## Basic statistics, and applications to X-ray spectral fitting

$\checkmark$ Normal error (Gaussian) distribution
$\rightarrow$ most important in statistical analysis of data, describes the distribution of random observations for many experiments
$\checkmark$ Poisson distribution
$\rightarrow$ generally appropriate for counting experiments related to random processes (e.g., radioactive decay of elementary particles)
$\checkmark$ Statistical tests: $\chi^{2}$ and F-test
$\checkmark$ Additional specific applications within XSPEC in the X-ray spectral analysis tutorial

## All measurements should be provided with errors

- Measurement $X \pm \boldsymbol{\delta} \mathbf{X}$ (units of measure)

$$
\downarrow
$$

Error associated with the measurement $X$

- Significant digits:
g (gravitational acceleration of an object in a vacuum near the Earth surface)= $=9.82 \pm .02385 \mathrm{~m} / \mathrm{s}^{2} \rightarrow 9.82 \pm 0.02 \mathrm{~m} / \mathrm{s}^{2}$
Another example: $v=100.2 \pm 30 \mathrm{~m} / \mathrm{s} \rightarrow 100 \pm 30 \mathrm{~m} / \mathrm{s}$
- Relative (fractionary) uncertainty: $\boldsymbol{\delta X} / \mathrm{X}$


## The Gaussian (normal error) distribution. I

Averages of random variables (sufficiently large in number ) independently drawn from independent distributions converge in distribution to the normal
Casual errors are above and below the "true" (most "common") value
$\rightarrow$ bell-shape distribution if systematic errors are negligible


## The Geussian probability function. II

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

$$
\pi
$$

Probability Density Function (centered on $\mu$ )
$\boldsymbol{\mu}=$ mean (expectation) value $\boldsymbol{\sigma}=$ standard deviation $\boldsymbol{\sigma}^{\mathbf{2}=\text { variance }}$


$$
e^{-x^{2} / 2 \sigma^{2}}
$$

## function centered on 0

## The Gaussian probability function. III




Probability within +/- no



$$
F(x)=\frac{1}{\sqrt{2 \pi}} \int_{-\infty}^{x} e^{-(x-\mu)^{2} / 2 \sigma^{2}} d x \quad \text { Cumulative Distribution Function }
$$

Value $\pm$ error at $1 \sigma$ confidence level: if we make a measurement N times, in $68.3 \%$ of the times we obtain such value. Every measurement should be reported and considered along its own error


## The Poisson distribution

Describes experimental results where events are counted and the uncertainty is not related to the measurement but reflects the intrinsically casual behavior of the process (e.g., radioactive decay of particles (Geiger counter), X-ray photons, etc.)

$$
\mathbf{P}(x)=e^{-\mu} \mu^{x} / x!\quad(x=0,1,2, \ldots)
$$

Probability of obtaining $x$ events when $\mu$ events are expected $\mathrm{x}=\mathrm{observed}$ number of events in a time interval (frequency of events)
average number of events

$$
\bar{x}=\sum_{x=0}^{\infty} x P(x)=\sum_{x=0}^{\infty} x e^{-\mu} \mu^{x} / x!=\mu
$$

$\rightarrow \mu=$ average number of expected events if the experiment is repeated many times

$$
\begin{aligned}
& \sigma^{2}=\left\langle(x-\mu)^{2}\right\rangle= \\
& =\sum_{x=0}^{\infty}(x-\mu)^{2} \frac{\mu^{x}}{x!} e^{-\mu}=\mu
\end{aligned}
$$

expectation value of the square of the deviations
the Poisson distribution with average counts $=\mu$ has standard deviation $V_{\mu}$

## Example: $\mathrm{N}_{\text {counts }} \pm \sqrt{ } \mathrm{N}$

High $\mu$ : the Poisson distribution is approximated by the Gaussian distribution

## defined by only one parameter $\mu$



## Statistical test: $x^{2}$

Test to compare the observed distribution of the results with that expected


It provides a measure on how much the data differ from the expectations (model), taking into account the errors associated with the measurement (e.g., datapoints)
$\mathrm{O}_{\mathrm{k}}=$ observed values (e.g., spectral datapoints)
$\mathrm{E}_{\mathrm{k}}=$ expected values (model, i.e. predicted distribution)
$\sigma_{\mathrm{k}}=$ error on the measured values (e.g., error on each spectral bin) $\mathrm{k}=$ number of datapoints (bins after rebinning)

$$
\chi^{2} / d o f \approx 1
$$

the observed and expected distributions are similar
dof=degrees of freedom = \#datapoints - \#free parameters


TABLE C. 4
$\chi^{2}$ distribution. Values of the reduced chi-square $\chi_{v}^{2}=\chi^{2} / v$ corresponding to the probability $P_{x}\left(\chi^{2} ; v\right)$ of exceeding $\chi^{2}$ versus the number of degrees of freedom $v, v=$ dof=\#datapoints - \#free parameters

| $P$ |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | 0.99 | 0.98 | 0.95 | 0.90 | 0.80 | 0.70 | 0.60 | 0.50 |
| 1 | 0.00016 | 0.00063 | 0.00393 | 0.0158 | 0.0642 | 0.148 | 0.275 | 0.455 |
| 2 | 0.0100 | 0.0202 | 0.0515 | 0.105 | 0.223 | 0.357 | 0.511 | 0.693 |
| 3 | 0.0383 | 0.0617 | 0.117 | 0.195 | 0.335 | 0.475 | 0.623 | 0.789 |
| 4 | 0.0742 | 0.107 | 0.178 | 0.266 | 0.412 | 0.549 | 0.688 | 0.839 |
| 5 | 0.111 | 0.150 | 0.229 | 0.322 | 0.469 | 0.600 | 0.731 | 0.870 |
| 6 | 0.145 | 0.189 | 0.273 | 0.367 | 0.512 | 0.638 | 0.762 | 0.891 |
| 7 | 0.177 | 0.223 | 0.310 | 0.405 | 0.546 | 0.667 | 0.785 | 0.907 |
| 8 | 0.206 | 0.254 | 0.342 | 0.436 | 0.574 | 0.691 | 0.803 | 0.918 |
| 9 | 0.232 | 0.281 | 0.369 | 0.463 | 0.598 | 0.710 | 0.817 | 0.927 |
| 10 | 0.256 | 0.306 | 0.394 | 0.487 | 0.618 | 0.727 | 0.830 | 0.934 |
| 11 | 0.278 | 0.328 | 0.416 | 0.507 | 0.635 | 0.741 | 0.840 | 0.940 |
| 12 | 0.298 | 0.348 | 0.436 | 0.525 | 0.651 | 0.753 | 0.848 | 0.945 |
| 13 | 0.316 | 0.367 | 0.453 | 0.542 | 0.664 | 0.764 | 0.856 | 0.949 |
| 14 | 0.333 | 0.383 | 0.469 | 0.556 | 0.676 | 0.773 | 0.863 | 0.953 |
| 15 | 0.349 | 0.399 | 0.484 | 0.570 | 0.687 | 0.781 | 0.869 | 0.956 |

## Staitistical test: F-test

If two statistics following the $\chi^{2}$ distribution have been determined, the ratio of the reduced chi-squares is distributed according to the $F$ distribution

$$
P_{f}\left(f ; v_{1}, v_{2}\right)=\frac{\chi_{1}^{2} / v_{1}}{\chi_{2}^{2} / v_{2}}
$$

## $\propto \Delta \chi^{2} / k$

with $\mathrm{k}=$ number of additional terms (parameters)

Example: Use the F-test to evaluate the improvement to a spectral fit due to the assumption of a different model, with additional terms
Conditions: (a) the simpler model is nested within the more complex model;
(b) the extra parameters have Gaussian distribution (not truncated by the parameter space boundaries)
$\rightarrow$ see the F-test tables for the corresponding probabilities (specific command in XSPEC)

## An application of the F-test within XSPEC





F-test probability in XSPEC: probability of exceeding F (see tabulated values)

## Fit (2) = Fit (1) + one component

xspec> ftest $X^{2}$ (best fit) dof (best fit) $X^{2}$ (previous fit) dof (previous fit)
xspec> ftest $90.810197 .2103 \rightarrow$ ftest=3.54 $\rightarrow$ prob=0.0328

$$
\begin{aligned}
& \quad F_{t}=\left(\frac{\chi^{2}(d o f)-\chi^{2}(d o f-k)}{d o f-(d o f-k)}\right) /\left(\chi^{2}(d o f-k) /(d o f-k)\right)= \\
& =\left(\Delta \chi^{2} / k\right) / \chi_{\nu}^{2} \\
& \text { Ex: } \chi^{2}(103)=97.23 \\
& \chi^{2}(101)=90.84 \\
& \rightarrow \Delta \chi^{2}=6.39, k=2 \rightarrow F_{t}=(6.39 / 2) /(90.84 / 101)=3.55
\end{aligned}
$$

$F_{\mathrm{t}}$ follows the F distribution with $\mathrm{v}_{1}=\mathrm{k}=\Delta(\mathrm{dof})$ and $\mathrm{v}_{2}=\operatorname{dof}-\mathrm{k}(-1)$
Search in the F-distribution tables for the probability of the null hypothesis $\left(\mathrm{H}_{0}\right)$ for $v_{1}=2$ and $v_{2} \sim 100$
The significance of the improvement is given by $P=1-p r o b=1-0.032=96.8 \%$ (i.e., not particularly significant)

Note of caution: F-test is an approximation (BUT quick); optimal solution would be running simulations (ses Protassov+2002)

You simulate N times (1000, 10000 trials) within XSPEC (command fakeit) data (source and background) of the same quality as that of your original data (including also response matrices ARF and RMF) and fit them with the same modeling without the line (e.g., a powerlaw); you then verify how many times your feature is found purely by chance

If you find it $X$ times, the significance of the line $=(1-X) /($ number of trials)

Percentage probability P within to: $P=\int_{X-t \sigma}^{X+t \sigma} G(x) d x$



