Lessons about Likelihood Functions from Nuclear Physics

Kenneth M. Hanson

T-16, Nuclear Physics; Theoretical Division
Los Alamos National Laboratory

Bayesian Inference and Maximum Entropy Workshop,
Saratoga Springs, NY, July 8-13, 2007

This presentation available at
http://www.lanl.gov/home/kmh/

LA-UR-07-5405

Overview

- Uncertainties in physics experiments
- Particle Data Group (PDG)
- Particle lifetime data
- Coping with outliers
- Uncertainty in the uncertainty
- Student t distributions – normal distribution
- Analysis of lifetime data using t distributions

Physics experiments

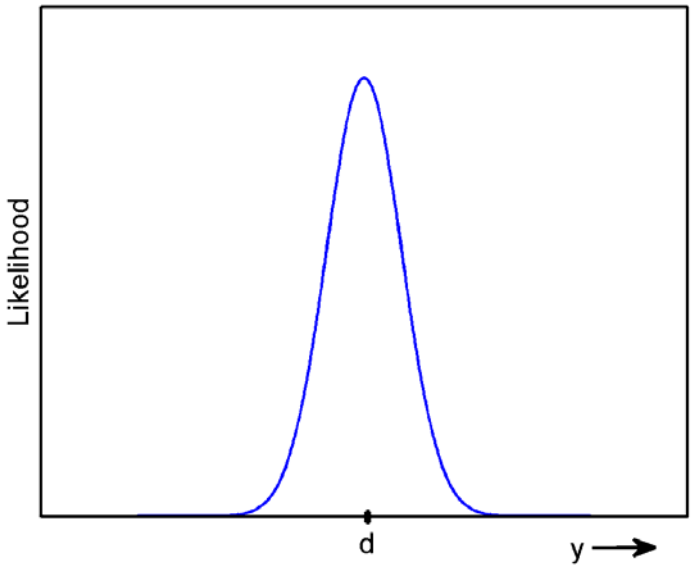
- Suppose experimenter states his/her measurement of physical quantity y as: measurement ± standard error or

  $$y = d \pm \sigma_d$$

  - σ_d represents experimenter's estimated uncertainty in d
- Experimental uncertainty composed of two components:
  - statistical (random) uncertainty
    - from noise in signal or event counting (Poisson distr.)
    - usually Type A – determined by repeated meas., frequentist methods
  - systematic uncertainty
    - may affect many or all of experimental results
    - from equipment calibration, experimental procedure, corrections
    - often Type B – determined by nonfrequentist methods
    - may be based on experimenter's judgment, hence subjective and uncertain
  - these usually added in quadrature (rms sum)

Physics experiments – likelihood functions

- Experimentalist's measurement
  $y = d \pm \sigma_d$ is interpreted
  probabilistically as likelihood function
  $$p(d \mid y \sigma_d I)$$
  - where I is background information, e.g. how experiment is performed
- Likelihood is a probability density function in d
  - normalized to unit area wrt. d
  - but usually viewed as function of y
  - not necessarily normalized wrt. y
- Likelihood usually taken to be normal distribution (Gaussian) with standard deviation σ_d



Physics experiments – inference

- Inference about the physical quantity y is obtained by Bayes law; the posterior distribution for y is

$$p(y \,|\, d\, \sigma_d I) \propto p(d \,|\, y\, \sigma_d I)\, p(y \,|\, I)$$

$$\text{posterior} \;\propto\; \text{likelihood} \times \text{prior}$$

  - where $p(y \mid I)$ is the prior on y, given background information I
- Prior often taken as flat, i.e., $p(y \mid I) = \text{const.}$
  - Bayesian analysis defaults to likelihood analysis
  - result is least-squares (or minimum χ^2) method

Physics experiments – least squares fitting

- Least-Squares (LS) analysis is based on assuming the likelihood is a normal (Gaussian) distribution

$$p(d\,|\,y\,\sigma) = \frac{1}{\sigma\sqrt{\pi}} \exp\left[-\frac{(d-y)^2}{2\sigma^2} \right]$$

- For data set with uncorrelated uncertainties, the likelihood is

$$p(\mathbf{d}\,|\,y\,\boldsymbol{\sigma}) \propto \prod_i \exp\left[-\frac{(d_i-y)^2}{2\sigma_i^{\,2}} \right] \propto \exp\left(-\tfrac{1}{2}\chi^2 \right) + \text{const.}$$

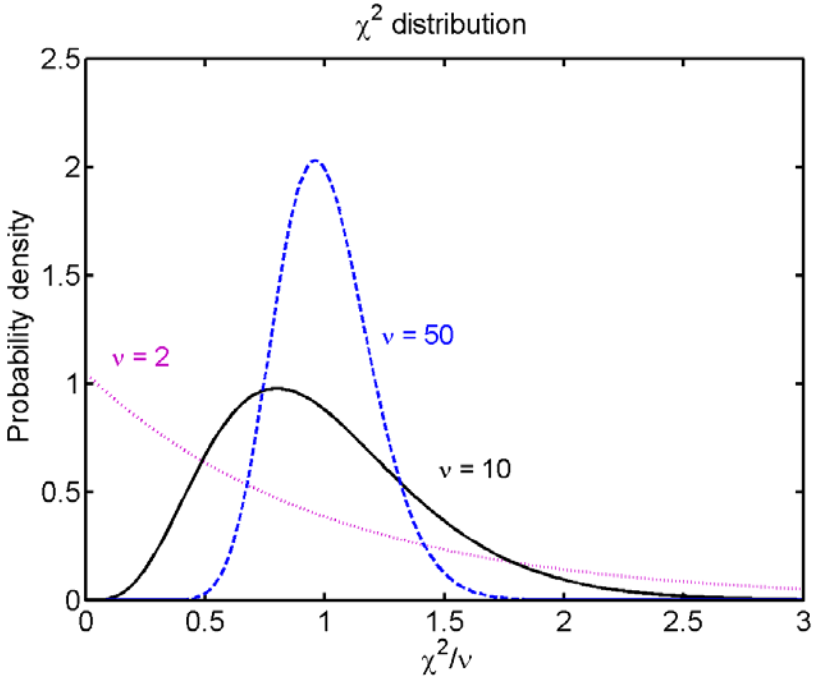
where χ^2 is

$$\chi^2 = \sum_i \frac{(d_i-y)^2}{\sigma_i^{\,2}}$$

- LS analysis – fit model for y by minimizing χ^2
- Check Goodness of Fit by comparing min χ^2 to
  # degrees of freedom = # data – # fit parameters

χ^2 distribution – Goodness of fit

- $\chi^2(\nu)$ distribution is sum of ν squared random numbers, drawn from unit-variance normal distr.
- Shown in graph for $\nu = 2, 10, 50$
  - rms width = $\sqrt{2/\nu}$
- After LS fit, compare min. χ^2 with ν (DOF) to check Goodness of Fit
  - assumes σ's correct, uncertainties independent, and normally distr.
  - quantitatively, calculate p value
- If χ^2 somewhat larger than DOF, analyst often multiplies LS std. error by $\sqrt{\chi^2/\mathrm{DOF}}$



Particle Data Group (PDG)

- Particle Data Group formed in 1957
  - annually summarizes measured properties of elementary particles
- For each particle property, committee:
  - lists all relevant experimental data
  - decides which data to include in final analysis
    - outliers often rejected
  - recommends value (least-squares average of accepted data) and its standard error
    - std. error often magnified by $\sqrt{\chi^2/(n-1)}$ (avg. factor of 2; 50% of time)
- PDG reports are excellent source of information about measurements of unambiguous physical quantities
  - available online; free
  - provide insight into how physicists interpret data

Five venerable elementary particles

- This study will include all measurements of the lifetimes of the following particles:

particle	discov.	mass (MeV)	lifetime, τ (s)	comments
$\mu^{\pm}$	1937	106	2.2×10^{-10}	lepton, cosmic rays
π^0	1950	135	8.4×10^{-17}	meson, nuclear force
K^0_s	1951	498	0.90×10^{-10}	strange meson, CP viol.
n	1931	940	886	baryon, nucleus constituent
Λ^0	1952	1116	2.63×10^{-10}	strange baryon

Lambda lifetime measurement in the 60s

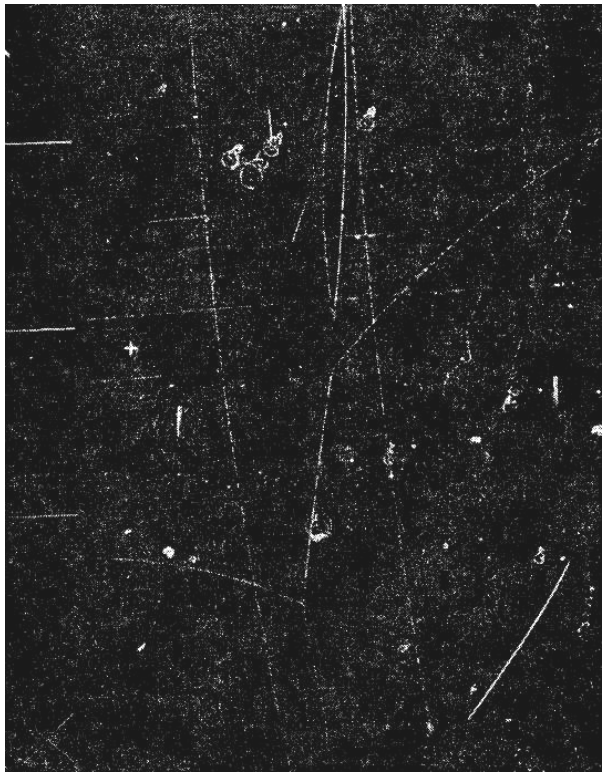
- Hydrogen bubble chambers used in 1950s and 60s to observe elementary particles
- Picture shows reaction sequence:

$$K^- + p \rightarrow \Xi^- + K^+$$
$$\Xi^- \rightarrow \Lambda^0 + \pi^-$$
$$\Lambda^0 \rightarrow p + \pi^-$$

- Track lengths and particle momenta, determined from curvature in magnetic field, yield survival time of Ξ^- and Λ
- Hubbard et al. observed 828 such events to obtain lifetimes:

$$\tau_\Xi = 1.69 \pm 0.06 \times 10^{-10} \text{ s}$$
$$\tau_\Lambda = 2.59 \pm 0.09 \times 10^{-10} \text{ s}$$

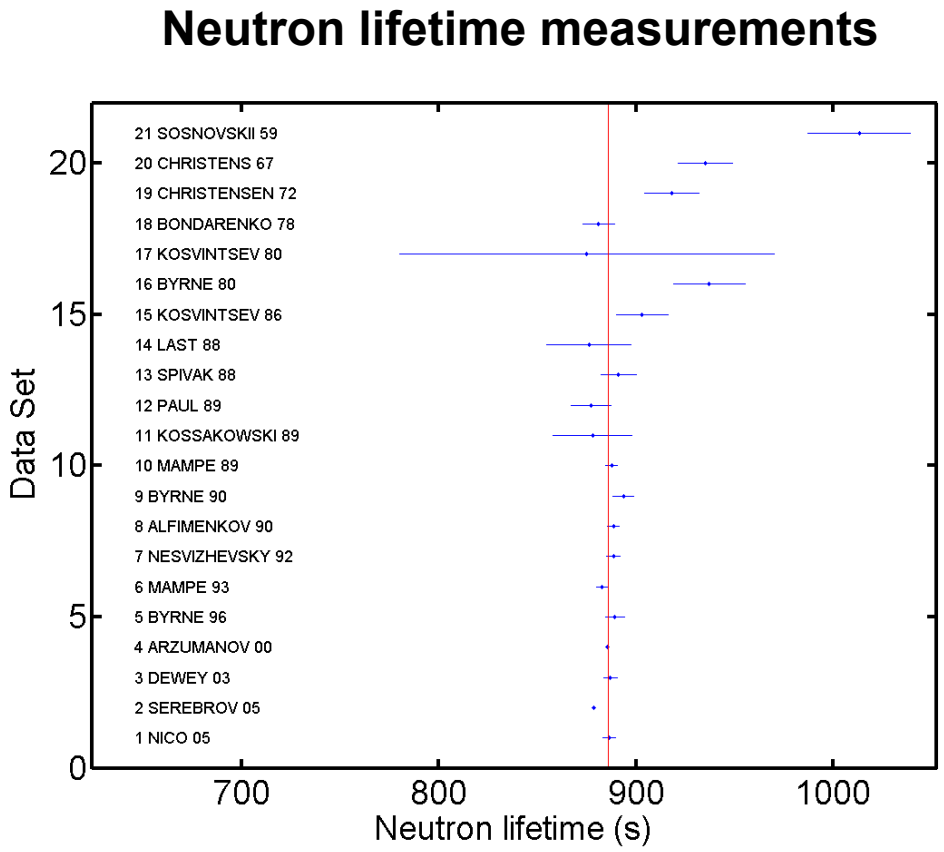
Hydrogen bubble-chamber photo



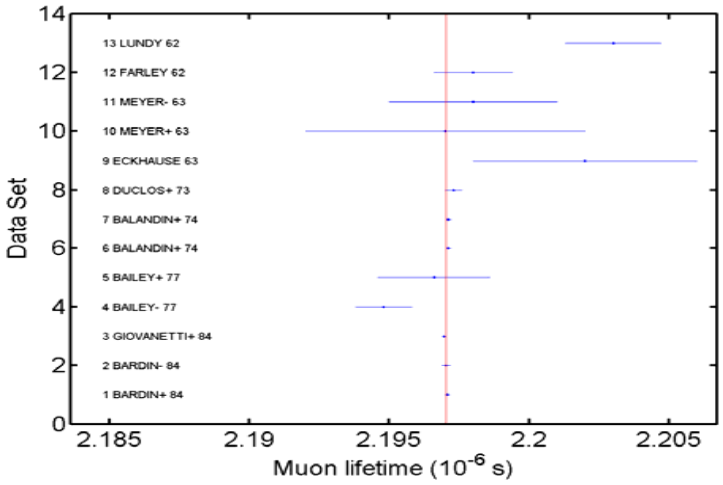
From J.R. Hubbard et al., *Phys.Rev.* **135B** (1964)

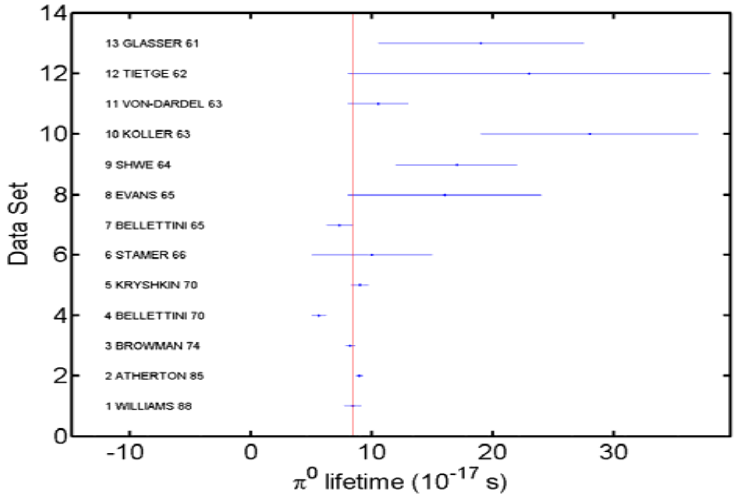
Measurements of neutron lifetime

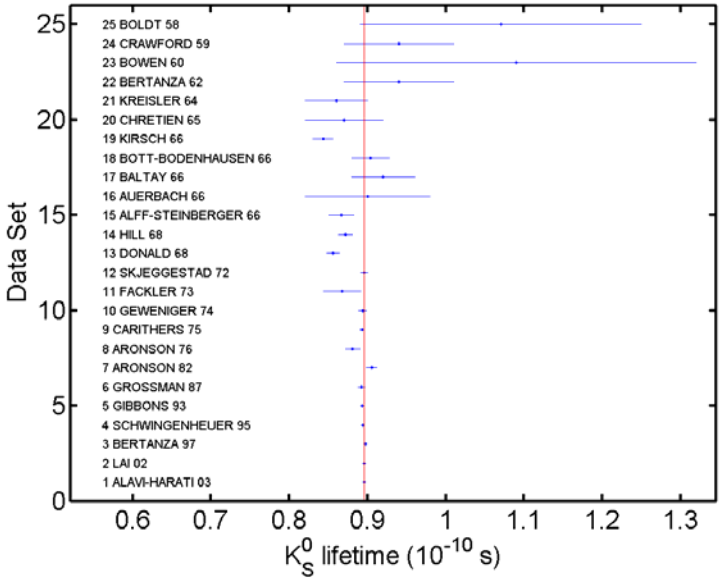
- Because n lifetime is so long, it is difficult to measure accurately without slowing neutrons or trapping them
- Plot shows all measurements of neutron lifetime
- Vertical line is PDG value, which includes 7 most recent measurements, except for #2 because it is highly discrepant
- χ^2 (PDG) = 149/21 pts.
- Several outliers exist

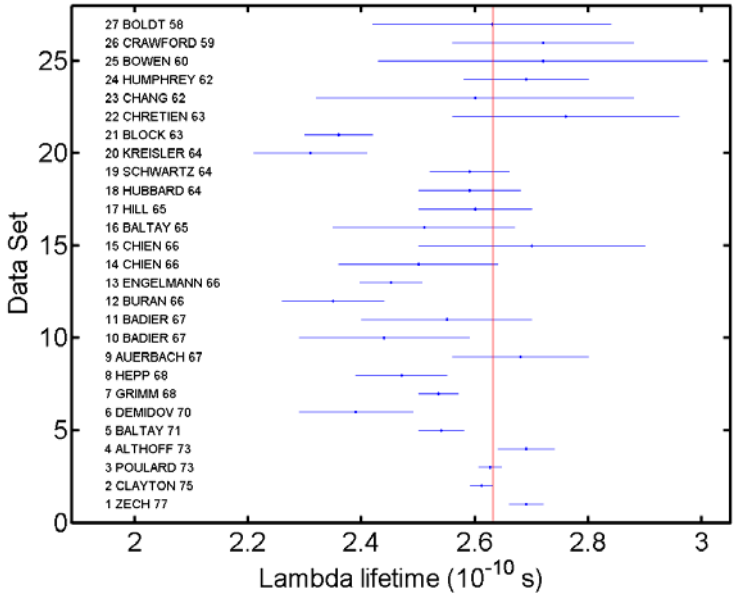


Some other lifetime measurements



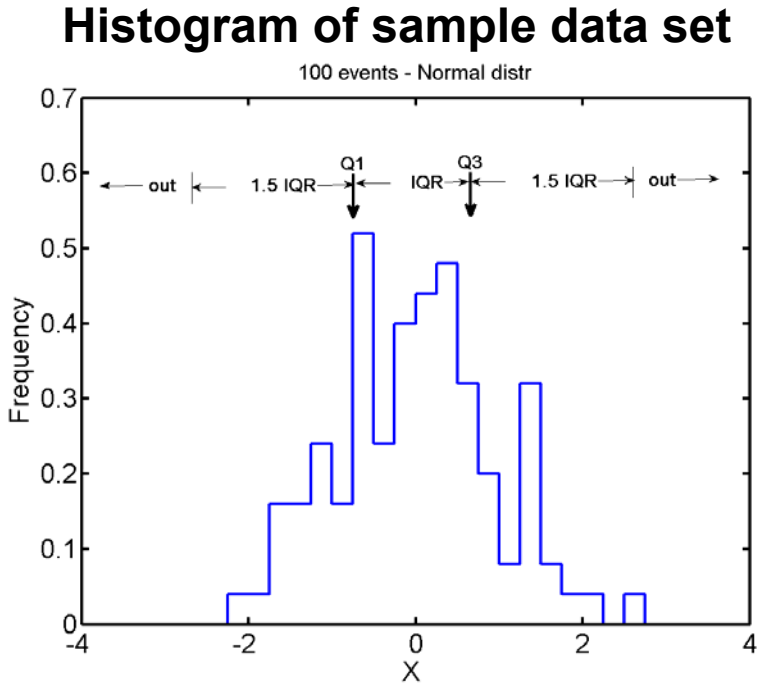






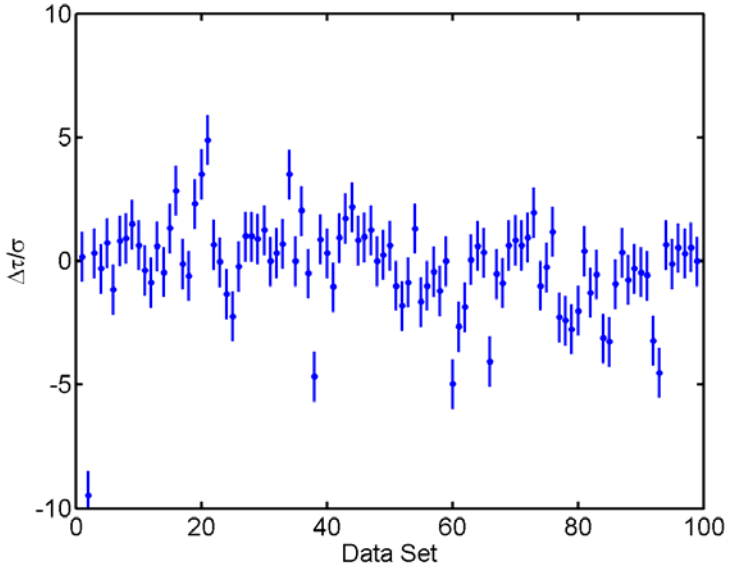
Exploratory data analysis – IRQ and SOF

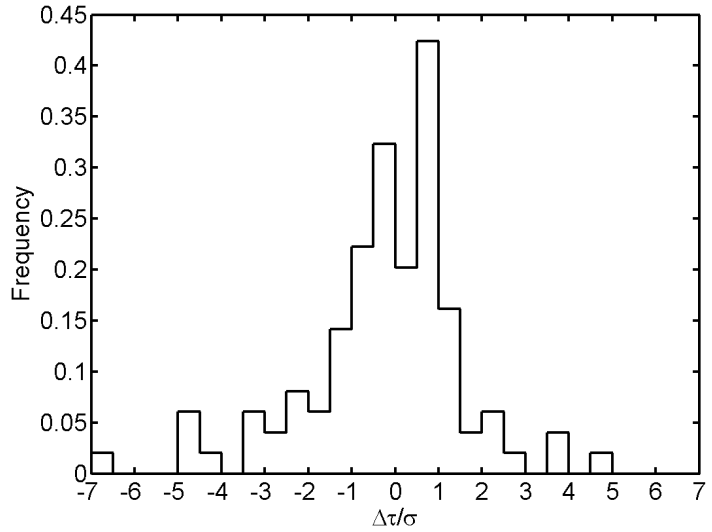
- John Tukey (1977) suggested each set of measurements be scrutinized
  - find quartile positions, Q1, Q2, Q3
  - calculate the inter-quartile range
    IQR = Q3 – Q1
  - calc. fraction of data in the intervals
    y < Q1 – 1.5 IQR; y > Q3 + 1.5 IQR
    called suspected outlier fraction (SOF)
- For normal distr.
  - IQR = 1.35 σ
  - SOF = 0.7% (outside 2.7 σ)
- Q2 (median) is good estimate of x
- IQR measures width of core
- SOF measures extent of tail



Composite of lifetime measurements

- Upper graph shows discrepancies of 99 lifetime meas. for 5 particles from PDG values, divided by their standard errors, i.e. $\Delta\tau/\sigma$
- Lower graph shows histogram
- $\chi^2(y = 0) = 367/99$ points
- IQR = 1.83 (1.35 for normal)
- Suspected outlier frac. = 6.1 %
- **Objective: characterize the distribution of discrepancies relative to their estimated uncertainties, $\Delta\tau/\sigma$**



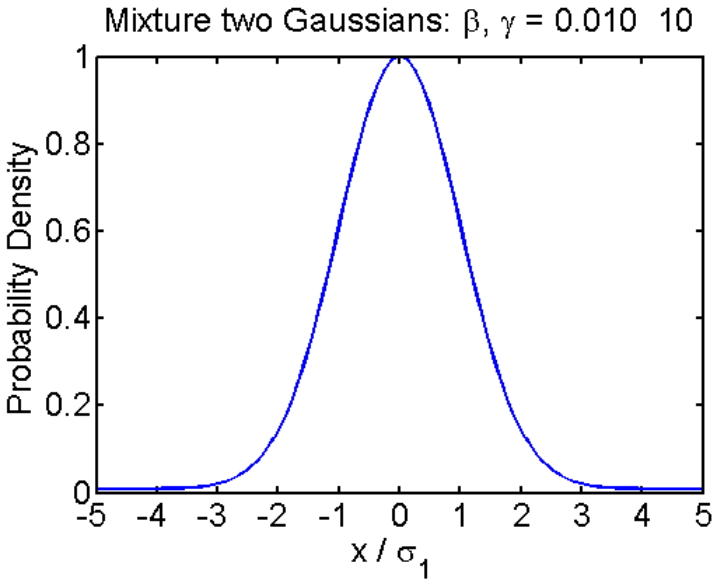


Coping with outliers

- Outliers disrupt analyses based on normal likelihood functions (e.g., LS)
- Outlier-tolerant likelihood functions generally have long (thick) tail
  - long tail admits possibility of large deviations from true value
  - exact form doesn't seem to matter
- A simple long-tail likelihood is mixture of two Gaussians:

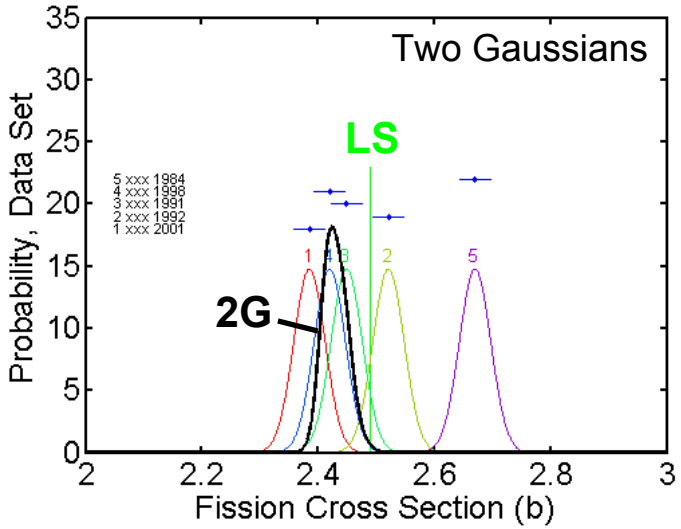
$$\propto (1-\beta)\exp\left\{-\frac{(y-d)^2}{2\sigma^2}\right\}+\frac{\beta}{\gamma}\exp\left\{-\frac{(y-d)^2}{2\gamma^2\sigma^2}\right\}$$

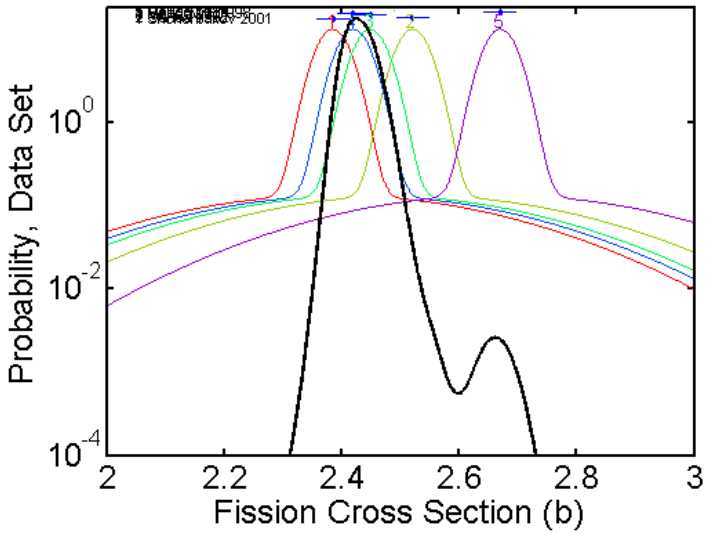
  - β is probability of long-tail Gaussian
  - typical values: $\beta = 0.01\text{-}0.05$, $\gamma = 5\text{-}20$



Outlier-tolerant likelihood functions

- Hypothetical data set with outlier
  - vertical green line show LS avg.
  - thick curve is likelihood of data set
- Plot shows posterior based on two-Gaussian likelihood
  - log scale shows tails of likelihood functions
  - long tail from outlier does not influence peak shape near cluster of three measurements
  - long tails from cluster allows outlier to produce a small secondary peak; has little effect on posterior mean



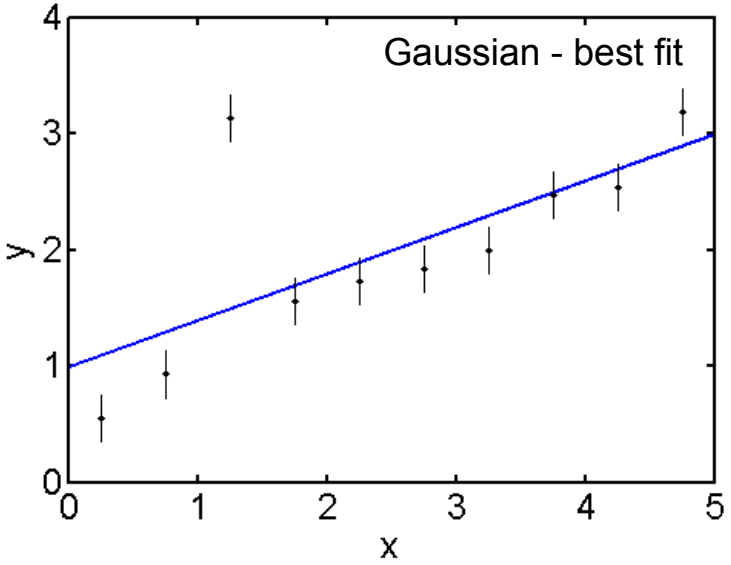


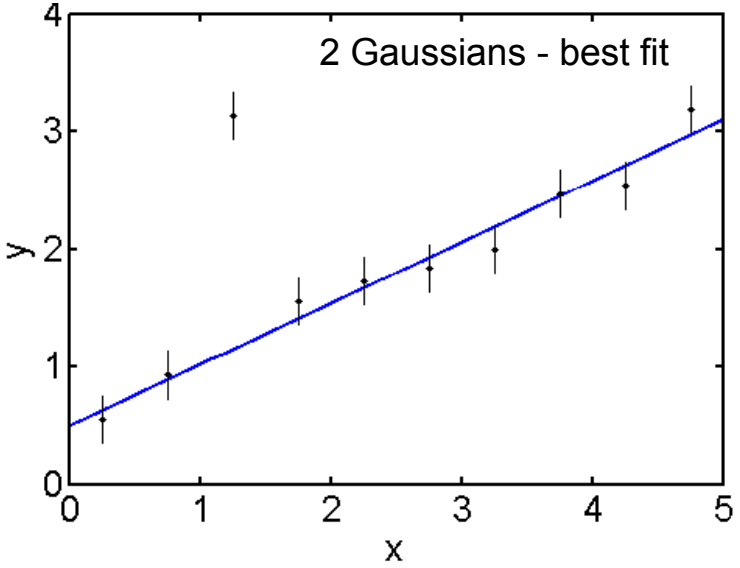
Effect of outlier on linear fit

- Outliers pose significant problem for LS algorithm, based on Gaussian likelihood
- Graph shows 10 data with outlier; error bars indicate known std. errs.
- LS (Gaussian like.) results in fitted line that disagrees with most data:

$$\chi^2_{\min} = 85.6/9\,\text{DOF}, \quad p = 10^{-15}$$

- Using two-Gaussian likelihood for all data gracefully handles outlier
  - fit is unchanged by outlier
- **All data treated in same way**
  - no need to identify outliers

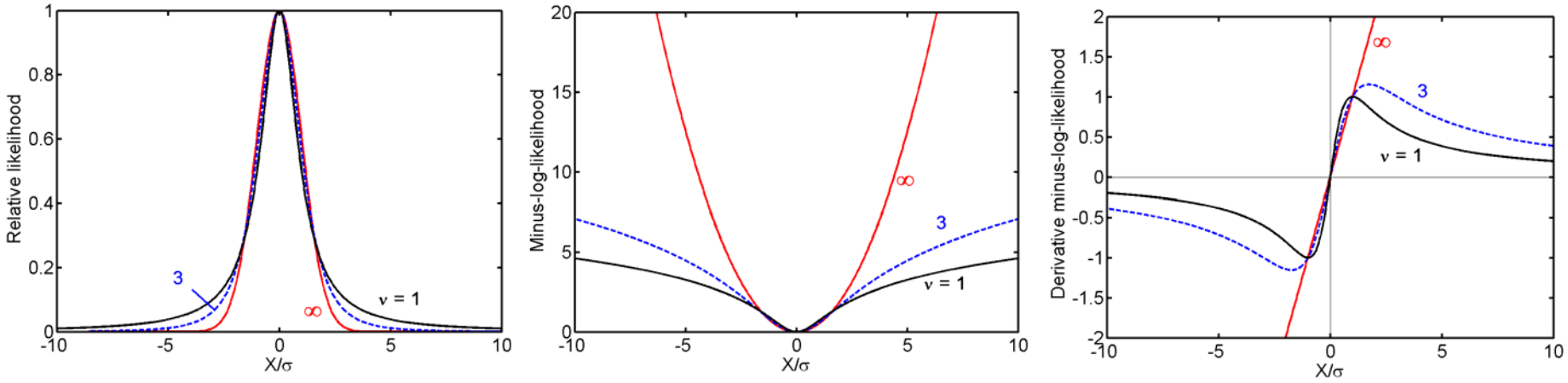




Physical analogy of probability

- $\varphi(\Delta x)$ = minus-log-likelihood is analogous to a physical potential
  - $\nabla\varphi$ is a force with which each datum pulls on model
- Outlier-tolerant likelihoods
  - generally have long tails
  - restoring force eventually decreases for large residuals

Examples of long-tailed distrs.: t distributions for ν = 1, 5, ∞ (normal)



Uncertainty in the uncertainty

- Suppose there is uncertainty in the stated standard error σ_0 for measurement *d*
- Dose and von der Linden (2000)* give plausible derivation:
  - assume likelihood has underlying normal distr.

$$p(d \mid y\,\sigma I) \propto \exp\left[-\frac{1}{2}\left(\frac{d-y}{\sigma}\right)^2\right]$$

  - assume uncertainty distr. for ω, where σ is scaled by $\sigma = \sigma_0 / \sqrt{\omega}$

$$p(\omega \mid I) = \Gamma_a(\omega) \propto \omega^{\frac{\nu}{2}-1} \exp\left[-\nu\omega/2\right]$$

  - marginalizing over ω, the likelihood is Student t distr.,

$$p(d \mid y\,\sigma_0 I) \propto \left[1+\frac{1}{\nu}\left(\frac{d-y}{\sigma_0}\right)^2\right]^{-\frac{\nu+1}{2}} \propto t_\nu\left(\frac{d-y}{\sigma_0}\right)$$

- **t distribution more appropriate for likelihood than normal**

*Other contributors: Box and Tiao, O'Hagan, Fröhner, Press, Sivia, Hanson and Wolf

Prior on standard error

- In derivation by Dose and von der Linden, prior on ω is:

$$p(\omega \mid I) = \Gamma_a(\omega) \propto \omega^{\frac{\nu}{2}-1} \exp\left[-\nu\omega/2\right]$$

where σ is scaled by

$$\sigma = \sigma_0 / \sqrt{\omega} = s\,\sigma_0$$

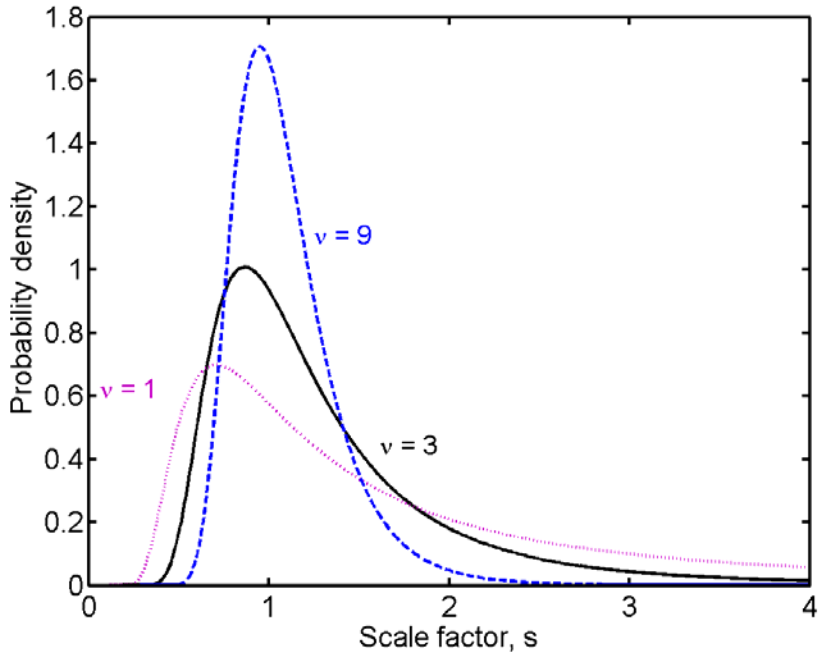
- Corresponding prior on s is

$$p(s \mid I) = \left[\left|\frac{d\omega}{ds}\right| p(\omega \mid I)\right]_{\omega = s^{-2}}$$

$$\propto s^{-(1+\nu)} \exp\left[-\nu\, s^{-2}/2\right]$$

- These are plausible distributions for representing uncertainty in σ
  - rms dev = 1.06, 0.69, 0.30 (ν =1,3,9)

Uncertainty distributions for scaling parameter *s*

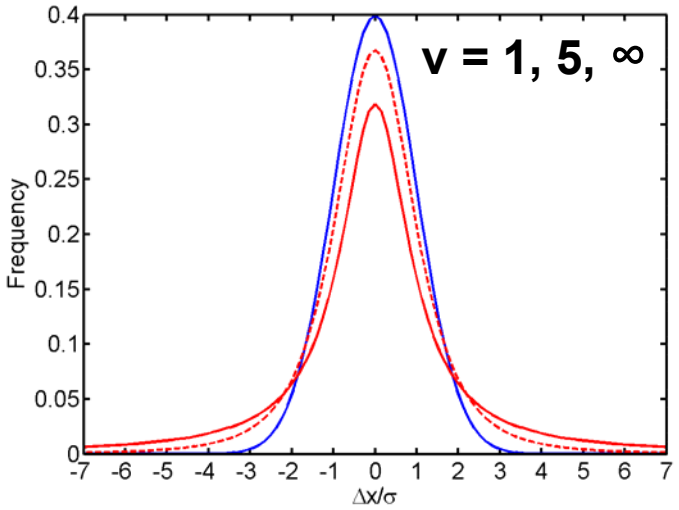


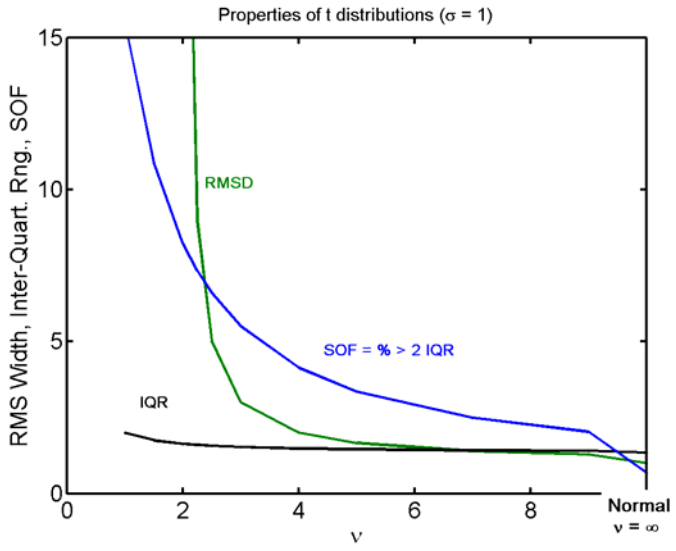
Student t distribution

- Student* t distribution

$$t_\nu(z) \propto \left[1 + \frac{z^2}{\nu}\right]^{-\left(\frac{\nu+1}{2}\right)}$$

  - long (thick) tail for $\nu < 9$ (SOF > 2%)
  - $\nu = 1$ is Cauchy distr. (solid red)
  - $\nu = \infty$ is normal distr. (solid blue)
- Lower graph shows ν dependence of
  - RMSD (square root of variance)
  - IQR (Intra-quartile range)
  - SOF (Suspected outlier fraction)

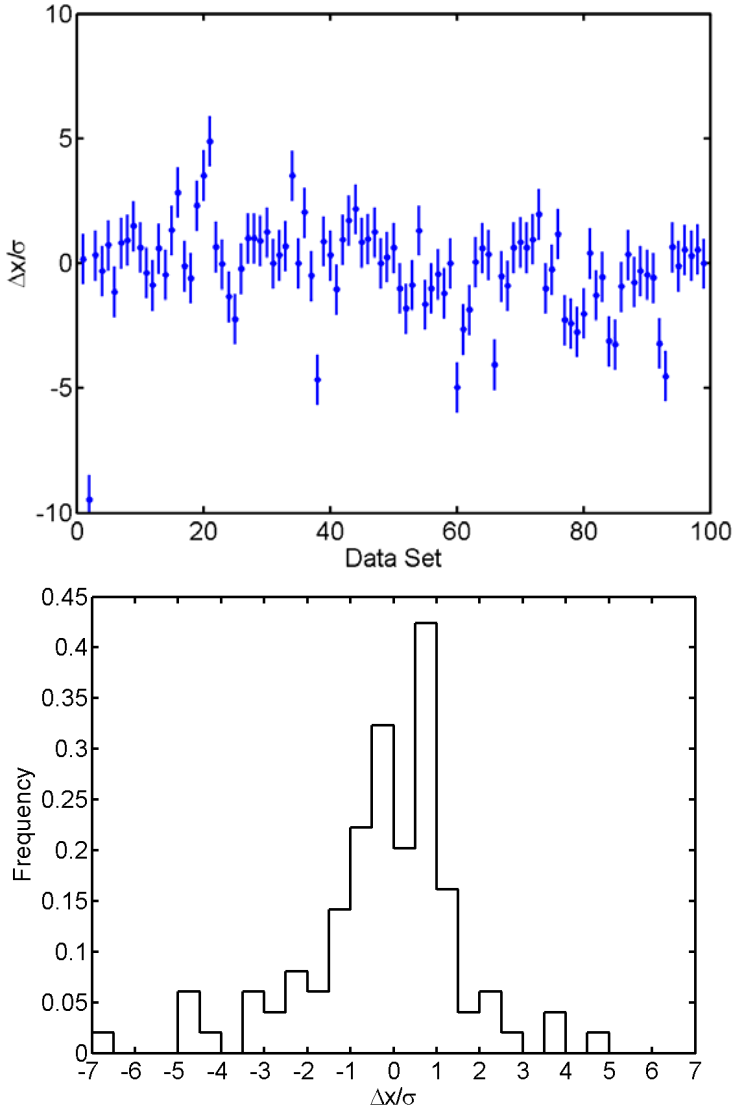




* Student (1908) was pseudonym for W.S. Gossett, who was not allowed to publish under his own name by his employer, Guinness brewery

Composite of lifetime measurements

- Recall data for discrepancies of 99 lifetime meas. for 5 particles from PDG values, divided by their standard errors, i.e. $\Delta\tau/\sigma$
- Lower graph shows histogram
- Objective: characterize the distribution of discrepancies relative to their estimated uncertainties, $\Delta\tau/\sigma$
  - **do data follow normal or t distribution?**



Bayesian model selection

- To select between two models, A and B, Bayes rule gives the odds ratio

$$\frac{p(\mathrm{A}\,|\,\mathbf{d}\,\boldsymbol{\sigma}\,I)}{p(\mathrm{B}\,|\,\mathbf{d}\,\boldsymbol{\sigma}\,I)} = \frac{p(\mathbf{d}\,|\,\mathrm{A}\boldsymbol{\sigma}\,I)}{p(\mathbf{d}\,|\,\mathrm{B}\boldsymbol{\sigma}\,I)}\,\frac{p(\mathrm{A}\,|\,I)}{p(\mathrm{B}\,|\,I)}$$

where $p(\mathrm{A}|I)/p(\mathrm{B}|I)$ is the prior odds ratio on the models and $p(\mathbf{d}\,|\,\mathrm{A}\boldsymbol{\sigma}\,I)$ is **evidence**, evaluated as the integral over the likelihood of the parameters τ and s for model A

$$p(\mathbf{d}\,|\,\mathrm{A}\boldsymbol{\sigma}\,I) = \int p(\mathbf{d}\,|\,\tau s\mathrm{A}\boldsymbol{\sigma}\,I)\,p(\tau s\,|\,\mathrm{A}I)\,d\tau\,ds$$

where $p(\mathbf{d}\mid \tau\ s\ \mathrm{A}\ \boldsymbol{\sigma}\ I)$ is the likelihood and $p(\tau\ s\mid \mathrm{A}\ I)$ is the prior on the lifetime and scale factor

Analysis of lifetime data set

- To calculate average value of data set, use the Student t distribution for likelihood of each measurement of lifetime, τ:

$$p(d_i \mid \tau\, \sigma_i\, s\, I) \propto \left[1+\frac{1}{\nu}\left(\frac{d_i-\tau}{s\,\sigma_i}\right)^2\right]^{-\frac{\nu+1}{2}} \propto t_\nu\left(\frac{d_i-\tau}{s\,\sigma_i}\right)$$

  where s is scaling factor of standard error for whole data set
- Select ν based on data using Bayesian model selection
- Scale factor s marginalized out of posterior

$$p(\tau \mid \mathbf{d}\,\boldsymbol{\sigma}) = \int p(\tau\, s \mid \mathbf{d}\,\boldsymbol{\sigma})\, ds = \int p(\mathbf{d} \mid \tau\, s\,\boldsymbol{\sigma})\, p(\tau)\, p(s)\, ds$$

  - $p(\tau)$ is prior on lifetime, τ (= const.)
  - $p(s_i)$ is prior on s_i (= const., although $1/s_i$ is often appropriate)
- Posterior for s determined by dispersion of data

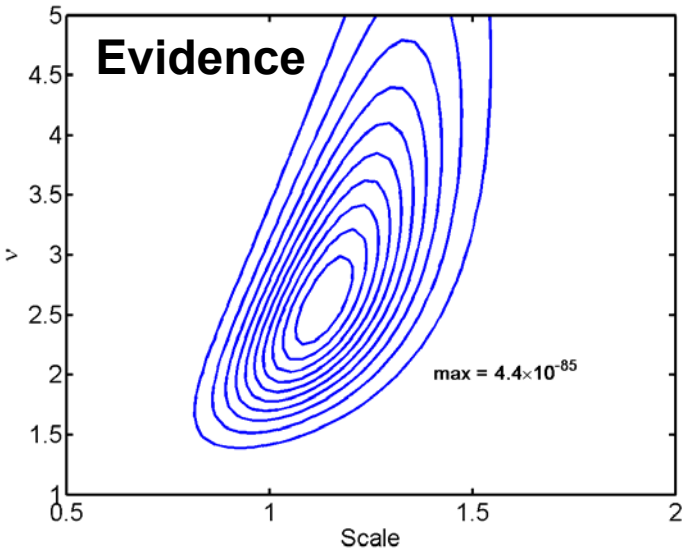
Model selection

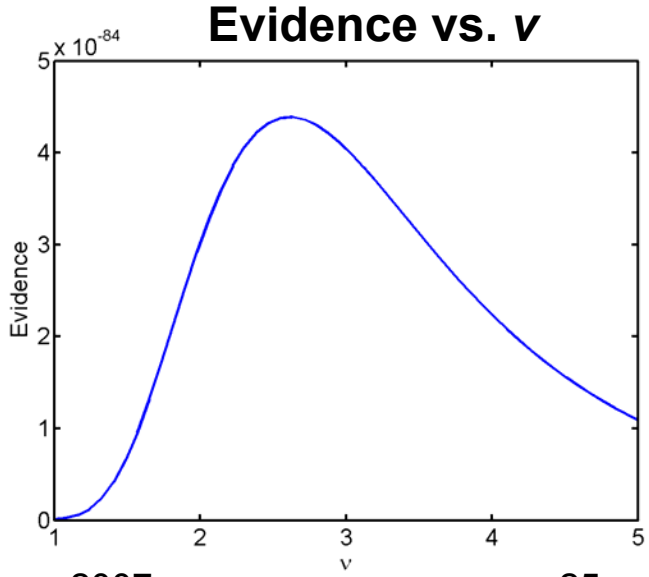
- Odds ratios of t distr. (T) to normal (N) is

$$\frac{p(\mathrm{T}\,|\,\mathbf{d}\,\boldsymbol{\sigma}\,I)}{p(\mathrm{N}\,|\,\mathbf{d}\,\boldsymbol{\sigma}\,I)} = \frac{p(\mathbf{d}\,|\,\nu{=}2.6\,\mathrm{T}\,\boldsymbol{\sigma}\,I)}{p(\mathbf{d}\,|\,\mathrm{N}\,\boldsymbol{\sigma}\,I)}\,\frac{p(\mathrm{T}\,|\,I)}{p(\mathrm{N}\,|\,I)}$$

$$= 1.3\times10^{-85}\,/2.2\times10^{-90} = 5.5\times10^{4}$$

  - assuming prior ratio on models = 1 and priors on parameters equal (~5)
  - evidence is integral over τ and s
  - priors on τ and s = constant
- **t distr. is strongly preferred by data to normal distr.**
  - $\nu \approx 2.6$ (maximizes evidence)
  - for normal: avg. $s = 1.95 \pm 0.14$





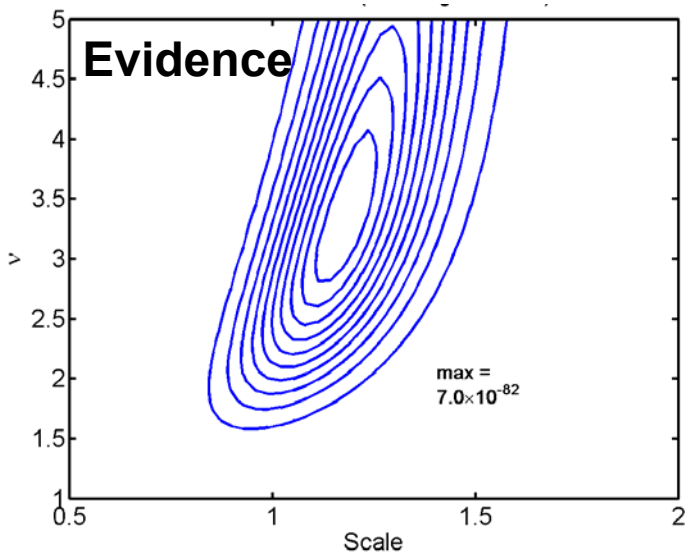
Model selection – excluding largest outlier

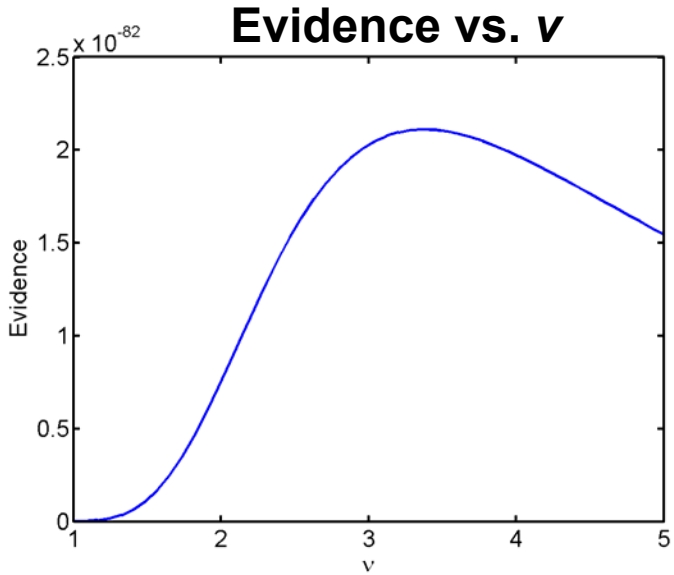
- Remove most discrepant datum (9.5 σ)
- Odds ratios of t distr. (T) to normal (N)

$$\frac{p(\mathrm{T}\,|\,\mathbf{d}\,\boldsymbol{\sigma}\,I)}{p(\mathrm{N}\,|\,\mathbf{d}\,\boldsymbol{\sigma}\,I)} = \frac{p(\mathbf{d}\,|\,\nu{=}3.3\,\mathrm{T}\,\boldsymbol{\sigma}\,I)}{p(\mathbf{d}\,|\,\mathrm{N}\,\boldsymbol{\sigma}\,I)}\,\frac{p(\mathrm{T}\,|\,I)}{p(\mathrm{N}\,|\,I)}$$

$$= 2.1\times10^{-82}/5.0\times10^{-84} = 42$$

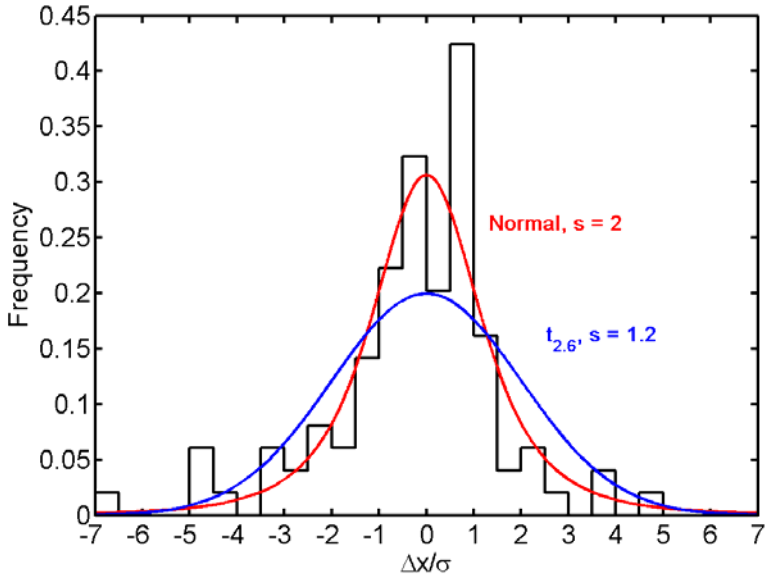
  - assuming prior ratio on models = 1 and priors on parameters equal (~5)
  - evidence is integral over τ and s
  - priors on τ and s = constant
- t distr. is still preferred by data to normal distr.
  - $\nu \approx 3.3$ (maximizes evidence)
  - for normal: avg. $s = 1.71 \pm 0.12$





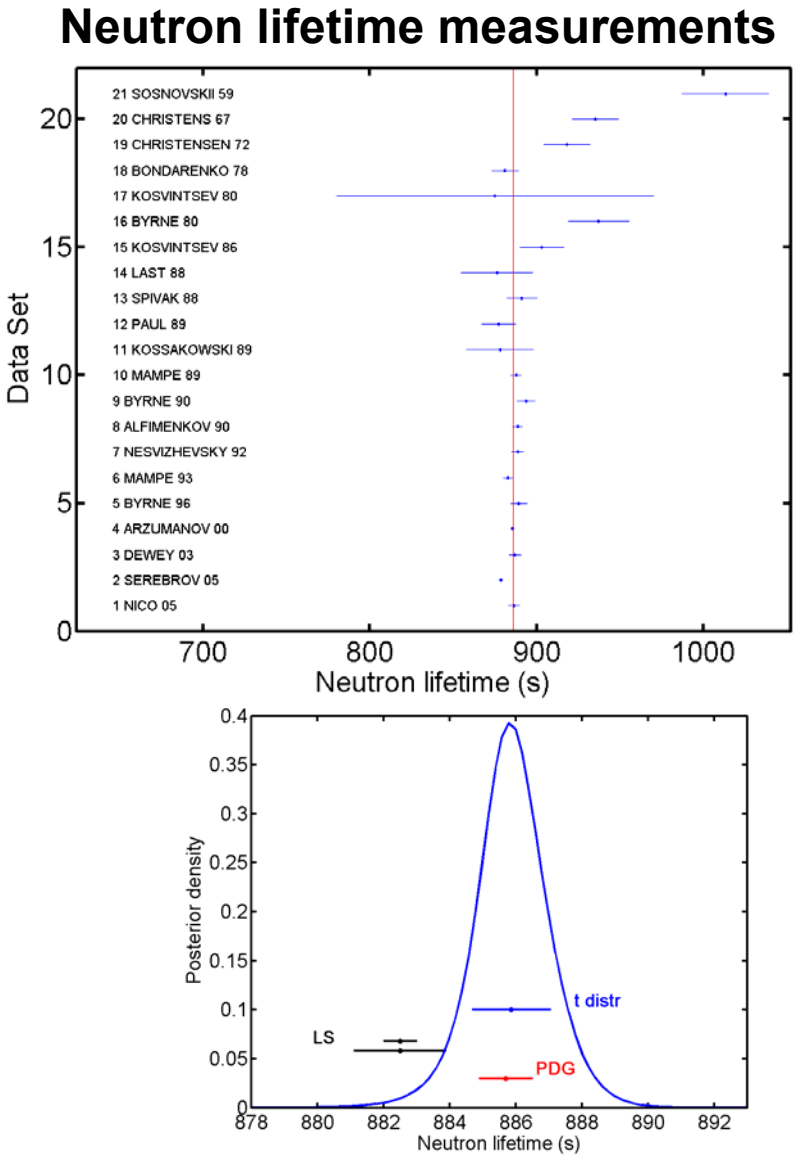
Bayesian model selection

- Graph shows best fits of likelihood functions to histogram of normalized residuals of lifetime data
- For normal: $s = 1.95 \pm 0.14$
- For t distr.: $s = 1.2 \pm 0.15$, $\nu = 2.6$
- Bayesian model-selection analysis indicates t distr. is 5.5×10^4 times more likely than normal distr.



Analysis of neutron lifetime data

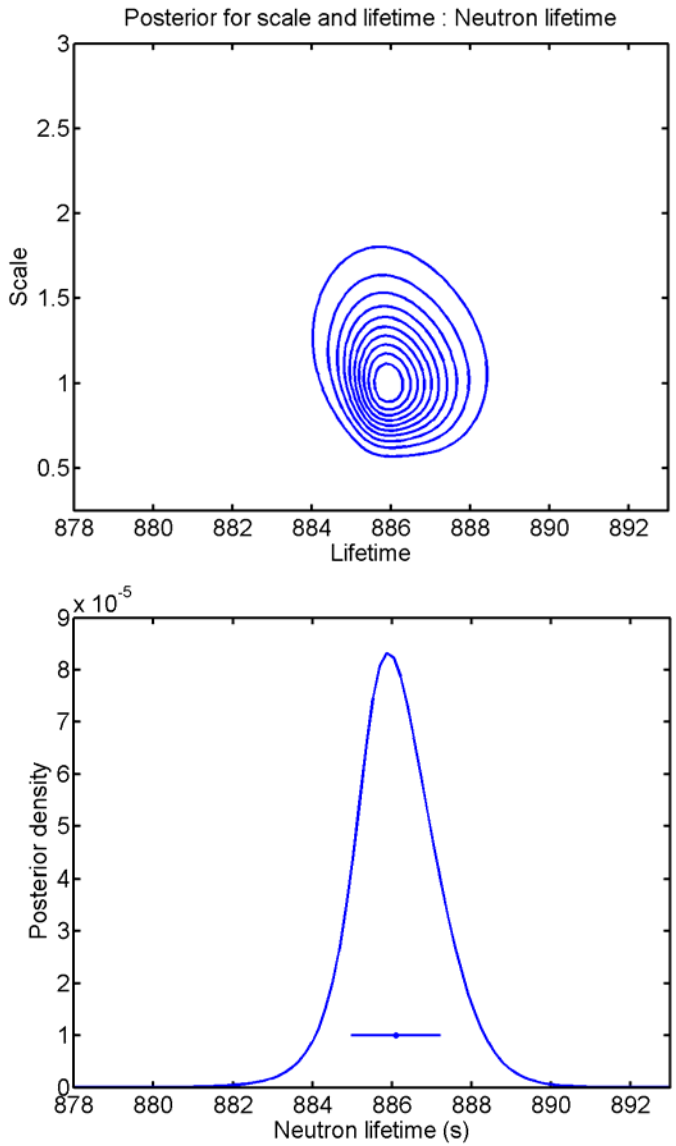
- Upper plot shows all measurements of neutron lifetime
- Lower plot shows results based on all 21 data points:
  - posterior for t-distr. analysis ($\nu = 2.6$, margin. over s; $\overline{s}=1.16$)
    - consistent with PDG
  - least-squares result (w/o and with χ^2 scaling, $s = 2.73$)
    - single outlier has large effect
- PDG value (using 7 most recent data points, excluding Serebrov; s = 1)



Analysis of neutron lifetime data

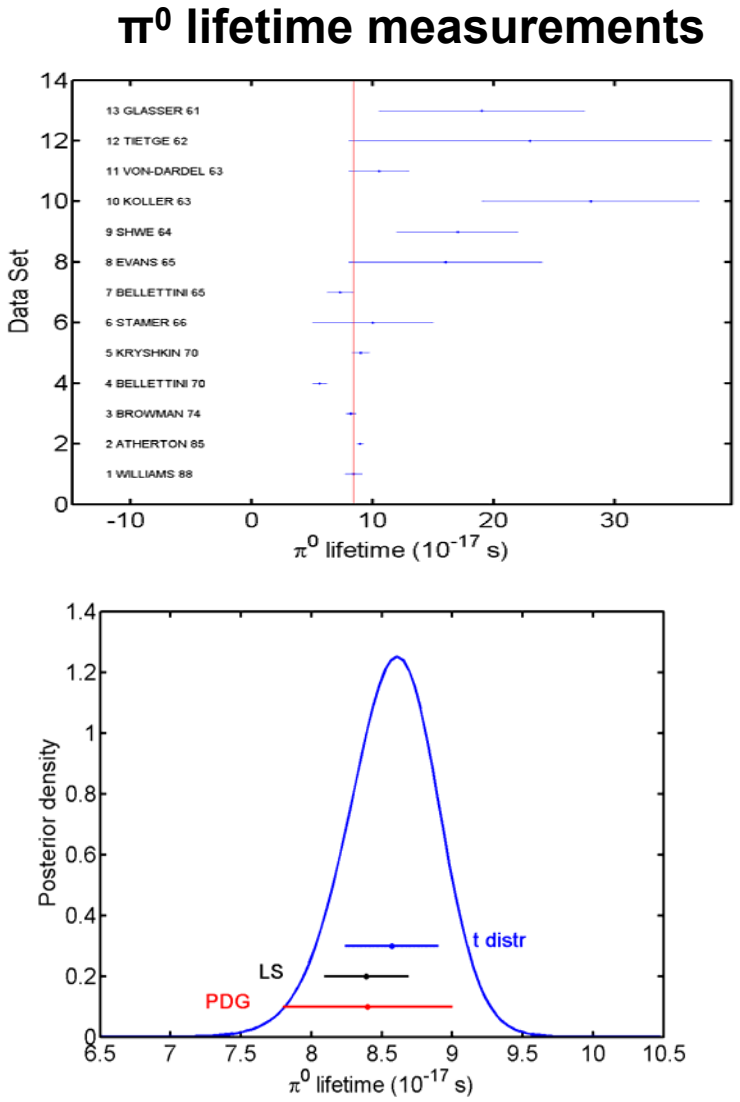
- Details of analysis of neutron data
  - t distr. with $\nu = 2.6$ (fixed)
- Upper plot shows joint posterior distr. for s and τ (lifetime)
  - priors for s and τ constant
- Lower plot:
  - posterior for lifetime (projection of joint distr. onto τ, i.e. marginalized over s)
  - lifetime estimate is posterior mean, standard error is rms dev.:

    $\tau = 886.1 \pm 1.1$ s



Analysis of π^0 lifetime data

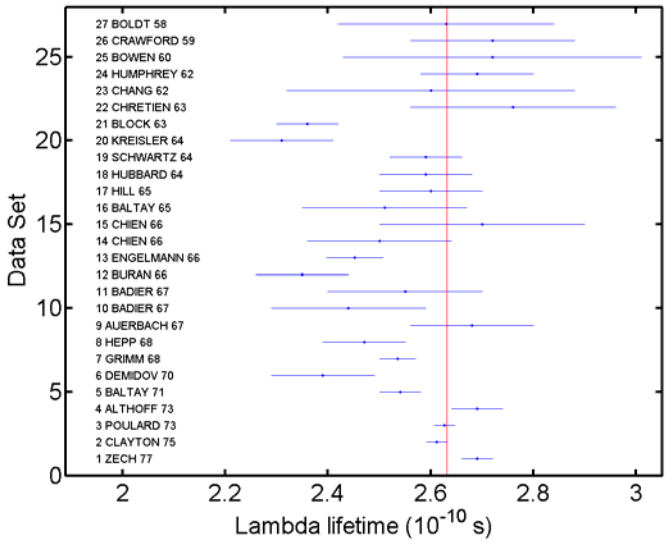
- Upper plot shows all measurements of π^0 lifetime
- Lower plot shows results based on all 13 data points:
  - posterior for t-distr. analysis ($\nu = 2.6$, margin. over s; $\overline{s}=1.55$)
  - least-squares result (with χ^2 scaling, $s = 1.58$)
- PDG values (using 4 selected data points, excluding latest one; σ scaled by $s = 3.0$)
- Results all consistent

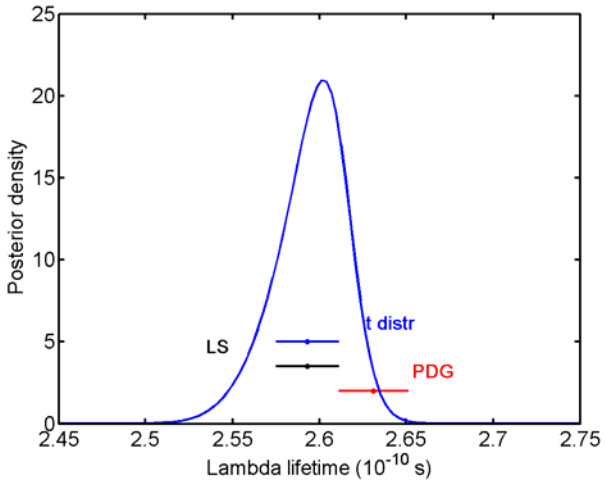


Analysis of Λ lifetime data

- Upper plot shows all measurements of Λ lifetime
- Lower plot shows results based on all 27 data points:
  - posterior for t-distr. analysis ($\nu = 2.6$, margin. over s $\overline{s}=1.59$)
  - least-squares result (with χ^2 scaling, $s = 1.81$)
- PDG value (using 3 latest data points, $s = 1.6$)
  - disagrees with LS and t-distr. results
  - ignores most data

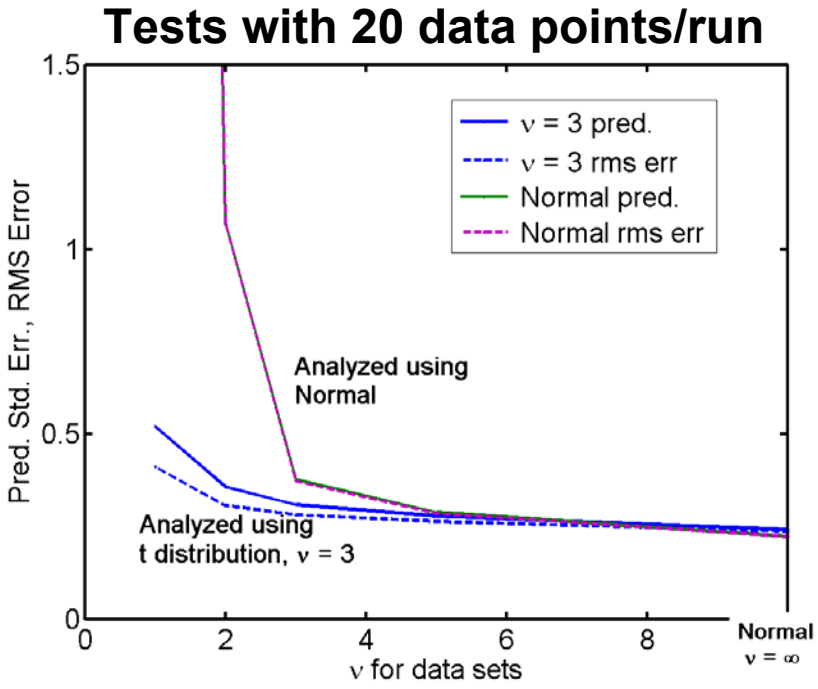
Lambda lifetime measurements





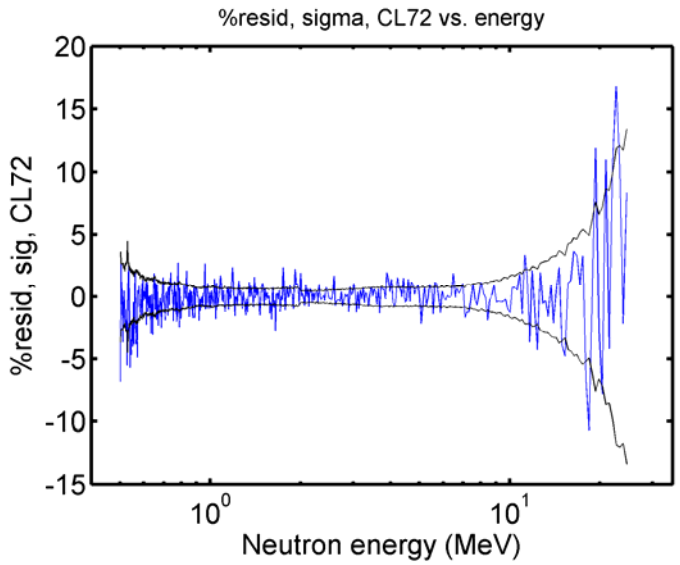
Robustness tests

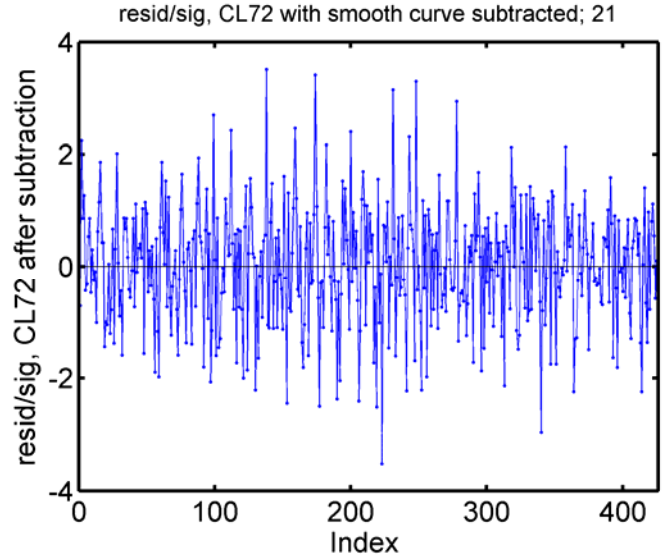
- How well does t-distr. analysis handle data from different distrs.?
- Analyze data using two likelihoods:
  a) t distr. with $\nu = 3$
  b) normal distr.
  - scale uncertainties by marginalizing over s distr.
  - results from 10,000 random trials
  - For each run, draw 20 data points from various t distributions
- Conclude
  - t distr. analysis well behaved
  - normal distr. analysis unstable when data have outliers



Statistical fluctuations: n-p scattering

- n-p cross sections measurements by Clement et al. (1972); 425 data pts.
- Compare to R-matrix calc. of G. Hale
- Upper graph shows residuals with specified statistical uncertainty
- Lower graph shows residuals (after subtracting running avg)/rms error
- $\chi^2 = 394/341$ points $= (1.07)^2$
- IQR = 1.55 (1.35 for normal)
- SOF = 1.2% (0.7% for normal)





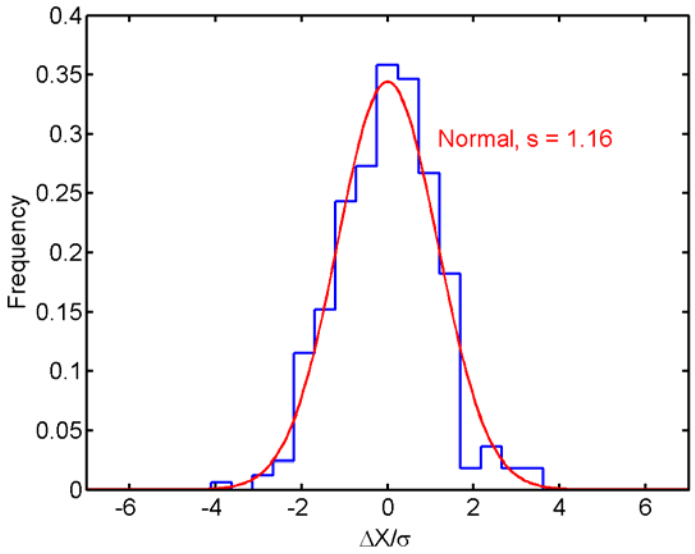
Statistical fluctuations: n-p scattering

- Histogram of Clement (1972) residuals, with running avg subtracted, normalized by statistical error
- Odds ratios of t distr. (t) to normal (N)

$$\frac{p(\mathrm{t}\,|\,\mathbf{d}\,\boldsymbol{\sigma}\,I)}{p(\mathrm{N}\,|\,\mathbf{d}\,\boldsymbol{\sigma}\,I)} = \frac{p(\mathbf{d}\,|\,\nu{=}3\,\mathrm{t}\,\boldsymbol{\sigma}\,I)}{p(\mathbf{d}\,|\,\nu{=}\infty\,\mathrm{t}\,\boldsymbol{\sigma}\,I)}\,\frac{p(\mathrm{t}\,|\,I)}{p(\mathrm{N}\,|\,I)}$$

$$= 2.8\times10^{-237}/8.7\times10^{-234} = 3.2\times10^{-4}$$

  - assuming prior ratio on models = 1 and priors on parameters equal (~5)
  - evidence is integral over τ and s
  - priors on τ and s = constant
- **Normal distr. is strongly preferred by data to t distr.**



Discussion

- Long-tailed likelihood functions
  - may result in posterior with multiple maxima
  - posterior mean is best estimator, but can be computationally costly
- Overall uncertainty may contain components that separately follow normal (or Poisson) and t distributions
  - likelihood is convolution of normal and t distrs.
  - can not be represented analytically
  - numerical computation of likelihood feasible
- Some outlier models are based on mixtures (good data – bad data)
  - likelihood is mixture of normal and t distributions:
    $(1 - \beta)\ \mathrm{N} + \beta\ \mathrm{S}$
    where N is normal and S is Student t distr.
  - assumes data follow either S (with prob. β) or N (with prob. $1 - \beta$)

Summary

- Variations in particle-lifetime data matched by t distribution with $\nu \approx 2.6$ to 3.0, not by normal distr.
- Likelihood or Bayesian analysis based on using Student t distribution
  - copes with outliers
  - treats each datum in same way – no need to identify outliers
  - produces stable results when outliers exist in data sets, whereas normal distr. does not
  - does not degrade results when outliers are not present
- These results for particle lifetimes do not represent all physical measurements, but are worth keeping in mind
- Repeat experiments are worthwhile to gain confidence and mitigate against outliers

Bibliography

- "A further look at robustness via Bayes;s theorem," G.E.P. Box and G.C. Tiao, *Biometrica* **49**, pp. 419-432 (1962)
- "On outlier rejection phenomena in Bayes inference," A. O'Hagan, *J. Roy. Statist. Soc.* **B 41**, 358–367 (1979)
- "Bayesian evaluation of discrepant experimental data," F.H. Fröhner, *Maximum Entropy and Bayesian Methods*, pp. 467–474 (Kluwer Academic, Dordrecht, 1989)
- "Estimators for the Cauchy distribution," K.M. Hanson and D.R. Wolf, *Maximum Entropy and Bayesian Methods*, pp. 157-164 (Kluwer Academic, Dordrecht, 1993)
- "Dealing with duff data," D. Sivia, *Maximum Entropy and Bayesian Methods*, pp. 157-164 (1996)
- "Understanding data better with Bayesian and global statistical methods," W.H. Press, *Unsolved Problems in Astrophysics*, pp. 49-60 (1997)
- "Outlier-tolerant parameter estimation," V. Dose and W. von der Linden, *Maximum Entropy and Bayesian Methods*, pp. 157-164 (AIP, 2000)
- "Lessons about likelihood functions from nuclear physics," K.M. Hanson, to appear in *Maximum Entropy and Bayesian Methods* (AIP, 2007)

This presentation available at http://www.lanl.gov/home/kmh/