How to Deal with Systematic Uncertainties

Rainer Wanke

Institut für Physik
Johannes Gutenberg-Universität Mainz

19 Feb 2016

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

PRISMA Cluster of Excellence

PHYSICS AT THE TERA SCALE Helmholtz Alliance

Usual situation at the end of a thesis (or any analysis):

- Finally data are taken, the selection is optimized, the Monte Carlo is produced, and the fit is implemented and working!

$$\Longrightarrow \quad \boxed{\textbf{Result} \; = \; \mathbf{x} \; \pm \; \sigma_{\text{stat}}}$$

Usual situation at the end of a thesis (or any analysis):

- Finally data are taken, the selection is optimized, the Monte Carlo is produced, and the fit is implemented and working!

$$\Longrightarrow \quad \boxed{\textbf{Result} \; = \; \mathbf{x} \; \pm \; \sigma_{\text{stat}}}$$

- The thesis/paper is of course overdue, but still the systematics have to be evaluated...

  And worse: No clear concept how to do it!

Usual situation at the end of a thesis (or any analysis):

- Finally data are taken, the selection is optimized, the Monte Carlo is produced, and the fit is implemented and working!

$$\Longrightarrow \boxed{\textbf{Result} \;=\; \mathbf{x} \;\pm\; \sigma_{\text{stat}}}$$

- The thesis/paper is of course overdue, but still the systematics have to be evaluated...

  And worse: No clear concept how to do it!

- Common solution in this situation:

  "Let's vary a few cuts, that's quickly done, and see if and how the result changes. We then call the variations of the result the ***systematic uncertainty!***"

Worst Method!

What is needed for estimating systematic uncertainties:

- **Some experience**
- Time
- Care
- Common sense
- Self confidence (but not too much!)

What is needed for estimating systematic uncertainties:

- **Some experience**
- Time
- Care
- Common sense
- Self confidence (but not too much!)

Unfortunately everything except the first item is out of the scope of this lecture...

What is needed for estimating systematic uncertainties:

- **Some experience**
- Time
- Care
- Common sense
- Self confidence (but not too much!)

Unfortunately everything except the first item is out of the scope of this lecture...

Disclaimer:

Unlike for statistical uncertainties, there are often no clear procedures and recipes for the determination of systematics.

You have to think yourself!

Outline

- **How to detect systematic uncertainties**
- **How to avoid systematic uncertainties**
- **How to estimate systematic uncertainties**
- **How to work with systematic uncertainties**

What are Systematic Uncertainties?

Definition used here:

> **Systematic uncertainties** $=$
>
> Measurement errors, which are not due to statistics in real or simulated data samples.

Side remark: Usually the term ***"uncertainty"*** is preferred.
Your analysis hopefully does not contain ***"errors"!***

Examples for systematics in HEP analyses: (not complete!)

- Badly known **detector acceptance** or **trigger efficiency.**

Examples for systematics in HEP analyses: (not complete!)

- Badly known **detector acceptance** or **trigger efficiency.**
- Wrong **detector calibrations.**

Examples for systematics in HEP analyses: (not complete!)

- Badly known **detector acceptance** or **trigger efficiency.**
- Wrong **detector calibrations.**
- Badly known **detector resolutions.**

Examples for systematics in HEP analyses: (not complete!)

- Badly known **detector acceptance** or **trigger efficiency.**
- Wrong **detector calibrations.**
- Badly known **detector resolutions.**
- **Time variations** of the experimental conditions.

Examples for systematics in HEP analyses: (not complete!)

- Badly known **detector acceptance** or **trigger efficiency.**
- Wrong **detector calibrations.**
- Badly known **detector resolutions.**
- **Time variations** of the experimental conditions.
- Badly known **background.**

Examples for systematics in HEP analyses: (not complete!)

- Badly known **detector acceptance** or **trigger efficiency.**
- Wrong **detector calibrations.**
- Badly known **detector resolutions.**
- **Time variations** of the experimental conditions.
- Badly known **background.**
- Uncertainties in the **simulation/theoretical model.**

Examples for systematics in HEP analyses: (not complete!)

- Badly known **detector acceptance** or **trigger efficiency.**
- Wrong **detector calibrations.**
- Badly known **detector resolutions.**
- **Time variations** of the experimental conditions.
- Badly known **background.**
- Uncertainties in the **simulation/theoretical model.**
- Uncertainties on **input parameters** (BR's, lifetimes, luminosity, . . .).

Examples for systematics in HEP analyses: (not complete!)

- Badly known **detector acceptance** or **trigger efficiency.**
- Wrong **detector calibrations.**
- Badly known **detector resolutions.**
- **Time variations** of the experimental conditions.
- Badly known **background.**
- Uncertainties in the **simulation/theoretical model.**
- Uncertainties on **input parameters** (BR's, lifetimes, luminosity, . . .).

- **Computational errors / program bugs / fit routines.**

Examples for systematics in HEP analyses: (not complete!)

- Badly known **detector acceptance** or **trigger efficiency.**
- Wrong **detector calibrations.**
- Badly known **detector resolutions.**
- **Time variations** of the experimental conditions.
- Badly known **background.**
- Uncertainties in the **simulation/theoretical model.**
- Uncertainties on **input parameters** (BR's, lifetimes, luminosity, . . .).

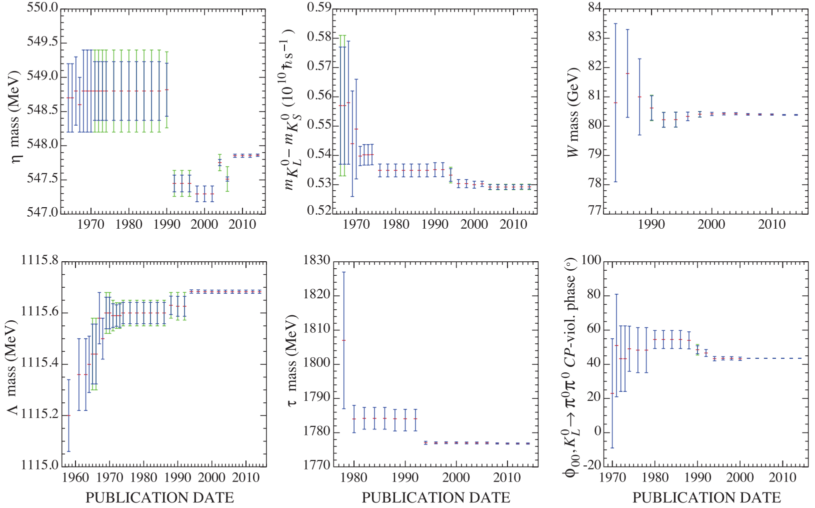
- **Computational errors / program bugs / fit routines.**
- **Biased experimentalist** (wants to measure "expected" result).

Examples for systematics in HEP analyses: (not complete!)

- Badly known **detector acceptance** or **trigger efficiency.**
- Wrong **detector calibrations.**
- Badly known **detector resolutions.**
- **Time variations** of the experimental conditions.
- Badly known **background.**
- Uncertainties in the **simulation/theoretical model.**
- Uncertainties on **input parameters** (BR's, lifetimes, luminosity, . . .).

- **Computational errors / program bugs / fit routines.**
- **Biased experimentalist** (wants to measure "expected" result).
- All other usually **unknown effects** on the measurement.

Variation of some constants with time (PDG 2016):



1. How to find systematic uncertainties

First: Think about ***any*** possible effect!

- Is every single **input number/parameter** well-known and understood? (efficiencies, calibrations, theory, external parameters, PDG, . . .)

1. How to find systematic uncertainties

First: Think about ***any*** possible effect!

- Is every single **input number/parameter** well-known and understood? (efficiencies, calibrations, theory, external parameters, PDG, . . .)
- Every possible **detector effect** considered? (geometrical acceptance, trigger efficiency, resolutions, detector inefficiencies, calibrations [energy scales!], . . .)

1. How to find systematic uncertainties

First: Think about ***any*** possible effect!

- Is every single **input number/parameter** well-known and understood? (efficiencies, calibrations, theory, external parameters, PDG, . . .)
- Every possible **detector effect** considered? (geometrical acceptance, trigger efficiency, resolutions, detector inefficiencies, calibrations [energy scales!], . . .)
- **Backgrounds** complete and well understood?

1. How to find systematic uncertainties

First: Think about ***any*** possible effect!

- Is every single **input number/parameter** well-known and understood? (efficiencies, calibrations, theory, external parameters, PDG, . . .)
- Every possible **detector effect** considered? (geometrical acceptance, trigger efficiency, resolutions, detector inefficiencies, calibrations [energy scales!], . . .)
- **Backgrounds** complete and well understood?
- **Fit routine** working correctly?

1. How to find systematic uncertainties

First: Think about ***any*** possible effect!

- Is every single **input number/parameter** well-known and understood? (efficiencies, calibrations, theory, external parameters, PDG, . . .)
- Every possible **detector effect** considered? (geometrical acceptance, trigger efficiency, resolutions, detector inefficiencies, calibrations [energy scales!], . . .)
- **Backgrounds** complete and well understood?
- **Fit routine** working correctly?
- **Theoretical inputs** correct?

1. How to find systematic uncertainties

First: Think about ***any*** possible effect!

- Is every single **input number/parameter** well-known and understood? (efficiencies, calibrations, theory, external parameters, PDG, . . .)
- Every possible **detector effect** considered? (geometrical acceptance, trigger efficiency, resolutions, detector inefficiencies, calibrations [energy scales!], . . .)
- **Backgrounds** complete and well understood?
- **Fit routine** working correctly?
- **Theoretical inputs** correct?

In case of any doubt:

Think of the cause of a possible effect!

$\Longrightarrow$ **Look at corresponding distributions**

Examples:

1. Background estimation correct?

DELPHI search for SUSY particles (Eur.Phys.J., C31, 421 (2004))

E.g.: stop-quark $\tilde{t}$ search in e^+e^- annihilation:

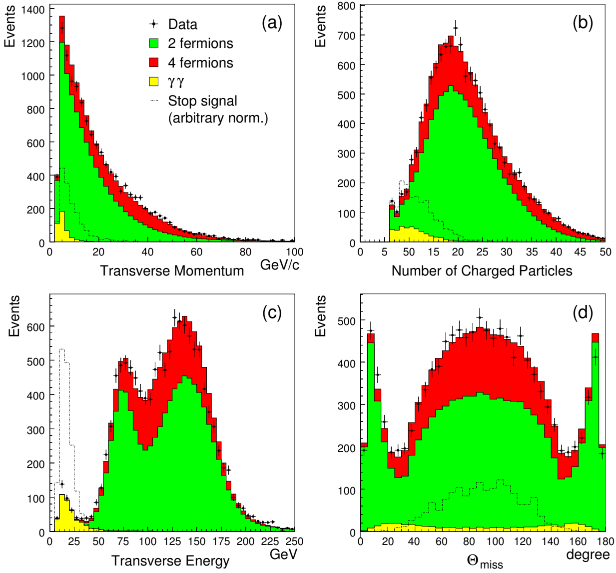
- $\tilde{t}$ should be pair-produced and to decay like $\tilde{t} \to c\tilde{\chi}^0_1$.
- Signature: Missing energy and two acoplanar jets.
- Main analysis problem (as usual for searches): Background suppression and estimation.
- Main backgrounds: SM 2-jet, 4-jet, two-photon events ($e^+e^- \to e^+e^-\gamma\gamma$).

How to make sure backgrounds are understood?

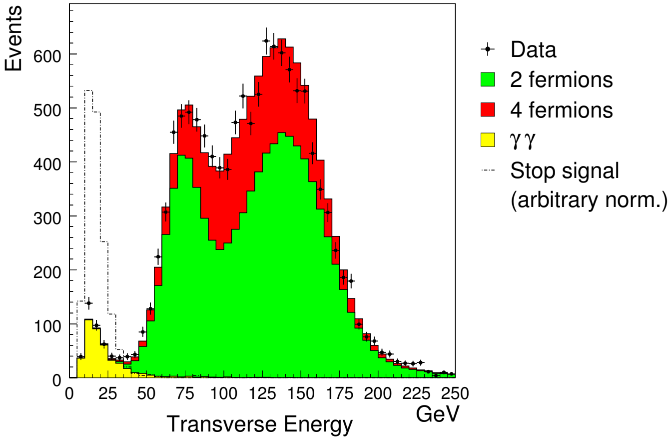
$\Longrightarrow$ Look for regions with large background contributions — are they well-described?

To enlarge backgrounds: Release corresponding cuts!

DELPHI stop analysis with loosened (preselection) cuts:



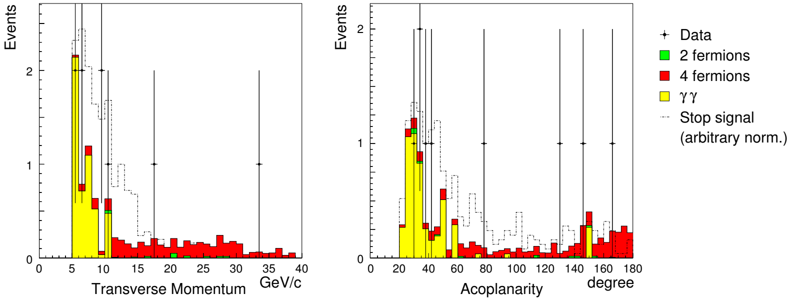
Look at one variable only (transverse energy):



⟹ Some discrepancy between data and $\gamma\gamma$ simulation ($\sim$ 15%).
(Unfortunately in the expected signal region.)

DELPHI stop analysis with final cuts:

After background evaluation, cuts are tightened:

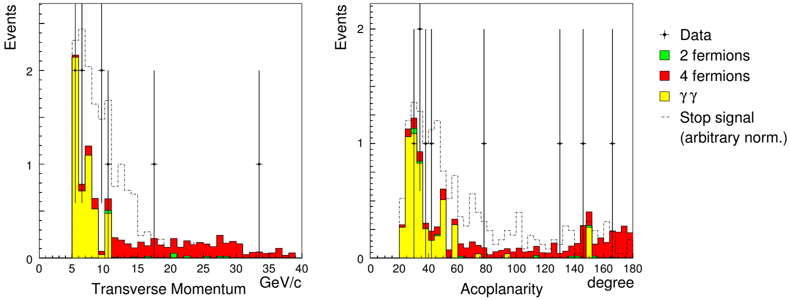


Systematic error estimation:

Vary background by the 15 % obtained above $\Longrightarrow$ limit changes.

DELPHI stop analysis with final cuts:

After background evaluation, cuts are tightened:



Systematic error estimation:

Vary background by the 15 % obtained above $\Longrightarrow$ limit changes.

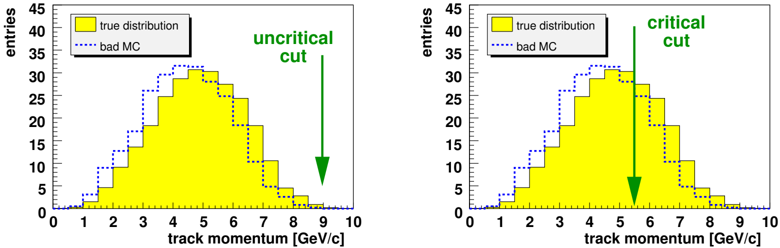
Caveat: Be careful when extrapolating from very many to very few events! Tails may not be well described and events in the tails may exhibit different behaviour than “normal” events.

2. Possible detector effects

(i.e.: poor MC description of e.g. inefficiencies)

$\Longrightarrow$ **Look at as many as possible data-MC comparisons, especially critical parameters!**

"critical parameter": Selection cuts hard into acceptance.



Acceptance:

true: $\epsilon_{\text{true}} = 99.8\,\%$

bad MC: $\epsilon_{\text{wrong}} = 99.9\,\%$

Acceptance:

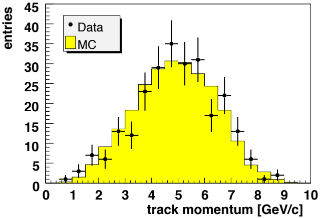
true: $\epsilon_{\text{true}} = 64.5\,\%$

bad MC: $\epsilon_{\text{wrong}} = 75.3\,\%$

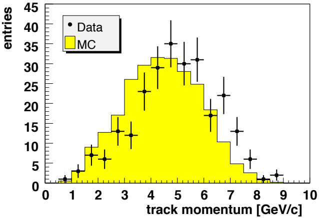
Data-MC comparison:

1. **Normalize MC sample to data and plot both.**

Good Monte Carlo:



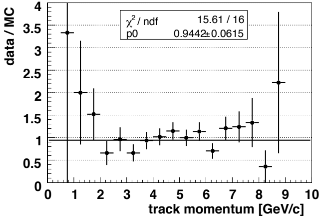
Bad Monte Carlo:



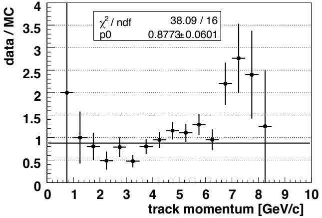
$\Longrightarrow$ Most problems can already be seen by eye.

2. **Divide data by *normalized* MC sample.**

Good Monte Carlo:



Bad Monte Carlo:



- Deviations usually immediately obvious by eye.
- Significance of disagreement:

  Fit a constant line $\Longrightarrow \chi^2/n_{dof}$
- However: still look at the original distributions!

  (deviations may e.g. fall into an unimportant region)

3. Plot result as function of analysis parameters

- **Possible time variations**
  - Usually the data sample consists of several sub-samples of more or less similar size.
  - Each sub sample has similar experimental conditions (detector status, trigger conditions, magnet polarities, collider performance, . . .), which may differ from sub-sample to sub-sample.
  - Normally some shut-down between the sub-samples.
  - Typical examples: LHC years; Tevatron Run I, IIa, IIb; . . .

Important check: Determine separate results for each sub-sample! Do they agree?

If not: 1. Why not?
2. Possibly discard sub-sample from analysis!

- **Other crucial parameters**

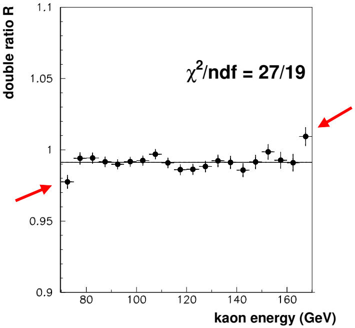
Example: Measurement of direct CP violation by NA48.

Measured parameter: Double ratio

$$R = \frac{\Gamma(K_L^0 \to \pi^0\pi^0)}{\Gamma(K_S^0 \to \pi^0\pi^0)} \Big/ \frac{\Gamma(K_L^0 \to \pi^+\pi^-)}{\Gamma(K_S^0 \to \pi^+\pi^-)} \approx 1 - 6 \times \mathrm{Re}\left(\frac{\epsilon'}{\epsilon}\right) \sim 0.99$$

Analysis is done in separate bins of kaon momentum (to be independent of the kaon spectra).

Disagreement at the borders of the momentum region! No obvious reason for such a behaviour found.



- **First question:**

  Is it really that bad?

  Rough estimation:

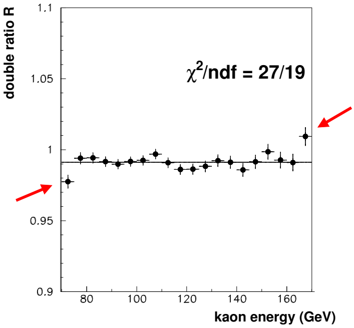
  $\sigma_{\chi^2 \text{ distribution}} = \sqrt{2\, n_{\text{dof}}} \approx 6.2$

  $\Longrightarrow$ about a $1.5\,\sigma$ effect

  Better:

  $\text{Probability}(27, 19) = 10.5\,\%$

  $\Longrightarrow$ once in ten times



- **First question:**

  Is it really that bad?

  Rough estimation:

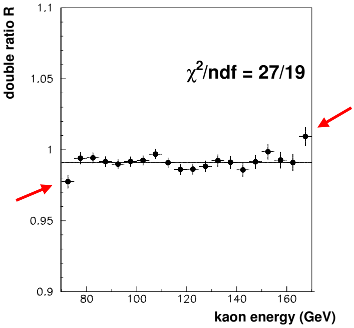
  $$\sigma_{\chi^2 \text{ distribution}} = \sqrt{2\, n_{\text{dof}}} \approx 6.2$$

  $\Longrightarrow$ about a $1.5\,\sigma$ effect

  Better:

  $$\text{Probability}(27, 19) = 10.5\,\%$$

  $\Longrightarrow$ once in ten times



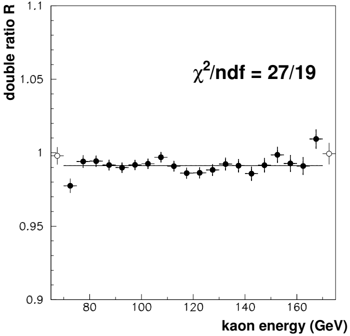
- **Second question:**

  How can we check?

  $\Longrightarrow$ Enlarge region!

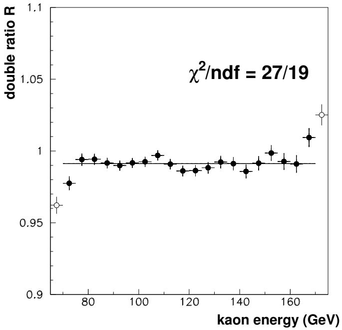
  Additional bins are ok!

  Decision: **No systematic!**



Hypothetical question:

What, if it would have looked like that?

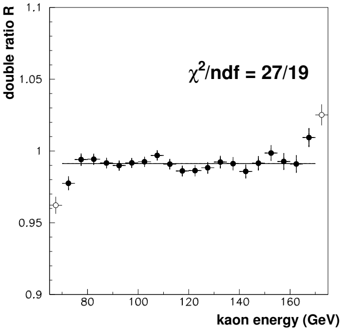


Hypothetical question:

What, if it would have looked like that?

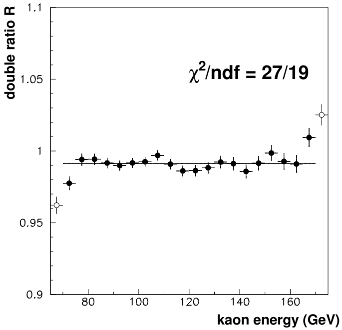
At first:

Need to understand the effect!



Hypothetical question:

What, if it would have looked like that?



At first:

Need to understand the effect!

Then:

Decide between two options:

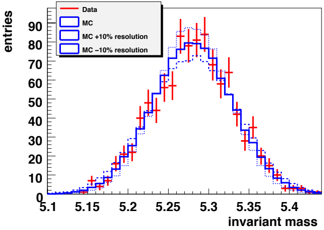
— Discard outer momentum bins.
(But base this decison on independent data/MC samples!)

— Determine a systematic error.
(Would probably not that large in this example.)

4. Detector resolutions

In general: Simulations are pretty bad in reproducing resolutions.

Example: Mass resolution



- MC not necessarily has same mass resolution as data.
- Simple way of changing resolution in the simulation:

  Modify reconstructed invariant mass m_{reco} to

  $$\mathbf{m}_{\text{reco}} + \mathbf{k} \times (\mathbf{m}_{\text{reco}} - \mathbf{m}_{\text{true}}),$$

  with: m_{true} = true invariant mass (e.g. PDG value).

- Parameter **k**: Relative change in resolution.

  Its reasonable range has to be found from "general considerations". ($\Longrightarrow$ "educated guess", see later)

- Similarly for other resolutions (e.g. shower widths)

5. Fit routines

Analyses usually involve more or less complex fit routines.

$\Longrightarrow$ Make sure they do, what they should!

- **Trivial test:**

  Run a full MC sample (as large as possible) with known input parameters instead of data.

  $\Longrightarrow$ Result as expected?

5. Fit routines

Analyses usually involve more or less complex fit routines.

$\Longrightarrow$ Make sure they do, what they should!

- **Trivial test:**

  Run a full MC sample (as large as possible) with known input parameters instead of data.

  $\Longrightarrow$ Result as expected?

- **Check for biases:**

  Repeat the above MC analysis (or possibly toy-MC) $N \sim 20 - 100$ times with MC sample size $\approx$ data sample size.

  $\Longrightarrow$ Do results follow a (gaussian) distribution around the input value with variance as expected?

5. Fit routines

Analyses usually involve more or less complex fit routines.

$\Longrightarrow$ Make sure they do, what they should!

- **Trivial test:**

  Run a full MC sample (as large as possible) with known input parameters instead of data.

  $\Longrightarrow$ Result as expected?

- **Check for biases:**

  Repeat the above MC analysis (or possibly toy-MC) $N \sim 20 - 100$ times with MC sample size $\approx$ data sample size.

  $\Longrightarrow$ Do results follow a (gaussian) distribution around the input value with variance as expected?

- Repeat analysis with **different binning.**

  (but distinguish statistical from systematic fluctuations!)

5. Fit routines

Analyses usually involve more or less complex fit routines.

$\Longrightarrow$ Make sure they do, what they should!

- **Trivial test:**

  Run a full MC sample (as large as possible) with known input parameters instead of data.

  $\Longrightarrow$ Result as expected?

- **Check for biases:**

  Repeat the above MC analysis (or possibly toy-MC) $N \sim 20 - 100$ times with MC sample size $\approx$ data sample size.

  $\Longrightarrow$ Do results follow a (gaussian) distribution around the input value with variance as expected?

- Repeat analysis with **different binning.**

  (but distinguish statistical from systematic fluctuations!)

- **Plot** $\log \mathcal{L}$ **distributions.**

  Do they look as expected? (correlations, other minima, ...?)

2. How to avoid systematic uncertainties

Usually most systematics cannot easily be avoided when doing data analysis (e.g. trigger inefficiencies, calibrations, backgrounds).

However: Some systematics may still be avoided or suppressed during planning/performing an analysis.

2. How to avoid systematic uncertainties

Usually most systematics cannot easily be avoided when doing data analysis (e.g. trigger inefficiencies, calibrations, backgrounds).

However: Some systematics may still be avoided or suppressed during planning/performing an analysis.

Examples:

- **Choice of selection cuts**
  - Every cut reduces the acceptance and may introduce errors, if acceptance is not well-known.
  - Usually uncritical: Geometrical cuts
    (exception: in inefficient detector regions)
  - More problematic: e.g. particle ID, jet energy, etc.

Therefore:

- Check every cut: is it really needed (in particular partially redundant cuts)? Often more but known background is better than unknown systematics.

- **Biased experimentalist**

  Everyone doing analysis usually has some prejudice on what will be the outcome (previous measurements, theory expectations, personal likes and dislikes, . . .).

  $\Longrightarrow$ Any deviation from the “expected” result is investigated, and possible causes are made up. No deviation, however, will not be investigated.

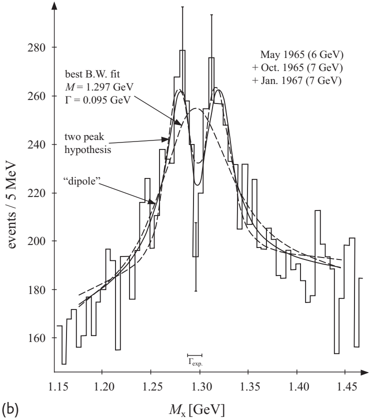
- **Biased experimentalist**

  Everyone doing analysis usually has some prejudice on what will be the outcome (previous measurements, theory expectations, personal likes and dislikes, . . .).

  $\Longrightarrow$ Any deviation from the "expected" result is investigated, and possible causes are made up. No deviation, however, will not be investigated.

Historical example: A_2 mass splitting

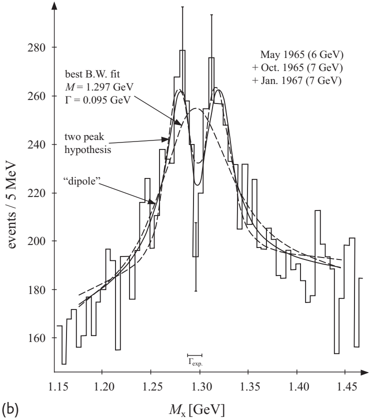
Observation of a narrow dip of six standard deviations in the center of the peak of the A_2 resonance in the reaction $\pi^- + p \to p + X$ by the CERN Missing-Mass-Spectrometer group (1967).



(b)

Results of fits to the data:

- Fit of a single Breit-Wigner: χ^2 probability $p = 0.1\,\%$ (3.1σ)
- Two independent Breit-Wigner's: $p = 15\,\%$
- Two Breit-Wigner's with same masses, widths, and destructive phase: $p = 70\,\%$



(b)

Results of fits to the data:

- Fit of a single Breit-Wigner: χ^2 probability $p = 0.1\,\%$ (3.1σ)
- Two independent Breit-Wigner's: $p = 15\,\%$
- Two Breit-Wigner's with same masses, widths, and destructive phase: $p = 70\,\%$

However:

Last two options would not fit into the quark-model of mesons!

Further developments after the observation:

- Several other experiments observed a split of the A_2 peak in several different reactions, with smaller significance $(< 3\sigma)$.
- After many more data were taken, the split disappeared.

Further developments after the observation:

- Several other experiments observed a split of the A_2 peak in several different reactions, with smaller significance ($< 3\sigma$).
- After many more data were taken, the split disappeared.

Lessons learned:

- Experimenters tend to prefer a “wanted or expected result”.
  - Biased choices of selection cuts.
  - Biased removal of “bad data”.
  - Looking for systematics, which could explain a difference to the expected result.

$\Longrightarrow$ Biased experimentalist!

Further developments after the observation:

- Several other experiments observed a split of the A_2 peak in several different reactions, with smaller significance ($< 3\sigma$).
- After many more data were taken, the split disappeared.

Lessons learned:

- Experimenters tend to prefer a "wanted or expected result".
  - Biased choices of selection cuts.
  - Biased removal of "bad data".
  - Looking for systematics, which could explain a difference to the expected result.

  $\Longrightarrow$ Biased experimentalist!

- Results are only published, if they support the expected result (*bandwagon effect* [Lindenbaum 1985]). Measurements are not published, if they don't see the expected effect.

  $\Longrightarrow$ Biased publication procedure!

Other historical examples:

- 17 keV/c^2 neutrino from tritium β-decay spectrum in 1985

  Only one experiment sees it, long discussion about systematic effects, refuted only in the early 90's.

- Pentaquark observations in the 00's

  Several experiments find 3-5 σ signals in several channels, which are attributed to "penta-quarks". Final refutal only about 2006.

Other historical examples:

- 17 keV/c^2 neutrino from tritium β-decay spectrum in 1985

  Only one experiment sees it, long discussion about systematic effects, refuted only in the early 90's.

- Pentaquark observations in the 00's

  Several experiments find 3-5 σ signals in several channels, which are attributed to "penta-quarks". Final refutal only about 2006.

How to avoid biases:

- Free yourself from expectations for the result!

  (But large deviations are still worrysome!)

Other historical examples:

- 17 keV/c^2 neutrino from tritium β-decay spectrum in 1985

  Only one experiment sees it, long discussion about systematic effects, refuted only in the early 90's.

- Pentaquark observations in the 00's

  Several experiments find 3-5 σ signals in several channels, which are attributed to "penta-quarks". Final refutal only about 2006.

How to avoid biases:

- Free yourself from expectations for the result!

  (But large deviations are still worrysome!)

- ***Never*** tune the selection by looking at the signal region!

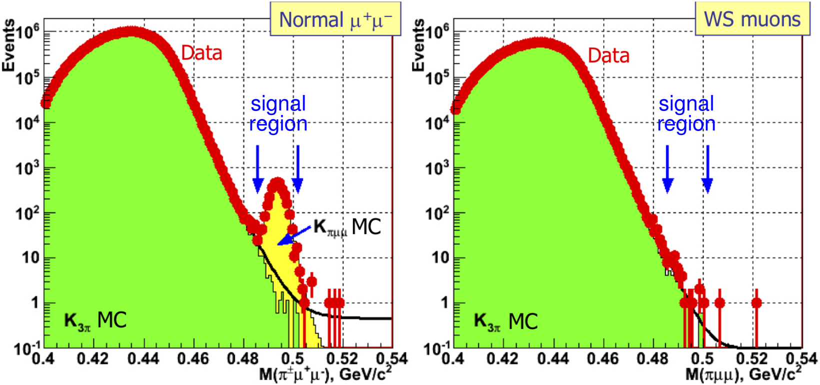
  No *"Look how my signal increases, when I change this cut!"*

  Instead: Tune the cuts ***solely*** on Monte Carlo or on a completely different data sample!

Another example:

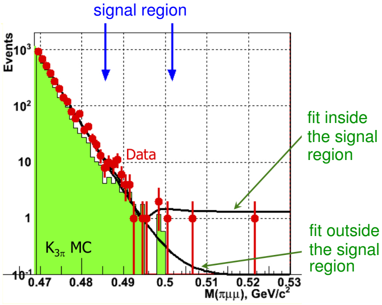
Search for the LFV decay $K^+ \to \pi^-\mu^+\mu^+$ in 2010.

- Main measurement of $K^+ \to \pi^+\mu^+\mu^-$ yields $\mathcal{O}(10^4)$ decays. Main background: $K^+ \to 3\pi$
- Analysis can easily be extented for a limit on $K^+ \to \pi^-\mu^+\mu^+$.



Zoom into the signal region:

Arrows show $\pm 3\sigma$ mass interval.



First attempt:

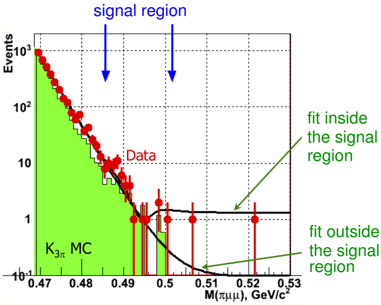
- Taking the signal region ($\pm 3\sigma$) results in 52 signal candidates.
- Background fit outside the signal region: 36.7 estimated bkg events.

Almost 3σ signal evidence!

- **But:** Background probably not well described.

Zoom into the signal region:

Arrows show $\pm 3\sigma$ mass interval.



Second attempt:

- Use Monte Carlo and fit it to the slope on the left side of the signal region:

  52.7 ± 22.0 estimated bkg events.

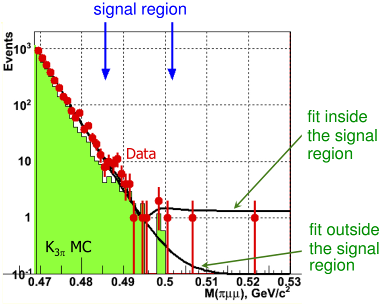
  (error from different methods)

  $N_{\text{sig}} < 17.3$ ev. (90 % CL).

- **But:** Signal region actually too large, one could also just fit the signal.

Zoom into the signal region:

Arrows show $\pm 3\sigma$ mass interval.



Third attempt:

- Fit the signal together with a "reasonable" background shape.

$N_{sig} = -8 \pm 6$ events

$N_{sig} < 8$ events (90 % CL).

What do we learn from this?:

- Results may vastly change, when changing ("optimizing") the analysis procedure.
- Even if the intentions are good: It is not possible to decide between different methods afterwards!
- You have to decide on your method ***before*** looking at the data!

⬇

At best a blind analysis is to be done!

- **Blind analysis**

  Signal region is covered (blinded) until the very end, when everything is fixed.

  Advantage: No bias possible by construction.

  Disadvantage: Systematics more difficult to discover
  (no simple cross-checks).

  Very useful with few statistics or in searches!

- **Blind analysis**

  Signal region is covered (blinded) until the very end, when everything is fixed.

  Advantage: No bias possible by construction.

  Disadvantage: Systematics more difficult to discover (no simple cross-checks).

  Very useful with few statistics or in searches!

  Also possible in other analyses:

  Example: $\sin 2\beta$ measurement from Belle and BaBar.

  Signal: Asymmetry in time-dependent $B^0\overline{B^0}$ mixing.

  Blinding: B flavor of tagging side randomized until end of analysis.

- **Blind analysis**

  Signal region is covered (blinded) until the very end, when everything is fixed.

  Advantage: No bias possible by construction.

  Disadvantage: Systematics more difficult to discover (no simple cross-checks).

  Very useful with few statistics or in searches!

  Also possible in other analyses:

  Example: $\sin 2\beta$ measurement from Belle and BaBar.

  Signal: Asymmetry in time-dependent $B^0\overline{B^0}$ mixing.

  Blinding: B flavor of tagging side randomized until end of analysis.

- **Talk to other people!**

  People not involved in the analysis usually have a much clearer view!

Conclusions on finding analysis problems

Finding problems is not a straight-forward task!

- **Try to think of every possible effect!**
- **Thoroughly check every possible input value!**
- **Look critically at every possible distribution!**
- **Talk to other people!**

3. How to estimate systematic uncertainties

- **Simplest case:**

  Input parameter x has well-known uncertainty σ_x

  Examples: BR's, lifetimes, luminosity, detector energy scale

  $\Longrightarrow$ vary x by $\pm\,\sigma_x$ $\Longrightarrow$ result varies by $\pm\,\sigma^{\text{syst}}_{\text{result}}$

  Independent systematics are added in quadrature.

3. How to estimate systematic uncertainties

- **Simplest case:**

  Input parameter x has well-known uncertainty σ_x

  Examples: BR's, lifetimes, luminosity, detector energy scale

  $\Longrightarrow$ vary x by $\pm\sigma_x$ $\Longrightarrow$ result varies by $\pm\sigma^{\text{syst}}_{\text{result}}$

  Independent systematics are added in quadrature.

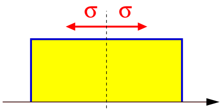
- **Sometimes:**

  Standard deviation σ_x is not known, but a tolerance (largest and smallest possible x)

  $\Longrightarrow$ assume uniform probability within $[x_{\text{low}}, x_{\text{high}}]$.

  $$\Longrightarrow \sigma_x = \frac{1}{\sqrt{12}}(x_{\text{high}} - x_{\text{low}}) \approx 0.29 \times (x_{\text{high}} - x_{\text{low}})$$

  (Gain of 60 % w.r.t. naive $\sigma_x = 0.5 \times (x_{\text{high}} - x_{\text{low}})$!)



- **Not as simple case:**
  - Some sort of problem was found by deeply looking into the data.

    The reason for the problem is more or less known.
    (It should better be!)
  - However: Found no way to remove it.

    Have to attach a systematic uncertainty.
  - Unfortunately: usually no clear and proper way to determine it.

- **Not as simple case:**
  - Some sort of problem was found by deeply looking into the data.

    The reason for the problem is more or less known.
    (It should better be!)
  - However: Found no way to remove it.

    Have to attach a systematic uncertainty.
  - Unfortunately: usually no clear and proper way to determine it.

⬇

Have to use a mixture of knowledge, reasoning, creativity, common sense, and — sometimes — believe.

“Educated guess”

“Educated guess”:

- **Example: Intensity of a radioactive source**

(see R.J. Barlow, “Statistics”)

Radioactive source ^{90}Sr, half life 28.5 a:

Labelled “10 μCi”, but purchase date unknown.

Question: What is the actual activity?

- Activity is measured in Bq since 5 years
  $\Longrightarrow$ source older than 5 years
- Laboratory was set up 10 years ago
  $\Longrightarrow$ source younger than 10 years

$\Longrightarrow$ **Age between 5 and 10 years.**

Original activity: 10 μCi = 370 000 Bq

- After 5 years: $\times 2^{-5/28.5} = 328\,000$ Bq
- After 10 years: $\times 2^{-10/28.5} = 290\,000$ Bq

$\sigma = 1/\sqrt{12}$ interval $\Longrightarrow$ **Activity: 309 000 ± 11 000 Bq**

▸ Example: Background estimation

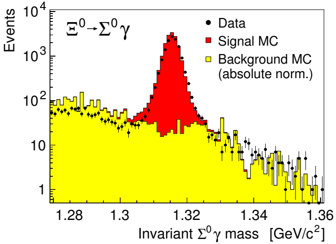
Decay $\Xi^0 \to \Sigma^0\gamma \to (\Lambda\gamma)\gamma$, **background** $\Xi^0 \to \Lambda\pi^0 \to \Lambda(\gamma\gamma)$

Background:

From MC simulation, normalized to total Ξ^0 flux.

$\Longrightarrow \approx 3\,\%$ under the signal.

But: Seems a factor of 2 too large in the left side-band.



▸ Example: Background estimation

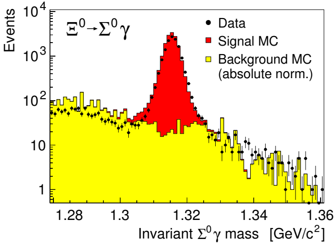
Decay $\Xi^0 \to \Sigma^0\gamma \to (\Lambda\gamma)\gamma$, **background** $\Xi^0 \to \Lambda\pi^0 \to \Lambda(\gamma\gamma)$

Background:

From MC simulation, normalized to total Ξ^0 flux.

$\Longrightarrow \approx 3\,\%$ under the signal.

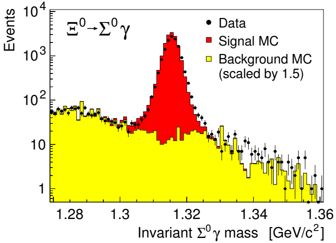
But: Seems a factor of 2 too large in the left side-band.



Alternative bkg normalization:

Scale to mass side-band.

$\Longrightarrow \approx 1.5\,\%$ background.



Good arguments for both methods.

- **Example: Background estimation**

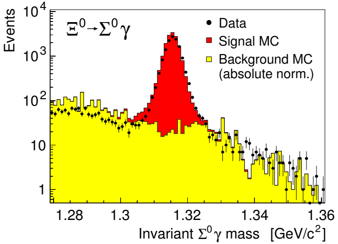
Decay $\Xi^0 \to \Sigma^0\gamma \to (\Lambda\gamma)\gamma$, **background** $\Xi^0 \to \Lambda\pi^0 \to \Lambda(\gamma\gamma)$

Background:

From MC simulation,
normalized to total Ξ^0 flux.

$\Longrightarrow \approx 3\,\%$ under the signal.

But: Seems a factor of 2 too large in the left side-band.

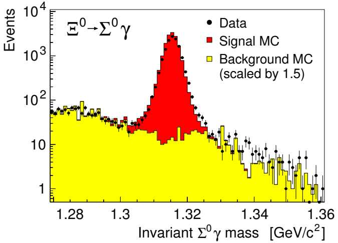


Alternative bkg normalization:

Scale to mass side-band.

$\Longrightarrow \approx 1.5\,\%$ background.

Good arguments for both methods.



Decision (conservative):

Background $= (\mathbf{1.5} \pm \mathbf{1.5})\%$

- **Example: Theory errors**

  Often more than one theoretical description available.

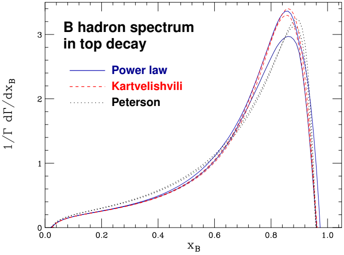
  Fragmentation function:

  - One function used as default, but one (or more) other available.

    ⬇

    Use discrepancy to other theory as $\pm\sigma_{\mathbf{syst}}$.

    (Not satisfying, but only possibility.)



  - More than one theory:
    Can take spread of different results as $\pm\sigma_{\mathbf{syst}}$.

- **Example: Theory errors**

Often more than one theoretical description available.

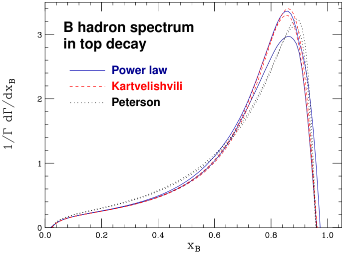
Fragmentation function:

- One function used as default, but one (or more) other available.

  ⬇

  Use discrepancy to other theory as $\pm\sigma_{\text{syst}}$.

  (Not satisfying, but only possibility.)



- More than one theory:
  Can take spread of different results as $\pm\sigma_{\text{syst}}$.

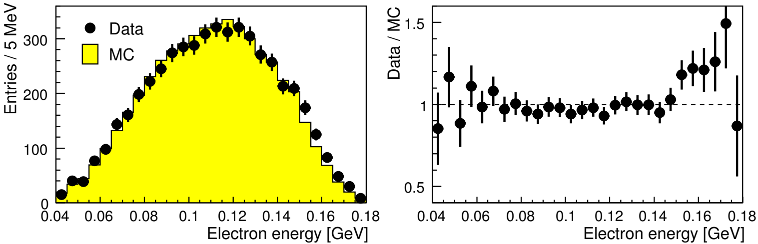
Radiative corrections:

- Two descriptions: 1. With radiative corr. (but not complete)
  2. No radiative corr. (surely wrong)
- Take a fraction of the difference as systematic.
  (which? → educated guess!)

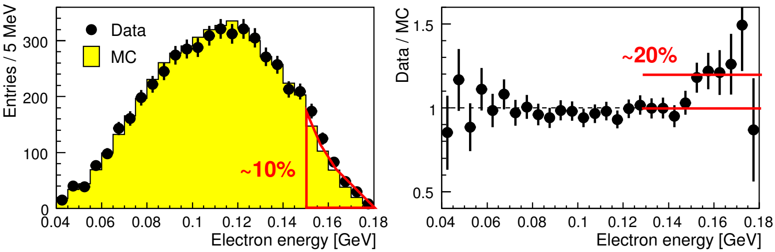
- **Example: Data-MC disagreement**

Rare decay $K^+ \to \pi^0 e^+ \nu\gamma$

Many data-MC comparisons, everything fine, except e^+ energy in kaon rest frame:



- $\chi^2/n_{dof} = 29.7/27 \implies$ quite ok!
  However: significant disagreement above 0.15 GeV.
  ("Normal" χ^2/n_{dof} due to extremely well agreement below 0.15 GeV!)
- No known source of the disagreement (additional background?).



How to estimate a systematic uncertainty?

- About 20 % more data than MC seen above 0.15 GeV.
- About 10 % of all data lie above 0.15 GeV.

$\Longrightarrow$ **Uncertainty of ± 2 % on decay rate.**

However:

- Would be largest single uncertainty in this analysis.
- Reason not understood – could also be too few data below 0.15 GeV!

$\Longrightarrow$ **More investigations needed!**

- **Analyses with little statistics**

  With too little statistics data-MC comparison not possible:

  - Statistical fluctuations could fake systematic problems.
  - Systematics may be hidden by statistical uncertainties.

  Possible solutions:

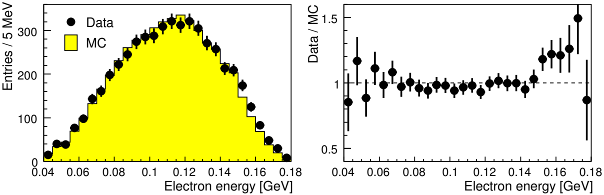
  - Enlarge data sample by releasing cuts (in a controlled way)

    $\Longrightarrow$ see e.g. DELPHI analysis on SUSY searches.

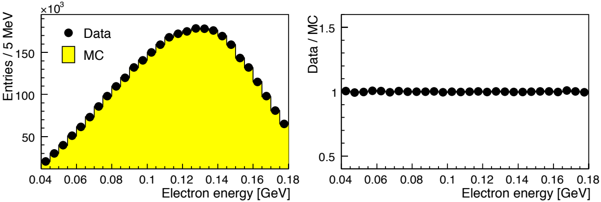
  - Use similar but more abundant control channel.

    $\Longrightarrow$ previous example:
    use $K^+ \to \pi^0 e^+ \nu$ instead of $K^+ \to \pi^0 e^+ \nu\gamma$.

$K^+ \to \pi^0 e^+ \nu\gamma$:



$K^+ \to \pi^0 e^+ \nu$:



$\Longrightarrow$ No discrepancy in control channel.

$\Longrightarrow$ No estimation of systematics possible this way.

Still learned something: Probably no detector problem!

Analogon in B physics: B_d^0 decays as control for B_s^0 decays.

- **Cut variations**

  Idea: Vary analysis cuts within a reasonable range and check, if the result stays stable.

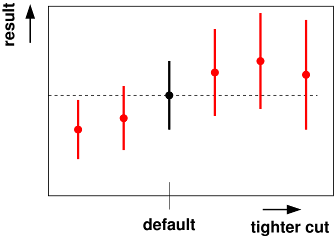
  - Very commonly used (mostly because of lack of better ideas).
  - However: **Not really recommended here.**

    Reason: The information from cut variations usually is very limited.

    In virtually all cases it is better to look at the underlying distributions for understanding of what is going on!

  - Still: in some cases it may be useful for systematics estimation.
    (If reason for variation is understood!)

    Problem:
    Errors are correlated.



Determination of uncorrelated errors:

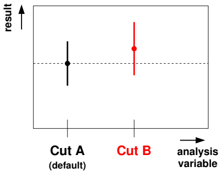
Default cut $\Longrightarrow$ sample A, result $x_A \pm \sigma_A$

Tighter cut $\Longrightarrow$ sample $B \subset A$, result $x_B \pm \sigma_B$

Question: Is x_B significant deviation from x_A?

Problem: σ_A, σ_B from partially same data

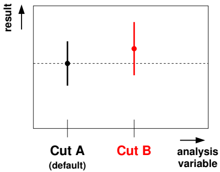
$\Longrightarrow$ **Correlated errors!**



Determination of uncorrelated errors:

Default cut $\Longrightarrow$ sample A, result $x_A \pm \sigma_A$

Tighter cut $\Longrightarrow$ sample $B \subset A$, result $x_B \pm \sigma_B$



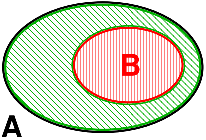
Question: Is x_B significant deviation from x_A?

Problem: σ_A, σ_B from partially same data

$\Longrightarrow$ **Correlated errors!**

Solution: Consider separate data samples

B and $C = (A \text{ without } B)$



Well-known formula for averaging results:

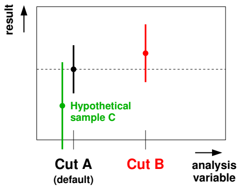
$$\bar{x} = x_A = \sum_i \frac{x_i}{\sigma_i^2} \Big/ \sum_i \frac{1}{\sigma_i^2} = \frac{x_B/\sigma_B^2 + x_C/\sigma_C^2}{1/\sigma_B^2 + 1/\sigma_C^2}$$

$$\sigma_{\bar{x}}^2 = \sigma_A^2 = 1 \Big/ \sum_i \frac{1}{\sigma_i^2} = \frac{1}{1/\sigma_B^2 + 1/\sigma_C^2}$$

Known: $x_A, x_B, \sigma_A, \sigma_B$

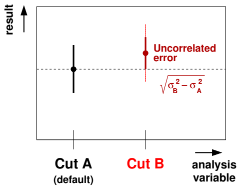
$\Longrightarrow$ determine x_C, σ_C!

$$\Longrightarrow \quad \frac{1}{\sigma_C^2} = \frac{1}{\sigma_A^2} - \frac{1}{\sigma_B^2}, \quad x_C = \frac{x_A/\sigma_A^2 - x_B/\sigma_B^2}{1/\sigma_A^2 - 1/\sigma_B^2}$$



Difference between uncorrelated x_B and x_C:

$$x_C - x_B = \cdots = \sigma_B^2 \frac{x_A - x_B}{\sigma_B^2 - \sigma_A^2}$$

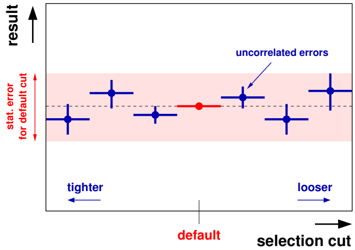


Statistical significance of the difference:

$$\frac{x_C - x_B}{\sqrt{\sigma_B^2 + \sigma_C^2}} = \cdots = \frac{x_A - x_B}{\sqrt{\sigma_B^2 - \sigma_A^2}}$$

$$\boxed{\sigma_{\text{uncorr}}^2 = |\sigma_B^2 - \sigma_A^2|}$$

Interpretation of variations:



Common result from cut variation:

- Shall we attach a systematic uncertainty?
- **First:**

  Is there any known reason for a possible variation?

  How do the original data & MC distributions agree?

  What happens with even tighter/looser cuts?

- Could attach variation ($\approx \frac{1}{2}\sigma_{\text{stat}}$) as systematic error.

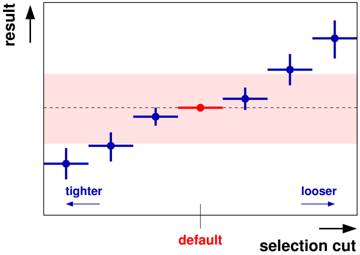
  $\Longrightarrow$ probably overcautious!

- Usually here: **No systematic uncertainty!**

Other possibility:

- **No way to attach a systematic uncertainty!**
- Variation has to be understood!

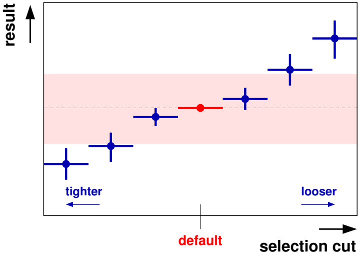
$\Longrightarrow$ *Look at underlying distributions!*



Other possibility:

- **No way to attach a systematic uncertainty!**
- Variation has to be understood!

$\Longrightarrow$ *Look at underlying distributions!*

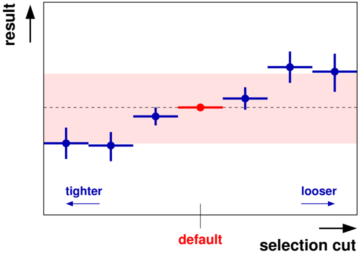


Similar case:

- Variation has to be understood!

$\Longrightarrow$ *Look at underlying distributions!*

$\Longrightarrow$ *What happens for looser/tighter cuts?*



- Only, if no other way: variation $=$ systematic uncertainty!

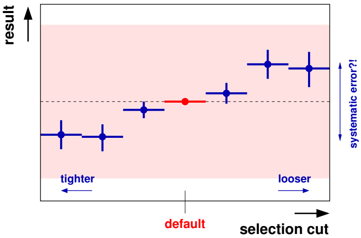
  Dangerous situation: systematic error $\geq \mathcal{O}$(statistical error)!

Similar, but not as worse:

- By far not as dangerous, since variation $\ll \sigma_{\text{stat}}$.
- Still: better be understood.

$\Longrightarrow$ Could attach variation as systematic.

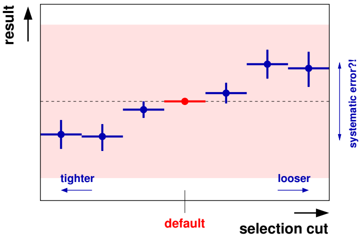
(if no other way found)



Similar, but not as worse:

- By far not as dangerous, since variation $\ll \sigma_{stat}$.
- Still: better be understood.

$\Longrightarrow$ Could attach variation as systematic.
(if no other way found)



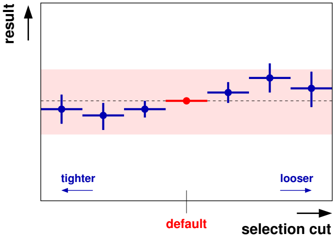
Yet another case:

- Could be fluctuation!

$\Longrightarrow$ *Look at underlying distributions!*

$\Longrightarrow$ *Any reason for a possible problem?*

- In case of doubt:
  systematics as above



No obvious reason for systematics: **Be bold!**

Special case:

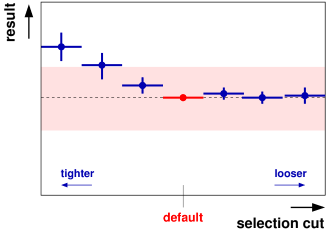
- Significant variation, but only to one side.
- Usual situation:

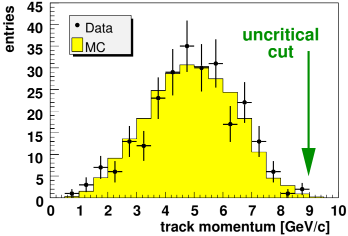
  MC description bad, but cut is almost fully efficient.

  $\Longrightarrow$ Disagreement does not matter.

No reason for a systematic uncertainty!

(As long as the bad MC does not have other impacts.)





4. How to work with systematic uncertainties

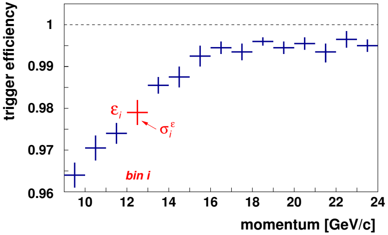
Example:

Trigger efficiency

Measured from data in several bins of e.g. momentum:

$$\epsilon_i \pm \sigma_i^\epsilon$$

Bins are independent from each other.



4. How to work with systematic uncertainties

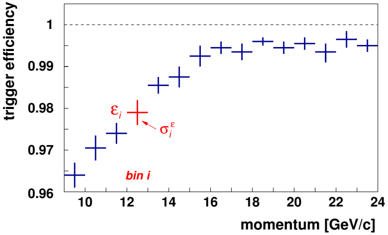
Example:

Trigger efficiency

Measured from data in several bins of e.g. momentum:

$$\epsilon_\mathbf{i} \pm \sigma^\epsilon_\mathbf{i}$$

Bins are independent from each other.



Option 1: Vary each bin i by $\pm\sigma_i^\epsilon \implies$ result varies by $\pm\sigma_i^{\text{result}}$

$$\implies \left(\sigma^{\text{result}}_{\text{total}}\right)^\mathbf{2} = \sum_\mathbf{i} \left(\sigma^{\text{result}}_\mathbf{i}\right)^\mathbf{2}$$

- Might be tedious for many bins.
- Problem, if σ_i^{result} asymmetric ($\sigma_i^+ \neq \sigma_i^-$).
- σ_i^ϵ have to be independent from each other (usually the case for trigger efficiencies).

Option 2: Toy Monte Carlo

- Generate simultaneously random trigger efficiencies for each bin i according to gaussians with mean ϵ_i and width σ_i^ϵ.

  $\Longrightarrow$ perform analysis with these modified values.

  $\Longrightarrow$ obtain new analysis result.

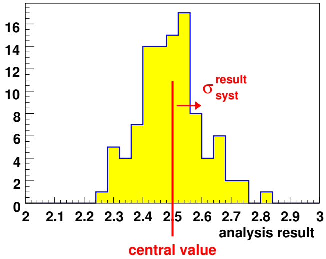
- Do this $N \sim 100$ times

  $\Longrightarrow$ distribution of different results around central value

  Take the variance

  (or fit a gaussian)

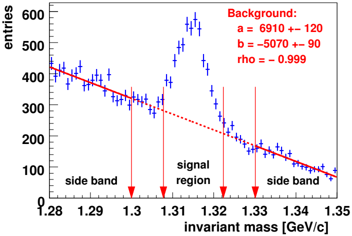
  $\Longrightarrow \quad \pm\, \sigma_{syst}^{result}$



- **Advantages:**
  - Total error in one go (but have to redo analysis quite often, too).
  - Non-gaussian effects automatically taken care of and immediately visible.
  - Correlations between variables can easily be included.

Correlations between external parameters

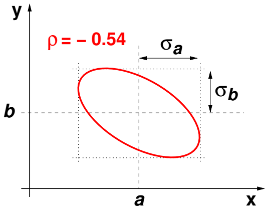
Example: Background shape



Signal over some combinatorial background, which is not known from MC simulation.

Common procedure:
Fit background in side-bands with 1. (or 2.) order polynomial.

$\Longrightarrow$ constant $= a \pm \sigma_a$
slope $= b \pm \sigma_b$
correlation coefficient $= \rho$



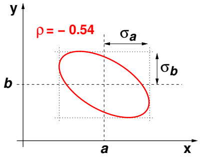
For systematics generate random a, b according to binormal distribution:

$$P(x,y) = \frac{1}{2\pi\sigma_a\sigma_b\sqrt{1-\rho^2}} \times \exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-a}{\sigma_a}\right)^2 + \left(\frac{y-b}{\sigma_b}\right)^2 - 2\rho\,\frac{x-a}{\sigma_a}\,\frac{y-b}{\sigma_b}\right]\right)$$

Adding correlated/uncorrelated systematics

- Consider a two-parameter fit, with results $a \pm \sigma_{\text{stat}}$, $b \pm \sigma_{\text{stat}}$, and correlation coeff. ρ.

  $\Longrightarrow$ confidence contours in the x, y plane.



- Systematics have been determined for a and b — but how to properly take them into account?

- **Figure out the correlations for each contribution to the systematics!**
  - Very often: systematic errors are fully or not at all correlated.
  - Often: Same correlation as statistical errors. (e.g. external parameters, which affect both variables the same way)
  - Sometimes: Correlation is not known.
    In this case: same correlation as for statistical errors is usually a good assumption. (But document it!)

- **Add covariance matrices** of single contributions:

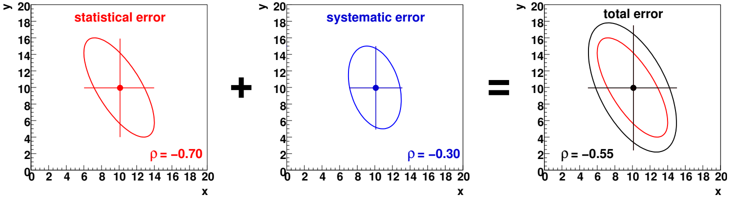
$$\mathbf{V}_{\text{syst}} = \begin{pmatrix} \sigma^2_{a,\text{uncorr}} & 0 \\ 0 & \sigma^2_{b,\text{uncorr}} \end{pmatrix}$$ **uncorrelated**

$$+ \begin{pmatrix} \sigma^2_{a,\text{corr}} & \sigma_{a,\text{corr}}\,\sigma_{b,\text{corr}} \\ \sigma_{a,\text{corr}}\,\sigma_{b,\text{corr}} & \sigma^2_{b,\text{corr}} \end{pmatrix}$$ **correlated**

$$+ \begin{pmatrix} \sigma^2_{a,\text{i}} & \rho_\text{i}\,\sigma_{a,\text{i}}\,\sigma_{b,\text{i}} \\ \rho_\text{i}\,\sigma_{a,\text{i}}\,\sigma_{b,\text{i}} & \sigma^2_{b,\text{i}} \end{pmatrix} + \cdots$$ **partially correlated**

- Finally **add statistics and systematics**

$$\mathbf{V}_{\text{stat}} + \mathbf{V}_{\text{syst}} = \mathbf{V}_{\text{total}}$$



Some final words

- **To be avoided** (if possible):

  One single dominant contribution to the systematics (in particular, if it is $\geq$ statistical error)

  - One single mistake/mis-estimation and the result is worthless.
  - Also: systematics usually have no gaussian behaviour

    (No problem when adding several small contributions

    $\Longrightarrow$ Central Limit Theorem ensures gaussian distribution)
  - Exception: largest systematic is of statistical nature (e.g. trigger efficiency).

Some final words

- **To be avoided** (if possible):

  One single dominant contribution to the systematics (in particular, if it is $\geq$ statistical error)

  - One single mistake/mis-estimation and the result is worthless.
  - Also: systematics usually have no gaussian behaviour

    (No problem when adding several small contributions

    $\Longrightarrow$ Central Limit Theorem ensures gaussian distribution)
  - Exception: largest systematic is of statistical nature (e.g. trigger efficiency).

- **Be cautious, but don't be too shy!**

  If there's no reason for quoting a conservative systematic or for quoting it at all, don't do it!

Conclusions

- **Be aware of *any* possible systematic**
  - Look at as many distributions as possible.
  - Try to look from “outside” on your analysis.
- **Avoid biases**
  - Free yourself from expections on the result.
  - Never look at the data when tuning cuts!
- **Do your best when estimating systematics**
  - Few cases: systematics are straight-forward.
  - Mostly: have to use some sort of “educated guess”