LECTURES ON

STATISTICS

@KyleCranmer
New York University
Department of Physics
Center for Data Science
Instead of choosing to accept/reject $H_0$ one can compute the p-value

$$p = \int_{T_0}^{\infty} f(T|H_0)$$

If the model for the data depends on parameters $\alpha$ the p-value also depends on $\alpha$. 

$$p(\alpha) = \int_{T_0}^{\infty} f(T|\alpha)dT = \int f(D|\alpha) \theta(T(D) - T_0) dD = P(T \geq T_0|\alpha)$$
When the model has nuisance parameters, only reject the null if \( p(\alpha) \) sufficiently small for all values of the nuisance parameters.

If the model for the data depends on parameters \( \alpha \) the p-value also depends on \( \alpha \).
THE PROFILE LIKELIHOOD RATIO

Consider our general model with a single parameter of interest $\mu$

- let $\mu=0$ be no signal, $\mu=1$ nominal signal

In the LEP approach the likelihood ratio is equivalent to:

$$Q_{\text{LEP}} = \frac{L(\mu = 1, \theta)}{L(\mu = 0, \theta)} = \frac{f(D|\mu = 1, \theta)}{f(D|\mu = 0, \theta)}$$

- but this variable is sensitive to uncertainty on $\theta$ and makes no use of auxiliary measurements $a$

Alternatively, one can define profile likelihood ratio

$$\lambda(\mu) = \frac{L(\mu, \hat{\theta}(\mu))}{L(\hat{\mu}, \hat{\theta})} = \frac{f(D, G|\mu, \hat{\theta}(\mu; D, G))}{f(D, G|\hat{\mu}, \hat{\theta})}$$

- where $\hat{\theta}(\mu; D, G)$ is best fit with $\mu$ fixed (the constrained maximum likelihood estimator, depends on data)
- and $\hat{\theta}$ and $\hat{\mu}$ are best fit with both left floating (unconstrained)
- Tevatron used $Q_{\text{Tev}} = \lambda(\mu=1)/\lambda(\mu=0)$ as generalization of $Q_{\text{LEP}}$
Essentially, you need to fit your model to the data twice: once with everything floating, and once with signal fixed to 0

\[
\lambda(\mu = 0) = \frac{L(\mu = 0, \hat{\theta}(\mu = 0))}{L(\hat{\mu}, \hat{\theta})} = \frac{f(\mathcal{D}, G|\mu = 0, \hat{\theta}(\mu = 0; D, G))}{f(\mathcal{D}, G|\hat{\mu}, \hat{\theta})}
\]

Figure 14 shows the fit to the signal candidates for a \(m_H = 120\) GeV Higgs with (a,c) and without (b,d) the signal contribution. It can be seen that the background shape and normalizations are trying to accommodate the excess near \(m_{\tau\tau} = 120\) GeV, but the control samples are constraining the variation.

Table 13 shows the significance calculated from the profile likelihood ratio for the \(ll\)-channel, the \(lh\)-channel, and the combined fit for various Higgs boson masses.
Properties of the Profile Likelihood Ratio

After a close look at the profile likelihood ratio

\[ \lambda(\mu) = \frac{L(\mu, \hat{\theta}(\mu))}{L(\hat{\mu}, \hat{\theta})} = \frac{f(D, G | \mu, \hat{\theta}(\mu; D, G))}{f(D, G | \hat{\mu}, \hat{\theta})} \]

one can see the function is independent of true values of \( \theta \)

- though its distribution might depend indirectly

Wilks’s theorem states that under certain conditions the distribution of

\(-2 \ln \lambda(\mu = \mu_0)\)

given that the true value of \( \mu \) is \( \mu_0 \) converges to a chi-square distribution

- more on this later, but the important points are:
  - “asymptotic distribution” is known and it is independent of \( \theta \)!
    - more complicated if parameters have boundaries (eg. \( \mu \geq 0 \))

Thus, we can calculate the p-value for the background-only hypothesis without having to generate Toy Monte Carlo!
TOY MONTE CARLO

Explicitly build distribution by generating “toys” / pseudo experiments assuming a specific value of \( \mu \) and \( \nu \).

- randomize both main measurements \( \mathcal{D}=\{x\} \) and auxiliary measurements \( \mathcal{G}=\{a\} \)
- fit the model twice for the numerator and denominator of profile likelihood ratio
- evaluate \(-2\ln \lambda(\mu)\) and add to histogram

Choice of \( \mu \) is straight forward: typically \( \mu=0 \) and \( \mu=1 \), but choice of \( \theta \) is less clear

- more on this later

This can be very time consuming. Plots below use millions of “toy” pseudo-experiments

![Graphs showing 2-channel and 5-channel results with significance levels 3.35\(\sigma\) and 4.4\(\sigma\).]
Recently we showed how to generalize this asymptotic approach

- generalize Wilks’s theorem when boundaries are present
- use Wald’s result for distribution for alternate hypothesis $f(-2\log \lambda(\mu) | \mu')$

Asymptotic formulae for likelihood-based tests of new physics

Glen Cowan, Kyle Cranmer, Eilam Gross, Ofer Vitells


http://arxiv.org/abs/1007.1727v2
**COMPARISON OF ASYMPTOTIC AND ENSEMBLES**

Compare asymptotic distributions to distributions obtained with large ensembles of pseudo-experiments generated with Monte Carlo techniques.

This is a significant development as building this distribution from Monte Carlo approaches can take 100,000 CPU hours for Higgs search!

G. Cowan, KC, E. Gross, O. Vitells
[arXiv:1007.1727]
So far this looks a bit like magic. How can you claim that you incorporated your systematic just by fitting the best value of your uncertain parameters and making a ratio? It won’t unless the the parametrization is sufficiently flexible. So check by varying the settings of your simulation, and see if the profile likelihood ratio is still distributed as a chi-square.

Here it is pretty stable, but it’s not perfect (and this is a log plot, so it hides some pretty big discrepancies).

For the distribution to be independent of the nuisance parameters your parametrization must be sufficiently flexible.
A VERY IMPORTANT POINT

If we keep pushing this point to the extreme, the physics problem goes beyond what we can handle practically.

The p-values are usually predicated on the assumption that the true distribution is in the family of distributions being considered.

- eg. we have sufficiently flexible models of signal & background to incorporate all systematic effects.
- but we don’t believe we simulate everything perfectly.
- ..and when we parametrize our models usually we have further approximated our simulation.

- nature -> simulation -> parametrization

At some point these approaches are limited by honest systematics uncertainties (not statistical ones). Statistics can only help us so much after this point. Now we must be physicists!
LOOK-ELSEWHERE EFFECT

Approximation best above $3\sigma$

Typically our signal model has some parameter (e.g. $m_H$), which does not affect the null (background only).

This modifies the distribution of the likelihood ratio test statistic we call this the “look-elsewhere effect”

Recently Gross & Vitells found the results of Rice, Davies, and Leadbetter for a fast asymptotic approximation for the global $p$-value

$$
    p_0^{global} \approx p_0^{local} + \left\langle N(q_{ref}) \right\rangle e^{-(q_{test}-q_{ref})/2}
$$


R. B. Davies, Hypothesis testing when a nuisance parameter is present only under the alternative, *Biometrika* 64 (1977); *Biometrika* 74 (1987).
Even if we fix the location of the signal, some systematic effects are equivalent to small uncertainty in the location (e.g. energy calibration).
A MORE SUBTLE EFFECT

Even if we fix the location of the signal some systematic effects are equivalent to small uncertainty in the location (e.g. energy calibration).

These parameters are slowing convergence to the asymptotic distribution and variance may not reduce with more data.

O. Vitells found exact solution by Leadbetter for the case of only one such nuisance parameter

\[ \mathbb{E}[N_u] = \sigma_2 \int \phi(u(M)) \left[ \phi \left( \frac{u'(M)}{\sigma_2} \right) + \frac{u'(M)}{\sigma_2} \left\{ \Phi \left( \frac{u'(M)}{\sigma_2} \right) - \frac{1}{2} \right\} \right] dM \]

\[ \sigma_M = 2 \]

\[ \text{simulation} \quad - \quad \chi^2/2 \quad - \quad \text{Leadbetter} \]

note: used in Higgs discovery papers
LIMITS & CONFIDENCE INTERVALS
**What do these plots mean?**

---

**Plot 1:**
- **$\Delta \chi^2$** against $m_H$ (GeV).
- The plot shows excluded regions and theoretical uncertainty bands.
- The theoretical uncertainty bands are indicated by different colors and line styles.
- The mH limit is 161 GeV.

**Plot 2:**
- **$m_W$** and **$m_t$** (GeV) axes.
- The plot compares LEP2 and Tevatron, LEP1 and SLD, and LEP1 and SLD with theoretical uncertainty.
- The plot includes low $Q^2$ data.
- The limit $m_H = 161$ GeV is shown.

**Plot 3:**
- **Tevatron Run II Preliminary, $L \leq 8.6$ fb$^{-1}$**
- LEP exclusion regions and Tevatron exclusion regions.
- The plot shows expected and observed limits.
- The SM=1 line is included.

**Plot 4:**
- **ATLAS Preliminary**
- Observed CLs and expected CLs.
- The plot is for 1.0-1.2 fb$^{-1}$ and $s = 7$ TeV.
- The 95% CL limit on $\sigma_{\text{SM}}$ is shown.

---

July 2011
OTHER EXAMPLES OF CONFIDENCE INTERVALS

Figure 11.2: Constraints on the $\bar{\rho}$, $\bar{\eta}$ plane. The shaded areas have 95% CL. Color version at end of book.

These values are obtained using the method of Refs. [6,95]. Using the prescription of Refs. [102,118] gives $\lambda = 0.2246 \pm 0.0011$, $A = 0.832 \pm 0.017$, $\bar{\rho} = 0.130 \pm 0.018$, $\bar{\eta} = 0.350 \pm 0.013$ [119]. The fit results for the magnitudes of all nine CKM elements are:

$$V_{\text{CKM}} = \begin{pmatrix}
0.97428 \pm 0.00015 & 0.2253 \pm 0.0007 & 0.00347 \pm 0.00016 - 0.00012 \\
0.2252 \pm 0.0007 & 0.97345 + 0.00015 - 0.00016 \\
0.00862 + 0.00026 - 0.00020 & 0.999152 + 0.000030 - 0.000045
\end{pmatrix},$$

(11.27)

and the Jarlskog invariant is

$$J = (2.91 + 0.19 - 0.11) \times 10^{-5}.$$
What is a “Confidence Interval?  

- you see them all the time: 

Want to say there is a 68% chance that the true value of \((m_W, m_t)\) is in this interval  

- but that’s \(P(\text{theory}|\text{data})!\)

Correct frequentist statement is that the interval **covers** the true value 68% of the time 

- remember, the contour is a function of the data, which is random. So it moves around from experiment to experiment 

Bayesian “credible interval” does mean probability parameter is in interval. The procedure is very intuitive:

\[
P(\theta \in V) = \int_V \pi(\theta|x) = \int_V d\theta \frac{f(x|\theta)\pi(\theta)}{\int d\theta f(x|\theta)\pi(\theta)}
\]
“Bayesians address the question everyone is interested in, by using assumptions no-one believes”

“Frequentists use impeccable logic to deal with an issue of no interest to anyone”

-L. Lyons
Discovery in pictures

Discovery: test b-only (null: s=0 vs. alt: s>0)
- note, **one-sided** alternative. larger N is “more discrepant”

![Graph showing probability distributions for b-only and s+b events with observed value marked as more discrepant.](image)
Upper limits in pictures

What is meant by “95% upper limit”? See the picture below?

- i.e. increase $s$, until the probability to have data “more discrepant” is < 5%

What is meant by “95% upper limit”? See the picture below?

- i.e. increase $s$, until the probability to have data “more discrepant” is < 5%
The sensitivity problem

The physicist’s worry about limits in general is that if there is a strong downward fluctuation, one might exclude arbitrarily small values of $s$

- with a procedure that produces proper frequentist 95% confidence intervals, one should expect to exclude the true value of $s$ 5% of the time, no matter how small $s$ is!
  - This is not a problem with the procedure, but an undesirable consequence of the Type I / Type II error-rate setup
To address the sensitivity problem, $CL_s$ was introduced

- common (misused) nomenclature: $CL_s = CL_{s+b}/CL_b$
- idea: only exclude if $CL_s<5\%$ (if $CL_b$ is small, $CL_s$ gets bigger)

$CL_s$ is known to be "conservative" (over-cover): expected limit covers with 97.5%

- Note: $CL_s$ is NOT a probability

http://inspirehep.net/record/599622
How do we generalize?

\[ V_{\text{CKM}} = \begin{pmatrix} 0.97428 \pm 0.00015 & 0.2253 \pm 0.0007 & 0.00347 \pm 0.00016 \\ 0.2252 \pm 0.0007 & 0.97345 \pm 0.00015 & -0.0410 \pm 0.0011 \\ 0.00862 \pm 0.00026 & -0.0403 \pm 0.0011 & 0.999152 \pm 0.000030 \end{pmatrix} \] (11.27)

and the Jarlskog invariant is

\[ J = (2.91 \pm 0.19 - 0.11) \times 10^{-5} \]

Fig. 11.2 illustrates the constraints on the $\bar{\rho}$, $\bar{\eta}$ plane from various measurements and the global fit result. The shaded 95% CL regions all overlap consistently around the global fit region, though the consistency of $|V_{ub}/V_{cb}|$ and $\sin 2\beta$ is not very good.
For each value of $\theta$ consider $f(x|\theta_o)$
Let's focus on a particular point $f(x|\theta_0)$
Let’s focus on a particular point $f(x|\theta_0)$

- we want a test of size $\alpha$
- equivalent to a $100(1 - \alpha)\%$ confidence interval on $\theta$
- so we find an acceptance region with $1 - \alpha$ probability
Let’s focus on a particular point $f(x|\theta_0)$

- No unique choice of an acceptance region
- here’s an example of a lower limit
Let’s focus on a particular point $f(x|\theta_0)$

- No unique choice of an acceptance region
- and an example of a central limit
Let’s focus on a particular point $f(x|\theta_0)$

- choice of this region is called an **ordering rule**
- In Feldman-Cousins approach, ordering rule is the likelihood ratio. Find contour of L.R. that gives size $\alpha$
Now make acceptance region for every value of $\theta$
This makes a **confidence belt** for $\theta$
This makes a **confidence belt** for $\theta$
the regions of **data** in the confidence belt can be considered as **consistent** with that value of $\theta$
Now we make a measurement $x_0$

the points $\theta$ where the belt intersects $x_0$ a part of the confidence interval in $\theta$ for this measurement

eg. $[\theta_-, \theta_+]$
A RESTATEMENT OF THE CONSTRUCTION

For every point \( \theta \), if it were true, the data would fall in its acceptance region with probability \( 1 - \alpha \).

If the data fell in that region, the point \( \theta \) would be in the interval [\( \theta_-, \theta_+ \)].

So the interval [\( \theta_-, \theta_+ \)] covers the true value with probability \( 1 - \alpha \).
A POINT ABOUT THE NEYMAN CONSTRUCTION

This is not Bayesian... it doesn’t mean the probability that the true value of \( \theta \) is in the interval is \( 1 - \alpha \)!
INVERTING HYPOTHESIS TESTS

There is a precise dictionary that explains how to move from hypothesis testing to confidence intervals

- **Type I error**: probability interval does not cover true value of the parameters (eg. it is now a function of the parameters)
- **Power** is probability interval does not cover a false value of the parameters (eg. it is now a function of the parameters)
  
  - We don’t know the true value, consider each point $\theta_0$ as if it were true

What about null and alternate hypotheses?

- when testing a point $\theta_0$ it is considered the null
- all other points considered “alternate”

So what about the Neyman-Pearson lemma & Likelihood ratio?

- as mentioned earlier, there are no guarantees like before
- a common generalization that has good power is:

$$\frac{f(x|H_0)}{f(x|H_1)} \quad \frac{f(x|\theta_0)}{f(x|\theta_{best}(x))}$$
Coverage is the probability that the interval covers the true value.

Methods based on the Neyman-Construction always cover.... by construction.
- sometimes they over-cover (eg. “conservative”)

Bayesian methods, do not necessarily cover
- but that’s not their goal.
- but that also means you shouldn’t interpret a 95% Bayesian “Credible Interval” in the same way

Coverage can be thought of as a calibration of our statistical apparatus. [explain under-/over-coverage]
**DISCRETE PROBLEMS**

In discrete problems (e.g., number counting analysis with counts described by a Poisson) one sees:

- discontinuities in the coverage (as a function of parameter)
- over-coverage (in some regions)
- Important for experiments with few events. There is a lot of discussion about this, not focusing on it here
Coverage can be different at each point in the parameter space

Example:

Poisson(+background), with a systematic uncertainty on efficiency:

\[ x \sim Pois(\varepsilon \mu + b) \quad e \sim G(\varepsilon, \sigma) \]

\( e \) is a measurement of the unknown efficiency \( \varepsilon \), with resolution \( \sigma \). \( \varepsilon \) is the efficiency (a “normalization factor”, can be larger than 1).
NEYMAN CONSTRUCTION WITH NUISANCE PARAMETERS

In the strict sense, one wants coverage for $\mu$ for all values of the nuisance parameters (here $\varepsilon$)

- The “full construction” one nuisance parameter

Challenge for full Neyman Construction is computational time (scan in 50-D isn’t practical) and to avoid significant over-coverage

- note: projection of nuisance parameters is a union (eg. set theory) not an integration (Bayesian)
**Profile Construction**

Gary Feldman presented an approximate Neyman Construction, based on the profile likelihood ratio as an ordering rule, but only performing the construction on a subspace (eg. their conditional maximum likelihood estimate)

The **profile construction** means that one does not need to scan each nuisance parameter (keeps dimensionality constant)

- easier computationally (in RooStats)

This approximation does not guarantee exact coverage, but

- tests indicate impressive performance
- one can expand about the profile construction to improve coverage, with the limiting case being the full construction
PROFILE CONSTRUCTION: PROFESSIONAL LITERATURE

While I have been calling it the "profile construction", it has been called a "hybrid resampling" technique by professional statisticians

› Note: ‘hybrid’ here has nothing to do with Bayesian-Frequentist Hybrid, but a connection to “boot-strapping”

Resampling methods for confidence intervals in group sequential trials

BY CHIN-SHAN CHUANG

Department of Statistics, University of Wisconsin at Madison, Madison, Wisconsin 53706, U.S.A.
cchuang@stat.wisc.edu

AND TZE LEUNG LAI

Department of Statistics, Stanford University, Stanford, California 94305, U.S.A.
lait@leland.stanford.edu

Quick Announcement

Luc Demortier has done first coverage study (that I have seen) of our standard approach (the profile construction) for dealing with nuisance parameters in the Neyman Construction when Asymptotics are not necessarily valid.

- Results are very good: no significant undercoverage even for small counts. Good news for SUSY and exotics.
ASYMPTOTIC PROPERTIES OF LIKELIHOOD BASED TESTS
&
LIKELIHOOD-BASED METHODS
**Likelihood-based Intervals**

Wilks’s theorem tells us how the profile likelihood ratio evaluated at $\theta$ is “asymptotically” distributed **when $\theta$ is true**

- asymptotically means there is sufficient data that the log-likelihood function is parabolic
- does NOT require the model $f(x|\theta)$ to be Gaussian

So we don’t really need to go to the trouble to build its distribution by using Toy Monte Carlo or fancy tricks with Fourier Transforms

We can go immediately to the threshold value of the profile likelihood ratio
**LIKELIHOOD-BASED INTERVALS**

Figure from R. Cousins, Am. J. Phys. 63 398 (1995)

And typically we only show the likelihood curve and don’t even bother with the implicit (asymptotic) distribution.
Wilks’s theorem gives a short-cut for the Monte Carlo procedure used to find threshold on test statistic ⇒ MINOS is asymptotic approximation of Feldman-Cousins

- With a physical constraint ($\mu > 0$) the confidence band changes

$$t_\mu = -2 \ln \lambda(\mu)$$

$$\tilde{t}_\mu = -2 \ln \tilde{\lambda}(\mu) = \begin{cases} 
-2 \ln \frac{L(\mu, \hat{\theta}(\mu))}{L(0, \hat{\theta}(0))} & \hat{\mu} < 0 , \\
-2 \ln \frac{L(\mu, \hat{\theta}(\mu))}{L(\hat{\mu}, \theta)} & \hat{\mu} \geq 0 . 
\end{cases}$$
MODIFIED TEST STATISTIC FOR 1-SIDED UPPER LIMITS

For 1-sided upper-limit the threshold on the test statistic is different

- and with physical boundaries, it is again more complicated

\[ q_{\mu} = \begin{cases} -2 \ln \lambda(\mu) & \hat{\mu} \leq \mu, \\ 0 & \hat{\mu} > \mu, \end{cases} \]

\[ \tilde{q}_{\mu} = \begin{cases} -2 \ln \frac{L(\mu, \hat{\theta}(\mu))}{L(0, \hat{\theta}(0))} & \hat{\mu} < 0 \\ -2 \ln \frac{L(\mu, \hat{\theta}(\mu))}{L(\hat{\mu}, \theta)} & 0 \leq \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu. \end{cases} \]
Some non-trivial tests: boundaries

Monte Carlo test of asymptotic formulae

\[ f(\tilde{q}_\mu | \mu') = \Phi \left( \frac{\mu' - \mu}{\sigma} \right) \delta(\tilde{q}_\mu) \]

Same message for test based on \( \tilde{q}_\mu \).

\( q_\mu \) and \( \tilde{q}_\mu \) give similar tests to the extent that asymptotic formulae are valid.

We now can describe effect of the boundary on the distribution of the test statistic.
2-sided boundaries on physical parameter

Asymptotic distribution for two–sided tests with lower and upper boundaries on the parameter of interest

Glen Cowan, Kyle Cranmer, Eilam Gross, Ofer Vitells

with

\[
 f_L(t_\mu|\mu') = \begin{cases} 
 \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( \sqrt{t_\mu - \frac{\hat{\mu}^2}{\sigma^2}} \right)^2 \right] & \hat{t}_\mu \leq \delta_L^2 \\
 \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( \frac{\hat{t}_\mu - (\mu - \mu_\pm)\sigma}{\sigma^2} \right)^2 \right] & \hat{t}_\mu > \delta_L^2 
\end{cases}
\]

(4)

and

\[
 f_R(t_\mu|\mu') = \begin{cases} 
 \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( \sqrt{t_\mu + \frac{\hat{\mu}^2}{\sigma^2}} \right)^2 \right] & \hat{t}_\mu \leq \delta_R^2 \\
 \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( \frac{\hat{t}_\mu + (\mu - \mu_\pm)\sigma}{\sigma^2} \right)^2 \right] & \hat{t}_\mu > \delta_R^2 
\end{cases}
\]

(5)

where the dimensionless variables \( \delta_- = (\mu - \mu_-)/\sigma, \delta'_- = (\mu' - \mu_-)/\sigma, \delta_+ = (\mu - \mu_+)/\sigma, \) and \( \delta'_+ = (\mu' - \mu_+)/\sigma \) are used to simplify the expressions.

arXiv:1210.6948
THUMBNAIL OF THE STATISTICAL PROCEDURE

\[ \lambda(\mu) = \frac{L(\mu, \hat{\theta}(\mu))}{L(\hat{\mu}, \hat{\theta})} \]

CL_s to test signal hypothesis

p_0 to test background hypothesis

\[ \hat{\mu} \] to estimate signal strength

Follow LHC-HCG Combination Procedures
BAYESIAN METHODS

Two centuries later (when this Book had become an official prayer book of the Church of England) Thomas Bayes was a non-conformist minister (Presbyterian) who refused to use Cranmer’s book
COVERAGE & LIKELIHOOD PRINCIPLE

Methods based on the Neyman-Construction always cover.... by construction.

‣ this approach violates the likelihood principle

Bayesian methods obey likelihood principle, but do not necessarily cover

‣ that doesn’t mean Bayesians shouldn’t care about coverage

Coverage can be thought of as a calibration of our statistical apparatus. [explain under-/over-coverage]

What should be the view today:

Objective Bayesian analysis is the best frequentist tool around.

-Jim Berger

Bayesian and Frequentist results answer different questions

‣ major differences between them may indicate severe coverage problems and/or violations of the likelihood principle
Here we see comparisons of explicit ensembles generated with Monte Carlo techniques, the asymptotic results, and Bayesian results using MCMC and nested sampling with a uniform prior on $\mu$. 

\[ \int_0^L dt = 1.04 - 4.9 \text{ fb}^{-1} \]

\[ \sqrt{s} = 7 \text{ TeV} \]
Bayesian “credible interval” $V$ does mean that there is a 95% that the probability parameter is in interval.

The procedure is very intuitive:

$$P(\theta \in V) = \int_V \pi(\theta | x) = \int_V d\theta \frac{f(x | \theta) \pi(\theta)}{\int d\theta f(x | \theta) \pi(\theta)}$$
Markov Chain Monte Carlo (MCMC) is a nice technique which will produce a sampling of a parameter space which is proportional to a posterior

- it works well in high dimensional problems
- **Metropolis-Hastings Algorithm**: generates a sequence of points \( \{ \vec{\alpha}^{(t)} \} \)
  
  - Given the likelihood function \( L(\vec{\alpha}) \) & prior \( P(\vec{\alpha}) \), the posterior is proportional to \( L(\vec{\alpha}) \cdot P(\vec{\alpha}) \)
  
  - propose a point \( \vec{\alpha}' \) to be added to the chain according to a proposal density \( Q(\vec{\alpha}'|\vec{\alpha}) \) that depends only on current point \( \vec{\alpha} \)
  
  - if posterior is higher at \( \vec{\alpha}' \) than at \( \vec{\alpha} \), then add new point to chain
  
  - else: add \( \vec{\alpha}' \) to the chain with probability
    \[
    \rho = \frac{L(\vec{\alpha}') \cdot P(\vec{\alpha}')}{L(\vec{\alpha}) \cdot P(\vec{\alpha})} \cdot \frac{Q(\vec{\alpha}|\vec{\alpha}')}{Q(\vec{\alpha}'|\vec{\alpha})}
    \]
    
    - (appending original point \( \vec{\alpha} \) with complementary probability)

- **RooStats** works with any \( L(\vec{\alpha}) , P(\vec{\alpha}) \)

- Can use any RooFit PDF as proposal function \( Q(\vec{\alpha}'|\vec{\alpha}) \)
Physicist Sir Harold Jeffreys had the clever idea that we can “objectively” create a flat prior uniform in a metric determined by \( I(\theta) \)

Adds “minimal information” in a precise sense, and results in:

\[
p(\hat{\theta}) \propto \sqrt{I(\hat{\theta})}.
\]

It has the key feature that it is invariant under reparameterization of the parameter vector \( \varphi \). In particular, for an alternate parameterization \( \theta \) we can derive

\[
p(\vec{\varphi}) = p(\vec{\theta}) \left| \det \left( \frac{\partial \theta_i}{\partial \varphi_j} \right) \right| \propto \sqrt{I(\vec{\theta})\det^2 \left( \frac{\partial \theta_i}{\partial \varphi_j} \right)}
\]

\[
= \sqrt{ \det \left( \frac{\partial \theta_k}{\partial \varphi_i} \right) \det \left( E \left[ \frac{\partial \ln L}{\partial \theta_k} \frac{\partial \ln L}{\partial \theta_l} \right] \right) \det \left( \frac{\partial \theta_l}{\partial \varphi_j} \right)}
\]

\[
= \sqrt{ \det \left( E \left[ \sum_{k,l} \frac{\partial \theta_k}{\partial \varphi_i} \frac{\partial \ln L}{\partial \theta_k} \frac{\partial \ln L}{\partial \theta_l} \frac{\partial \theta_l}{\partial \varphi_j} \right] \right) }
\]

\[
= \sqrt{ \det \left( E \left[ \frac{\partial \ln L}{\partial \varphi_i} \frac{\partial \ln L}{\partial \varphi_j} \right] \right) } = \sqrt{I(\vec{\varphi})}.
\]

Unfortunately, the Jeffreys prior in multiple dimensions causes some problems, and in certain circumstances gives undesirable answers.
Reference priors are another type of “objective” priors, that try to save Jeffreys’ basic idea.

Noninformative priors have been studied for a long time and most of them have been found defective in more than one way. Reference analysis arose from this study as the only general method that produces priors that have the required invariance properties, deal successfully with the marginalization paradoxes, and have consistent sampling properties.

Ideally, such a method should be very general, applicable to all kinds of measurements regardless of the number and type of parameters and data involved. It should make use of all available information, and coherently so, in the sense that if there is more than one way to extract all relevant information from data, the final result will not depend on the chosen way. The desiderata of generality, exhaustiveness and coherence are satisfied by Bayesian procedures, but that of objectivity is more problematic due to the Bayesian requirement of specifying prior probabilities in terms of degrees of belief. Reference analysis\(^2\), an objective Bayesian method developed over the past twenty-five years, solves this problem by replacing the question “what is our prior degree of belief?” by “what would our posterior degree of belief be, if our prior knowledge had a minimal effect, relative to the data, on the final inference?”

See Luc Demortier’s PhyStat 2005 proceedings

http://physics.rockefeller.edu/luc/proceedings/phystat2005_refana.ps
JEFFREYS’S PRIOR

Jeffreys’s Prior is an “objective” prior based on formal rules (it is related to the Fisher Information and the Cramér-Rao bound)

\[ \pi(\tilde{\theta}) \propto \sqrt{\det \mathcal{I}(\tilde{\theta})}. \]

\[ (\mathcal{I}(\theta))_{i,j} = -E \left[ \frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln f(X;\theta) \bigg| \theta \right]. \]

Eilam, Glen, Ofer, and I showed in arXiv:1007.1727 that the Asimov data provides a fast, convenient way to calculate the Fisher Information

\[ V_{jk}^{-1} = -E \left[ \frac{\partial^2 \ln L}{\partial \theta_j \partial \theta_k} \right] = -\frac{\partial^2 \ln L_{\Lambda}}{\partial \theta_j \partial \theta_k} = \sum_{i=1}^{N} \frac{\partial \nu_i}{\partial \theta_j} \frac{\partial \nu_i}{\partial \theta_k} \frac{1}{\nu_i} + \sum_{i=1}^{M} \frac{\partial u_i}{\partial \theta_j} \frac{\partial u_i}{\partial \theta_k} \frac{1}{u_i} \]

Use this as basis to calculate
Jeffreys’s prior for an arbitrary PDF!

RooWorkspace w("w");
w.factory("Uniform::u(x[0,1])");
w.factory("mu[100,1,200]");
w.factory("ExtendPdf::p(u,mu)");

w.defineSet("poi","mu");
w.defineSet("obs","x");
// w.defineSet("obs2","n");

RooJeffreysPrior pi("jeffreys","jeffreys",*w.pdf("p"),*w.set("poi"),*w.set("obs"));
THE BAYESIAN SOLUTION

Bayesian solution generically have a prior for the parameters of interest as well as nuisance parameters
- 2010 recommendations largely echoes the PDG’s stance.

**Recommendation:** When performing a Bayesian analysis one should separate the objective likelihood function from the prior distributions to the extent possible.

**Recommendation:** When performing a Bayesian analysis one should investigate the sensitivity of the result to the choice of priors.

**Warning:** Flat priors in high dimensions can lead to unexpected and/or misleading results.

**Recommendation:** When performing a Bayesian analysis for a single parameter of interest, one should attempt to include Jeffreys's prior in the sensitivity analysis.
RooFit’s Workspace now provides the ability to save in a ROOT file the full likelihood model, any priors you might want, and the minimal data necessary to reproduce likelihood function.

Workspace has all the necessary information for both frequentist and Bayesian methods.
THE END

THANK YOU!