start with basics, gradually build up to complexity

**Model building**
- Counting models
- Modeling distributions
- Signal parameterization strategies
- Models with nuisance parameters, joint models, modeling systematic uncertainties
- Diagnosing inference on complex models

**Statistical methods**
- Statistical tests with counting experiments
- Test statistics for models describing distributions
- Parameter estimation, confidence intervals & limits
- Inference with nuisance parameters
Roadmap of this course

- Start with basics, gradually build up to complexity

Model building

- Counting models
- Modeling distributions
- Signal parameterization strategies
- Models with nuisance parameters, joint models, modeling systematic uncertainties
- Diagnosing inference on complex models

Statistical methods

- Statistical tests with counting experiments
- Supervised Machine Learning (sig/bkg)
- Parameter estimation, confidence intervals & limits
- Inference with nuisance parameters

Day 1

Day 2
Roadmap of this course

• Start with basics, gradually build up to complexity

Model building
- Counting models
- Modeling distributions
- Signal parameterization strategies
- Models with nuisance parameters, joint models, modeling systematic uncertainties
- Diagnosing inference on complex models

Statistical methods
- Statistical tests with counting experiments
- Test statistics for models describing distributions
- Parameter estimation, confidence intervals & limits
- Inference with nuisance parameters
Counting models

- Central concept in statistics is the ‘probability model’
- A probability model assigns a probability to each possible experimental outcome.
- Example: a HEP counting experiment
  - Count number of ‘events’ in a fixed time interval → Poisson distribution
  - Given the expected event count, the probability model is fully specified

$$P(N | \mu) = \frac{\mu^N e^{-\mu}}{N!}$$

\(\mu=3\) (“bkg only”) \hspace{1cm} \(\mu=7\) (“bkg+signal”)
Probabilities vs conditional probabilities

• Note that probability models strictly give conditional probabilities (with the condition being that the underlying hypothesis is true)

\[ P(N) \rightarrow P(N \mid H_{bkg}) \quad P(N) \rightarrow P(N \mid H_{\text{sig+bkg}}) \]

• Suppose we measure \( N=7 \) then can calculate

\[ L(N=7 \mid H_{bkg})=2.2\% \quad L(N=7 \mid H_{\text{sig+bkg}})=14.9\% \]

• Data is more likely under sig+bkg hypothesis than bkg-only hypo

• Is this what we want to know? Or do we want to know \( L(H_{s+b} \mid N=7) \)?

Wouter Verkerke, NIKHEF
Inverting the conditionality on probabilities

- Do $L(7|H_b)$ and $L(7|H_{sb})$ provide you enough information to calculate $P(H_b|7)$ and $P(H_{sb}|7)$
- No!
- Image the ‘whole space’ and two subsets $A$ and $B$

\[ P(A|B) \neq P(B|A) \]

\[ P(7|H_b) \neq P(H_b|7) \]
Inverting the conditionality on probabilities

\[ P(A|B) \neq P(B|A) \]

but you can deduce their relation

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

Wouter Verkerke, NIKHEF
Inverting the conditionality on probabilities

- This conditionality inversion relation is known as **Bayes Theorem**

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

*Essay “Essay Towards Solving a Problem in the Doctrine of Chances” published in Philosophical Transactions of the Royal Society of London in 1764*

- And choosing \( A = \text{data} \) and \( B = \text{theory} \)

\[ P(\text{theo}|\text{data}) = P(\text{data}|\text{theo}) \times P(\text{theo}) / P(\text{data}) \]

- **Return to original question:**

Do you \( L(7|H_b) \) and \( L(7|H_{sb}) \) provide you enough information to calculate \( P(H_b|7) \) and \( P(H_{sb}|7) \)

- **No! \( \rightarrow \)** Need \( P(A) \) and \( P(B) \) \( \rightarrow \) Need \( P(H_b) \), \( P(H_{sb}) \) and \( P(7) \)

Wouter Verkerke, NIKHEF
Inverting the conditionality on probabilities

- **What is \( P(\text{data}) \)?**
  
  \[ P(\text{theo}|\text{data}) = \frac{P(\text{data}|\text{theo}) \times P(\text{theo})}{P(\text{data})} \]

- It is the probability of the data under *any* hypothesis
  
  - For example for two competing hypotheses \( H_b \) and \( H_{sb} \)
    
    \[ P(N) = L(N|H_b)P(H_b) + L(N|H_{sb})P(H_{sb}) \]
    
    and generally for \( N \) hypotheses
    
    \[ P(N) = \sum_i P(N|H_i)P(H_i) \]

- Bayes theorem reformulated using law of total probability
  
  \[ P(\text{theo}|\text{data}) = \frac{L(\text{data}|\text{theo}) \times P(\text{theo})}{\sum_i L(\text{data}|\text{theo}-i)P(\text{theo}-i)} \]

- *Return to original question:* Do you \( L(7|H_b) \) and \( L(7|H_{sb}) \) provide you enough information to calculate \( P(H_b|7) \) and \( P(H_{sb}|7) \)?
  
  *No!* → Still need \( P(H_b) \) and \( P(H_{sb}) \)
Prior probabilities

• What is the **meaning** of $P(H_b)$ and $P(H_{sb})$?
  – They are the probability assigned to hypothesis $H_b$ *prior to the experiment*.

• What are the **values** of $P(H_b)$ and $P(H_{sb})$?
  – Can be result of an earlier measurement
  – Or more generally (e.g. when there are no prior measurement)
    they quantify a *prior degree of belief* in the hypothesis

• **Example** – suppose prior belief $P(H_{sb})=50\%$ and $P(H_b)=50\%$

$$P(H_{sb}|N=7) = \frac{P(N=7|H_{sb}) \times P(H_{sb})}{[P(N=7|H_{sb})P(H_{sb})+P(N=7|H_b)P(H_b)]}$$

$$= \frac{0.149 \times 0.50}{[0.149 \times 0.5 + 0.022 \times 0.5]} = 87\%$$

• Observation $N=7$ strengthens belief in hypothesis $H_{sb}$
  (and weakens belief in $H_b \rightarrow 13\%$)
Interpreting probabilities

- We have seen

probabilities assigned observed experimental outcomes
(probability to observed 7 events under some hypothesis)

probabilities assigned to hypotheses
(prior probability for hypothesis $H_{sb}$ is 50%)

which are conceptually different.

- How to interpret probabilities – two schools

  **Bayesian probability** = (subjective) degree of belief
  $P(\text{data}|\text{theo})$

  **Frequentist probability** = fraction of outcomes in
  future repeated identical experiments
  $P(\text{theo}|\text{data})$  

  “If you’d repeat this experiment identically many times,
in a fraction $P$ you will observe the same outcome”
Interpreting probabilities

• **Frequentist:**
  Constants of nature are fixed – you cannot assign a probability to these. Probability are restricted to observable experimental results
  – “The Higgs either exists, or it doesn’t” – you can’t assign a probability to that
  – Definition of $P(\text{data}|\text{hypo})$ is objective (and technical)

• **Bayesian:**
  Probabilities can be assigned to constants of nature
  – Quantify your *belief* in the existence of the Higgs – can assign a probability
  – But is can very difficult to assign a meaningful number (e.g. Higgs)

• **Example of weather forecast**

  Bayesian: “*The probability it will rain tomorrow is 95%*”
  – Assigns probability to constant of nature (“rain tomorrow”)
    $P(\text{rain-tomorrow}|\text{satellite-data}) = 95\%$

  Frequentist: “*If it rains tomorrow, 95% of time satellite data looks like what we observe now*”
  – Only states $P(\text{satellite-data}|\text{rain-tomorrow})$
Back to H\textsubscript{b}/H\textsubscript{sb} - Formulating evidence for discovery of H\textsubscript{sb}

- Given a scenario with exactly two competing hypotheses
- In the Bayesian school you can cast evidence as an odd-ratio

\[
O_{\text{prior}} \equiv \frac{P(H_{sb})}{P(H_b)} = \frac{P(H_{sb})}{1 - P(H_{sb})} \quad \text{If } p(H_{sb}) = p(H_b) \rightarrow \text{Odds are 1:1}
\]

\[
O_{\text{posterior}} \equiv \frac{L(x \mid H_{sb})P(H_{sb})}{L(x \mid H_b)P(H_b)} = \frac{L(x \mid H_{sb})}{L(x \mid H_b)} O_{\text{prior}}
\]

If \( P(\text{data} \mid H_b) = 10^{-7} \)
\( P(\text{data} \mid H_{sb}) = 0.5 \)
\( K = 2.000.000 \rightarrow \text{Posterior odds are } 2.000.000 : 1 \)
Formulating evidence for discovery

• In the frequentist school you restrict yourself to $P(\text{data}|\text{theory})$ and there is no concept of ‘priors’
  - But given that you consider (exactly) 2 competing hypothesis, very low probability for data under $H_b$ lends credence to ‘discovery’ of $H_{sb}$ (since $H_b$ is ‘ruled out’). Example

  $P(\text{data}|H_b)=10^{-7}$
  $P(\text{data}|H_{sb})=0.5$

  “$H_b$ ruled out” $\rightarrow$ “Discovery of $H_{sb}$”

• Given importance to interpretation of the lower probability, it is customary to quote it in “physics intuitive” form: Gaussian $\sigma$.
  - E.g. ‘5 sigma’ $\rightarrow$ probability of 5 sigma Gaussian fluctuation = $2.87 \times 10^{-7}$

• No formal rules for ‘discovery threshold’
  - Discovery also assumes data is not too unlikely under $H_{sb}$. If not, no discovery, but again no formal rules (“your good physics judgment”)
  - NB: In Bayesian case, both likelihoods low $\rightarrow$ reduces Bayes factor $K$ to $O(1)$

Wouter Verkerke, NIKHEF
Taking decisions based on your result

- What are you going to do with the results of your measurement?
- Usually basis for a *decision*
  - **Science**: declare discovery of Higgs boson (or not), make press release, write new grant proposal
  - **Finance**: buy stocks or sell
- Suppose you believe $P(\text{Higgs}|\text{data})=99\%$.
- **Should** declare discovery, make a press release?
  *A: Cannot be determined from the given information!*
- Need in addition: the utility function (or cost function),
  - The cost function specifies the relative costs (to You) of a Type I error (declaring model false when it is true) and a Type II error (not declaring model false when it is false).
Taking decisions based on your result

• Thus, your *decision*, such as where to invest your time or money, requires two subjective inputs:

  Your prior probabilities, and

  the relative costs to You of outcomes.

• Statisticians often focus on decision-making; in HEP, the tradition thus far is to communicate experimental results (well) short of formal decision calculations.

• Costs can be difficult to quantify in science.
  – What is the cost of declaring a false discovery?
  – Can be high (“Fleischman and Pons”), but hard to quantify
  – What is the cost of missing a discovery (“Nobel prize to someone else”), but also hard to quantify
Summary on statistical test with simple hypotheses

• So far we considered simplest possible experiment we can do: counting experiment

• For a set of 2 or more completely specified (i.e. simple) hypotheses

  \[ P(N|\text{bkg}), \quad P(N|\text{sig}) \]

  \[ P(N_{\text{obs}}|H_x) \]

  Given probability models \( P(N|\text{bkg}) \), and \( P(N|\text{sig}) \)

  we can calculate \( P(N_{\text{obs}}|H_x) \) under either hypothesis

  → With additional information on \( P(H_i) \) we can also calculate \( P(H_x|N_{\text{obs}}) \)

• In principle, *any potentially complex measurement (for Higgs, SUSY, top quarks) can ultimately take this a simple form.* But there is some ‘pre-work’ to get here – examining (multivariate) discriminating distributions → Now try to incorporate that

Wouter Verkerke, NIKHEF
Model building 2

Modelling distributions –
template based models or
analytical models
Roadmap of this course

- Start with basics, gradually build up to complexity

**Model building**
- Counting models
- Modeling distributions
- Signal parameterization strategies
- Models with nuisance parameters, joint models, modeling systematic uncertainties
- Diagnosing inference on complex models
- Advanced signal modeling techniques

**Statistical methods**
- Statistical tests with counting experiments
- Test statistics for models describing distributions
- Parameter estimation, confidence intervals & limits
- Inference with nuisance parameters
Discriminating observables & counting experiments

- HEP experimental data usually has many discriminating observables that carry information that can distinguish signal from background hypothesis.

- In principle can use them all directly in an elaborate hypothesis test.
  - But would need to formulate a model that describe the expected distribution of all of these \( \rightarrow \) Complicated.
  - If expectations are uncertain (from simulation or theory) process of modeling becomes even more complex.

- A pragmatic solution to reduce complexity is to split task in two
  - Define empirical selection of events enriched in signal using one or more observable properties of the event (invariant masses, distributions, angles etc).
  - Perform statistical test (hypothesis test, parameter estimation etc) on sample that reduced in size and in dimensionality of discriminating observables that are modeled.
  - Most extreme reduction of dimensionality is to zero \( \rightarrow \) counting experiment.
Example 1 – Discrimination in selection stage only

NB1: All discriminating power in selection step, none in inference step. *This is a design choice!*

NB2: Selection must be tuned on a ‘figure of merit’ usually a simplified statistical inference test

**Event selection:** reduce sample size and dimensionality

**Formulation of probability model of reduced sample:** 
$\text{Poisson}(N|s+b)$

**Statistical inference:**
$L(15|5) = 1.5 \times 10^{-4}$
Modeling discriminating observables

- **Example 2 – Discrimination in inference stage**

Event selection: reduce sample size and dimensionality

Formulation of probability model of reduced sample:

\[ N_{bkg} \times \text{Uniform}(x) + N_{\text{sig}} \times \text{Gaussian}(x) \]

Statistical inference:

\[ L(\text{data}|\text{hypo}) = \text{something} \]

- **NB1**: Most discrimination power in inference step. *This is again design choice!*
- **NB2**: Optimal selection less critical
- **NB3**: Correct description of selected sample more complex
Modeling discriminating observables

- Example 2 – full dataset has one discriminating observable: $x$

Formulation of probability model of reduced sample:

$$\text{Nbkg} \times \text{Uniform}(x) + \text{Nsig} \times \text{Gaussian}(x)$$

Statistical inference:

$$L(\text{data}|\text{hypo}) = \text{something}$$

Q: Which strategy is better?
A: Depends on how ‘better’ is defined?

For hypothesis testing ‘discovery of a new particle’
the ‘power’ of the test can be the same, but doesn’t need to be

Choice is real life largely dictated by practicalities
- How easy is it to formulate a description of the observables?
- How many observables are important?

NB1: Most discrimination power in inference step. This is again design choice!

NB2: Optimal selection less critical

NB3: Correct description of selected sample more complex
Formulating probability models for discriminating observables

- For **counting experiments** could derive \( \text{Poisson}(N|\mu) \) from first principles (‘random discrete events measured in fixed time interval)

- For **experiments with discriminating observables**, description should ideally **also derive from underlying (physics) hypothesis/theory**
  - In many cases this is possible, but not always without assumptions.
  - Assumptions lead to uncertainties in predictions \( \rightarrow \) we’ll revisit later how to deal with those.

- Example: common underlying principle in (signal) model is that discriminating observable is sum/average of many components
  - E.g. light collected by photomultiplier has contributions from \( \gg 1 \) photons
  - Tracks reconstructed in detector have contributions \( \gg 1 \) hits
  - Central Limit Theorem: for large \( N \) \( \rightarrow \) Can be analytically described by Gaussian

- In case there is no easy analytical solution \( \rightarrow \) empirical models (polynomial) or numerical solution (simulation-based histogram)
Empirical probability models

- In case no description from first principles exists for a differential distribution, empirical or simulation-based models can be deployed.

**Empirical models**

\[ B(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 \ldots \]

**Simulation-based models**

\[ B(x) = \text{histogram} \]

**Drawbacks:**
- **Arbitrariness in parameterization**, e.g. which order to choose for a polynomial
- **Quantization** of model prediction in bins
- **Poor modeling in regions** with low simulation statistics
Modeling low-statistics simulation predictions

- For low-statistics simulation predictions, **kernel estimation techniques** can improve modeling substantially.

- Procedure:
  - Assign a **Gaussian probability density distribution** to each simulated event.
  - Sum Gaussian probability densities of all events.
  - Started from unbinned data \(\rightarrow\) no binning effects.

Wouter Verkerke, NIKHEF
Modeling low-statistics simulation predictions

- Technique does *not* require that all Gaussian kernels have same width
- Improved procedure: ‘adaptive kernel’
  - Adjust width of Gaussian kernels depending on local event density
  - High density $\rightarrow$ narrow kernels $\rightarrow$ preserve more detail
  - Low density $\rightarrow$ wide kernels $\rightarrow$ promote smoothness
Statistical methods 2

Adapting statistical methods to use with distributions: test statistics as ordering principle, likelihood ratios, contrast with Bayesian methods, the likelihood principle. Practical aspects of toy MC sampling
Roadmap of this course

- Start with basics, gradually build up to complexity

**Model building**
- Counting models
- Modeling distributions
- Signal parameterization strategies
- Models with nuisance parameters, joint models, modeling systematic uncertainties
- Diagnosing inference on complex models
- Advanced signal modeling techniques

**Statistical methods**
- Statistical tests with counting experiments
- Test statistics for models describing distributions
- Parameter estimation, confidence intervals & limits
- Inference with nuisance parameters
Working with Likelihood functions for distributions

- How do the statistical inference procedures change for Likelihoods describing distributions?
- Bayesian calculation of $P(\text{theo}|\text{data})$ they are exactly the same.
  - Simply substitute counting model with binned distribution model

\[
P(H_{s+b} \mid \vec{N}) = \frac{L(\vec{N} \mid H_{s+b})P(H_{s+b})}{L(\vec{N} \mid H_{s+b})P(H_{s+b}) + L(\vec{N} \mid H_b)P(H_b)}
\]

Simply fill in new Likelihood function
Calculation otherwise unchanged

\[
P(H_{s+b} \mid \vec{N}) = \prod_i \text{Poisson}(N_i \mid \tilde{s}_i + \tilde{b}_i)P(H_{s+b})
\]

\[
= \frac{\prod_i \text{Poisson}(N_i \mid \tilde{s}_i + \tilde{b}_i)P(H_{s+b})}{\prod_i \text{Poisson}(N_i \mid \tilde{s}_i + \tilde{b}_i)P(H_{s+b}) + \prod_i \text{Poisson}(N_i \mid \tilde{b}_i)P(H_b)}
\]

Wouter Verkerke, NIKHEF
Working with Likelihood functions for distributions

- Frequentist calculation of $P(\text{data}|\text{hypo})$ also unchanged, but question arises if $P(\text{data}|\text{hypo})$ is still relevant?

- $L(N|H)$ is probability to obtain exactly the histogram observed.

- $L(\tilde{N} | H_b) = \prod_i \text{Poisson}(N_i | \tilde{b}_i)$

- $L(\tilde{N} | H_{s+b}) = \prod_i \text{Poisson}(N_i | \tilde{s}_i + \tilde{b}_i)$

- $L(N|H)$ is probability to obtain exactly the histogram observed.

- *Is that what we want to know?* Not really.. We are interested in probability to observe any ‘similar’ dataset to given dataset, or in practice dataset ‘similar or more extreme’ that observed data

- Need a way to quantify ‘similarity’ or ‘extremity’ of observed data

Wouter Verkerke, NIKHEF
Working with Likelihood functions for distributions

- **Definition**: a test statistic $T(x)$ is *any* function of the data $x$
- We need a test statistic that will **classify** (‘order’) all possible observations in terms of ‘extremity’ (definition to be chosen by physicist)
- **NB**: For a counting measurement the count itself is already a useful test statistic for such an ordering (i.e. $T(x) = x$)

Test statistic $T(N) = N_{\text{obs}}$ orders observed events count by estimated signal yield

Low $N \rightarrow$ low estimated signal
High $N \rightarrow$ large estimated signal
P-values for counting experiments

- Now make a measurement $N = N_{\text{obs}}$ (example $N_{\text{obs}} = 7$)

- **Definition: p-value:** probability to obtain the observed data, or more extreme in future repeated identical experiments
  - Example: p-value for background-only hypothesis

\[
p_b = \int_{N_{\text{obs}}}^{\infty} \text{Poisson}(N; b + 0) dN
\]

(= 0.23)
Ordering distributions by ‘signal-likeness’ aka ‘extremity’

- How to define ‘extremity’ if observed data is a distribution

Observation

\[ N_{\text{obs}} = 7 \]

Median expected by hypothesis

\[ N_{\text{exp}}(s=0) = 5 \]
\[ N_{\text{exp}}(s=5) = 10 \]

Predicted distribution of observables

Which histogram is more ‘extreme’?
The Likelihood Ratio as a test statistic

- Given two hypotheses $H_b$ and $H_{s+b}$, the ratio of likelihoods is a useful test statistic

$$\lambda(\vec{N}) = \frac{L(\vec{N} | H_{s+b})}{L(\vec{N} | H_b)}$$

- Intuitive picture:

  $\rightarrow$ If data is likely under $H_b$, $L(\vec{N} | H_b)$ is **large**, $L(\vec{N} | H_{s+b})$ is smaller

  $\lambda(\vec{N}) = \frac{\text{small}}{\text{large}} = \text{small}$

  $\rightarrow$ If data is likely under $H_{s+b}$, $L(\vec{N} | H_{s+b})$ is **large**, $L(\vec{N} | H_b)$ is smaller

  $\lambda(\vec{N}) = \frac{\text{large}}{\text{small}} = \text{large}$
Visualizing the Likelihood Ratio as ordering principle

• The Likelihood ratio as ordering principle

- Frequentist solution to ‘relevance of P(data|theory’) is to order all observed data samples using a (Likelihood Ratio) test statistic
  - Probability to observe ‘similar data or more extreme’ then amounts to calculating ‘probability to observe test statistic $\lambda(N)$ as large or larger than the observed test statistic $\lambda(N_{\text{obs}})$

\[ L(N|H_{s+b}) = \text{small} \quad L(N|H_b) = \text{large} \]
\[ L(N|H_{s+b}) = \text{soso} \quad L(N|H_b) = \text{soso} \]
\[ L(N|H_{s+b}) = \text{large} \quad L(N|H_b) = \text{small} \]
\[ \lambda(N) = 0.0005 \quad \lambda(N) = 0.47 \quad \lambda(N) = 5000 \]
The distribution of the test statistic

- Distribution of a test statistic is *generally not known*
- Use toy MC approach to approximate distribution
  - Generate many toy datasets \( N \) under \( H_b \) and \( H_{s+b} \) and evaluate \( \lambda(N) \) for each dataset

\[
p - value = \int_{\lambda_{obs}}^{\infty} f(\lambda \mid H_b) \ log(\lambda)\]

Wouter Verkerke, NIKHEF
The distribution of the test statistic

- **Definition: p-value:**
  probability to obtain the observed data, or more extreme in future repeated identical experiments
  (extremity define in the precise sense of the (LR) ordering rule)

\[
p - value = \int_{\lambda_{\text{obs}}}^{\infty} f(\lambda | H_b) \, d\lambda
\]
Likelihoods for distributions - summary

• Bayesian inference unchanged

→ simply insert $L$ of distribution to calculate $P(H|data)$

$$P(H_{s+b} \mid \bar{N}) = \frac{L(\bar{N} \mid H_{s+b}) P(H_{s+b})}{L(\bar{N} \mid H_{s+b}) P(H_{s+b}) + L(\bar{N} \mid H_b) P(H_b)}$$

• Frequentist inference procedure *modified*

→ Pure $P(data|hypo)$ not useful for non-counting data
→ Order all possible data with a (LR) test statistic in ‘extremity’
→ Quote $p(data|hypo)$ as ‘p-value’ for hypothesis

  Probability to obtain observed data, or more extreme, is $X\%$

‘Probability to obtain 13 or more 4-lepton events under the no-Higgs hypothesis is $10^{-7}$’

‘Probability to obtain 13 or more 4-lepton events under the SM Higgs hypothesis is 50%’
The likelihood principle

- Note that ‘ordering procedure’ introduced by test statistic also has a profound implication on interpretation
- Bayesian inference only uses the Likelihood of the observed data

\[
P(H_{s+b} \mid \vec{N}) = \frac{L(\vec{N} \mid H_{s+b})P(H_{s+b})}{L(\vec{N} \mid H_{s+b})P(H_{s+b}) + L(\vec{N} \mid H_b)P(H_b)}
\]

- While the observed Likelihood Ratio also only uses likelihood of observed data.

\[
\lambda(\vec{N}) = \frac{L(\vec{N} \mid H_{s+b})}{L(\vec{N} \mid H_b)}
\]

- Distribution \( f(\lambda \mid N) \), and thus \( p \)-value, also uses likelihood of non-observed outcomes (in fact Likelihood of every possible outcome is used)
Likelihood Principle

• In Bayesian methods and likelihood-ratio based methods, the probability (density) for obtaining the data at hand is used (via the likelihood function), but probabilities for obtaining other data are not used!

• In contrast, in typical frequentist calculations (e.g., a p-value which is the probability of obtaining a value as extreme or more extreme than that observed), one uses probabilities of data not seen.

• This difference is captured by the Likelihood Principle*:

If two experiments yield likelihood functions which are proportional, then Your inferences from the two experiments should be identical.

Wouter Verkerke, NIKHEF
Generalizing to multiple dimensions

- Can also generalize likelihood models to distributions in *multiple* observables

\[
L(\bar{x}) = \prod f(x_i)
\]

\[
L(\bar{x}, \bar{y}) = \prod f(x_i, y_i)
\]

- Neither generalization (binned→continuous, one→multiple observables) has any further consequences for Bayesian or Frequentist inference procedures
The Likelihood Ratio test statistic as tool for event selection

- Note that hypothesis testing with two simple hypotheses for observable distributions, exactly describes ‘event selection’ problem

- In fact we have already ‘solved’ the optimal event selection problem! Given two hypothesis \(H_{s+b}\) and \(H_b\) that predict an complex multivariate distribution of observables, you can always classify all events in terms of ‘signal-likeness’ (a.k.a ‘extremity’) with a likelihood ratio

\[
\lambda(x, y, z, \ldots) = \frac{L(x, y, z, \ldots \mid H_{s+b})}{L(x, y, z, \ldots \mid H_b)}
\]

- So far we have exploited \(\lambda\) to calculate a frequentist p-value now explore properties ‘cut on \(\lambda\)’ as basis of (optimal) event selection
Roadmap of this course

- Start with basics, gradually build up to complexity

**Model building**
- Counting models
- Modeling distributions
- Signal parameterization strategies
- Models with nuisance parameters, joint models, modeling systematic uncertainties
- Diagnosing inference on complex models
- Advanced signal modeling techniques

**Statistical methods**
- Statistical tests with counting experiments
- Test statistics for models describing distributions
- Relation of test statistics to event selection
- Parameter estimation, confidence intervals & limits
- Inference with nuisance parameters
Deciding on a split

- HEP data analysis often a 2-step process:
  - first selection, then inference

- Focus in this course on inference, but Likelihood Ratio as test statistics shows that there is a general optimal solution for any event selection problem: the ratio will order all events by signal-likeness

\[
\lambda(\bar{x}, \bar{y}, \bar{z}, \ldots) = \frac{L(\bar{x}, \bar{y}, \bar{z}, \ldots | H_{s+b})}{L(\bar{x}, \bar{y}, \bar{z}, \ldots | H_b)}
\]

- Hence if we can construct \(\lambda\), a selection defined by \(\lambda > \lambda_c\) will always be optimal for some stated level of desired purity

Wouter Verkerke, NIKHEF
Event selection

• The event selection problem:
  – Input: Two classes of events “signal” and “background”
  – Output: Two categories of events “selected” and “rejected”

• Goal: select as many signal events as possible, reject as many background events as possible

• Note that optimization goal as stated is ambiguous.
  – But can choose a well-defined by optimization goal by e.g. fixing desired background acceptance rate, and then choose procedure that has highest signal acceptance.

• Relates to “classical hypothesis testing”
  – Two competing hypothesis (traditionally named ‘null’ and ‘alternate’)
  – Here null = background, alternate = signal
Terminology of classical hypothesis testing

- **Definition of terms**
  - Rate of type-I error = $\alpha$
  - Rate of type-II error = $\beta$
  - Power of test is $1 - \beta$

- **Treat hypotheses asymmetrically**
  - Null hypo is usually special $\Rightarrow$ Fix rate of type-I error
  - Criminal convictions: Fix rate of unjust convictions
  - Higgs discovery: Fix rate of false discovery
  - Event selection: Fix rate of background that is accepted

- **Now can define a well stated goal for optimal testing**
  - Maximize the power of test (minimized rate of type-II error) for given $\alpha$
  - Event selection: Maximize fraction of signal accepted
The Neyman-Pearson lemma

- In 1932-1938 Neyman and Pearson developed a theory in which one must consider competing hypotheses
  - Null hypothesis ($H_0$) = Background only
  - Alternate hypotheses ($H_1$) = e.g. Signal + Background

and proved that

- The region $W$ that minimizes the rate of the type-II error (not reporting true discovery) is a contour of the Likelihood Ratio

\[
\frac{P(x | H_1)}{P(x | H_0)} > k_\alpha
\]

- Any other region of the same size will have less power

Wouter Verkerke, NIKHEF
The Neyman-Pearson lemma

- Example of application of NP-lemma with two observables

- Cut-off value $c$ controls type-I error rate (‘size’ = bkg rate)
  Neyman-Pearson: LR cut gives best possible ‘power’ = signal eff.

- So why don’t we *always* do this? (instead of training neural networks, boosted decision trees etc)

Wouter Verkerke, NIKHEF
Why Neyman-Pearson doesn’t always help

- The problem is that we usually don’t have explicit formulae for the pdfs $f(\tilde{x}|s), f(\tilde{x}|b)$.
- Instead we may have Monte Carlo samples for signal and background processes
  - Difficult to reconstruct analytical distributions of pdfs from MC samples, especially if number of dimensions is large
- If physics problem has only few observables can still estimate pdfs with histograms or kernel estimation,
  - But in such cases one can also forego event selection and go straight to hypothesis testing / parameter estimation with all events

![Diagram showing approximations of true pdfs](image.png)

Approximation of true $f(x|s)$

Approximation of true $f(x|b)$

Wouter Verkerke, NIKHEF
Hypothesis testing with a large number of observables

- When number of observables is large follow different strategy
- Instead of aiming at approximating p.d.f.s \( f(x|s) \) and \( f(x|b) \) aim to approximate decision boundary with an empirical parametric form

\[
A_\alpha(\bar{x}) = \left[ \frac{f(\bar{x} \mid s)}{f(\bar{x} \mid s + b)} > \alpha \right] \quad \Rightarrow \quad A_\alpha(\bar{x}) = c(\bar{x}, \bar{\theta})
\]
Empirical parametric forms of decision boundaries

- Can in principle choose any type of Ansatz parametric shape

  *Rectangular cut*
  \[ t(x) = \theta(x_j - c_j)\theta(x_i - c_i) \]

  *Linear cut*
  \[ t(x) = a_j \cdot x_j + a_i \cdot x_i \]

  *Non-linear cut*
  \[ t(x) = \bar{a} \cdot \bar{x} + \bar{x}A\bar{x} + \ldots \]

- Goal of Ansatz form is estimate of a ‘signal probability’ for every event in the observable space \( x \) (just like the LR)

- Choice of desired type-I error rate (selected background rate), can be set later by choosing appropriate cut on Ansatz test statistic.
Machine learning and all that

- A wide range of modern tools exist to perform supervised learning of a multivariate discriminant with the aim to approximate the optimal Neyman-Pearson discriminant.
  - Deep Learning, Boosted Decision Trees, GAN’s etc etc.

- Variation in
  - Ansatz (empirical parametric form of discriminant)
  - Learning process (error back propagation, Bayesian)

- Commonality in
  - Input (labeled simulation samples)
  - Output (single function that maps signal probability)

- In all cases output functions is functionally comparable to likelihood ratio discriminant (modulo some trivial transformations)
Classification with Machine Learning
Machine Learning

• What is Machine Learning?
  – Giving computers the ability to learn without explicitly programming them (Arthur Samuel, 1959)
  – Mathematical models learnt from data that characterize the patterns, regularities, and relationships amongst variables in the system

• Huge variety of choices in goals, formulations, training procedures

• Mathematical structure of model: (deep) neural networks, convolutional networks, transformer models, (boosted) decision trees, etc etc

• Input data: supervised learning (learn from simulation with truth-labels), unsupervised learning (learn from data without truth labels)

• Learning goal: classification, regression

• Scope of model: discrimination or generative

Wouter Verkerke, NIKHEF
Machine Learning in HEP

- ML used in HEP in many places in many ways

- Dominant use case: Supervised learning for classification
  - Signal/background separation, object tagging – trained on simulation samples
  - Will largely focus on this use case today

- But also many other uses
  - Unsupervised learning: anomaly detection
  - Regression: improving mass estimate of e.g. jets in events

Wouter Verkerke, NIKHEF
Multivariate Discriminants – the simplest case: the linear discriminant

- A **linear discriminant** constructs $t(x)$ from a linear combination of the variables $x_i$
  
  $$t(\vec{x}) = \sum_{i=1}^{N} a_i x_i = \vec{a} \cdot \vec{x}$$

  - Optimize discriminant by choosing $a_i$ to maximize separation between signal and background

- Most common form of the linear discriminant is the **Fisher discriminant**

  $$F(\vec{x}) = (\vec{\mu}_S - \vec{\mu}_B)^T V^{-1} \vec{x}$$

  Mean values in $x_i$ for sig,bkg

  Inverse of variance matrix of signal/background (assumed to be the same)

---

R.A. Fisher  
Ann. Eugen. 7(1936) 179.
Ansatz test statistics – The Fisher discriminant

\[ F(\bar{x}) = (\bar{\mu}_S - \bar{\mu}_B)^T V^{-1} \bar{x} \]

R.A. Fisher
Ann. Eugen. 7(1936) 179.

- Advantage of Fisher Discriminant:
  - Ingredients \( \mu_S, \mu_B, V \) can all be calculated directly from data or simulation samples. No ‘training’ or ‘tuning’

- Disadvantages of Fisher Discriminant
  - Fisher discriminant only exploits difference in means.
  - If signal and background have different variance, this information is not used.
Example of Fisher discriminant

• The “CLEO” Fisher discriminant
  – **Goal**: distinguish between $e^+e^\rightarrow Y4s \rightarrow \bar{b}b$ and $\bar{u}u, \bar{d}d, \bar{s}s, \bar{c}c$
  – **Method**: Measure energy flow in 9 concentric cones around direction of B candidate

![Energy flow in $\bar{b}b$](image)

![Energy flow in $u,d,s,c$](image)

Energy flows in $\bar{b}b$ and $u,d,s,c$ cone 1 to cone 9
When is Fisher discriminant the optimal discriminant?

• A very simple dataset

\[
S = \prod_{i} \text{Gauss}(x_i; \mu_i^S, \sigma_i) \\
B = \prod_{i} \text{Gauss}(x_i; \mu_i^B, \sigma_i)
\]

\[
\text{Multivariate Gaussian distributions with different means but same width for signal and background}
\]

• Fisher is optimal discriminant for this case
  - In this case we can also directly correlate \( F(x) \) to absolute signal probability

\[
P(F') = \frac{1}{1 + e^{-F}}
\]

‘Logistic sigmoid function’
Multivariate data selection – Neural networks

- Neural networks are used in neurobiology, pattern recognition, financial forecasting (and also HEP)

\[
N(\vec{x}) = s\left(a_0 + \sum_i a_i x_i\right)
\]

- This formula corresponds to the ‘single layer perceptron’
  - Visualization of single layer network topology

\[
s(t) = \frac{1}{1 + e^{-t}}
\]

Since activation function \(s(t)\) is monotonic, the single layer \(N(\mathbf{x})\) is equivalent to the Fisher discriminant \(F(\mathbf{x})\)
Neural networks – general structure

- The single layer model and easily be generalized to a **multilayer** perceptron

  - Easy to generalize to arbitrary number of layers
  - **Feed-forward net**: values of a node depend only on earlier layers (usually only on preceding layer) ‘the network architecture’
  - More nodes bring \( N(x) \) close to optimal \( t(x) = S(x)/B(x) \) but with much more parameters to be determined

\[
N(\vec{x}) = s(a_0 + \sum_{i=1}^{m} a_i h_i(\vec{x}))
\]

with \( h_i(\vec{x}) = s(w_{i0} + \sum_{j=1}^{n} w_{ij} x_j) \)

with \( a_i \) and \( w_{ij} \) weights (connection strengths)
Deep Neural Networks

- Availability of much more computing power, and notably GPUs have, and intense efforts worldwide outside HEP on algorithm and architecture development have increased ambitions (and results) of ML by many orders of magnitude

- Generally networks are labeled ‘deep’ if they reach a level of complexity where specialized nodes inside the network serve to extract specific ‘features’ of the data

Wouter Verkerke, NIKHEF
Example deep NN - convolutional NLL

- Convolutions NNs primarily designed to process ‘image data’
  - Scan for features defined by convolutions of local image data with a specific kernel function. Network designed to be insensitive to spatial location of feature
  - Structure allows for capturing local structure in early convolutions, and long range structure in later stage convolutions and in fully connected layers
Example deep NN - convolutional NLL

- Example here on generic image data, but many HEP problems are similar to image data (e.g. jet particles projected on a calorimeter surface)

VGGNet

(Simonyan and Zisserman, 2014)
Many network structures possible for many purposes

- For more information see specialized lectures in e.g. IN2P3 school of statistics, Terascale school of statistics, CERN academic lectures by M. Kagan...

- Generally, continuous rapid developments in ML/AI community
Activation functions

- So far only discussion logistic sigmoid as activation function

\[ s(t) = \frac{1}{1 + e^{-t}} \]

- Other functions are generally possible and useful. In particular the Rectified Linear Unit (ReLU) activation function improves Deep NN training as its derivative is constant (rather than vanishing) is therefore often used
Training (Deep) Neural Networks

- Training Deep Networks computationally very challenging
- Why did this take off in the last decade (or so)?
- Big data $\rightarrow$ large training sets
- Wide availability of cheap GPUs has increased computational power by orders of magnitude
- Many assorted improvement in many areas: Improved optimization algorithms, new regularization techniques, new activations functions
Supervised learning – general procedure

1. Design a (D)NN with adjustable parameters
2. Design a Loss function
3. Find best parameters which minimize loss

Wouter Verkerke, NIKHEF
Supervised learning – Loss function

- General form

$$\arg \min_w \frac{1}{N} \sum_{i=1}^{N} L(h(x_i; w), y_i) + \lambda \Omega(w)$$

compares MVA prediction $h()$ with (training) target data $y$

Additional term $\Omega(w)$ penalizes certain values of parameters $w$ with the aim to regularize the minimization process.

- Specific choice of loss function depends among others on training goal

Square error loss

$$L(h(x; w), y) = (h(x; w) - y)^2$$

(often used in regression)

Cross-entropy loss

$$L(h(x; w), y) = -y \log h(x; w) - (1 - y) \log(1 - h(x; w))$$

(often used in classification)
Supervised learning – minimization of loss function

- Minimize loss function using **back-propagation**
  - Compute gradient on each training set or batch
    \[
    \nabla_{w_j} L = \frac{\partial L}{\partial f} \frac{\partial f}{\partial g_n} \frac{\partial g_n}{\partial g_{n-1}} \ldots \frac{\partial g_{k+1}}{\partial g_k} \frac{\partial g_k}{\partial w_j}
    \]

- Update weights with **gradient descent**
  - where $\alpha$ is the learning rate
    \[
    w_j \leftarrow w_j - \alpha \nabla_{w_j} L
    \]

- Often advantageous to compute gradient only on subset of data (mini batch)
  ‘Mini Batch Gradient Descent’
    - Less computation required
    - Noise in gradient descent helps to avoid local minima
Supervised learning – Loss function

- General form
  \[ \arg \min_w \frac{1}{N} \sum_{i=1}^{N} L(h(x_i; w), y_i) + \lambda \Omega(w) \]

- Add regularization term to penalize overly complex models
  \[ L' = L + \frac{1}{2} \sum_j w_j^2 \]
Some notes on model complexity

- **Simple models under-fit**: will deviate from data (high bias) but will not be influenced by peculiarities of data (low variance).

- **Complex models over-fit**: will not deviate systematically from data (low bias) but will be very sensitive to data (high variance).

\[
\text{generalization error} = \text{systematic error} + \text{sensitivity of prediction}
\]
\[
= \text{bias} + \text{variance}
\]
\[
E[(y - h(x))^2] = E[(y - \bar{y})^2] + (\bar{y} - \bar{h}(x))^2 + E[(h(x) - \bar{h}(x))^2]
\]

Regularization term can help to control model complexity
Modeling unordered data

- So far (implicitly) assumed input data is structured, i.e. similar to a fixed-length vector where each element has same meaning for every event.
  - Maps well to classical ML approach in HEP → inputs are precalculated quantities, often using some physics input (invariant masses, highest-pT of lepton in event etc etc)

- But with ever increasing power and success of automatic feature extraction by deep networks on HEP data, question arises, why not simply give all event information to the DNN?
  - Apart from scale of problem, presents a small logistical challenge in the data format: full event reconstructed event record of events is not ‘structured’ in the sense above: 4-vectors are not ordered in a particular way, nor is the data set fixed size.
  - Other network structures can help here, notable Graph Neural Networks are very suitable for this type of data
Graph Neural Networks

- Modeling of *ordered* data as matrix or vector

<table>
<thead>
<tr>
<th>set of inputs with N constituents, M features</th>
</tr>
</thead>
<tbody>
<tr>
<td>{..., (p_T, \eta, \phi, particle ID), ...}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>feature matrix (N, M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>p_T (GeV)</td>
</tr>
<tr>
<td>12.3</td>
</tr>
<tr>
<td>11.8</td>
</tr>
<tr>
<td>10.4</td>
</tr>
<tr>
<td>9.8</td>
</tr>
<tr>
<td>6.4</td>
</tr>
<tr>
<td>5.3</td>
</tr>
</tbody>
</table>

- Modeling of graph data

<table>
<thead>
<tr>
<th>graph = set of nodes/vertices/elements + edges between them</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Edges represented as a index pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>edges = [(1,4), (1,3), (2,5), (6,5)]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Or as an N x N adjacency matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_{ij} =$</td>
</tr>
</tbody>
</table>
| \[
\begin{bmatrix}
0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 \\
\end{bmatrix}
\] |

<table>
<thead>
<tr>
<th>Where do we get this graph structure?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. All-to-all connections, in case of small input sets.</td>
</tr>
<tr>
<td>2. From physics priors: connect “nearby” elements in advance</td>
</tr>
<tr>
<td>3. Optimize as a part of the learning process (Graph Structure Learning)</td>
</tr>
</tbody>
</table>

Wouter Verkerke, NIKHEF
Graph Neural Networks

- Graph structures common in HEP data

GNNs very successful in flavor tagging for LHC experiments
ML Classification in HEP – some summarizing thoughts

• Tremendous progress over past decade in better, more powerful techniques to use ML for classification, regression.
  – Only discussed in the briefest possible terms at the conceptual level here
  – Many practical online courses available for many of these tools
  – Tendency to move away from letting ML deal with ‘pre-cooked’ physics observables, to letting ML analyse complete event records

• Many other aspects of ML not discussed at all here (unsupervised learning, generative models (not discussed), decision trees)
  – Also with many use cases other than sig/bkg classification (regression, fast simulation)

• It is quite likely that in a few years even newer ML techniques in will replace the current best performing ones..
ML Classification in HEP – some summarizing thoughts

• But keep in mind that ML/AI techniques are not magic → their performance is also bound by the Neyman-Pearson limit
  – There is a well-defined upper bound on the reachable performance by any algorithm. This bound is not calculable for many complex models, but it is nevertheless there

• ML/AI techniques take their inputs quite literally → simulation samples used in supervised training are known to be subject to uncertainties.
  – If simulation differs from data, ML-based results may be suboptimal, or even wrong, depending on how ML was used
  – Impact of ML training on imperfect simulation depends on analysis design → more on this later
  – There is also room to take some uncertainty on input samples into account, sometimes this is trivial (e.g. if the total cross-section is uncertain)
  – But in other cases this is exceeding difficult if the specification of what is uncertain is fuzzy or incomplete. (e.g. hadronization uncertainties that are only expressed as different outcomes for two different generators)

• Bottom line – if systematic effects are non-negligible, great care must be taken in the use of ML discriminants in the analysis

Wouter Verkerke, NIKHEF
Event selection as dimensionality reduction

- In the limit of an optimal discriminant – the (ML) event selection step is effectively (and only) a reduction of dimensionality of the data without loss of information (in the optimal case)

- In case the full discriminant distribution is tested → no loss of information
  - But need for pdf that model distribution

- But can also select high-signal region and perform simplified inference
  - e.g. counting model in that region

\[
\frac{P(x|H_1)}{P(x|H_0)} > k_\alpha
\]
The simplest analysis design: *cut-and-count*

- A common scenario for searches in a low-statistics regime is to perform a simplified analysis
  1. Train MVA to obtain discriminant D
  2. Apply a *cut* on D
  3. Perform only a *counting* analysis

- And a common question is then – what is the ‘optimal cut on D’?
  - To answer question, a ‘figure of merit’ (FOM) must be chosen that quantifies the optimality of the selection.
  - The FOM for a search is usually the *expected signal significance*.
  - A ‘traditional’ choice is FOM=$s/\sqrt{b}$. For low-statistic searches $s/\sqrt{b}$ is a bad choice! It assumes Gaussian distribution, whereas the true distribution is Poisson, which is quite unlike Gaussian especially in the tails at low N
    - A better, and equally easy to use, equation exists based on a Poisson calculation
  - NB: the question arise due to choice for simplified counting in step 3).
  - If a *probability density model* is used for the analysis of the selected data, then the answer is always ‘the full range of the discriminant’
A better FOM for discovery - the ‘Expected Poisson Z’

- The expected counting significance for a Poisson process is analytically calculable: \( \sqrt{2((s + b) \ln(1 + s/b) - s)} \).

- For discovery, the traditional FOM \( s/\sqrt{b} \) shows significant deviations from the ‘exact’ expected Poisson significance at low \( b \).

\[
\sqrt{q_{0,A}} = \sqrt{2((s + b) \ln(1 + s/b) - s)}.
\]

\[
= \frac{s}{\sqrt{b}} \left(1 + \mathcal{O}(s/b)\right).
\]
Model building 3

Models with parameters I -
analytical parametric models,
template morphing approach for
histogram-based models
Roadmap of this course

• Start with basics, gradually build up to complexity

**Model building**
- Counting models
- Modeling distributions
- Signal parameterization strategies
- Models with nuisance parameters, joint models, modeling systematic uncertainties
- Diagnosing inference on complex models
- Advanced signal modeling techniques

**Statistical methods**
- Statistical tests with counting experiments
- Test statistics for models describing distributions
- Parameter estimation, confidence intervals & limits
- Inference with nuisance parameters
Introduce concept of composite hypotheses

- In most cases in physics, a hypothesis is not “simple”, but “composite”
- **Composite hypothesis** = Any hypothesis which does *not* specify the population distribution completely
- Example: counting experiment with signal and background, that leaves signal expectation unspecified

\[
L = \text{Poisson}(N \mid \tilde{s} + \tilde{b})
\]

\[
L(s) = \text{Poisson}(N \mid s + \tilde{b})
\]

(My) notation convention: all symbols with ~ are constants

Wouter Verkerke, Nikhef
A common convention in the meaning of model parameters

- A common convention is to recast signal rate parameters into a normalized form (e.g. w.r.t the Standard Model rate)

\[ L = \text{Poisson}(N \mid \tilde{s} + \tilde{b}) \]

\[ L(s) = \text{Poisson}(N \mid s + \tilde{b}) \]

\[ L(\mu) = \text{Poisson}(N \mid \mu \cdot \tilde{s} + \tilde{b}) \]

'Universal' parameter interpretation makes it easier to work with your models

- \(\mu = 0\) → no signal
- \(\mu = 1\) → expected signal
- \(\mu > 1\) → more than expected signal
Model building for measurements $\rightarrow$ shape parameter

- Beyond discovery/rate measurements, can also build models to measure properties of particles (e.g. mass) $\rightarrow$ introduce shape parameters
- Often trivial for analytical models, less so for simulation-based models

$$F(x|m) = \text{Gaussian}(x, m, \sigma) + \text{bkg}$$

$$F(x|m) = ??$$
Modeling of shape variations in the likelihood

- If underlying simulation has free parameter $\theta$, can assess impact on reconstructed shapes by rerunning simulation at different values
  - Obtain histogram templates for distributions at $'+1\sigma'$ and $'\pm 1\sigma'$ settings of systematic effect

- Challenge: construct an empirical response function based on the interpolation of the shapes of these three templates.

Wouter Verkerke, NIKHEF
Need to interpolate between template models

- Need to define ‘morphing’ algorithm to define distribution $s(x)$ for each value of $\alpha$
Piecewise linear interpolation

- Simplest solution is piece-wise linear interpolation for each bin

Piecewise linear interpolation response model for a one bin

Extrapolation to $|\alpha| > 1$

Kink at $\alpha = 0$

Ensure $s_i(\alpha) \geq 0$

Wouter Verkerke, NIKHEF
Visualization of bin-by-bin linear interpolation of distribution

Wouter Verkerke, NIKHEF

'ex15_build binned morphing.C'
Other morphing strategies – ‘horizontal morphing’

- Other template morphing strategies exist that are less prone to unintended side effects
- A ‘horizontal morphing’ strategy was invented by Alex Read.
  - Interpolates the cumulative distribution function instead of the distribution
  - Especially suitable for shifting distributions
  - Here shown on a continuous distribution, but also works on histograms
  - Drawback: computationally expensive, algorithm only worked out for 1 NP
Yet another morphing strategy – ‘Moment morphing’

- Given two template model $f_-(x)$ and $f_+(x)$ the strategy of moment morphing considers first two moment of template models (mean and variance)

\[
\mu_- = \int x \cdot f_-(x) dx \\
V_- = \int (x - \mu_-)^2 \cdot f_-(x) dx
\]

\[
\mu_+ = \int x \cdot f_+(x) dx \\
V_+ = \int (x - \mu_+)^2 \cdot f_+(x) dx
\]

- The goal of moment morphing is to construct an interpolated function that has linearly interpolated moments

\[
\mu(\alpha) = \alpha \mu_- + (1 - \alpha) \mu_+ \\
V(\alpha) = \alpha V_- + (1 - \alpha) V_+ \quad [1]
\]

- It constructs this morphed function as combination of linearly transformed input models

\[
f(x, \alpha) \rightarrow \alpha f_-(ax + b) + (1 - \alpha) f_+(cx - d)
\]

- Where constants a,b,c,d are chosen such so that $f(x,\alpha)$ satisfies conditions [1]

\[\text{Wouter Verkerke, NIKHEF}\]
There are other morphing algorithms to choose from

- Vertical Morphing
- Horizontal Morphing
- Moment Morphing

Gaussian varying width

Gaussian varying mean

Gaussian to Uniform (this is conceptually ambiguous!)

n-dimensional morphing?

✓ ✔ ✗ ✔
Statistical methods 3

Inference with parameters: maximum likelihood, confidence intervals, upper limits, likelihood ratio and asymptotic formulae
Roadmap of this course

- Start with basics, gradually build up to complexity

**Model building**
- Counting models
- Modeling distributions
- Signal parameterization strategies
- Models with nuisance parameters, joint models, modeling systematic uncertainties
- Diagnosing inference on complex models
- Advanced signal modeling techniques

**Statistical methods**
- Statistical tests with counting experiments
- Test statistics for models describing distributions
- Parameter estimation, confidence intervals & limits
- Inference with nuisance parameters
What can we do with composite hypothesis

• With simple hypotheses – inference is restricted to making statements about $P(D|hypo)$ or $P(hypo|D)$

• With composite hypotheses – many more options

  1 Parameter estimation and variance estimation
  - What is value of $s$ for which the observed data is most probable?
  - What is the variance (std deviation squared) in the estimate of $s$?

  $s = 5.5 \pm 1.3$

• 2 Confidence intervals
  - Statements about model parameters using frequentist concept of probability
  - $s < 12.7$ at 95% confidence level
  - $4.5 < s < 6.8$ at 68% confidence level

• 3 Bayesian credible intervals
  - Bayesian statements about model parameters
  - $s < 12.7$ at 95% credibility
Parameter estimation using Maximum Likelihood

- Likelihood is high for values of $p$ that result in distribution similar to data

- Define the maximum likelihood (ML) estimator to be the procedure that finds the parameter value for which the likelihood is maximal.

Wouter Verkerke, NIKHEF
Parameter estimation – Maximum likelihood

- Practical estimation of maximum likelihood performed by minimizing the negative log-Likelihood

\[ L(\bar{p}) = \prod_{i} f(\bar{x}_i; \bar{p}) \]

\[ -\ln L(\bar{p}) = -\sum_{i} \ln F(\bar{x}_i; \bar{p}) \]

- Advantage of log-Likelihood is that contributions from events can be summed, rather than multiplied (computationally easier)

- In practice, find point where derivative of $-\log L$ is zero

\[ \frac{d \ln L(\bar{p})}{d\bar{p}} \bigg|_{\bar{p}_i = \hat{\bar{p}}_i} = 0 \]

- Standard notation for ML estimation of $p$ is $\hat{p}$
Example of Maximum Likelihood estimation

- Illustration of ML estimate on Poisson counting model

\[ L(N \mid s) = \text{Poisson}(N \mid s + \tilde{b}) \]

-\log L(N\mid s) \text{ versus } N \quad [s=0,5,10,15] \quad \text{-log } L(N\mid s) \text{ versus } s \quad [N=7]

- Note that Poisson model is discrete in N, but continuous in s!
Properties of Maximum Likelihood estimators

- In general, Maximum Likelihood estimators are
  - Consistent (gives right answer for $N \to \infty$)
  - Mostly unbiased (bias $\propto 1/N$, may need to worry at small $N$)
  - Efficient for large $N$ (you get the smallest possible error)
  - Invariant: (a transformation of parameters will not change your answer, e.g. $\left(\hat{p}\right)^2 = \left(p^2\right)$)

- MLE efficiency theorem: the MLE will be unbiased and efficient if an unbiased efficient estimator exists
  - Proof not discussed here
  - Of course this does not guarantee that any MLE is unbiased and efficient for any given problem
Relation between Likelihood and $\chi^2$ estimators

- Properties of $\chi^2$ estimator follow from properties of ML estimator using *Gaussian probability density functions*

\[
F(x_i, y_i, \sigma_i; \bar{p}) = \prod_i \exp \left[ -\left( \frac{y_i - f(x_i; \bar{p})}{\sigma_i} \right)^2 \right]
\]

The Likelihood function in $p$ for given points $x_i(s_i)$ and function $f(x; p)$

\[
-\ln L(\bar{p}) = \frac{1}{2} \sum_i \left( \frac{y_i - f(x_i; \bar{p})}{\sigma_i} \right)^2 = \frac{1}{2} \chi^2
\]

- The $\chi^2$ estimator follows from ML estimator, i.e it is
  - Efficient, consistent, bias 1/N, invariant,
  - But only in the limit that the error on $x_i$ is truly Gaussian
Estimating parameter variance

- Note that ‘uncertainty’ on a parameter estimate is an ambiguous statement.
- Can either mean an interval with a stated confidence or credible level (e.g. 68%), or simply assume it is the square-root of the variance of a distribution.

For a Gaussian distribution, mean and variance map to parameters for mean and \( \sigma^2 \) and interval defined by \( \sqrt{\text{V}} \) contains 68% of the distribution (= ‘1 sigma’ by definition).

Thus for Gaussian distributions, all common definitions of ‘error’ work out to the same numeric value.

Wouter Verkerke, NIKHEF
Estimating parameter variance

• Note that ‘uncertainty’ on a parameter estimate is an ambiguous statement

• Can either mean an interval with a stated confidence or credible level (e.g. 68%), or simply assume it is the square-root of the variance of a distribution

For other distributions intervals by $\sqrt{V}$ do not necessarily contain 68% of the distribution
Estimating variance on parameters

- Variance on a parameter can also be estimated from Likelihood using the variance estimator

\[
\hat{\sigma}(p)^2 = \hat{V}(p) = \left(\frac{d^2 \ln L}{d^2 p}\right)^{-1}
\]

- Valid if estimator is efficient and unbiased!

- Illustration of Likelihood Variance estimate on a Gaussian distribution

\[
f(x \mid \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right]
\]

\[
\ln f(x \mid \mu, \sigma) = -\ln \sigma - \ln \sqrt{2\pi} + \frac{1}{2} \left(\frac{x - \mu}{\sigma}\right)^2
\]

\[
\left.\frac{d \ln f}{d \sigma}\right|_{x=\mu} = -\frac{1}{\sigma} \quad \Rightarrow \quad \left.\frac{d^2 \ln f}{d^2 \sigma}\right|_{x=\mu} = \frac{1}{\sigma^2}
\]

From Rao-Cramer-Frechet inequality

\[
V(\hat{p}) \geq 1 + \frac{db}{dp} \left(\frac{d^2 \ln L}{d^2 p}\right)
\]

\(b = \text{bias as function of } p, \text{ inequality becomes equality in limit of efficient estimator}\)
Bayesian parameter estimation

- Bayesian parameter estimate is the posterior mean
- Bayesian variance is the posterior variance

\[
\hat{\mu} = \int \mu P(\mu | N) d\mu
\]

\[
\hat{V} = \int (\hat{\mu} - \mu)^2 P(\mu | N) d\mu
\]
What can we do with composite hypothesis

• With simple hypotheses – inference is restricted to making statements about $P(D|hypo)$ or $P(hypo|D)$

• With composite hypotheses – many more options

• 1 Parameter estimation and variance estimation
  – What is value of $s$ for which the observed data is most probable?
  – What is the variance (std deviation squared) in the estimate of $s$?
  \[
  s = 5.5 \pm 1.3
  \]

• 2 Confidence intervals
  – Statements about model parameters using frequentist concept of probability
  – $s < 12.7$ at 95% confidence level
  – $4.5 < s < 6.8$ at 68% confidence level

• 3 Bayesian credible intervals
  – Bayesian statements about model parameters
  – $s < 12.7$ at 95% credibility

Wouter Verkerke, NIKHEF
Interval estimation with fundamental methods

- Can also construct parameters intervals using ‘fundamental’ methods explored earlier (Bayesian or Frequentist).
- Construct **Confidence Intervals** or **Credible Intervals** with defined probabilistic meaning, independent of assumptions on normality of distribution (Central Limit Theorem) → “95% C.L.”
- With fundamental methods you **greater flexibility in types of interval**. E.g. when no signal observed → usually wish to set an upper limit (construct ‘upper limit interval’).

Wouter Verkerke, NIKHEF
Reminder - Frequentist test statistics and p-values

- Definition of ‘p-value’: *Probability to observe this outcome or more extreme in future repeated measurements is x%, if hypothesis is true*

- Note that the definition of p-value assumes an explicit ordering of possible outcomes in the ‘or more extreme’ part

\[
p_b = \int_{N_{obs}}^{\infty} \text{Poisson}(N; b + 0) dN \quad (= 0.23)
\]
P-values with a likelihood ratio test statistic

- With the introduction of a (likelihood ratio) test statistic, hypothesis testing of models of arbitrary complexity is now reduced to the same procedure as the Poisson example.

\[
\lambda(\vec{N}) = \frac{L(\vec{N} \mid H_{s+b})}{L(\vec{N} \mid H_b)}
\]

- Except that we generally don’t know distribution \(f(\lambda)\)…
A different Likelihood ratio for composite hypothesis testing

- On composite hypotheses, where both null and alternate hypothesis map to values of \( \mu \), we can define an alternative likelihood-ratio test statistics that has better properties.

\[
\lambda(\vec{N}) = \frac{L(\vec{N} \mid H_0)}{L(\vec{N} \mid H_1)}
\]

\[
\lambda_{\mu}(\vec{N}_{\text{obs}}) = \frac{L(\vec{N} \mid \mu)}{L(\vec{N} \mid \hat{\mu})}
\]

- Advantage: distribution of new \( \lambda_{\mu} \) has known asymptotic form.

- Wilks theorem: distribution of \(-\log(\lambda_{\mu})\) is asymptotically distribution as a \( \chi^2 \) with \( N_{\text{param}} \) degrees of freedom.*

*Some regularity conditions apply

- \( \rightarrow \) Asymptotically, we can directly calculate p-value from \( \lambda_{\mu}^{\text{obs}} \)
What does a $\chi^2$ distribution look like for $n=1$?

- Note that it for $n=1$, it does not peak at 1, but rather at 0…
Composite hypothesis testing in the asymptotic regime

- For ‘histogram example’: what is p-value of null-hypothesis

\[ t_0 = -2 \ln \frac{L(data | \mu = 0)}{L(data | \hat{\mu})} \]

\( \hat{\mu} \) is best fit value of \( \mu \)

'likelihood assuming zero signal strength'

'likelihood of best fit'

\( -\log \mu \)

On signal-like data \( t_0 \) is large

\[ t_0 = 34.77 \]

Distribution of test statistic value for data obtained under \( s=0 \) hypothesis

\( f(\lambda | s = 0) \)

Test statistic value for observed data

\( \lambda(\hat{N}_{\text{obs}}) \)

Wilks: \( f(\lambda | 0) \rightarrow \chi^2 \) distribution

P-value = TMath::Prob(34.77,1)

\[ = 3.7 \times 10^{-9} \]
Composite hypothesis testing in the asymptotic regime

- For ‘histogram example’: what is p-value of null-hypothesis

\[ t_0 = -2 \ln \frac{L(data | \mu = 0)}{L(data | \hat{\mu})} \]

\( \hat{\mu} \) is best fit value of \( \mu \)

‘likelihood assuming zero signal strength’

‘likelihood of best fit’

On signal-like data \( t_0 \) is large

On background-like data \( t_0 \) is small

\[ t_0 = 34.77 \]

\[ t_0 = 0.02 \]

P-value = TMath::Prob(34.77,1) 
\[ = 3.7 \times 10^{-9} \]

P-value = TMath::Prob(0.02,1) 
\[ = 0.88 \]

Use Wilks Theorem
How quickly does \( f(\lambda | \mu) \) converge to its asymptotic form

- Pretty quickly –

Here is an example of likelihood function for 10-bin distribution with 200 events

Here is an example for event counting at various \( s, b \)

\[
\sqrt{q_{0,A}} = \sqrt{2\left((s+b) \ln(1+s/b) - s\right)}.
\]

Wouter Verkerke, NIKHEF
From hypothesis testing to confidence intervals

- Next step for composite hypothesis is to go from p-values for a hypothesis defined by fixed value of $\mu$ to *an interval statement on $\mu$*

- Definition: *A interval on $\mu$ at X% confidence level is defined such that the true value of $\mu$ is contained X% of the time in the interval.*
  - Note that the output is *not* a probabilistic statement on the true $\mu$ value
  - The true $\mu$ is fixed but unknown – each observation will result in an estimated interval $[\mu_-, \mu_+]$. X% of those intervals will contain the true value of $\mu$
  - Coverage = guarantee that probabilistic statements is true (i.e. repeated future experiments do reproduce results in X% of cases)

- Definition of confidence intervals does not make any assumption on shape of interval
  
  -> Can choose one-sided intervals (‘limits’),
  two-sided intervals (‘measurements’),
  or even disjoint intervals (‘complicated measurements’)
Exact confidence intervals – the Neyman construction

- Simplest experiment: one measurement (x), one theory parameter (θ)
- For each value of parameter θ, determine distribution in observable x
How to construct a Neyman Confidence Interval

• Focus on a slice in \( \theta \)
  
  – For a \( 1 - \alpha \)% confidence Interval, define *acceptance interval* that contains \( 100\% - \alpha \)% of the distribution

pdf for observable \( x \)
given a parameter value \( \theta_0 \)

\[
f(x | \theta_0)
\]
How to construct a Neyman Confidence Interval

- Definition of acceptance interval is not unique → Choose shape of interval you want to set here.
  - Algorithm to define acceptance interval is called ‘ordering rule’

![Diagram showing Neyman Confidence Interval](image)

- pdf for observable \( x \) given a parameter value \( \theta_0 \)
- Lower Limit
- Central Interval
- Other options, are e.g. ‘symmetric’ and ‘shortest’
How to construct a Neyman Confidence Interval

- Now make an acceptance interval in observable $x$ for each value of parameter $\theta$
How to construct a Neyman Confidence Interval

- This makes the confidence belt
How to construct a Neyman Confidence Interval

- This makes the confidence belt
How to construct a Neyman Confidence Interval

- The confidence belt can be constructed in advance of any measurement, it is a property of the model, not the data.
- Given a measurement $x_0$, a confidence interval $[\theta_-, \theta_+]$ can be constructed as follows.
- The interval $[\theta_-, \theta_+]$ has a 68% probability to cover the true value.

![Diagram showing the Neyman Confidence Interval](attachment:diagram.png)
What confidence interval means & concept of coverage

- A confidence interval is an interval on a parameter that contains the true value X% of the time.

- This is a property of the procedure, and should be interpreted in the concept of repeated identical measurements:

  Each future measurement will result in a confidence interval that has somewhat different limits every time ('confidence interval limits are a random variable').

  But procedure is constructed such that true value is in X% of the intervals in a series of repeated measurements (this calibration concept is called ‘coverage’. The Neyman constructions guarantees coverage).

- It is explicitly not a probability statement on the true value you are trying to measure. In the frequentist the true value is fixed (but unknown).
The confidence interval – Poisson counting example

- Given the probability model for Poisson counting example: for every hypothesized value of $s$, plot the expected distribution $N$

Confidence belt for 68% and 90% central intervals

Confidence belt for 68% and 90% lower limit
The confidence interval – Poisson counting example

- Given confidence belt and observed data, confidence interval on parameter is defined by belt intersection.

Central interval on $s$ at 68% C.L.

Lower limit on $s$ at 90% C.L.
Confidence intervals using the Likelihood Ratio test statistic

- Neyman Construction on Poisson counting looks like ‘textbook’ belt.
- In practice we’ll use the Likelihood Ratio test statistic to summarize the measurement of a (multivariate) distribution for the purpose of hypothesis testing.
- Procedure to construct belt with LR is identical: obtain distribution of $\lambda$ for every value of $\mu$ to construct confidence belt

$$x=3.2$$

$$\lambda_{\mu}(x,\mu)$$
The asymptotic distribution of the likelihood ratio test statistic

- Given the likelihood ratio

\[ t_\mu = -2 \log \lambda_\mu (x) = -2 \log \frac{L(x | \mu)}{L(x | \hat{\mu})} \]

Q: What do we know about asymptotic distribution of \( \lambda(\mu) \)?

- A: Wilks theorem \( \Rightarrow \) Asymptotic form of \( f(t|\mu) \) is a \( \chi^2 \) distribution

\[ f(t_\mu | \mu) = \chi^2(t_\mu, n) \]

Where

- \( \mu \) is the hypothesis being tested and
- \( n \) is the number of parameters (here 1: \( \mu \))

- **Note that \( f(t_\mu | \mu) \) is independent of \( \mu \)!**
  \( \Rightarrow \) Distribution of \( t_\mu \) is the same for every ‘horizontal slice’ of the belt
Confidence intervals using the Likelihood Ratio test statistic

- Procedure to construct belt with LR is identical: obtain distribution of $\lambda$ for every value of $\mu$ to construct belt.

Confidence belt now range in LR.
What does the observed data look like with a LR?

• Note that while belt is (asymptotically) independent of parameter $\mu$, observed quantity now is dependent of the assumed $\mu$.

$\mu(x, \mu)$

Measurement $= t_{\mu}(x_{\text{obs}}, \mu)$ is now a function of $\mu$.
Connection with likelihood ratio intervals

- If you assume the asymptotic distribution for $t_\mu$,
  - Then the confidence belt is exactly a box
  - And the constructed confidence interval can be simplified to finding the range in $\mu$ where $t_\mu = \frac{1}{2}Z^2$

→ This is exactly the MINOS error
Recap on confidence intervals

- Confidence intervals on parameters are constructed to have precisely defined probabilistic meaning
  - This calibration is called “coverage”
    The Neyman Construction has coverage by construction
  - This is different from parameter variance estimates (or Bayesian methods) that don’t have (a guaranteed) coverage
  - For most realistic models confidence intervals are calculated using (Likelihood Ratio) test statistics to define the confidence belt

- Asymptotic properties
  - In the asymptotic limit (Wilks theorem), Likelihood Ratio interval converges to a Neyman Construction interval (with guaranteed coverage) “Minos Error”
    
    NB: the likelihood does not need to be parabolic for Wilks theorem to hold
  - Separately, in the limit of normal distributions the likelihood becomes exactly parabolic and the ML Variance estimate converges to the Likelihood Ratio interval
Bayesian inference with composite hypothesis

- With change $L \rightarrow L(\mu)$ the prior and posterior model probabilities become probability density functions

$$P(H_{s+b} \mid \tilde{N}) = \frac{L(\tilde{N} \mid H_{s+b})P(H_{s+b})}{L(\tilde{N} \mid H_{s+b})P(H_{s+b}) + L(\tilde{N} \mid H_b)P(H_b)}$$

$$P(\mu \mid \tilde{N}) \propto L(\tilde{N} \mid \mu)P(\mu)$$

**NB:** Likelihood is not a probability density
Bayesian credible intervals

- From the posterior density function, a credible interval can be constructed through integration.

- Note that Bayesian interval estimation require *no minimization* of \( -\log L \), just integration.

Wouter Verkerke, NIKHEF
Bayesian parameter estimation

- Bayesian parameter estimate is the posterior mean
- Bayesian variance is the posterior variance

\[ \hat{\mu} = \int \mu P(\mu | N) d\mu \]

\[ \hat{V} = \int (\hat{\mu} - \mu)^2 P(\mu | N) d\mu \]
Choosing Priors

- As for simple models, Bayesian inference always involves a prior \( \rightarrow \) now a prior probability density on your parameter

- When there is clear prior knowledge, it is usually straightforward to express that knowledge as prior density function
  - Example: prior measurement of \( \mu = 50 \pm 10 \)

- Posterior represents updated belief \( \rightarrow \) It incorporates information from measurement and prior belief
  - But sometimes we only want to publish result of this experiment, or there is no prior information. What to do?
Choosing Priors

- Common but thoughtless choice: a flat prior
  - Flat implies choice of metric. Flat in x, is not flat in $x^2$

- Flat prior implies choice on of metric
  - A prior that is flat in $\mu$ is not flat in $\mu^2$
  - ‘Preferred metric’ has often no clear-cut answer.
    (E.g. when measuring neutrino-mass-squared, state answer in m or m$^2$)
  - In multiple dimensions even complicated (prior flat in x,y or is prior flat in r,φ?)

Wouter Verkerke, NIKHEF
Is it possible to formulate an ‘objective’ prior?

• Can one define a prior $p(\mu)$ which contains as little information as possible, so that the posterior pdf is dominated by the likelihood?
  – A bright idea, vigorously pursued by physicist Harold Jeffreys in in mid-20th century:
  – This is a really really thoughtless idea, recognized by Jeffreys as such, but dismayingly common in HEP: just choose $p(\mu)$ uniform in whatever metric you happen to be using!

• “Jeffreys Prior” answers the question using a prior uniform in a metric related to the Fisher information.

$$I(\theta) = -E \left[ \frac{\partial^2}{\partial \theta^2} \log f(x | \theta) \bigg| \theta \right]$$
  – Unbounded mean $\mu$ of gaussian: $p(\mu) = 1$
  – Poisson signal mean $\mu$, no background: $p(\mu) = 1/\sqrt{\mu}$

• Many ideas and names around on non-subjective priors
  – Advanced subject well beyond scope of this course.
  – Many ideas (see e.g. summary by Kass & Wasserman), but very much an open/active in area of research.
Sensitivity Analysis

• Since a Bayesian result depends on the prior probabilities, which are either personalistic or with elements of arbitrariness, it is widely recommended by Bayesian statisticians to study the sensitivity of the result to varying the prior.

• Sensitivity generally decreases with precision of experiment

• Some level of arbitrariness – what variations to consider in sensitivity analysis
Summary

- **Maximum Likelihood**
  - Point and variance estimation
  - Variance estimate assumes normal distribution. No upper/lower limits

- **Frequentist confidence intervals**
  - Extend hypothesis testing to composite hypothesis
  - Neyman construction provides exact “coverage” = calibration of quoted probabilities
  - Strictly $p(\text{data}|\text{theory})$
  - Asymptotically identical to likelihood ratio intervals (MINOS errors, *does not assume parabolic L*)

- **Bayesian credible intervals**
  - Extend $P(\text{theo})$ to p.d.f. in model parameters
  - Integrals over posterior density $\rightarrow$ credible intervals
  - Always involves prior density function in parameter space