Systematic uncertainties and nuisance parameters

In general our model of the data is not perfect:

\[
\begin{align*}
  \text{model:} & \quad L(x|\theta) = \theta x \\
  \text{truth:} & \quad L(x|\theta) = \theta x + \alpha x^2 + \beta x^3 + \cdots
\end{align*}
\]

Can improve model by including additional adjustable parameters.

\[L(x|\theta) \rightarrow L(x|\theta, \nu)\]

Nuisance parameter ↔ systematic uncertainty. Some point in the parameter space of the enlarged model should be “true”.

Presence of nuisance parameter decreases sensitivity of analysis to the parameter of interest (e.g., increases variance of estimate).
\textit{p}-values in cases with nuisance parameters

Suppose we have a statistic $q_\theta$ that we use to test a hypothesized value of a parameter $\theta$, such that the $p$-value of $\theta$ is

$$ p_\theta = \int_{q_\theta, \text{obs}}^{\infty} f(q_\theta | \theta, \nu) \, dq_\theta $$

But what values of $\nu$ to use for $f(q_\theta | \theta, \nu)$?

Fundamentally we want to reject $\theta$ only if $p_\theta < \alpha$ for all $\nu$.

$\rightarrow$ “exact” confidence interval

Recall that for statistics based on the profile likelihood ratio, the distribution $f(q_\theta | \theta, \nu)$ becomes independent of the nuisance parameters in the large-sample limit.

But in general for finite data samples this is not true; one may be unable to reject some $\theta$ values if all values of $\nu$ must be considered, even those strongly disfavoured by the data (resulting interval for $\theta$ “overcovers”).
Profile construction ("hybrid resampling")

oai:cds.cern.ch:1021125, cdsweb.cern.ch/record/1099969.

Approximate procedure is to reject $\theta$ if $p_{\theta} \leq \alpha$ where the $p$-value is computed assuming the value of the nuisance parameter that best fits the data for the specified $\theta$:

$$\nu(\theta)$$

"double hat" notation means value of parameter that maximizes likelihood for the given $\theta$.

The resulting confidence interval will have the correct coverage for the points $(\theta, \hat{\nu}(\theta))$.

Elsewhere it may under- or overcover, but this is usually as good as we can do (check with MC if crucial or small sample problem).
“Hybrid frequentist-Bayesian” method

Alternatively, suppose uncertainty in $\nu$ is characterized by a Bayesian prior $\pi(\nu)$.

Can use the marginal likelihood to model the data:

$$L_m(x|\theta) = \int L(x|\theta, \nu)\pi(\nu)\,d\nu$$

This does not represent what the data distribution would be if we “really” repeated the experiment, since then $\nu$ would not change.

But the procedure has the desired effect. The marginal likelihood effectively builds the uncertainty due to $\nu$ into the model.

Use this now to compute (frequentist) $p$-values $\rightarrow$ the model being tested is in effect a weighted average of models.
Example of treatment of nuisance parameters: fitting a straight line

Data: \((x_i, y_i, \sigma_i), i = 1, \ldots, n\).

Model: \(y_i\) independent and all follow \(y_i \sim \text{Gauss}(\mu(x_i), \sigma_i)\)

\[\mu(x; \theta_0, \theta_1) = \theta_0 + \theta_1 x,\]

assume \(x_i\) and \(\sigma_i\) known.

Goal: estimate \(\theta_0\)

Here suppose we don’t care about \(\theta_i\) (example of a “nuisance parameter”)

![Graph showing data points and model line](image.png)
**Maximum likelihood fit with Gaussian data**

In this example, the $y_i$ are assumed independent, so the likelihood function is a product of Gaussians:

$$L(\theta_0, \theta_1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i}} \exp \left[ -\frac{1}{2} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} \right],$$

Maximizing the likelihood is here equivalent to minimizing

$$\chi^2(\theta_0, \theta_1) = -2 \ln L(\theta_0, \theta_1) + \text{const} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2}.$$

i.e., for Gaussian data, ML same as Method of Least Squares (LS)
\( \theta_1 \) known a priori

\[
L(\theta_0) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i}} \exp \left[ -\frac{1}{2} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} \right].
\]

\[
\chi^2(\theta_0) = -2 \ln L(\theta_0) + \text{const} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2}.
\]

For Gaussian \( y_i \), ML same as LS

Minimize \( \chi^2 \rightarrow \) estimator \( \hat{\theta}_0 \).

Come up one unit from \( \chi^2_{\text{min}} \) to find \( \sigma_{\hat{\theta}_0} \).
ML (or LS) fit of $\theta_0$ and $\theta_1$

$$
\chi^2(\theta_0, \theta_1) = -2 \ln L(\theta_0, \theta_1) + \text{const} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2}.
$$

Standard deviations from tangent lines to contour

$$
\chi^2 = \chi^2_{\text{min}} + 1.
$$

Correlation between $\hat{\theta}_0$, $\hat{\theta}_1$ causes errors to increase.
If we have a measurement $t_1 \sim \text{Gauss} \left( \theta_1, \sigma_{t_1} \right)$

$$\chi^2(\theta_0, \theta_1) = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} + \frac{(\theta_1 - t_1)^2}{\sigma_{t_1}^2}.$$ 

The information on $\theta_1$ improves accuracy of $\hat{\theta}_0$.
Bayesian method

We need to associate prior probabilities with $\theta_0$ and $\theta_1$, e.g.,

$$\pi(\theta_0, \theta_1) = \pi_0(\theta_0) \pi_1(\theta_1)$$

‘non-informative’, in any case much broader than $L(\theta_0)$

$$\pi_0(\theta_0) = \text{const.}$$

$$\pi_1(\theta_1) = \frac{1}{\sqrt{2\pi \sigma_{t_1}^2}} e^{-(\theta_1-t_1)^2/2\sigma_{t_1}^2}$$

← based on previous measurement

Putting this into Bayes’ theorem gives:

$$p(\theta_0, \theta_1 | \bar{y}) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi \sigma_i}} e^{-(y_i-\mu(x_i; \theta_0, \theta_1))^2/2\sigma_i^2} \pi_0 \frac{1}{\sqrt{2\pi \sigma_{t_1}^2}} e^{-(\theta_1-t_1)^2/2\sigma_{t_1}^2}$$

posterior $\propto$ likelihood $\times$ prior
Bayesian method (continued)

We then integrate (marginalize) \( p(\theta_0, \theta_1 | x) \) to find \( p(\theta_0 | x) \):

\[
p(\theta_0|x) = \int p(\theta_0, \theta_1|x) \, d\theta_1.
\]

In this example we can do the integral (rare). We find

\[
p(\theta_0|x) = \frac{1}{\sqrt{2\pi}\sigma_{\theta_0}} e^{-\frac{(\theta_0-\hat{\theta}_0)^2}{2\sigma_{\theta_0}^2}} \quad \text{with}
\]

\[
\hat{\theta}_0 = \text{same as ML estimator}
\]

\[
\sigma_{\theta_0} = \sigma_{\hat{\theta}_0} \quad \text{(same as before)}
\]

Usually need numerical methods (e.g. Markov Chain Monte Carlo) to do integral.
Digression: marginalization with MCMC

Bayesian computations involve integrals like

\[ p(\theta_0|x) = \int p(\theta_0, \theta_1|x) \, d\theta_1. \]

often high dimensionality and impossible in closed form, also impossible with ‘normal’ acceptance-rejection Monte Carlo.

Markov Chain Monte Carlo (MCMC) has revolutionized Bayesian computation.

MCMC (e.g., Metropolis-Hastings algorithm) generates correlated sequence of random numbers:

- cannot use for many applications, e.g., detector MC;
- effective stat. error greater than if all values independent.

Basic idea: sample multidimensional \( \vec{\theta} \), look, e.g., only at distribution of parameters of interest.
MCMC basics: Metropolis-Hastings algorithm

Goal: given an $n$-dimensional pdf $p(\vec{\theta})$, generate a sequence of points $\vec{\theta}_1, \vec{\theta}_2, \vec{\theta}_3, \ldots$

1) Start at some point $\vec{\theta}_0$
2) Generate $\vec{\theta} \sim q(\vec{\theta}; \vec{\theta}_0)$
3) Form Hastings test ratio $\alpha = \min \left[ 1, \frac{p(\vec{\theta})q(\vec{\theta}_0; \vec{\theta})}{p(\vec{\theta}_0)q(\vec{\theta}; \vec{\theta}_0)} \right]$
4) Generate $u \sim \text{Uniform}[0, 1]$
5) If $u \leq \alpha$, $\vec{\theta}_1 = \vec{\theta}$, move to proposed point
    else $\vec{\theta}_1 = \vec{\theta}_0$, old point repeated
6) Iterate
Metropolis-Hastings (continued)

This rule produces a *correlated* sequence of points (note how each new point depends on the previous one).

For our purposes this correlation is not fatal, but statistical errors larger than if points were independent.

The proposal density can be (almost) anything, but choose so as to minimize autocorrelation. Often take proposal density symmetric: \( q(\vec{\theta}; \vec{\theta}_0) = q(\vec{\theta}_0; \vec{\theta}) \)

Test ratio is *(Metropolis-Hastings)*: \[ \alpha = \min \left[ 1, \frac{p(\vec{\theta})}{p(\vec{\theta}_0)} \right] \]

I.e. if the proposed step is to a point of higher \( p(\vec{\theta}) \), take it; if not, only take the step with probability \( p(\vec{\theta}) / p(\vec{\theta}_0) \). If proposed step rejected, hop in place.
Example: posterior pdf from MCMC

Sample the posterior pdf from previous example with MCMC:

Summarize pdf of parameter of interest with, e.g., mean, median, standard deviation, etc.

Although numerical values of answer here same as in frequentist case, interpretation is different (sometimes unimportant?)
Bayesian method with alternative priors

Suppose we don’t have a previous measurement of $\theta_1$ but rather, e.g., a theorist says it should be positive and not too much greater than 0.1 "or so", i.e., something like

$$\pi_1(\theta_1) = \frac{1}{\tau} e^{-\theta_1/\tau}, \quad \theta_1 \geq 0, \quad \tau = 0.1.$$  

From this we obtain (numerically) the posterior pdf for $\theta_0$:

This summarizes all knowledge about $\theta_0$.

Look also at result from variety of priors.
A typical fitting problem

Given measurements: 
\[ y_i \pm \sigma^\text{stat}_i \pm \sigma^\text{sys}_i, \quad i = 1, \ldots, n, \]

and (usually) covariances: 
\[ V^\text{stat}_{ij}, V^\text{sys}_{ij}. \]

Predicted value: \( \mu(x_i; \theta) \), expectation value \[ E[y_i] = \mu(x_i; \theta) + b_i \]

control variable parameters bias

Often take: \[ V_{ij} = V^\text{stat}_{ij} + V^\text{sys}_{ij} \]

Minimize \[ \chi^2(\theta) = (\bar{y} - \bar{\mu}(\theta))^T V^{-1} (\bar{y} - \bar{\mu}(\theta)) \]

Equivalent to maximizing \( L(\theta) \sim e^{-\chi^2/2} \), i.e., least squares same as maximum likelihood using a Gaussian likelihood function.
Its Bayesian equivalent

Take

\[ L(\vec{y}|\vec{\theta}, \vec{b}) \sim \exp \left[ -\frac{1}{2} (\vec{y} - \vec{\mu}(\theta) - \vec{b})^T V_{\text{stat}}^{-1} (\vec{y} - \vec{\mu}(\theta) - \vec{b}) \right] \]

\[ \pi_b(\vec{b}) \sim \exp \left[ -\frac{1}{2} \vec{b}^T V_{\text{sys}}^{-1} \vec{b} \right] \]

\[ \pi_\theta(\theta) \sim \text{const.} \]

Joint probability for all parameters

and use Bayes’ theorem:

\[ p(\theta, \vec{b}|\vec{y}) \propto L(\vec{y}|\theta, \vec{b}) \pi_\theta(\theta) \pi_b(\vec{b}) \]

To get desired probability for \( \theta \), integrate (marginalize) over \( b \):

\[ p(\theta|\vec{y}) = \int p(\theta, \vec{b}|\vec{y}) \, d\vec{b} \]

→ Posterior is Gaussian with mode same as least squares estimator, \( \sigma_\theta \) same as from \( \chi^2 = \chi^2_{\text{min}} + 1 \). (Back where we started!)
The error on the error

Some systematic errors are well determined

Error from finite Monte Carlo sample

Some are less obvious

Do analysis in $n$ ‘equally valid’ ways and extract systematic error from ‘spread’ in results.

Some are educated guesses

Guess possible size of missing terms in perturbation series;
vary renormalization scale \( \mu/2 < Q < 2\mu \) ?

Can we incorporate the ‘error on the error’?

(cf. G. D’Agostini 1999; Dose & von der Linden 1999)
Motivating a non-Gaussian prior $\pi_b(b)$

Suppose now the experiment is characterized by

$$y_i, \sigma_i^{\text{stat}}, \sigma_i^{\text{sys}}, s_i, \quad i = 1, \ldots, n,$$

where $s_i$ is an (unreported) factor by which the systematic error is over/under-estimated.

Assume correct error for a Gaussian $\pi_b(b)$ would be $s_i \sigma_i^{\text{sys}}$, so

$$\pi_b(b_i) = \int \frac{1}{\sqrt{2\pi s_i \sigma_i^{\text{sys}}}} \exp \left[ -\frac{1}{2} \frac{b_i^2}{(s_i \sigma_i^{\text{sys}})^2} \right] \pi_s(s_i) \, ds_i$$

Width of $\sigma_s(s_i)$ reflects ‘error on the error’.
Error-on-error function $\pi_s(s)$

A simple unimodal probability density for $0 < s < 1$ with adjustable mean and variance is the Gamma distribution:

$$\pi_s(s) = \frac{a (a s)^{b-1} e^{-a s}}{\Gamma(b)}$$

mean $= b/a$

variance $= b/a^2$

Want e.g. expectation value of 1 and adjustable standard Deviation $\sigma_s$, i.e., $a = b = 1/\sigma_s^2$

In fact if we took $\pi_s(s) \sim \text{inverse Gamma}$, we could integrate $\pi_b(b)$ in closed form (cf. D’Agostini, Dose, von Linden). But Gamma seems more natural & numerical treatment not too painful.
Prior for bias $\pi_b(b)$ now has longer tails

$$\pi_b(b_i) = \int \frac{1}{\sqrt{2\pi s_i \sigma_{sys}^i}} \exp \left[ -\frac{1}{2} \frac{b_i^2}{(s_i \sigma_{sys}^i)^2} \right] \pi_s(s_i) \, ds_i$$

Gaussian ($\sigma_s = 0$)  \[ P(|b| > 4\sigma_{sys}) = 6.3 \times 10^{-5} \]

$\sigma_s = 0.5$  \[ P(|b| > 4\sigma_{sys}) = 0.65\% \]
A simple test

Suppose fit effectively averages four measurements.

Take $\sigma_{\text{sys}} = \sigma_{\text{stat}} = 0.1$, uncorrelated.

Case #1: data appear compatible

Posterior $p(\mu|y)$:

Usually summarize posterior $p(\mu|y)$ with mode and standard deviation:

$\sigma_s = 0.0 : \quad \hat{\mu} = 1.000 \pm 0.071$

$\sigma_s = 0.5 : \quad \hat{\mu} = 1.000 \pm 0.072$
Simple test with inconsistent data

Case #2: there is an outlier

→ Bayesian fit less sensitive to outlier.

→ Error now connected to goodness-of-fit.

Posterior $p(\mu | y)$:
Goodness-of-fit vs. size of error

In LS fit, value of minimized $\chi^2$ does not affect size of error on fitted parameter.

In Bayesian analysis with non-Gaussian prior for systematics, a high $\chi^2$ corresponds to a larger error (and vice versa).

2000 repetitions of experiment, $\sigma_s = 0.5$, here no actual bias.

$\sigma_\mu$ from least squares
Is this workable in practice?

Should to generalize to include correlations

Prior on correlation coefficients $\sim \pi(\rho)$

(Myth: $\rho = 1$ is “conservative”)

Can separate out different systematic for same measurement

Some will have small $\sigma_s$, others larger.

Remember the “if-then” nature of a Bayesian result:

We can (should) vary priors and see what effect this has on the conclusions.
Bayesian model selection (‘discovery’) 

The probability of hypothesis $H_0$ relative to its complementary alternative $H_1$ is often given by the posterior odds:

$$\frac{P(H_0|x)}{P(H_1|x)} = \frac{P(x|H_0)}{P(x|H_1)} \times \frac{\pi(H_0)}{\pi(H_1)}$$

The Bayes factor is regarded as measuring the weight of evidence of the data in support of $H_0$ over $H_1$.

Interchangeably use $B_{10} = 1/B_{01}$
Assessing Bayes factors

One can use the Bayes factor much like a \( p \)-value (or \( Z \) value).

The Jeffreys scale, analogous to HEP's \( 5\sigma \) rule:

\[
B_{10} \quad \text{Evidence against } H_0
\]

<table>
<thead>
<tr>
<th>( B_{10} )</th>
<th>\text{Evidence against } H_0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 to 3</td>
<td>Not worth more than a bare mention</td>
</tr>
<tr>
<td>3 to 20</td>
<td>Positive</td>
</tr>
<tr>
<td>20 to 150</td>
<td>Strong</td>
</tr>
<tr>
<td>&gt; 150</td>
<td>Very strong</td>
</tr>
</tbody>
</table>

Rewriting the Bayes factor

Suppose we have models \( H_i, i = 0, 1, \ldots, \)
each with a likelihood \( p(x | H_i, \theta_i) \)
and a prior pdf for its internal parameters \( \pi_i(\theta_i) \)
so that the full prior is \( \pi(H_i, \theta_i) = p_i \pi_i(\theta_i) \)
where \( p_i = P(H_i) \) is the overall prior probability for \( H_i \).

The Bayes factor comparing \( H_i \) and \( H_j \) can be written

\[
B_{ij} = \frac{P(H_i | \bar{x})}{P(H_i)} \frac{P(H_j | \bar{x})}{P(H_j)}
\]
Bayes factors independent of $P(H_i)$

For $B_{ij}$ we need the posterior probabilities marginalized over all of the internal parameters of the models:

\[
P(H_i|\vec{x}) = \int P(H_i, \vec{\theta}_i|\vec{x}) \, d\vec{\theta}_i
\]

\[
= \int L(\vec{x}|H_i, \vec{\theta}_i) p_i \pi_i(\vec{\theta}_i) \, d\vec{\theta}_i
\]

\[
= \frac{\int L(\vec{x}|H_i, \vec{\theta}_i) p_i \pi_i(\vec{\theta}_i) \, d\vec{\theta}_i}{P(x)}
\]

So therefore the Bayes factor is

\[
B_{ij} = \frac{\int L(\vec{x}|H_i, \vec{\theta}_i) \pi_i(\vec{\theta}_i) \, d\vec{\theta}_i}{\int L(\vec{x}|H_j, \vec{\theta}_j) \pi_j(\vec{\theta}_j) \, d\vec{\theta}_j}
\]

The prior probabilities $p_i = P(H_i)$ cancel.
Numerical determination of Bayes factors

Both numerator and denominator of $B_{ij}$ are of the form

$$m = \int L(\bar{x}|\bar{\theta}) \pi(\bar{\theta}) \, d\bar{\theta} \quad \text{‘marginal likelihood’}$$

Various ways to compute these, e.g., using sampling of the posterior pdf (which we can do with MCMC).

- Harmonic Mean (and improvements)
- Importance sampling
- Parallel tempering (~thermodynamic integration)
- Nested Sampling (MultiNest), ...


Priors for Bayes factors

Note that for Bayes factors (unlike Bayesian limits), the prior cannot be improper. If it is, the posterior is only defined up to an arbitrary constant, and so the Bayes factor is ill defined.

Possible exception allowed if both models contain *same* improper prior; but having same parameter name (or Greek letter) in both models does not fully justify this step.

If improper prior is made proper e.g. by a cut-off, the Bayes factor will retain a dependence on this cut-off.

In general or Bayes factors, all priors must reflect “meaningful” degrees of uncertainty about the parameters.
Harmonic mean estimator

E.g., consider only one model and write Bayes theorem as:

\[
\frac{\pi(\theta)}{m} = \frac{p(\theta|x)}{L(x|\theta)}
\]

\(\pi(\theta)\) is normalized to unity so integrate both sides,

\[
m^{-1} = \int \frac{1}{L(x|\theta)} p(\theta|x) \, d\theta = E_p[1/L]
\]

Therefore sample \(\theta\) from the posterior via MCMC and estimate \(m\) with one over the average of \(1/L\) (the harmonic mean of \(L\)).

Improvements to harmonic mean estimator

The harmonic mean estimator is numerically very unstable; formally infinite variance (!). Gelfand & Dey propose variant:

Rearrange Bayes thm; multiply both sides by arbitrary pdf $f(\theta)$:

$$\frac{f(\theta)p(\theta|x)}{L(x|\theta)\pi(\theta)} = \frac{f(\theta)}{m}$$

Integrate over $\theta$:

$$m^{-1} = \int \frac{f(\theta)}{L(x|\theta)\pi(\theta)}p(\theta|x) = E_p \left[ \frac{f(\theta)}{L(x|\theta)\pi(\theta)} \right]$$

Improved convergence if tails of $f(\theta)$ fall off faster than $L(x|\theta)\pi(\theta)$

Note harmonic mean estimator is special case $f(\theta) = \pi(\theta)$.

Importance sampling

Need pdf $f(\theta)$ which we can evaluate at arbitrary $\theta$ and also sample with MC.

The marginal likelihood can be written

$$m = \int \frac{L(x|\theta)\pi(\theta)}{f(\theta)} f(\theta) \, d\theta = E_f \left[ \frac{L(x|\theta)\pi(\theta)}{f(\theta)} \right]$$

Best convergence when $f(\theta)$ approximates shape of $L(x|\theta)\pi(\theta)$.

Use for $f(\theta)$ e.g. multivariate Gaussian with mean and covariance estimated from posterior (e.g. with MINUIT).
The nested sampling algorithm

An algorithm originally aimed primarily at the Bayesian evidence computation (Skilling, 2006):

\[ X(\lambda) = \int_{\mathcal{L}(\theta) > \lambda} P(\theta) \, d\theta \]

\[ P(d) = \int d\theta \mathcal{L}(\theta) P(\theta) = \int_0^1 X(\lambda) \, d\lambda \]

Extra slides
The Look-Elsewhere Effect

Suppose a model for a mass distribution allows for a peak at a mass $m$ with amplitude $\mu$.

The data show a bump at a mass $m_0$.

How consistent is this with the no-bump ($\mu = 0$) hypothesis?
Local $p$-value

First, suppose the mass $m_0$ of the peak was specified a priori. Test consistency of bump with the no-signal ($\mu = 0$) hypothesis with e.g. likelihood ratio

$$t_{\text{fix}} = -2 \ln \frac{L(0, m_0)}{L(\hat{\mu}, m_0)}$$

where “fix” indicates that the mass of the peak is fixed to $m_0$. The resulting $p$-value

$$p_{\text{local}} = \int_{t_{\text{fix}, \text{obs}}}^{\infty} f(t_{\text{fix}}|0) \, dt_{\text{fix}}$$

gives the probability to find a value of $t_{\text{fix}}$ at least as great as observed at the specific mass $m_0$ and is called the local $p$-value.
Global $p$-value

But suppose we did not know where in the distribution to expect a peak.

What we want is the probability to find a peak at least as significant as the one observed anywhere in the distribution.

Include the mass as an adjustable parameter in the fit, test significance of peak using

$$t_{\text{float}} = -2 \ln \frac{L(0)}{L(\hat{\mu}, \hat{\mu})}$$

(Note $m$ does not appear in the $\mu = 0$ model.)

$$p_{\text{global}} = \int_{t_{\text{float,obs}}}^{\infty} f(t_{\text{float}} | 0) \, dt_{\text{float}}$$
Distributions of $t_{\text{fix}}$, $t_{\text{float}}$

For a sufficiently large data sample, $t_{\text{fix}} \sim$ chi-square for 1 degree of freedom (Wilks’ theorem).

For $t_{\text{float}}$ there are two adjustable parameters, $\mu$ and $m$, and naively Wilks theorem says $t_{\text{float}} \sim$ chi-square for 2 d.o.f.

In fact Wilks’ theorem does not hold in the floating mass case because one of the parameters ($m$) is not-defined in the $\mu = 0$ model.

So getting $t_{\text{float}}$ distribution is more difficult.
Approximate correction for LEE

We would like to be able to relate the $p$-values for the fixed and floating mass analyses (at least approximately).

Gross and Vitell show the $p$-values are approximately related by

$$ p_{global} \approx p_{local} + \langle N(c) \rangle $$

where $\langle N(c) \rangle$ is the mean number “upcrossings” of $t_{fix} = -2\ln \lambda$ in the fit range based on a threshold

$$ c = t_{fix, obs} = Z_{local}^2 $$

and where $Z_{local} = \Phi^{-1}(1 - p_{local})$ is the local significance.

So we can either carry out the full floating-mass analysis (e.g. use MC to get $p$-value), or do fixed mass analysis and apply a correction factor (much faster than MC).
Upcrossings of $-2\ln L$

The Gross-Vitells formula for the trials factor requires $\langle N(c) \rangle$, the mean number “upcrossings” of $t_{\text{fix}} = -2\ln \lambda$ above a threshold $c = t_{\text{fix,obs}}$ found when varying the mass $m_0$ over the range considered.

$\langle N(c) \rangle$ can be estimated from MC (or the real data) using a much lower threshold $c_0$:

$$\langle N(c) \rangle \approx \langle N(c_0) \rangle e^{-(c-c_0)/2}$$

In this way $\langle N(c) \rangle$ can be estimated without need of large MC samples, even if the the threshold $c$ is quite high.
Multidimensional look-elsewhere effect

Generalization to multiple dimensions: number of upcrossings replaced by expectation of Euler characteristic:

$$E[\varphi(A_u)] = \sum_{d=0}^{n} \mathcal{N}_d \rho_d(u)$$

- Number of disconnected components minus number of `holes`

Applications: astrophysics (coordinates on sky), search for resonance of unknown mass and width, ...
Remember the Look-Elsewhere Effect is when we test a single model (e.g., SM) with multiple observations, i.e., in multiple places.

Note there is no look-elsewhere effect when considering exclusion limits. There we test specific signal models (typically once) and say whether each is excluded.

With exclusion there is, however, the analogous issue of testing many signal models (or parameter values) and thus excluding some even in the absence of signal (“spurious exclusion”)

Approximate correction for LEE should be sufficient, and one should also report the uncorrected significance.

“There's no sense in being precise when you don't even know what you're talking about.” — John von Neumann
Why 5 sigma?

Common practice in HEP has been to claim a discovery if the $p$-value of the no-signal hypothesis is below $2.9 \times 10^{-7}$, corresponding to a significance $Z = \Phi^{-1}(1 - p) = 5$ (a $5\sigma$ effect).

There a number of reasons why one may want to require such a high threshold for discovery:

- The “cost” of announcing a false discovery is high.
- Unsure about systematics.
- Unsure about look-elsewhere effect.
- The implied signal may be a priori highly improbable (e.g., violation of Lorentz invariance).
Why 5 sigma (cont.)?

But the primary role of the $p$-value is to quantify the probability that the background-only model gives a statistical fluctuation as big as the one seen or bigger.

It is not intended as a means to protect against hidden systematics or the high standard required for a claim of an important discovery.

In the processes of establishing a discovery there comes a point where it is clear that the observation is not simply a fluctuation, but an “effect”, and the focus shifts to whether this is new physics or a systematic.

Providing LEE is dealt with, that threshold is probably closer to $3\sigma$ than $5\sigma$. 
Jackknife, bootstrap, etc.

To estimate a parameter we have various tools such as maximum likelihood, least squares, etc.

Usually one also needs to know the variance (or the full sampling distribution) of the estimator – this can be more difficult.

Often use asymptotic properties, e.g., sampling distribution of ML estimators becomes Gaussian in large sample limit; std. dev. from curvature of log-likelihood at maximum.

The jackknife and bootstrap are examples of “resampling” methods used to estimate the sampling distribution of statistics.

In HEP we often do this implicitly by using Toy MC to determine sampling properties of statistics (e.g., Brazil plot for 1σ, 2σ bands of limits).
The Jackknife

Invented by Quenouille (1949) and Tukey (1958).

Suppose data sample consists of $n$ events: $x = (x_1, \ldots, x_n)$.

We have an estimator $\theta(x)$ for a parameter $\theta$.

Idea is to produce pseudo data samples $x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ by leaving out the $i$th event.

Let $\hat{\theta}_{-i}$ be the estimator obtained from the data sample $x_{-i}$.

Suppose the estimator has a nonzero bias: $b = E[\hat{\theta}] - \theta$

The jackknife estimator of the bias is

$$\hat{b}_j = \frac{n - 1}{n} \sum_{i=1}^{n} (\hat{\theta}_{-i} - \hat{\theta})$$

See, e.g., Notes on Jackknife and Bootstrap by G. J. Babu:

www.iiap.res.in/astrostat/School10/LecFiles/JBabu_JackknifeBootstrap_notes.pdf
The Bootstrap (Efron, 1979)

Idea is to produce a set of “bootstrapped” data samples of same size as the original (real) one by sampling from some distribution that approximates the true (unknown) one.

By evaluating a statistic (such as an estimator for a parameter $\theta$) with the bootstrapped-samples, properties of its sampling distribution (often its variance) can be estimated.

If the data consist of $n$ events, one way to produce the bootstrapped samples is to randomly select from the original sample $n$ events with replacement (the non-parametric bootstrap).

That is, some events might get used multiple times, others might not get used at all.

In other cases could generate the bootstrapped samples from a parametric MC model, using parameter values estimated from real data in the MC (parametric bootstrap).
Call the data sample $x = (x_1,...,x_n)$, observed data are $x_{obs}$, and the bootstrapped samples are $x_1^*, x_2^*, ...$

Idea is to use the distribution of

$$\hat{\theta}(x^*) - \hat{\theta}(x_{obs})$$

as an approximation for the distribution of

$$\hat{\theta}(x) - \theta$$

In the first quantity everything is known from the observed data plus bootstrapped samples, so we can use its distribution to estimate bias, variance, etc. of the estimator $\hat{\theta}$. 