

Part III

- Bias
- Confidence limits
- Pearson's chi-squared
- Nuisance parameters
- More details of curve-fitting

Bias

- It's only in the large-sample limit that maximum likelihood is unbiased.
- An elementary example of bias is what happens if you use a data set to first estimate the mean $\bar{x}$ and then estimate the variance σ^2 of that data set.
- It can be shown (let's use the well-known result and derive it later) that the minimum-variance estimate of the variance $\widehat{\sigma^2} = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$, with $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$ tends to underestimate σ^2 .
- The average bias can be shown to be corrected by replacing $N \rightarrow N - 1$: $\widehat{\sigma^2} = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2$
- which stops being an issue when $N \gg 10$

Bias

- Another way to think about this is that we have «used up» 1 degree of freedom to calculate $\bar{x}$
- This is analogous to calculating the χ^2 after a fit of n data points to m parameters (e.g. 2 for a line with slope and intercept).
- The $\chi^2 = \sum_{i=1}^n \left(\frac{y_i - f(x_i, \vec{\mu})}{\sigma_i} \right)^2$ should be compared to the χ^2 -distribution for $n - m$ degrees of freedom since we have «used up» m dof to identify the best-fit curve.

BONUS: Derivation of bias in S^2

- $S^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$, with $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$
- $\frac{1}{N} \sum_i (x_i - \bar{x})^2 = \frac{1}{N} \sum_i x_i^2 - \bar{x}^2$
- Use $x_i = \mu + \sigma \delta_i$ (fluctuations about model parameters, also for $\bar{x}$)
- $S^2 = \frac{1}{N} \sum_i (\mu + \sigma \delta_i)^2 - \frac{1}{N^2} [\sum_i (\mu + \sigma \delta_i)] [\sum_j (\mu + \sigma \delta_j)]$
- $= \frac{1}{N} \sum_i (\mu^2 + \sigma^2 \delta_i^2 + 2\mu\sigma \delta_i) - \frac{1}{N^2} [N\mu + \sigma \sum_i \delta_i] [N\mu + \sigma \sum_j \delta_j]$

BONUS: Derivation of bias in S^2

- So far $S^2 = \frac{1}{N} \sum_i (\mu^2 + \sigma^2 \delta_i^2 + 2\mu\sigma\delta_i) - \frac{1}{N^2} [N^2\mu^2 + 2N\mu\sigma \sum_i \delta_i + \sigma^2 \sum_i \delta_i^2 + \sigma^2 \sum_i \sum_{j \neq i} \delta_i \delta_j]$
- Average over datasets:
$$\langle S^2 \rangle = \frac{1}{N} (N\mu^2 + N\sigma^2 + 0) - \frac{1}{N^2} [N^2\mu^2 + 0 + N\sigma^2 + 0] = \sigma^2 \left(1 - \frac{1}{N}\right) = \sigma^2 (N-1)/N$$
- So $\langle S^2 \rangle < \sigma^2$
- To correct for the bias, we should use $\widehat{\sigma^2} = S^2 \frac{N}{N-1} = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2$
- which we can use further in the error on the mean $\sigma_m = \sigma/\sqrt{N}$

BONUS: Deriving errors

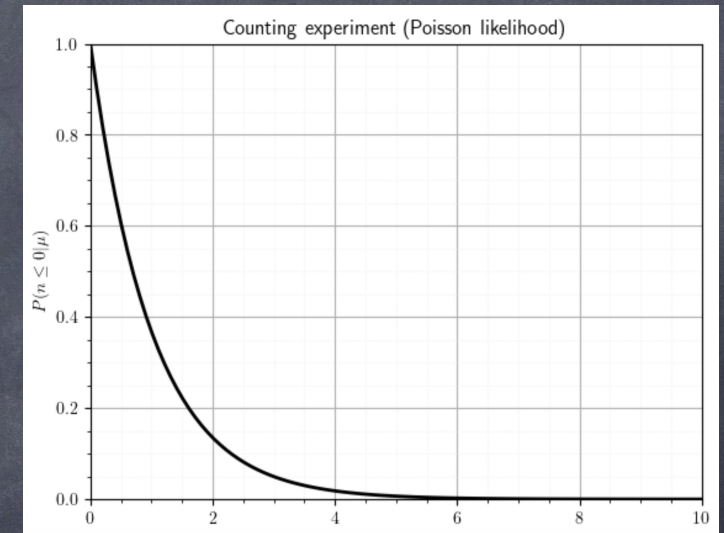
- This “trick” of expressing the observables in terms of the model parameters and the expected statistical fluctuations is extremely powerful.
- You can derive expectation values and variances even in situations with tricky correlations.
- A nice example is estimating the uncertainties on counts in a histogram bin if there are, e.g., 2 entries per event and both entries are expected to have the same pdf.
- The naïve uncertainty is $\sqrt{p_i 2N}$, where p_i is probability per entry to end up in histogram bin i , and N is the number of events.
- Due to correlations between the two entries (both tend to end up in the bins with the largest values of p_i), the uncertainty turns out to be $\sqrt{p_i 2N(1 + p_i)}$.
- The limiting case of a single bin ($p_i = 1$) is an uncertainty of $2\sqrt{N}$, which makes sense since you are in fact measuring N and counting each event twice.

Confidence limits

- In HEP and NP there are many examples of experiments with null results, e.g., Higgs boson searches before 2012 and neutrinoless double- β decays
- Suppose one searches for an exotic nucleus or a new particle, say a charged Higgs boson of a particular mass, and observes *zero* events. How high can the predicted counting rate be before it becomes very unlikely to not observe candidates in the data?

Confidence limits

- The probability of observing zero events from a Poisson process $P(n|\mu) = \frac{e^{-\mu}\mu^n}{n!}$ is $P(n=0|\mu) = e^{-\mu} \frac{\mu^0}{0!} = e^{-\mu}$
- To set a 95% confidence level upper limit on the counting rate we solve for $e^{-\mu} = 1 - 0.95 = 0.05$, i.e. we solve for the rate that gives a probability of seeing zero events of 5%.
- In other words, we most likely (95%) would have observed at least 1 exotic nucleus or $H^\pm$ if it were a true hypothesis.
- Another way to think about it, is that if nature does happen to be *at* the limit on μ , we are guaranteeing that on average 1 out of 20 experiments will falsely exclude it (i.e. observe zero $H^\pm$).

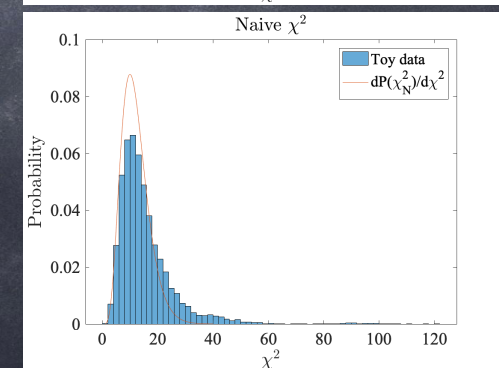
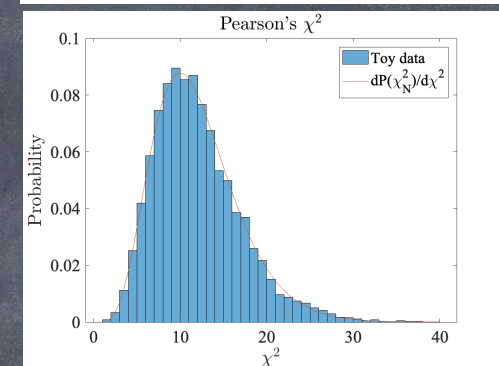
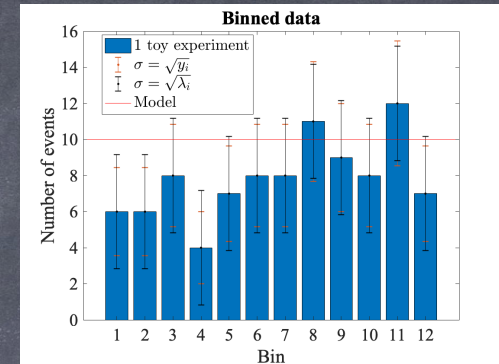


Confidence limits

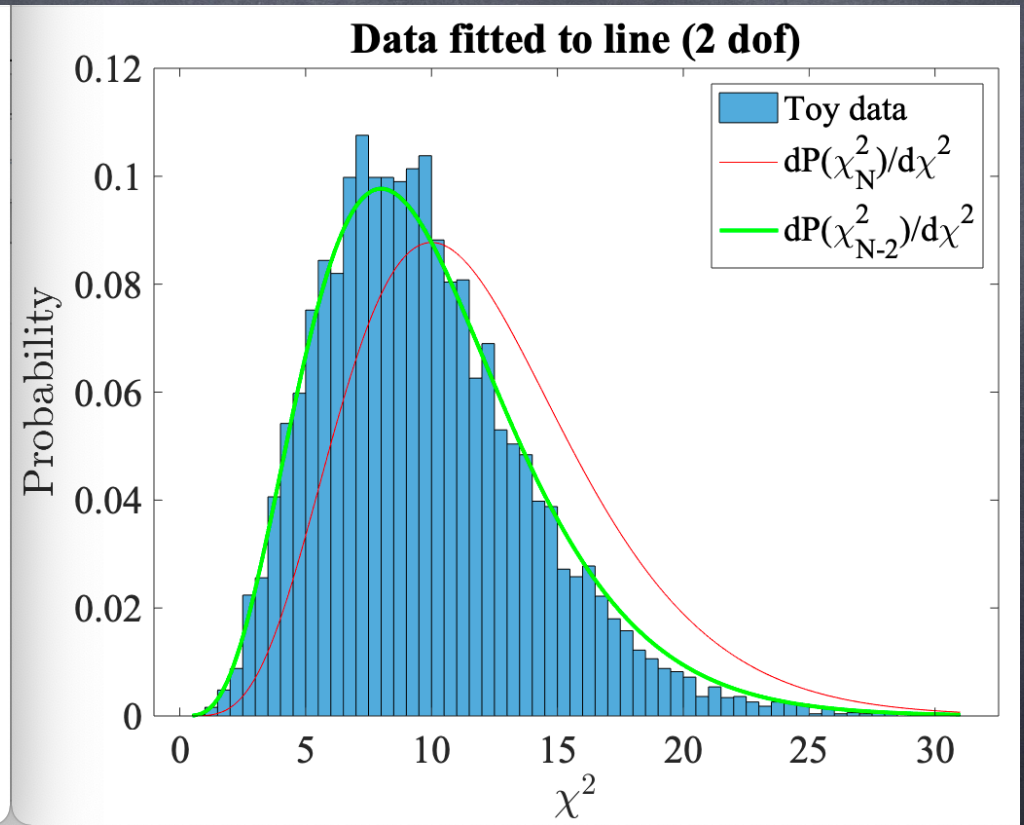
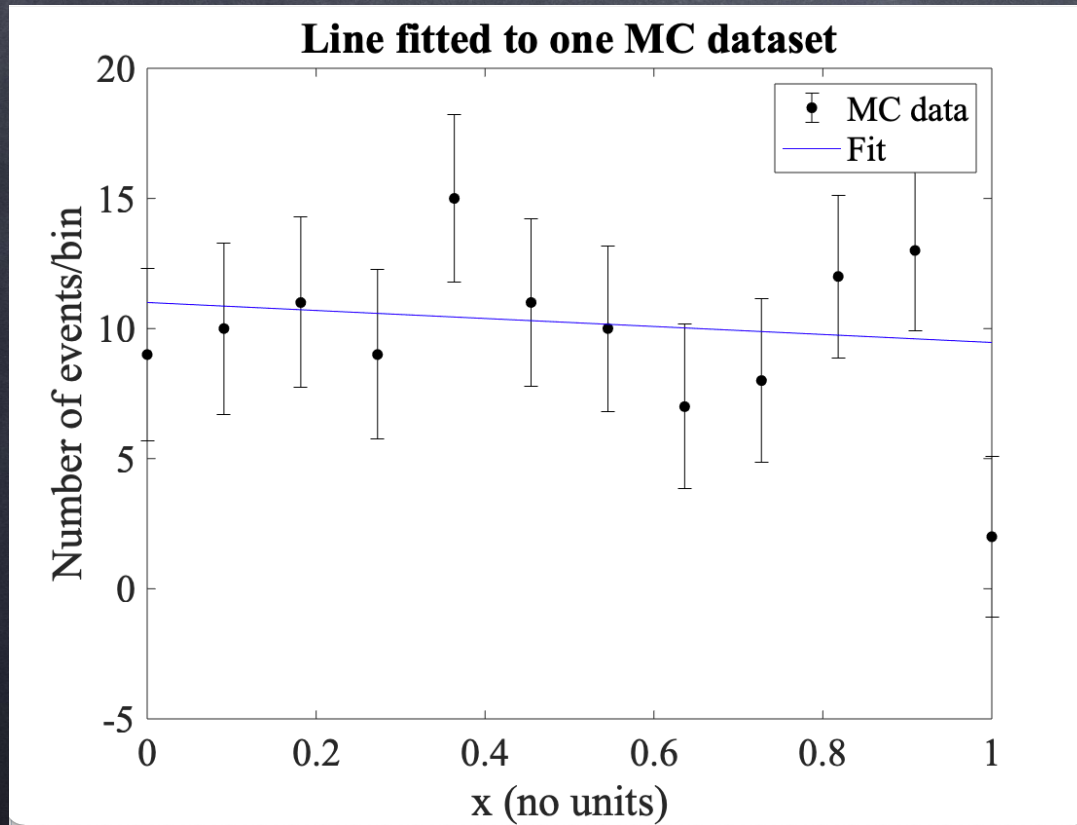
- It turns out that $\mu \approx 3$ for $e^{-\mu} = 5\%$.
- In HEP μ often consists of a product of integrated luminosity, detection efficiency, and a predicted cross-section, so we can interpret the upper limit on μ as an **upper limit on the cross-section**.
- If the predicted rate is higher than 3, then the probability of observing zero is less than 5%.
 - Question: What does $\sigma \geq 2.5$ fb excluded at the 95% CL mean? The cross section is most likely small (smaller than 2.5 fb). If the cross section turns out to be 2.5 fb, then we have been unlucky and falsely excluded it – this should happen in only $1-0.95=5\%$ of experiments.
- Similar considerations can be used to set **lower limits** on nuclear or particle (e.g. proton) lifetimes if no decays are observed during a certain exposure time of the experiment (see Leo).
- Life gets very interesting when the estimated background is of the same order as the hypothetical signal being searched for and *more than zero events are observed!* (see, e.g., the pdg review of statistics or «wikipedia CLs method»)

Binned data – Pearson's χ^2

- Very often we fit curves to binned data (histograms), where the error naively should be the uncertainty on the number of counts in that bin.
- It can be shown that Pearson's $\chi^2 = \sum_{i=1}^N \frac{(y_i - f(x_i, \mu))^2}{f(x_i, \mu)}$ gives a χ^2 distribution with $N - m$ degrees of freedom, where μ has dimension m ,
- whereas $\chi^2 = \sum_{i=1}^N \frac{(y_i - f(x_i, \mu))^2}{y_i}$ does not.
- Note that the number of counts in every bin should be $y_i > \sim 5$ for this to be a good approximation.
- This example is without a fit, so N_{bins} degrees of freedom
- Question: What would we have to change if we fitted a line (with slope and intercept) ? We would have to subtract two degrees of freedom when evaluating the chi-squared probability.

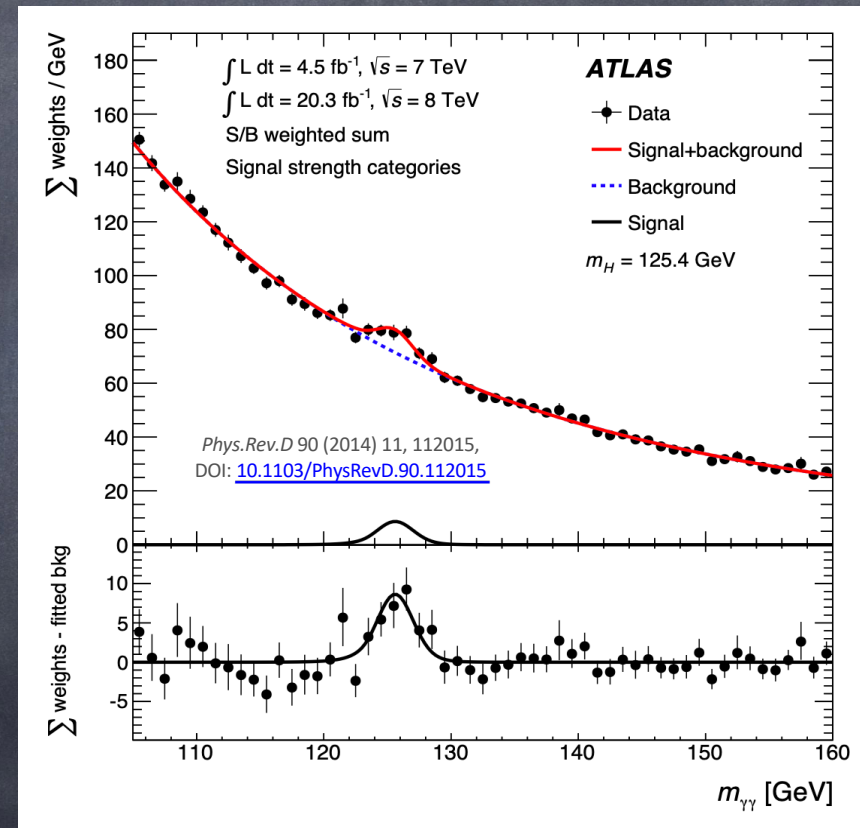


Pearson's χ^2 with fitted line (2 dof)



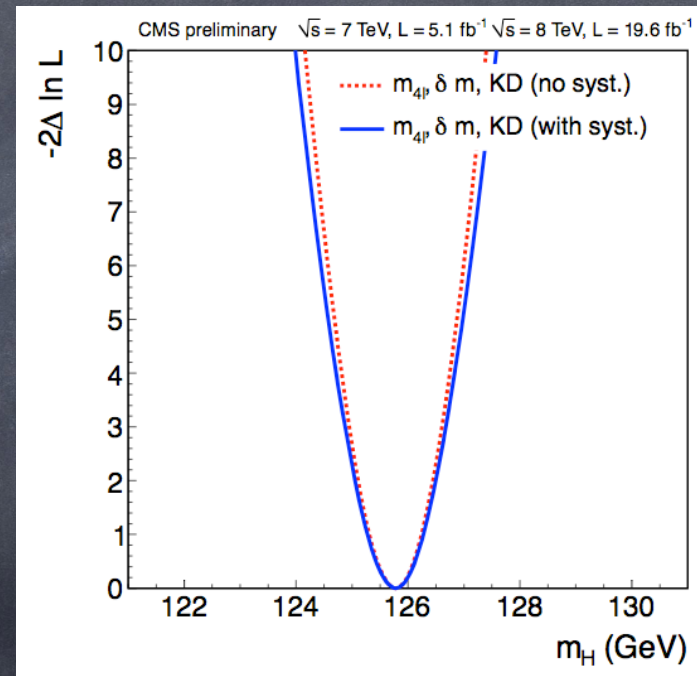
Nuisance parameters (NP)

- Parameters of a model that need to be fitted but that are not of primary interest: NP
- Example: $H \rightarrow \gamma\gamma$
 - Amount of signal (Gaussian-like peak) is of primary interest – POI
 - Amount and distribution of background (exponential or polynomial: 2–3 parameters) is of necessary but secondary interest – NP



Nuisance parameters

- NP's often associated to systematic uncertainties (constraint term in likelihood)
- NP's broaden the log-likelihood profile
- Multiplicative NP's introduce non-linearity (need numerical solution)
- Additive NP's can sometimes be treated analytically



Nuisance parameters

- Simple additive example:

$$\chi^2 = \frac{(x - (\mu + \delta))^2}{\sigma^2} + \frac{\delta^2}{\sigma_s^2}$$

- μ is parameter of interest (POI)
- δ is nuisance parameter (NP) of systematic uncertainty σ_s
- $\frac{\delta^2}{\sigma_s^2}$ constrains the NP (here Gaussian pdf)

$$\begin{aligned} \frac{\partial \chi^2}{\partial \mu} = \frac{\partial \chi^2}{\partial \delta} = 0 &= 2 \frac{(x - (\mu + \delta))}{\sigma^2} \\ &= -2 \frac{(x - (\mu + \delta))}{\sigma^2} + \frac{2\delta}{\sigma_s^2} \end{aligned}$$

- $\Rightarrow \hat{\mu} = x, \hat{\delta} = 0$ (global minimum)

- Away from minimum $\delta = \delta(\mu)$:

$$\frac{x - \mu}{\sigma^2} = \left(\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2} \right) \delta$$

- Solve for $\delta(\mu)$ and replace in χ^2

- One page later... $\chi^2 = \frac{(x - \mu)^2}{\sigma^2 + \sigma_s^2}$
i.e. $\sigma^2 \rightarrow \sigma^2 + \sigma_s^2$

- In other words, the introduction of the additive systematic uncertainty on μ gives

$$\sigma^2 \rightarrow \sigma^2 + \sigma_s^2$$

Nuisance parameters (details)

$$\bullet \quad \delta = \frac{(x-\mu)}{\sigma^2} \frac{\sigma^2 \sigma_s^2}{\sigma^2 + \sigma_s^2}$$

$$\bullet \quad \chi^2 = \frac{[(x-\mu)-\delta]^2}{\sigma^2} + \frac{\delta^2}{\sigma_s^2}$$

$$\bullet \quad \chi^2 = \frac{(x-\mu)^2}{\sigma^2} + \frac{\delta^2}{\sigma^2} - \frac{2\delta(x-\mu)}{\sigma^2} + \frac{\delta^2}{\sigma_s^2}$$

$$\bullet \quad \chi^2 = \frac{(x-\mu)^2}{\sigma^2} + \delta^2 \frac{\sigma^2 + \sigma_s^2}{\sigma^2 \sigma_s^2} - 2 \frac{(x-\mu)}{\sigma^2} \frac{(x-\mu)}{\sigma^2} \frac{\sigma^2 \sigma_s^2}{\sigma^2 + \sigma_s^2}$$

$$\bullet \quad \chi^2 = \frac{(x-\mu)^2}{\sigma^2} + \frac{\sigma^2 + \sigma_s^2}{\sigma^2 \sigma_s^2} \frac{(x-\mu)^2}{\sigma^4} \frac{\sigma^4 \sigma_s^4}{(\sigma^2 + \sigma_s^2)^2} - 2 \frac{(x-\mu)^2}{\sigma^4} \frac{\sigma^2 \sigma_s^2}{\sigma^2 + \sigma_s^2}$$

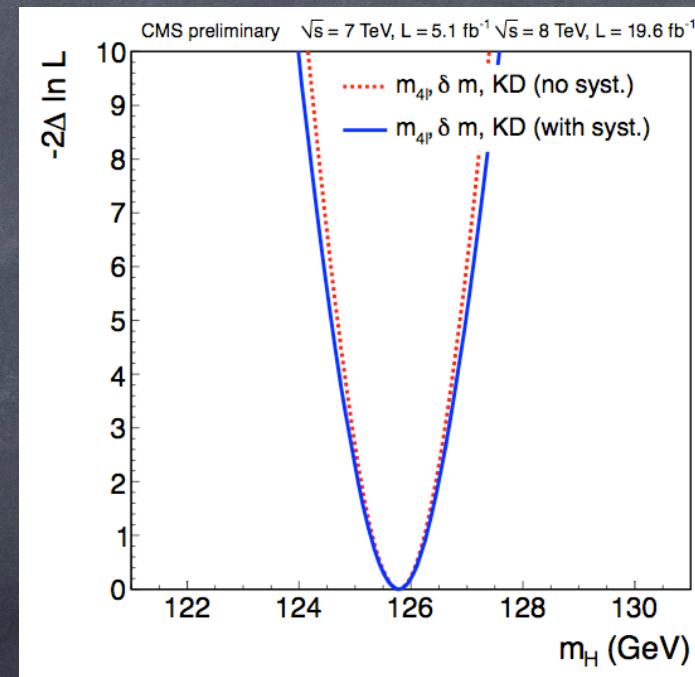
$$\bullet \quad \chi^2 = \frac{(x-\mu)^2}{\sigma^2} \left[1 + \frac{\sigma_s^2}{\sigma^2 + \sigma_s^2} - 2 \frac{\sigma_s^2}{\sigma^2 + \sigma_s^2} \right]$$

$$\bullet \quad \chi^2 = \frac{(x-\mu)^2}{\sigma^2} \left[\frac{\sigma^2 + \sigma_s^2 + \sigma_s^2 - 2\sigma_s^2}{\sigma^2 + \sigma_s^2} \right]$$

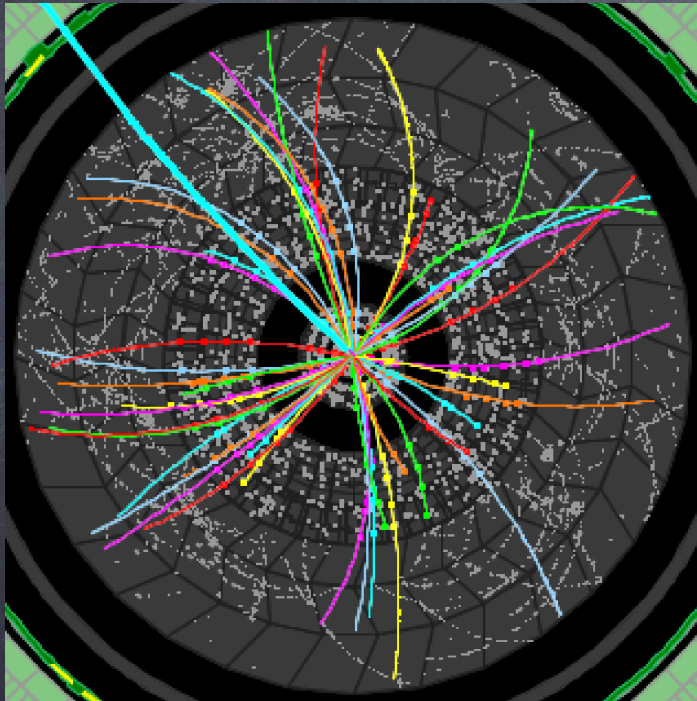
$$\bullet \quad \chi^2 = \frac{(x-\mu)^2}{\sigma^2 + \sigma_s^2}$$

Nuisance parameters

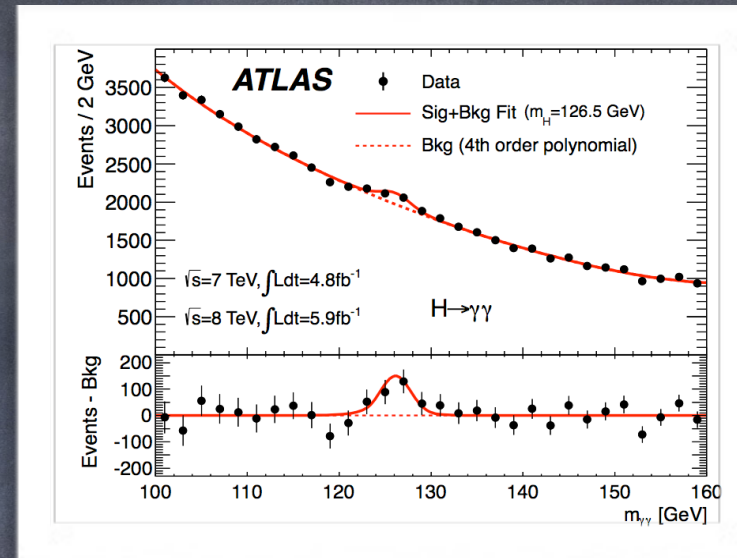
- In other words, the introduction of the additive systematic uncertainty on μ gives $\sigma^2 \rightarrow \sigma^2 + \sigma_s^2$
- Question: What could we do if there were several independent sources of additive uncertainty on μ ?
- Add them in quadrature!
- $\sigma^2 \rightarrow \sigma^2 + \sigma_1^2 + \sigma_2^2 + \dots$



Curve-fitting (examples)

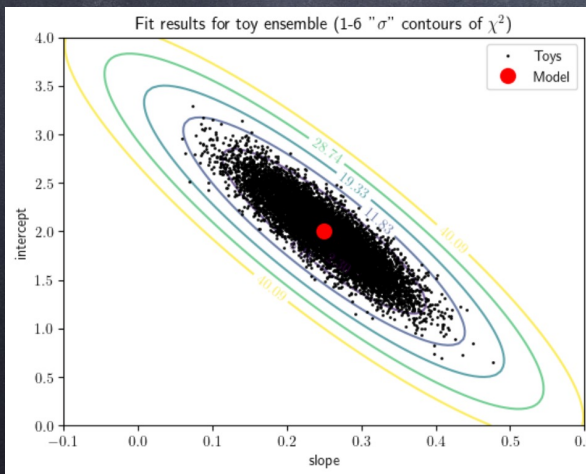
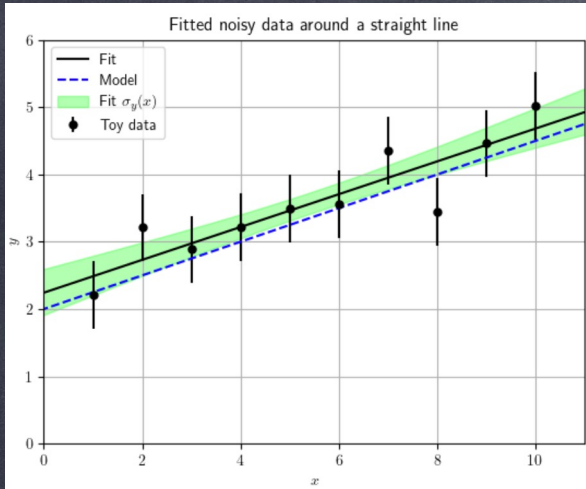


- Following particles through an experiment (charged particles in a magnetic field)



- Extracting information about a signal from a fit of data to a parametric statistical model

Curve fitting



- How is the green band calculated?
- How are the contours of constant χ^2 found?
- How do we **expect** the points (m, b) to be distributed? This determines the uncertainties we quote for a single observed dataset.
- This is directly related to the previous question and two additional questions:
 - What are the variances on m and b ?
 - What is the covariance between m and b ? It's clearly neither zero nor maximal covariance.

Curve fitting – the green band

- The green band answers the question about how much the fitted line could vary at any given x .

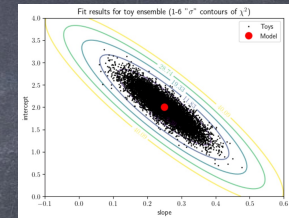
- The fit is our best estimate of where the line is, so:

$$\hat{y} = f(x) = \hat{m}x + \hat{b}$$

- We have to propagate uncertainty, but we can't neglect the covariance:

$$\sigma_y^2 = \left(\frac{\partial f}{\partial m}\right)^2 \sigma_m^2 + \left(\frac{\partial f}{\partial b}\right)^2 \sigma_b^2 + 2 \left(\frac{\partial f}{\partial m}\right) \left(\frac{\partial f}{\partial b}\right) \text{cov}(m, b)$$

$$\sigma_y^2 = x^2 \sigma_m^2 + \sigma_b^2 + 2x \cdot \text{cov}(m, b)$$



- The fitting-software will typically return information about the covariance matrix:

$$\Sigma = \begin{bmatrix} \sigma_m^2 & \rho \sigma_m \sigma_b \\ \rho \sigma_m \sigma_b & \sigma_b^2 \end{bmatrix}, \text{ where } \rho \equiv \frac{\text{cov}(m, b)}{\sigma_m \sigma_b}$$

113

Curve fitting – the green band

- The fitting-software will typically return information about the covariance matrix:

$$C = \begin{bmatrix} \sigma_m^2 & \rho\sigma_m\sigma_b \\ \rho\sigma_m\sigma_b & \sigma_b^2 \end{bmatrix}, \text{ where } \rho \equiv \frac{\text{cov}(m,b)}{\sigma_m\sigma_b}$$

- P.S. Some software may also return or display the correlation matrix, where all the elements of the covariance matrix are divided by the relevant $\sigma_i\sigma_j$. The diagonal terms of the correlation matrix are all 1 and the $\rho_{i,j}$ are all between ± 1 , e.g.

$$\rho(3 \times 3) = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{21} & 1 & \rho_{23} \\ \rho_{31} & \rho_{32} & 1 \end{bmatrix}$$

115

Question: What is the relationship between ρ_{ij} and ρ_{ji} ? $\rho_{ij} = \rho_{ji}$

Curve fitting by hand

- To gain understanding, let's do the χ^2 -minimization by hand...
- $\chi^2 = \sum_{i=1}^n \left(\frac{y_i - f(x_i, \mu)}{\sigma_i} \right)^2$ and minimize: $\frac{\partial \chi^2}{\partial \mu} = 0$, i.e. $\frac{\partial \chi^2}{\partial \mu_j} = 0$ for all j
- $\chi^2 = \sum_{i=1}^n \left(\frac{y_i - (mx_i + b)}{\sigma_i} \right)^2 = \sum_{i=1}^n \frac{1}{\sigma_i^2} (y_i^2 + m^2 x_i^2 + b^2 + 2mx_i b - 2mx_i y_i - 2y_i b)$
- $\chi^2 = \Lambda_{yy} + m^2 \Lambda_{xx} + b^2 \Lambda_1 + 2mb \Lambda_x - 2m \Lambda_{xy} - 2\Lambda_y b$, where $\Lambda_z \equiv \sum_{i=1}^n \frac{z_i^2}{\sigma_i^2}$
- $\frac{\partial \chi^2}{\partial m} = 2\hat{m} \Lambda_{xx} + 2\hat{b} \Lambda_x - 2\Lambda_{xy} = 0 = \frac{\partial \chi^2}{\partial b} = 2\hat{b} \Lambda_1 + 2\hat{m} \Lambda_x - 2\Lambda_y$
- 2 equations, 2 unknowns, solve for $[\hat{m}, \hat{b}]$ with linear matrix algebra

Curve fitting by hand

- Solve $[\hat{m}, \hat{b}] \begin{bmatrix} \Lambda_{xx} & \Lambda_x \\ \Lambda_x & \Lambda_1 \end{bmatrix} = [\Lambda_{xy}, \Lambda_y]$
- $[\hat{m}, \hat{b}] = [\Lambda_{xy}, \Lambda_y] \begin{bmatrix} \Lambda_{xx} & \Lambda_x \\ \Lambda_x & \Lambda_1 \end{bmatrix}^{-1}$
- We still need to find the uncertainties, including the covariance, and relate these to contours of constant χ^2 and the bivariate distribution of $\hat{m}$ and $\hat{b}$...

Aside: Multivariate gaussian (normal) distribution

- 1-D: $\mathcal{L} = \frac{1}{N} \frac{dN}{dx} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-x_0)^2}{2\sigma^2}}$
- 2-D: $\mathcal{L} = \frac{1}{N} \frac{d^2N}{dx dy} = \frac{1}{\sqrt{2\pi\sigma_x^2}} \frac{1}{\sqrt{2\pi\sigma_y^2}} \frac{1}{\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)} \left(\frac{(x-x_0)^2}{2\sigma_x^2} + \frac{(y-y_0)^2}{2\sigma_y^2} - 2\rho \frac{(x-x_0)(y-y_0)}{2\sigma_x\sigma_y} \right)}$
- n-D: $\mathcal{L} = \frac{1}{N} \frac{d^n N}{\prod_{i=1}^n dx_i} = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}$, where $\mathbf{x}$ and $\boldsymbol{\mu}$ are vectors of length n and Σ is the $n \times n$ covariance matrix $\Sigma_{i,j} = \langle (x_i - \mu_i)(x_j - \mu_j) \rangle$
- If we can identify Σ^{-1} or Σ for our curve fit, we have everything we need.
- $-2 \ln \mathcal{L} = (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) + \text{constants} = \chi^2 + \text{constants}$

Parameters of multi-dimensional parabola

- In 2-D a general parabola is
$$Z^2 = a(x - x_0)^2 + b(y - y_0)^2 + c(x - x_0)(y - y_0) + d(x - x_0) + e(y - y_0) + Z_0^2$$
- We are at the minimum $[x_0, y_0]$, where $\frac{\partial Z^2}{\partial x} = \frac{\partial Z^2}{\partial y} = 0$, so $d = e = 0$
- We also see that both a and b must be positive.
- $\frac{\partial^2 Z^2}{\partial x^2} = 2a, \frac{\partial^2 Z^2}{\partial y^2} = 2b, \frac{\partial^2 Z^2}{\partial x \partial y} = c$
- In other words, we can find the parameters of the 2-D parabola from the second derivatives! It can be shown that this applies to n -D for $n \geq 1$.

Parameters of multi-dimensional parabola

- Let's take the second derivatives of $\chi^2 = \Lambda_{yy} + m^2\Lambda_{xx} + b^2\Lambda_1 + 2mb\Lambda_x - 2m\Lambda_{xy}$
- $\frac{\partial^2 \chi^2}{\partial m^2} = 2\Lambda_{xx}, \frac{\partial^2 \chi^2}{\partial b^2} = 2\Lambda_1, \frac{\partial^2 \chi^2}{\partial m \partial b} = 2\Lambda_x$
- So, we can write $\chi^2 = [m - \hat{m}, b - \hat{b}] \begin{bmatrix} \Lambda_{xx} & \Lambda_x \\ \Lambda_x & \Lambda_1 \end{bmatrix} \begin{bmatrix} m - \hat{m} \\ b - \hat{b} \end{bmatrix}$
- And identify $\begin{bmatrix} \Lambda_{xx} & \Lambda_x \\ \Lambda_x & \Lambda_1 \end{bmatrix} \equiv \Sigma^{-1}$

Parameters of multi-dimensional parabola

P.S. For this problem
 Σ doesn't depend on y !!

- We take the inverse of both sides:

$$\Sigma = \begin{bmatrix} \Lambda_{xx} & \Lambda_x \\ \Lambda_x & \Lambda_1 \end{bmatrix}^{-1} = \frac{1}{\Lambda_{xx}\Lambda_1 - \Lambda_x^2} \begin{bmatrix} \Lambda_1 & -\Lambda_x \\ -\Lambda_x & \Lambda_{xx} \end{bmatrix} \equiv \begin{bmatrix} \sigma_m^2 & \rho\sigma_m\sigma_b \\ \rho\sigma_m\sigma_b & \sigma_b^2 \end{bmatrix}$$

- We have everything we need! Now we can draw contours of constant χ^2 in the (m, b) plane!

Alternate calculation

- A more direct, but complicated, approach is to start from
$$\chi^2 = \Lambda_{yy} + m^2 \Lambda_{xx} + b^2 \Lambda_1 + 2mb\Lambda_x - 2m\Lambda_{xy}$$
- Substitute with the equations we got from the minimization:
$$2\hat{m}\Lambda_{xx} + 2\hat{b}\Lambda_x - 2\Lambda_{xy} = 0 = 2\hat{b}\Lambda_1 + 2\hat{m}\Lambda_x - 2\Lambda_y$$

-or-

$$\Lambda_{xy} = \hat{m}\Lambda_{xx} + \hat{b}\Lambda_x, \Lambda_y = \hat{b}\Lambda_1 + \hat{m}\Lambda_x$$
- Complete the squares to get terms proportional to $(m - \hat{m})^2$, $(b - \hat{b})^2$, and $(m - \hat{m})(b - \hat{b})$
- One can show that all the other terms cancel and what is left is the same result:
$$\chi^2 = \Lambda_{xx}(m - \hat{m})^2 + \Lambda_1(b - \hat{b})^2 + 2\Lambda_x(m - \hat{m})(b - \hat{b})$$

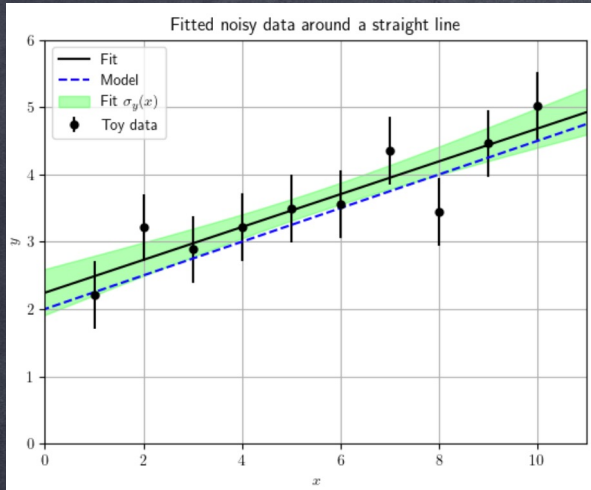
Probability density of $\hat{m}, \hat{b}$

- We can integrate the 2-D normal distribution between elliptical contours of constant density given by

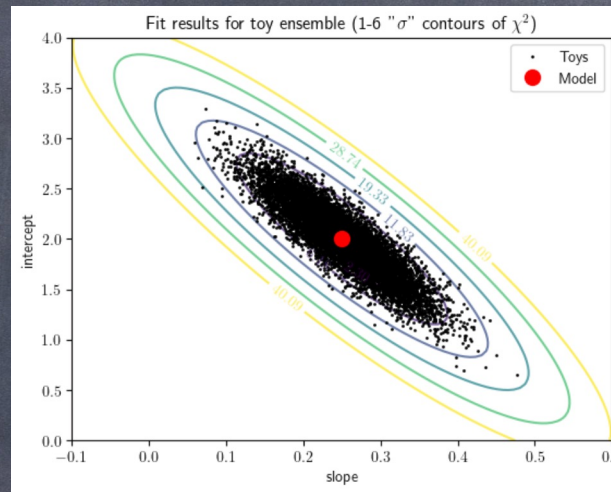
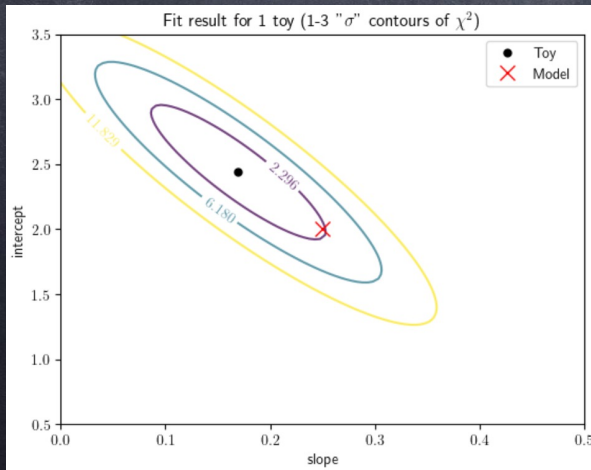
$$\frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)} \equiv \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\chi^2/2}$$

- This is mathematically equivalent to integrating the χ^2 for 2 dof (m and b)!

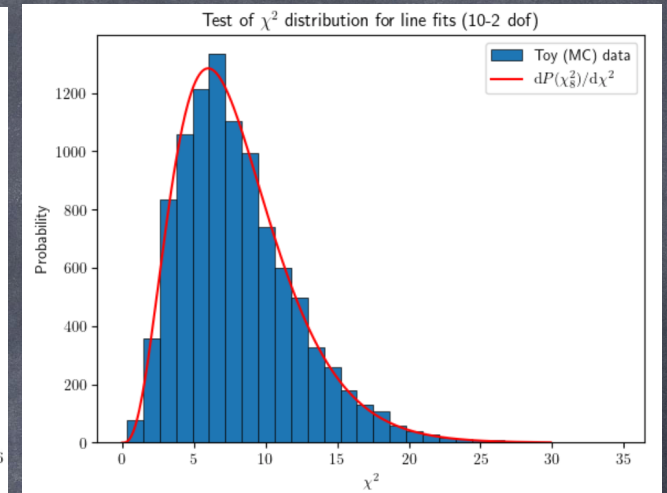
Curve-fitting



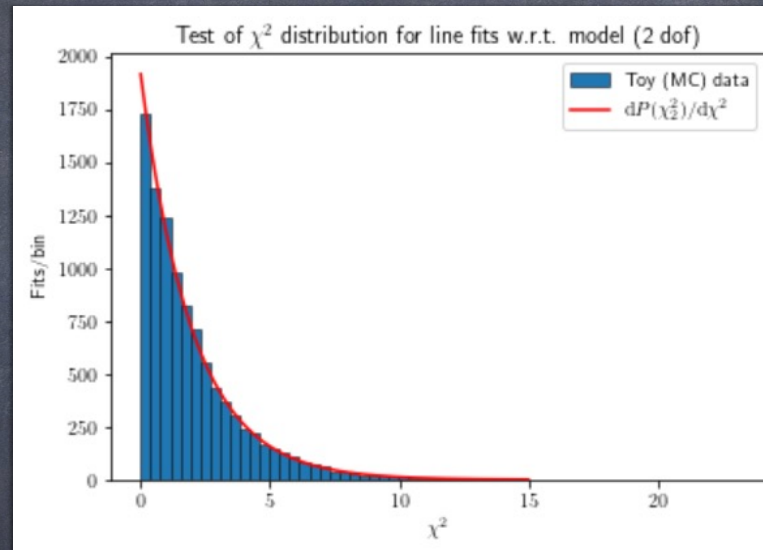
fit to 1 «toy» data set



an ensemble of «toys»

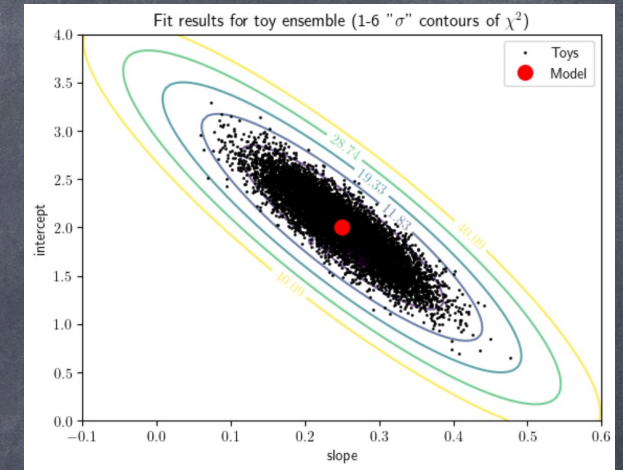
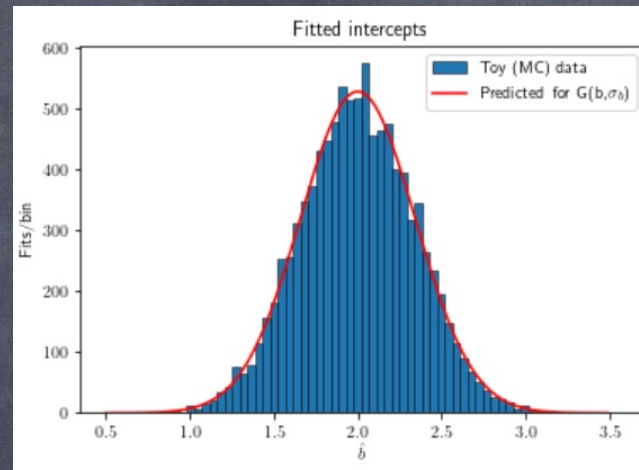
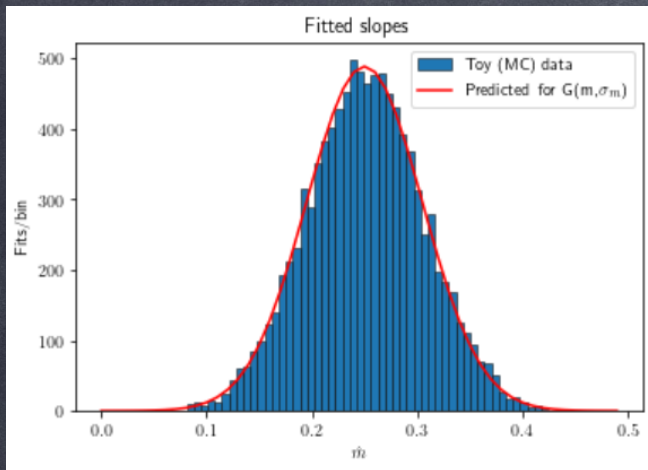


χ^2 probability for 2 dof



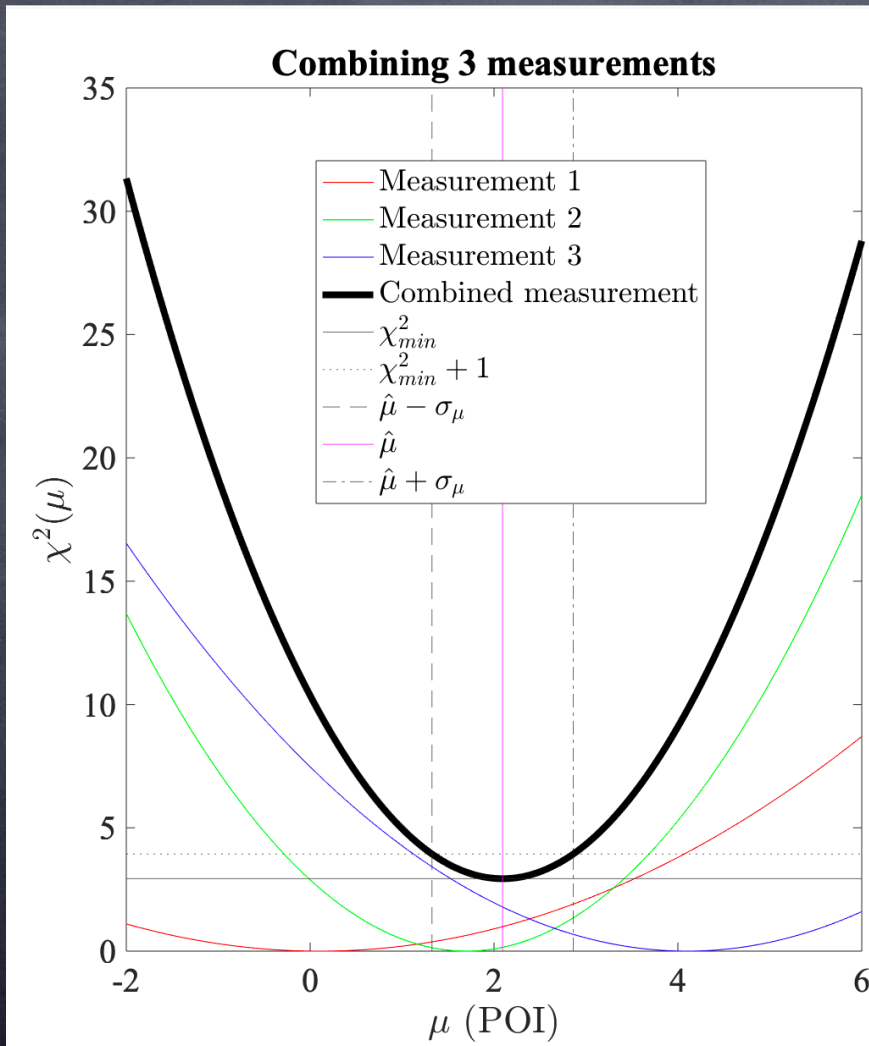
```
chi-squared > 2.296 , fraction = 3.14e-01 +- 4.64e-03 , theory = 3.17e-01
chi-squared > 6.180 , fraction = 4.50e-02 +- 2.07e-03 , theory = 4.55e-02
chi-squared > 11.829 , fraction = 3.20e-03 +- 5.65e-04 , theory = 2.70e-03
chi-squared > 19.334 , fraction = 3.00e-04 +- 1.73e-04 , theory = 6.33e-05
chi-squared > 28.744 , fraction = 0.00e+00 +- 0.00e+00 , theory = 5.73e-07
```

Projections



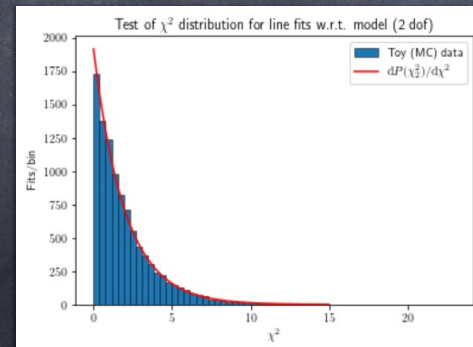
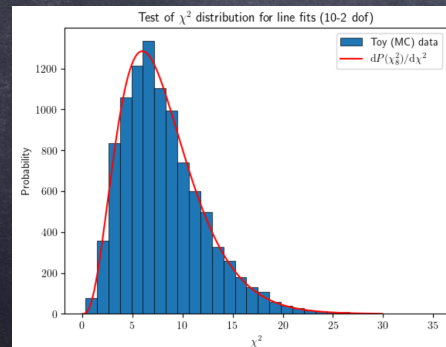
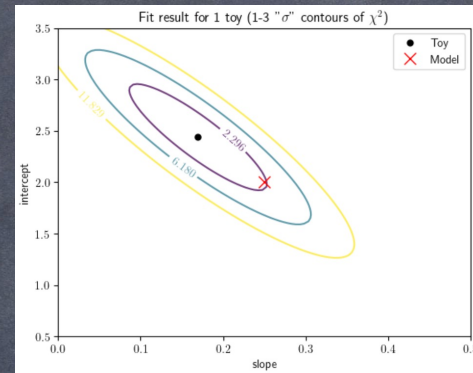
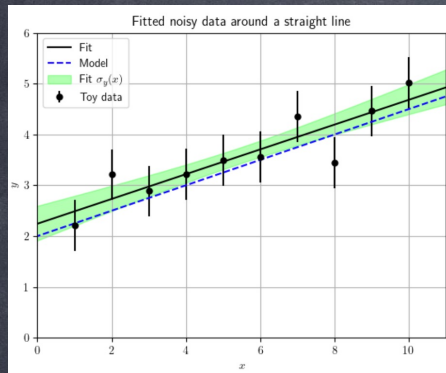
- If we project the pdf of the multivariate normal distribution on 1 axis at a time, we get the 1-D normal distribution with mean μ_i and variance σ_i^2 for each axis i .
- It's only when we plot 2 dimensions at a time that we see the effect of the covariance, i.e., the size of the covariance (value of correlation ρ) doesn't affect the projections.

Interpreting $\chi^2_{\min}$ and $\chi^2(\mu)$ - review



- The black $\chi^2(\mu)$ is the sum of the red, green, and blue $\chi^2(\mu)$'s: $\chi^2 = \sum_{i=1}^3 (x_i - \mu)^2 / \sigma_i^2$
- The value of μ at the minimum of the black curve ($\hat{\mu}$) is our **best estimate** that combines the 3 individual measurements.
- The value of the black curve at the minimum ($\chi^2_{\min}$) is called the **goodness-of-fit** and tells us how compatible the distribution of measurements is with our hypothesis that all 3 have a common mean. A very large value of $\chi^2_{\min}$ would speak against this hypothesis. The observed value of ~ 3 , which is of order $3-1=2$ (dof), speaks in its favor.
- The **curvature of the black curve determines the uncertainty on $\hat{\mu}$** . We can either find out where the black curve has increased by 1 above the minimum, $\chi^2 = \chi^2_{\min} + 1$, or exploit that $\frac{\partial^2 \chi^2}{\partial \mu^2} = \frac{2}{\sigma_{\mu}^2}$ (for normally distributed data and approximately for many other 1-D likelihoods).
- Finally, we could calculate a **p-value (χ^2 probability) for some specific prediction**. As an example, let's say some theory predicts $\mu = 6$, where $\chi^2 \approx 28$. The p-value is $P(\chi^2 > 28 | 2 \text{ dof}) = 8.3 \times 10^{-7}$, which tells us that it is exceedingly unlikely to observe such a high χ^2 (or higher) if the prediction is to be correct.

Question



- Given some model we wish to test, how could we use the two different χ^2 with two different pdfs to characterise such results for a single experiment?
- The χ^2 of the fit at the global minimum (with 10-2 dof) is the "goodness-of-fit". Use this (χ^2 probability) to decide if the hypothesis that the data come from a straight line should be excluded.
- The χ^2 in the slope-intercept plane (with 2 dof) shows the "tension" w.r.t. the best fit of e.g. a model point. Use the second (χ^2 probability) to decide whether the model can be excluded.

Summary of covariance matrix

- The **inverse** of the covariance matrix Σ is the expectation value of the second derivatives of the negative log-likelihood. The χ^2 comes from $-2\log\mathcal{L}$ and ignoring the constant terms.
- In other words $\Sigma_{i,j}^{-1} = \frac{1}{2} \left\langle \frac{\partial^2 \chi^2}{\partial \mu_i \partial \mu_j} \right\rangle$
- For our straight-line fit the $\frac{\partial^2 \chi^2}{\partial \mu_i \partial \mu_j}$ is independent of the data (y_i), so the average is trivial.
- Even for most cases with data-dependent derivatives, a single $\frac{\partial^2 \chi^2}{\partial \mu_i \partial \mu_j}$ with the observed data is an acceptable approximation of the expectation value.
- We found $\frac{\partial^2 \chi^2}{\partial m^2} = 2\Lambda_{xx}$, $\frac{\partial^2 \chi^2}{\partial b^2} = 2\Lambda_1$, $\frac{\partial^2 \chi^2}{\partial m \partial b} = 2\Lambda_x$ for the line-fit, giving $\Sigma^{-1} = \begin{bmatrix} \Lambda_{xx} & \Lambda_x \\ \Lambda_x & \Lambda_1 \end{bmatrix}$