Basic statistics, and applications to X-ray spectral fitting

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

All measurements should be provided with errors

• Measurement $X \pm \delta X$ (units of measure)



Error associated with the measurement X

- Significant digits:
- g (gravitational acceleration of an object in a vacuum near the Earth surface)= $=9.82\pm0.02385$ m/s² \rightarrow 9.82 ± 0.02 m/s²

Another example: $v=100.2 \pm 30 \text{ m/s} \rightarrow 100 \pm 30 \text{ m/s}$

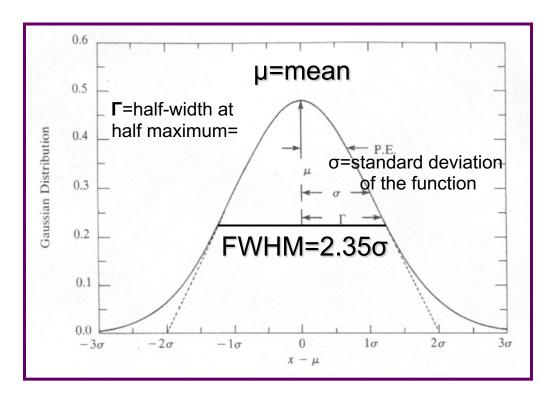
Relative (fractionary) uncertainty: δX/X

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

→ bell-shape distribution if systematic errors are negligible



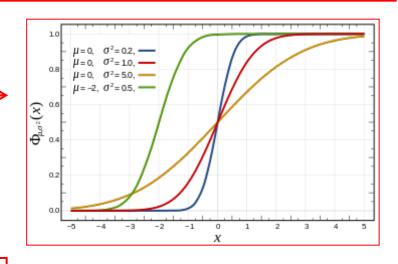
The Gaussian probability function. II

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

normalization factor, so that $\int f(x) dx = 1$

Probability Density Function (centered on μ)

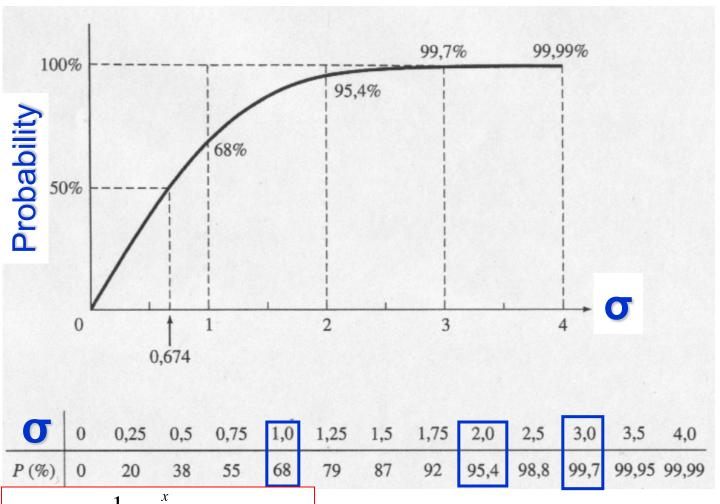
μ=mean (expectation) value
 σ=standard deviation
 σ²=variance



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

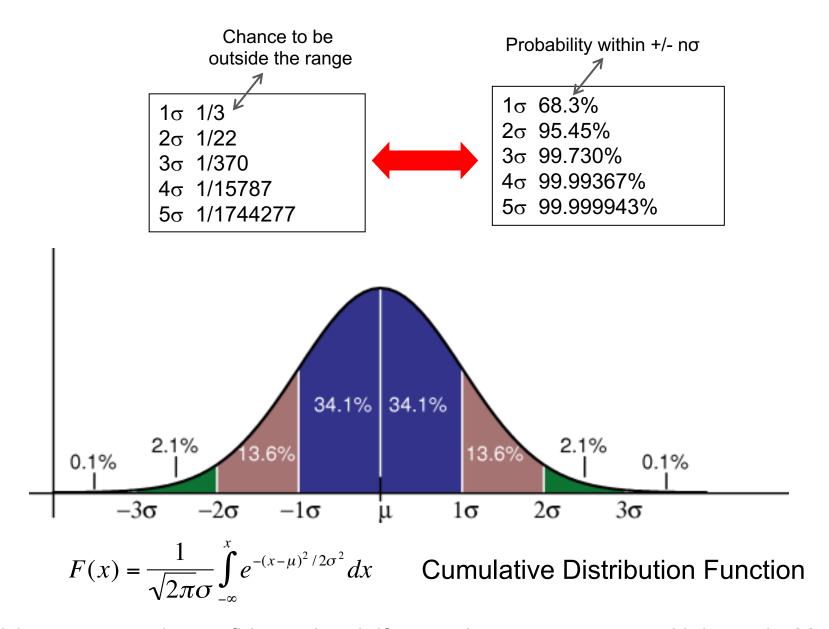
function centered on 0

The Gaussian probability function. III



 $F(x) = \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{x} e^{-(x-\mu)^{2}/2\sigma^{2}} dx$

Cumulative Distribution Function



Value±error at 1σ 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

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

						X - 1	σ	х	$X + t\sigma$	
ı	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00	0.80	1.60	2.39	3.19	3.99	4.78	5.58	6.38	7.17
0.1	7.97	8.76	9.55	10.34	11.13	11.92	12.71	13.50	14.28	15.07
0.2	15.85	16.63	17.41	18.19	18.97	19.74	20.51	21.28	22.05	22.82
0.3	23.58	24.34	25.10	25.86	26.61	27.37	28.12	28.86	29.61	30.35
0.4	31.08	31.82	32.55	33.28	34.01	34.73	35.45	36.16	36.88	37.59
0.5	38.29	38.99	39.69	40.39	41.08	41.77	42.45	43.13	43.81	44.48
0.6	45.15	45.81	46.47	47.13	47.78	48.43	49.07	49.71	50.35	50.98
0.7	51.61	52.23	52.85	53.46	54.07	54.67	55.27	55.87	56.46	57.05
0.8	57.63	58.21	58.78	59.35	59.91	60.47	61.02	61.57	62.11	62.65
0.9	63.19	63.72	64.24	64.76	65.28	65.79	66.29	66.80	67.29	67.78
1.0	68.27	68.75	69.23	69.70	70.17	70.63	71.09	71.54	71.99	72.43
1.1	72.87	73.30	73.73	74.15	74.57	74.99	75.40	75.80	76.20	76.60
1.2	76.99	77.37	77.75	78.13	78.50	78.87	79.23	79.59	79.95	80.29
1.3	80.64	80.98	81.32	81.65	81.98	82.30	82.62	82.93	83.24	83.55
1.4	83.85	84.15	84.44	84.73	85.01	85.29	85.57	85.84	86.11	86.38
1.5	86.64	86.90	87.15	87.40	87.64	87.89	88.12	88.36	88.59	88.82
1.6	89.04	89.26	89.48	89.69	89.90	90.11	90.31	90.51	90.70	90.90
1.7	91.09	91.27	91.46	91.64	91.81	91.99	92.16	92.33	92.49	92.65
1.8	92.81	92.97	93.12	93.28	93.42	93.57	93.71	93.85	93.99	94.12
1.9	94.26	94.39	94.51	94.64	94.76	94.88	95.00	95.12	95.23	95.34
2.0	95.45	95.56	95.66	95.76	95.86	95.96	96.06	96.15	96.25	96.34
2.1	96.43	96.51	96.60	96.68	96.76	96.84	96.92	97.00	97.07	97.15
2.2	97.22	97.29	97.36	97.43	97.49	97.56	97.62	97.68	97.74	97.80
2.3	97.86	97.91	97.97	98.02	98.07	98.12	98.17	98.22	98.27	98.32
2.4	98.36	98.40	98.45	98.49	98.53	98.57	98.61	98.65	98.69	98.72
2.5	98.76	98.79	98.83	98.86	98.89	98.92	98.95	98.98	99.01	99.04
2.6	99.07	99.09	99.12	99.15	99.17	99.20	99.22	99.24	99.26	99.29
2.7	99.31	99.33	99.35	99.37	99.39	99.40	99.42			99.47
2.8		99.50	99.52	99.53	99.55	99.56	99.58			99.61
2.9	99.63	99.64	99.65	99.66	99.67	99.68	99.69	99.70	99.71	99.72
						>	00 -	700/	. 2	400

 3σ=99.73%: in 1000 experiments you can get results outside this ±3σ range three times

 5σ =99.99994%: 6 cases out of 10⁶

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.)

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

Probability of obtaining x events when μ events are expected x=observed number of events in a time interval (frequency of events)

average number of events

$$\frac{-}{x} = \sum_{x=0}^{\infty} xP(x) = \sum_{x=0}^{\infty} xe^{-\mu} \mu^{x} / x! = \mu$$

→ µ=average number of expected events if the experiment is repeated many times

$$\sigma^{2} = \langle (x - \mu)^{2} \rangle =$$

$$= \sum_{x=0}^{\infty} (x - \mu)^{2} \frac{\mu^{x}}{x!} e^{-\mu} = \mu$$

expectation value of the square of the deviations



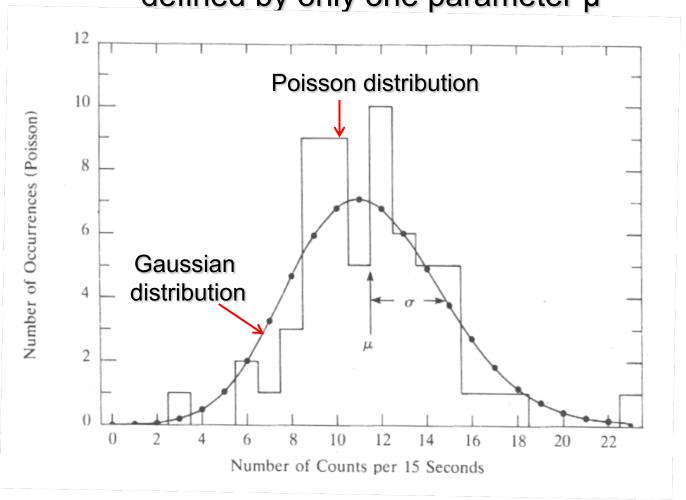
the Poisson distribution with average counts= μ has standard deviation $\sqrt{\mu}$



Example: N_{counts}±√N

High µ: the Poisson distribution is approximated by the Gaussian distribution

defined by only one parameter µ



Statistical test: χ^2

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

$$\chi^2 = \sum_{k=1}^n \frac{(O_k - E_k)^2}{\sigma_k^2}$$

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)

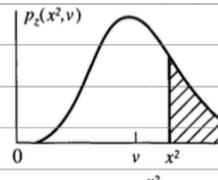
 O_k =observed values (e.g., spectral datapoints) E_k =expected values (model, i.e. predicted distribution) σ_k =error on the measured values (e.g., error on each spectral bin) k=number of datapoints (bins after rebinning)

$$\chi^2 / dof \approx 1$$

t t

the observed and expected distributions are similar

dof=degrees of freedom = #datapoints - #free parameters



This table gives the probability that a random sample of data, when compared to its parent distribution, would yield values of X²/v as large as (or larger than) the observed value

 x^2

TABLE C.4

 χ^2 distribution. Values of the reduced chi-square $\chi^2_{\nu} = \chi^2/\nu$ corresponding to the probability $P_{\chi}(\chi^2; \nu)$ of exceeding χ^2 versus the number of degrees of freedom v_v=dof=#datapoints - #free parameters

	P										
v	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			

Statistical test: F-test

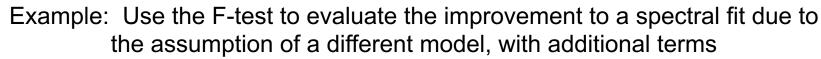
If two statistics following the χ^2 distribution have been determined, the ratio of the reduced chi-squares is distributed according to the F distribution

$$P_f(f;v_1,v_2) = \frac{\chi_1^2/v_1}{\chi_2^2/v_2}$$



with k=number of additional terms (parameters)





- 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)
 - → see the F-test tables for the corresponding probabilities (specific command in XSPEC)

An application of the F-test within XSPEC

```
Model phabs<1>*powerlaw<2> Source No.: 1
                                           Active/On
Model Model Component Parameter Unit
                                           Value
     comp
par
           phabs
                                  10^22
                                           1.59000E-02
                       nH
                                                         frozen
                                                                        counts
           powerlaw
                       PhoIndex
                                           2.72811
           powerlaw
                                           1.51490E-04
                       norm
                                                                           0.01
                                                  Model1
  Using energies from responses.
                                                                                    powerlaw model
Chi-Squared =
                       97.23 using 105 PHA bins.
Reduced chi-squared =
                               0.9440 for
                                             103 degrees of freedom
Null hypothesis probability = 6.417127e-01
Model phabs<1>(laor<2> + powerlaw<3>) Source No.: 1
                                                      Active/On
Model Model Component Parameter Unit
                                           Value
par
     comp
                                                                                                 Energy (keV)
           phabs
                                                         frozen
                       nH
                                  10^22
                                           1.59000F-02
                       lineE
                                           5.23582
           laor
                                  keV
                                                         +/- 0.0
                       Index
           laor
                                           3.00000
                                                         frozen
       2
           laor
                       Rin(G)
                                           1.23500
                                                         frozen
                       Rout (G)
           laor
                                           400.000
                                                         frozen
                       Incl
           laor
                                  deg
                                           30.0000
                                                         frozen
       2
           laor
                                           6.83065E-06
                                                              0.0
                       norm
                       PhoIndex
           powerlaw
                                           2.77137
           powerlaw
                                           1.48123E-04
                       norm
                                                             0.0
                                                                           0.01
                                                                                            powerlaw
  Using energies from response Model1+extra component
                                                                                        +iron line model
                                                                           10^{-3}
Chi-Squared =
                        90.84 using 105 PHA bins.
Reduced chi-squared =
                               0.8994 for
                                             101 degrees of freedom
Null hypothesis probability = 7.557789e-01
Current data and model not fit | www. F value ⇒ low significance
SEECLO ftest 90 84 101 97 2 102 of the added component
                                                                                   0.5
  statistic value = 3.53567 and probability 0.0327981
                                                                                                  Energy (keV)
```

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

xspec> ftest χ^2 (best fit) dof (best fit) χ^2 (previous fit) dof (previous fit)

xspec> ftest 90.8 101 97.2 103 → ftest=3.54 → prob=0.0328

$$F_t = (\frac{\chi^2(dof) - \chi^2(dof - k)}{dof - (dof - k)}) / (\chi^2(dof - k)/(dof - k)) =$$

$$= (\Delta \chi^2/k) / \chi_{\nu}^2$$
Ex: $\chi^2(103) = 97.23$

$$\chi^2(101) = 90.84$$

$$\to \Delta \chi^2 = 6.39, k = 2 \to F_t = (6.39/2)/(90.84/101) = 3.55$$

 F_t follows the F distribution with $v_1=k=\Delta(dof)$ and $v_2=dof-k(-1)$

Search in the F-distribution tables for the probability of the null hypothesis (H_0) for v_1 =2 and v_2 ~100

The significance of the improvement is given by P=1-prob=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:
$$P = \int_{x-t\sigma}^{x+t\sigma} G(x) dx$$

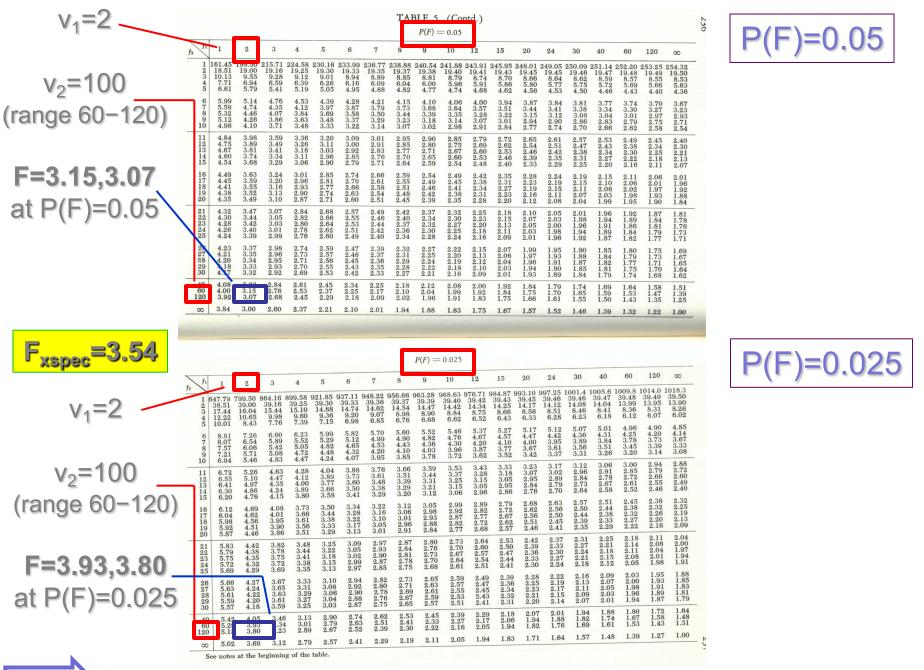
					_					
						X-to	Γ	Х	X+tσ	
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
	0.00	0.80	1.60	2.39	3.19	3.99	4.78	5.58	6.38	7.17
	7.97	8.76	9.55	10.34	11.13	11.92	12.71	13.50	14.28	15.07
	15.85	16.63	17.41	18.19	18.97	19.74	20.51	21.28	22.05	22.82
	23.58	24.34	25.10	25.86	26.61	27.37	28.12	28.86	29.61	30.35
	31.08	31.82	32.55	33.28	34.01	34.73	35.45	36.16	36.88	37.59
	38.29	38.99	39.69	40.39	41.08	41.77	42.45	43.13	43.81	44.48
	45.15	45.81	46.47	47.13	47.78	48.43	49.07	49.71	50.35	50.98
	51.61	52.23	52.85	53.46	54.07	54.67	55.27	55.87	56.46	57.05
	57.63	58.21	58.78	59.35	59.91	60.47	61.02	61.57	62.11	62.65
	63.19	63.72	64.24	64.76	65.28	65.79	66.29	66.80	67.29	67.78
	68.27	68.75	69.23	69.70	70.17	70.63	71.09	71.54	71.99	72.43
	72.87	73.30	73.73	74.15	74.57	74.99	75.40	75.80	76.20	76.60
	76.99	77.37	77.75	78.13	78.50	78.87	79.23	79.59	79.95	80.29
	80.64	80.98	81.32	81.65	81.98	82.30	82.62	82.93	83.24	83.55
	83.85	84.15	84.44	84.73	85.01	85.29	85.57	85.84	86.11	86.38
	86.64	86.90	87.15	87.40	87.64	87.89	88.12	88.36	88.59	88.82
	89.04	89.26	89.48	89.69	89.90	90.11	90.31	90.51	90.70	90.90
	91.09	91.27	91.46	91.64	91.81	91.99	92.16	92.33	92.49	92.65
	92.81	92.97	93.12	93.28	93.42	93.57	93.71	93.85	93.99	94.12
	94.26	94.39	94.51	94.64	94.76	94.88	95.00	95.12	95.23	95.34
	95.45	95.56	95.66	95.76	95.86	95.96	96.06	96.15	96.25	96.34
	96.43	96.51	96.60	96.68	96.76	96.84	96.92	97.00	97.07	97.15
	97.22	97.29	97.36	97.43	97.49	97.56	97.62	97.68	97.74	97.80
	97.86	97.91	97.97	98.02	98.07	98.12	98.17	98.22	98.27	98.32
+	98.36	98.40	98.45	98.49	98.53	98.57	98.61	98.65	98.69	98.72
;	98.76	98.79	98.83	98.86	98.89	98.92	98.95	98.98	99.01	99.04
i	99.07	99.09	99.12	99.15	99.17	99.20	99.22	99.24	99.26	99.29
,	99.31	99.33	99.35	99.37	99.39	99.40	99.42	99.44	99.46	99.47
;	99.49	99.50	99.52	99.53	99.55	99.56	99.58	99.59	99.60	99.61
)	99.63	99.64	99.65	99.66	99.67	99.68	99.69	99.70	99.71	99.72
)	99.73									
;	99.95									
)	99.994									
;	99.9993									
)	99.99994									

shaded region

X between -tσ and +tσ

Compute the significance of the improvement in terms of σ given P=0.0328, hence (1-P)=0.968

P=96.8% **→**≈2.1σ



Probability intermediate intermediate between 0.05 and 0.025 (actually, **0.0323**)