

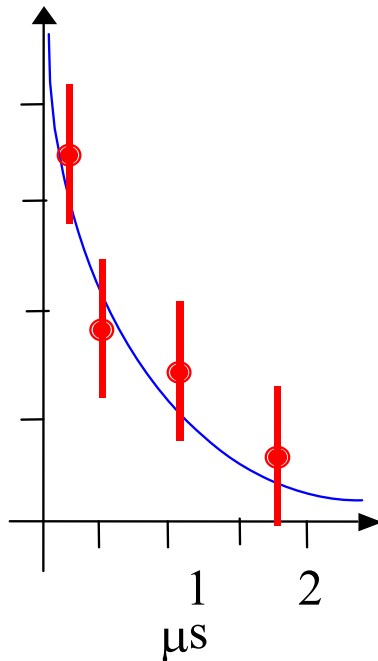
# Refresher on data analysis

Based on material from  
Profs. Becker and Chuang,  
and from Bevington and Robinson

# How good is the data?

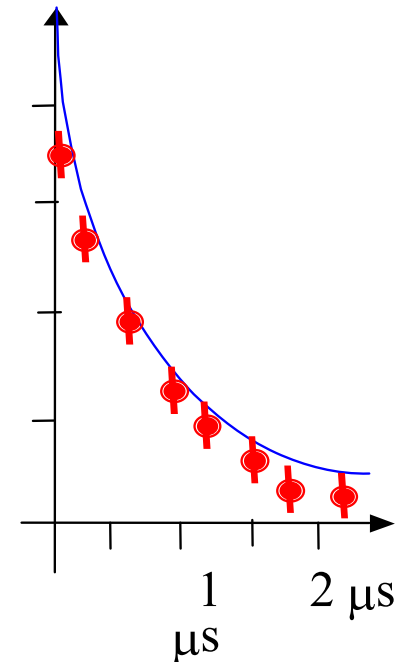
- Accuracy

- Closeness to the “truth”



- Precision

- Quality of the data (with no reference to the “true” value)
- Can be improved with more data (as  $1/\sqrt{N}$ )



# Types of Error

- Systematic errors
  - Reproducible deviation in the
    - Apparatus
    - Environment
    - Calibration
    - Fit hypothesis
  - Same from measurement to measurement
- Random errors
  - Due to finite statistics
  - Statistical fluctuations
    - Reduce by repeating measurements
    - $1/\sqrt{N}$
  - Independent measurements
    - Do not depend on previous operations or history

# Some measures of data

- Make  $N$  measurements of  $x$

- Mean of sample

$$\bar{x} \equiv \frac{1}{N} \sum_{i=1}^N x_i$$

- Median

- Half the observations are less than the median, and half are greater

- Mode

- Most probable value

- Standard deviation,  $s$

$$s^2 \equiv \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2$$

- Average of square of deviations from the mean value

- Variance,  $s^2$

# Significant figures

- Definition
  - Number of digits between leftmost nonzero digit and rightmost nonzero digit (or any rightmost digit if after decimal point)
- Examples
- Cardinal rule

Number	# of sig. figs.
1020	3
120.0	4
0120	2
0.0120	3

**Number of significant figures is never greater than precision**

No !  $\rightarrow 120.07 \pm 10 \rightarrow 120 \pm 10$   $\leftarrow$  Cool

# Distributions

# Distributions

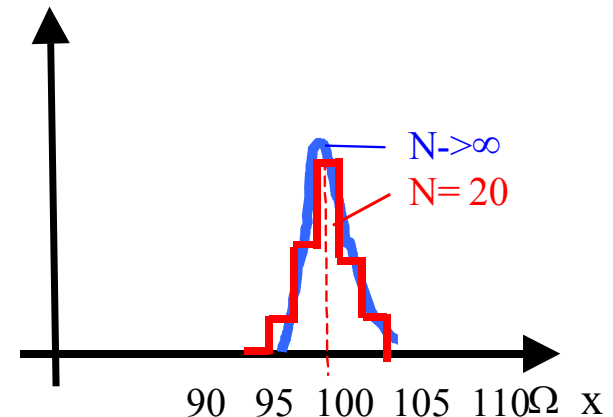
- Make  $N$  measurements  $\rightarrow$  sample distribution
- $N \rightarrow \infty \rightarrow$  parent distribution

	N measured samples	Parent distribution
Graph	histogram	smooth curve
Average $\rightarrow$ Mean	$\langle x \rangle = \frac{1}{N} \sum_{i=1}^N x_i$	$\mu = \lim_{N \rightarrow \infty} \left( \frac{1}{N} \sum_{i=1}^N x_i \right)$
Variance $\rightarrow$ standard deviation	$s_x^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \langle x \rangle)^2$	$\sigma_x^2 = \lim_{N \rightarrow \infty} \left( \frac{1}{N-1} \sum_{i=1}^N (x_i - \mu)^2 \right)$
Result $\pm$ error	$\langle x \rangle \pm s_x$	$\mu \pm 0$

If parent distribution is not known, then best guess is

$$\text{Mean} \pm \sqrt{\text{Variance}}$$

$$\bar{x} \pm s_x$$



# Some common distributions

## Binomial

$$P(x; n, p) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

$$\text{Mean} = \mu = np$$

$$\text{Variance} = \sigma = \sqrt{np(1-p)}$$

Probability to get 'yes'  $x$  times out of  $n$  tries if  $p$  is the Prob('yes') for a single try

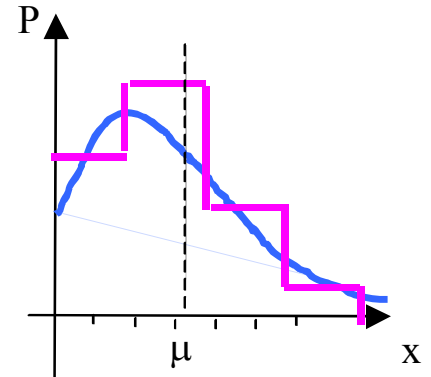
# Some common distributions

## Poisson

$$P(x; \mu) = \frac{\mu^x}{x!} \exp(-\mu)$$

$$\text{Mean} = \mu$$

$$\text{Variance} = \sqrt{\mu}$$



- Approx. to binomial distr. for  $p \ll 1$  and  $np = \mu$
- Useful for expts. with low counting rates

# Some common distributions

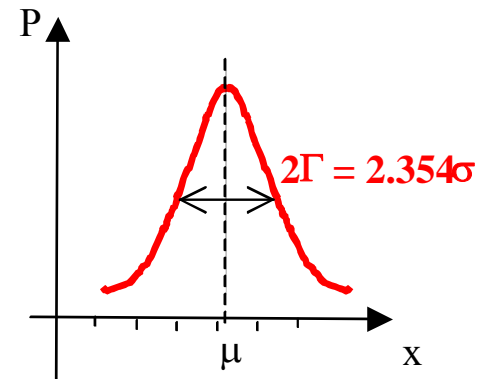
## Gaussian

$$P(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

$$\text{Mean} = \mu$$

$$\text{Variance} = \sigma^2 = \sqrt{\mu}$$

- Approx. to Poisson distr. for  $np \gg 1$
- Univeral law of large numbers
- Two Gaussians gives another Gaussian  
→ errors add in quadrature



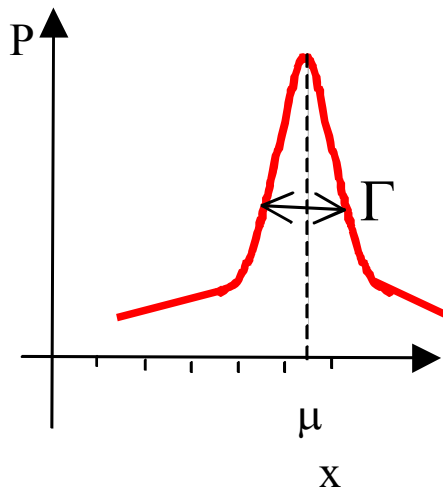
# Another common distribution

## Lorentzian (aka Breit-Wigner)

$$P(x; \mu, \Gamma) = \frac{1}{\pi} \frac{\frac{\Gamma}{2}}{(x - \mu)^2 + (\frac{\Gamma}{2})^2}$$

$$\text{Mean} = \mu$$

$$\text{Linewidth} = \Gamma$$



## Resonance

- Fourier transform of an exponentially decaying sinusoid is a Lorentzian
- Recall damped, driven harmonic oscillator from 8.03

$$x(t) = x_0 \exp(-\gamma t/2) \cos(\omega t)$$

↓

$$x^2(\Omega) \propto \frac{1}{(\omega_0^2 - \omega^2)^2 + (\omega\gamma)^2}$$

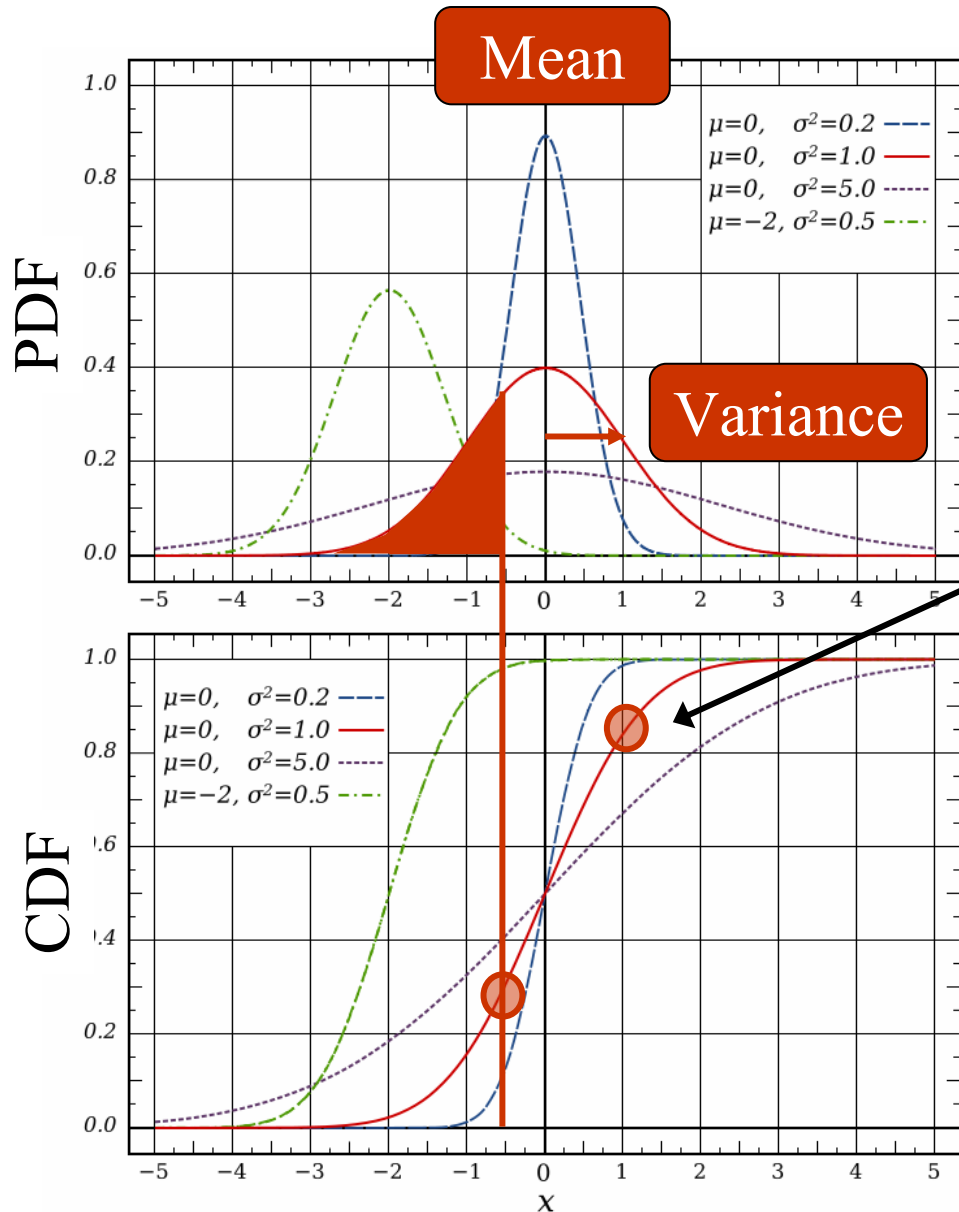
# Cumulative probability distribution

$$\begin{aligned} \text{Prob}[a < X < b] &= \int_a^b P(x) dx \\ CDF(x) &= \int_{-\infty}^x P(x') dx' \end{aligned}$$

Probability that random variable takes on a value less than or equal to  $x$

Probability that random variable is within infinitesimal interval  $[x, x + dx]$

# Gaussian PDF and CDF



68% + 16% = 84%

value of the *CDF*  
at  $x$  is area under  
the *PDF* up to  $x$

# Significance of the results

## Confidence level

- A *confidence interval* gives an estimated range of values which is likely to include the unknown population parameter
  - The estimated range is calculated from a given set of sample data
- Degree of confidence is related to the width of the confidence interval
- For Gaussian (normal) distribution the confidence interval is the area under the curve

$$\pm\sigma \rightarrow 68\%$$

$$\pm 2\sigma \rightarrow 95\%$$

Probability of observing a value outside of this area is  $< 0.05$

# Significance of the results

## Confidence level

- Best estimate of the mean  $\mu$  with uncertainty  $\sigma_\mu$  is  $\bar{x} \pm \sigma_\mu$
- But how successful were we in determining the parent parameters?

- Relate uncertainty to a Gaussian probability
- E.g. 68% of measurements in a Gaussian distribution fall within  $\pm 1\sigma$ , i.e. 68% of measurements of  $x$  would fall in the range
- $N$  repeated determinations of  $\bar{x}$  give 68% probability that the true value of the mean  $\mu$  lies in the range

$$(\bar{x} - s) < \bar{x} < (\bar{x} + s)$$

$$(\bar{x} - s_\mu) < \mu < (\bar{x} + s_\mu) \quad s_\mu = \frac{s}{\sqrt{N}}$$

- Typically results are significant at the 95% confidence level  
→ approx.  $\pm 2\sigma$  for Gaussian OR  $\pm 4\sigma$  for Student distribution

# Propagation of errors

# Error propagation

- Suppose we measure  $u = u \pm \sigma_u$  and  $v = v \pm \sigma_v$
- But we are interested in evaluating  $x = f(u, v)$   
e.g.  $x = au + bv$  or  $x = au/v$  or  $x = (uv)^p$
- What are the errors on  $x$  in each case?
- Taylor expand a general  $x = f(u, v)$   
to get variance of  $x$

$$\sigma_x^2 = \sigma_u^2 \left( \frac{\partial x}{\partial u} \right)^2 + \sigma_v^2 \left( \frac{\partial x}{\partial v} \right)^2 + 2\sigma_{uv}^2 \left( \frac{\partial x}{\partial u} \right) \left( \frac{\partial x}{\partial v} \right) + \dots$$

$$\sigma_{uv}^2 = \langle (u - \bar{u})(v - \bar{v}) \rangle$$

Covariance between  $u$  and  $v \rightarrow 0$   
if fluctuations in  $u$  and  $v$  are uncorrelated

# Error propagation

- Sums and differences → Absolute errors add

$$x = u \pm v \Rightarrow \frac{\partial x}{\partial u} = \frac{\partial x}{\partial v} = 1$$

$$\sigma_x^2 = \sigma_u^2 + \sigma_v^2 + 2\sigma_{uv}$$

- Products and quotients → Fractional errors add

$$x = uv \Rightarrow \frac{\partial x}{\partial u} = v, \frac{\partial x}{\partial v} = u$$

$$\frac{\sigma_x^2}{x^2} = \frac{\sigma_u^2}{u^2} + \frac{\sigma_v^2}{v^2} + 2\frac{\sigma_{uv}}{uv}$$

# Error propagation

- Powers  $\rightarrow$  fractional (relative) errors multiply by power

$$x = au^p \Rightarrow \frac{\partial x}{\partial u} = a p u^{p-1} = p \frac{x}{u}$$

$$\frac{\sigma_x}{x} = p \frac{\sigma_u}{u}$$

# Some cautionary reminders

- Make a series of  $N$  measurements  
Find the mean value of the quantity
- What is the error on the mean value?

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i \pm \frac{\sigma_x}{\sqrt{N}} \quad \text{where} \quad \sigma_x^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2$$

Not  $\sigma_x$ . Error improves with more measurements

- If errors on each measurement are different

$$\bar{x} = \frac{\sum_{i=1}^N x_i / \sigma_i^2}{\sum_{i=1}^N 1 / \sigma_i^2} \pm \sqrt{\frac{1}{\sum_{i=1}^N 1 / \sigma_i^2}}$$

Weighted mean  $\pm$  combined error

# Curve fitting

# Why fit curves to data?

- To find the functional form of a hypothesis that describes the data (with errors)
- Enables experimenter to
  - Extract physical parameters from the data
  - Test validity of the model or hypothesis
  - Interpolate/extrapolate data

# How to fit data?

- Vary parameters of the fitting (hypothesis) function to find a global goodness-of-fit criterion
- Goodness-of-fit criteria
  - Chi-squared
  - Maximum likelihood
  - Many others (e.g. Kolmogorov-Smirnov test)
- Chi-squared is the most commonly used criterion
  - Numerical minimization of chi-squared (using Levenberg-Marquadt method for Jr. Lab matlab routines)

# The *art* of fitting – the $\chi^2$ distribution

- We make  $N$  measurements of  $x_i$  each with random error  $\sigma_i$
- We have a theoretical hypothesis for the true values  $\mu_i$

■ Define

$$\chi^2 \equiv \sum_{i=1}^N \left( \frac{x_i - \mu_i}{\sigma_i} \right)^2$$

Measures spread in observed value  $x_i$

Measures expected spread

- This is a distribution itself, since the next random sample of measurements will have a different  $\chi^2$

# $\chi^2$ distribution

- Question: “How well do the data fit the theory (hypothesis)?”
- The  $\chi^2$  **distribution** gives the probability that the measurement has **this** particular value of  $\chi^2$
- Define

- Degrees of freedom =  $\nu = N - N_c$   
 $N_c$  = # of constraints or # of “tweakable” parameters

- Reduced chi-squared  $\chi_r^2 \equiv \chi^2 / \nu$

- Distribution 
$$P(\chi^2; \nu) \equiv \frac{(\chi^2)^{(\nu/2-1)} e^{-\chi^2/2}}{2^{(\nu/2)} \Gamma\left(\frac{\nu}{2}\right)}$$

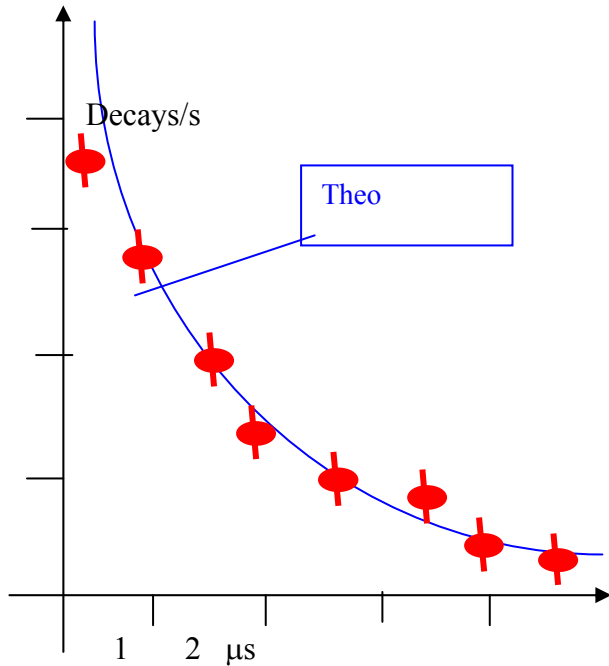
# Interpreting $\chi^2$ distribution

- Even if estimate of errors ( $\sigma_i$ ) is correct, and the choice of model is correct
  - $\chi_r^2$  will fluctuate between  $\chi_r^2 < 1$  and  $\chi_r^2 > 1$
  - The shape of the  $\chi_r^2$  distribution depends only on  $\nu$
  - The distribution allows us to calculate the probability that data and the model agree
- Chi-squared probability
  - Percentage of all measurements that you would expect to have worse  $\chi^2$  than you see

# Interpreting $\chi^2$ distribution

- Rules of thumb
  - If  $\chi_r^2 \sim 1 \rightarrow$  good fit
  - If  $\chi_r^2 \ll 1 \rightarrow$  bad model was bad (too many dof)
  - If  $\chi_r^2 \gg 1 \rightarrow$  bad fit
  - Vary parameters to change  $\chi_r^2 \rightarrow$  significant if changes  $\chi_r^2$  by 1/2

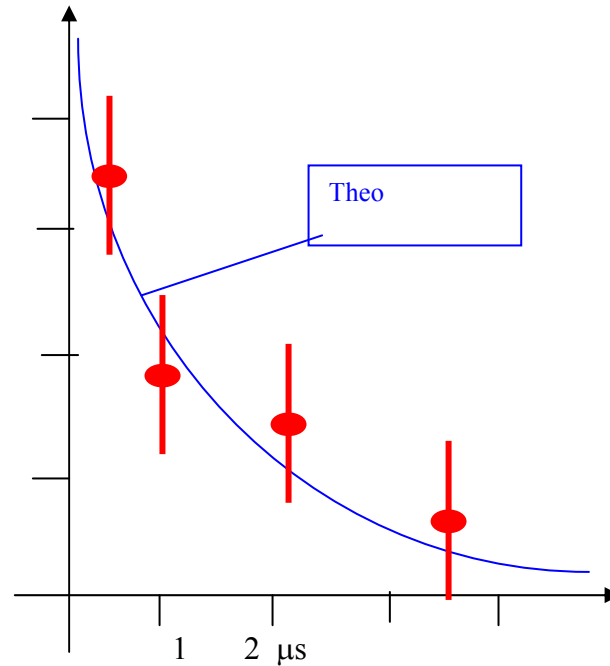
# Throwing away data?



$\chi^2_{\nu} = 18.1$  for  $\nu = 8 - 2$ ,  
Hypothesis (exponential) or data doubtful.  
Notice: the first point is  $>4\sigma$  off the curve  
 $\sim 16$  to  $\chi^2$ . Measurement was started too  
late, delete (since explained).

$\rightarrow \tau(\mu) = 2.05 \pm 0.03 \mu\text{s}$

Good 😊



$\chi^2_{\nu} = 0.3$  for  $\nu = 4 - 2$ ,  
which is an okay fit, but the data  
isn't very good at all (e.g. intervals  
too large).

$\rightarrow \tau(\mu) = 2.3 \pm 0.6 \mu\text{s}$

Not accurate 😞