Statistical analysis

Amy Lien
Goddard Space Flight Center

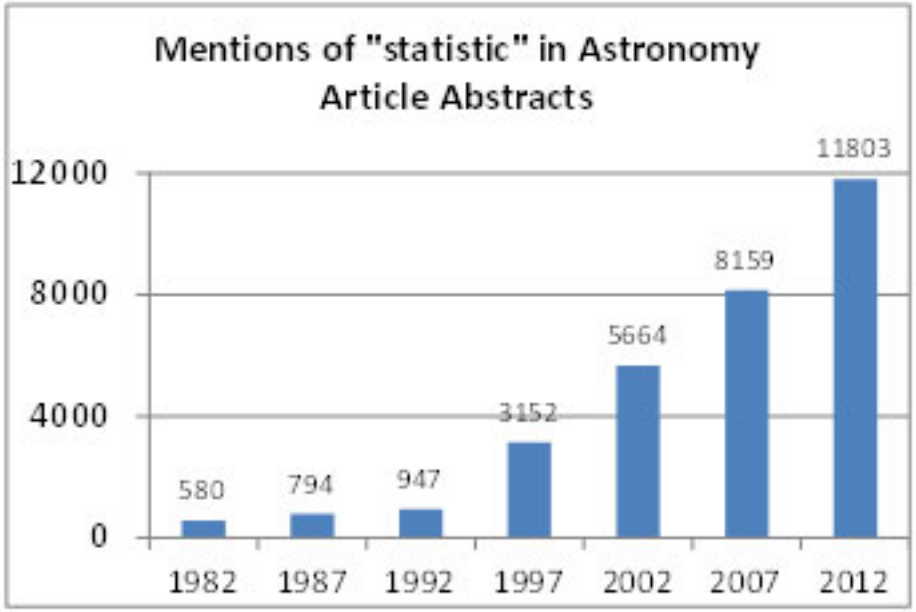
ASTR 288C, Lecture 8, 2017/10/23

Acknowledgement

Class material from:

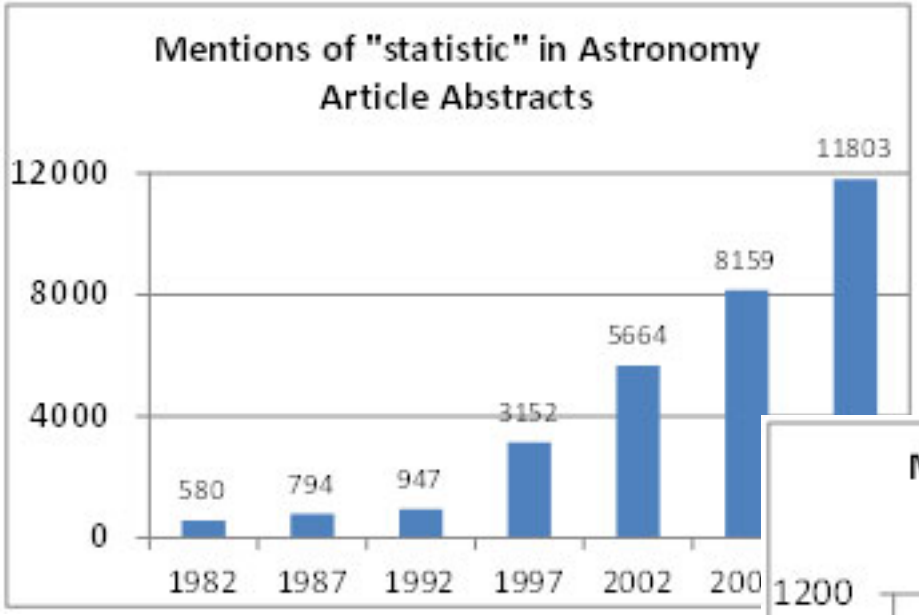
- **Slides from Prof. M.A. Thomson (Univ. of Cambridge)**
  http://www.hep.phy.cam.ac.uk/~thomson/lectures/lectures.html
- **Prof. Sylvain Guiriec (George Washington Univ.)**

Why statistics?



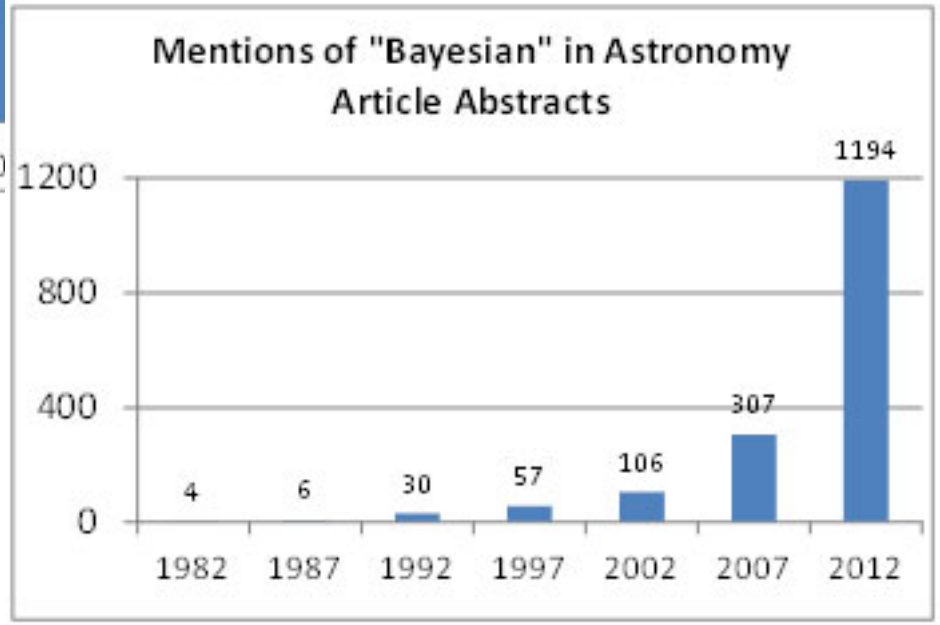
Pierson et al. (2013)

Why statistics?



Pierson et al. (2013)

Pierson et al. (2013)



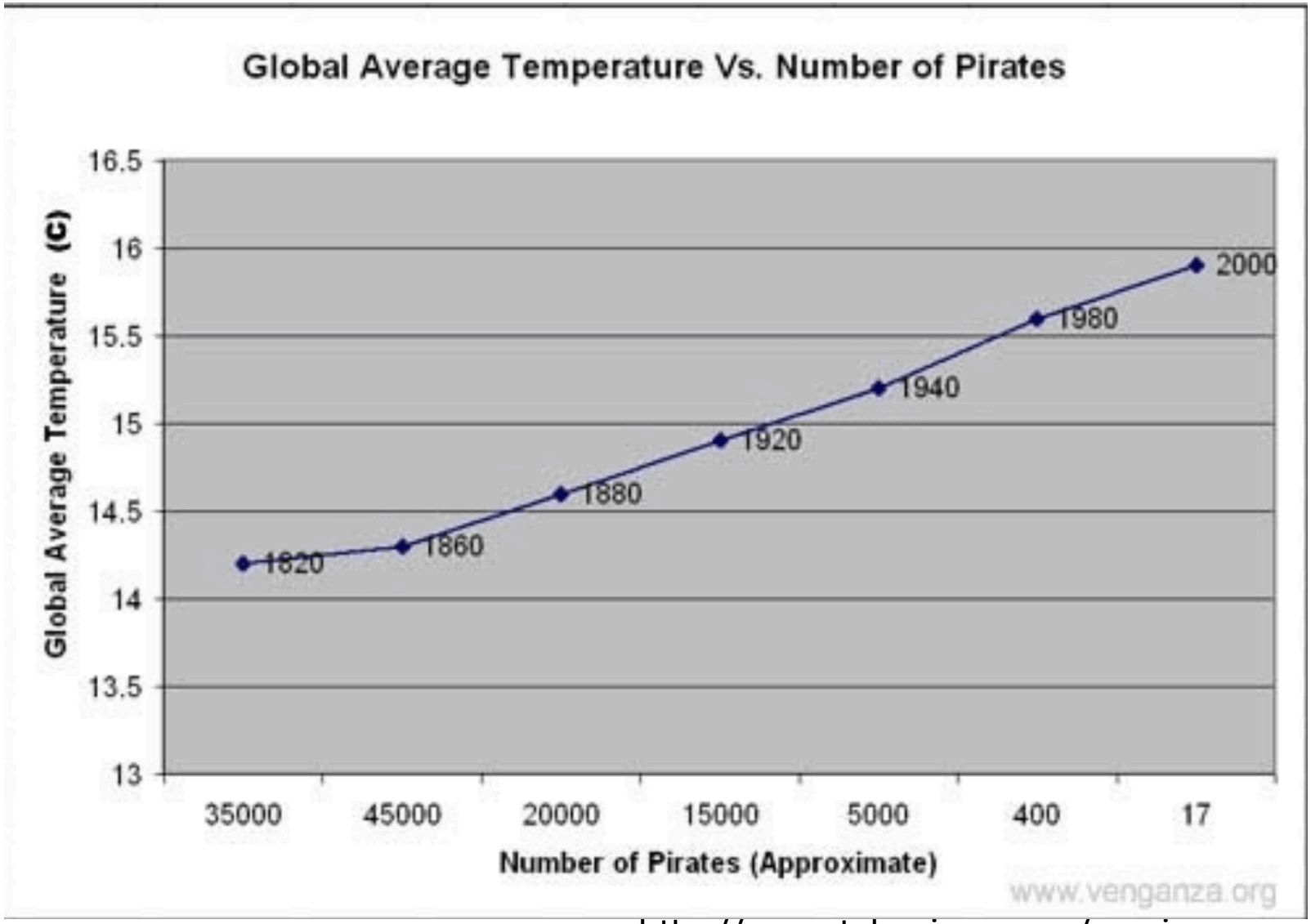
Most astronomers and physicists don't understand statistics

Suppose you derive a 90% confidence region on some parameter of interest. What does this mean ?

Does it tell you that there is a probability of 0.9 that the true value of the parameter lies in the range calculated ?

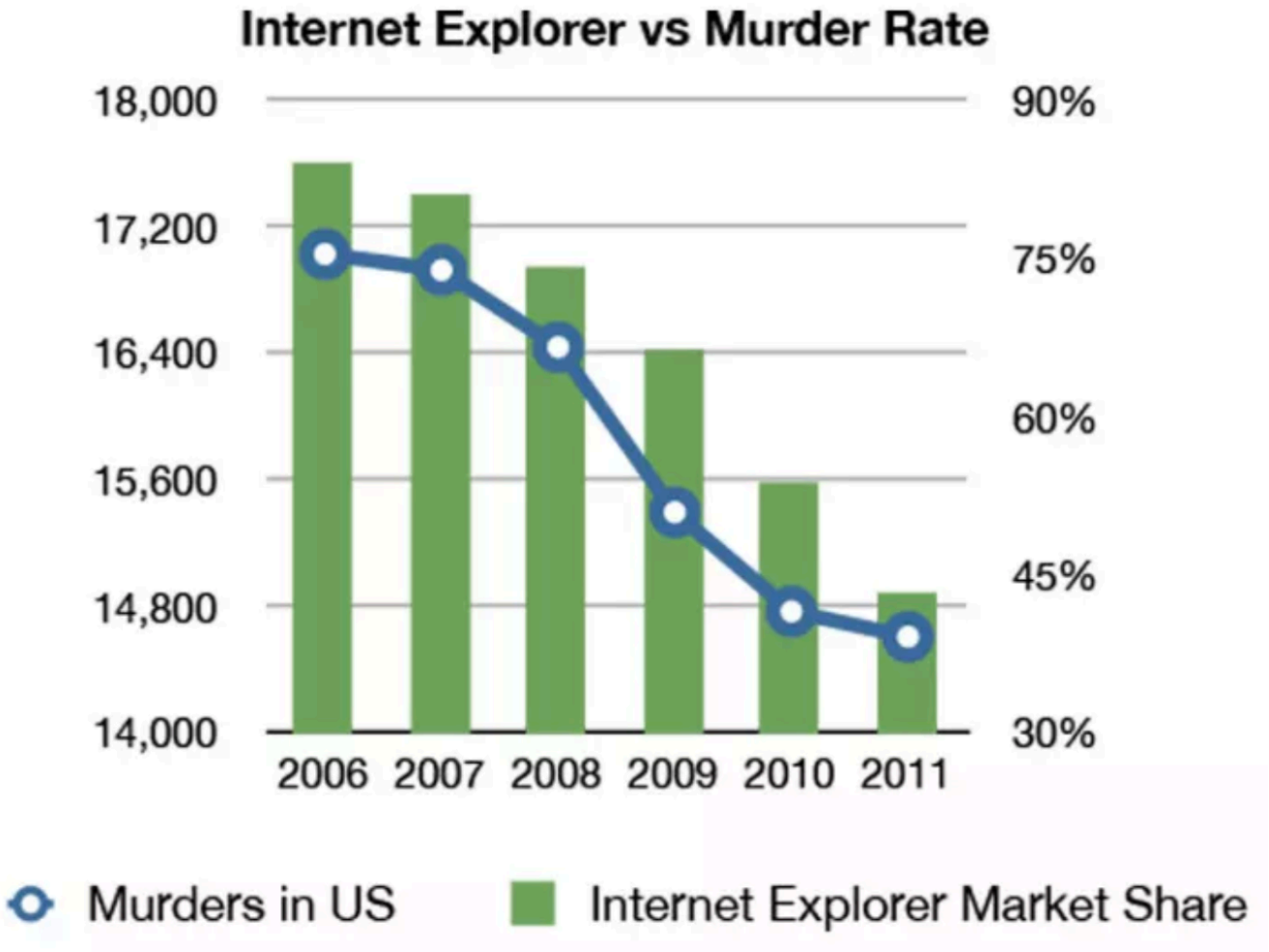
Slides from Keith Arnaud
Source: http://slideplayer.com/slide/9361524/

Be careful with statistics analysis



http://www.tylervigen.com/spurious-correlations

Be careful with statistics analysis



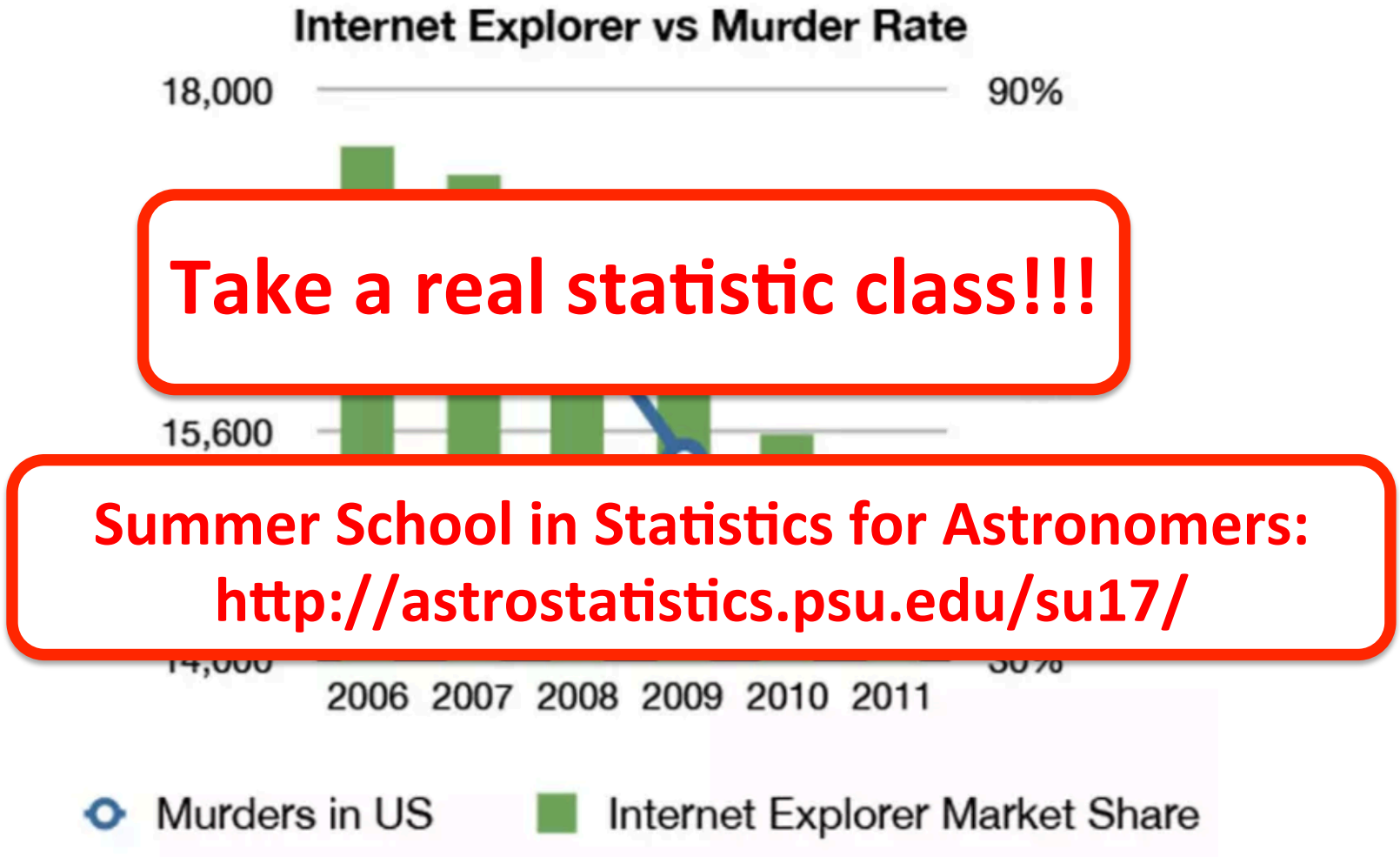
http://www.tylervigen.com/spurious-correlations

Be careful with statistics analysis

Take a real statistic class!!!

http://www.tylervigen.com/spurious-correlations

Be careful with statistics analysis



Take a real statistic class!!!

Summer School in Statistics for Astronomers: http://astrostatistics.psu.edu/su17/

http://www.tylervigen.com/spurious-correlations

Basic statistical terms

- Mean (average): $\mu \equiv \langle x \rangle = \int x P(x) \mathrm{d}x$
- Mean of squares: $\langle x^2 \rangle = \int x^2 P(x) \mathrm{d}x$
- Variance: $Var(x) \equiv \sigma^2 \equiv \langle (x-\mu)^2 \rangle = \int (x-\mu)^2 P(x) \mathrm{d}x$

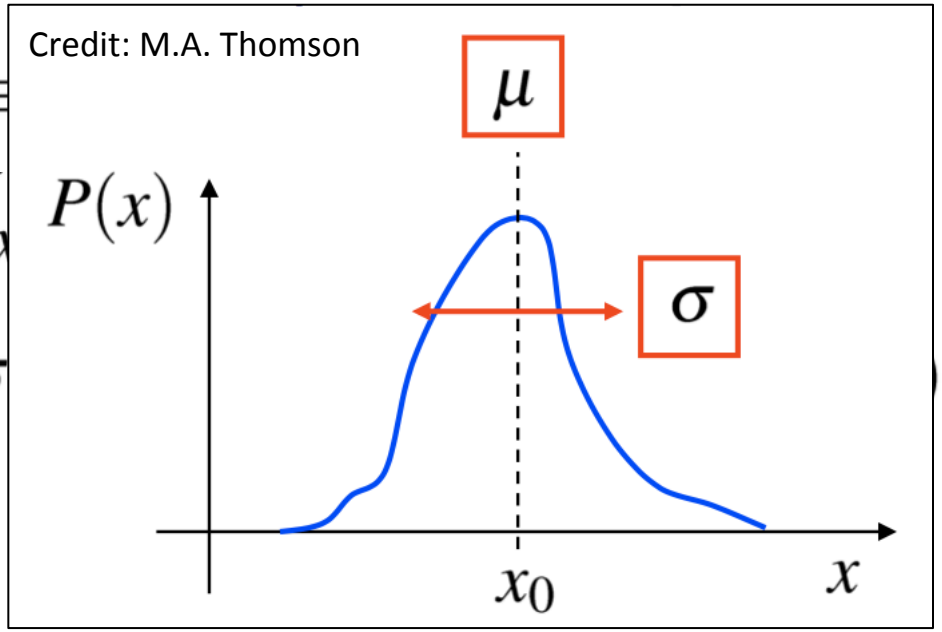
$$= \langle x^2 \rangle - \mu^2$$

- Standard deviation σ = sqrt(variance)
- Degrees of freedom

  = number of data − number of variables

Basic statistical term

- Mean (average): $\mu \equiv$
- Mean of squares: $\langle x$
- Variance: $Var(x) \equiv \sigma$ $\quad \ldots \, dx$



- Standard deviation σ = sqrt(variance)
- Degrees of freedom

  = number of data − number of variables

Poisson distribution

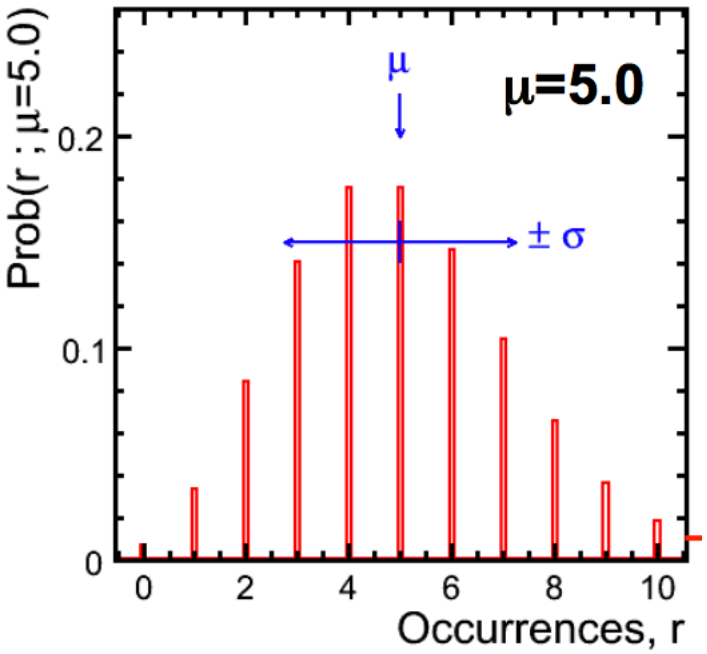
- Most of the high-energy instruments simply count photons → Poisson distribution
- Poisson distribution:

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

$$\langle x \rangle = \mu$$

$$\sigma^2 = \mu$$

μ: Average of the distribution
σ: Standard deviation



Poisson distribution

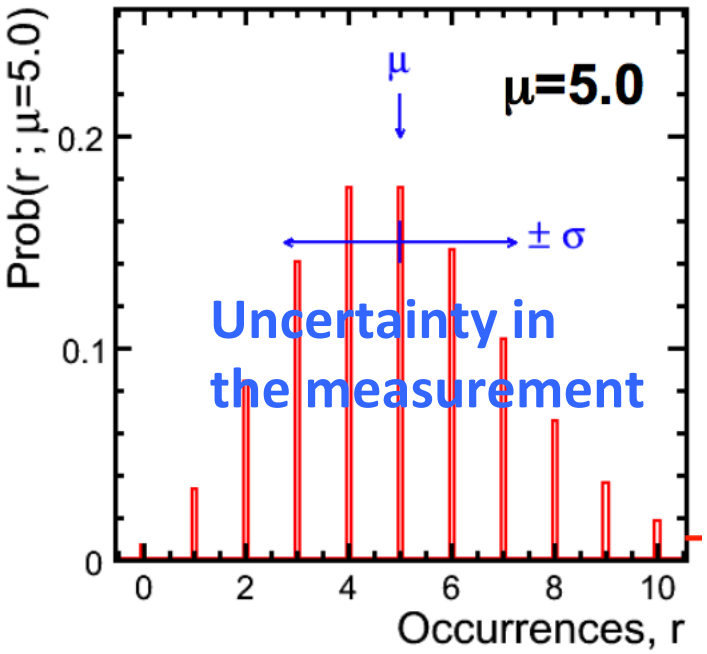
- Most of the high-energy instruments simply count photons → Poisson distribution
- Poisson distribution:

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

$$\langle x \rangle = \mu$$

$$\sigma^2 = \mu$$

μ: Average of the distribution
σ: Standard deviation



Gaussian distribution

- When the number of data (μ) is large,

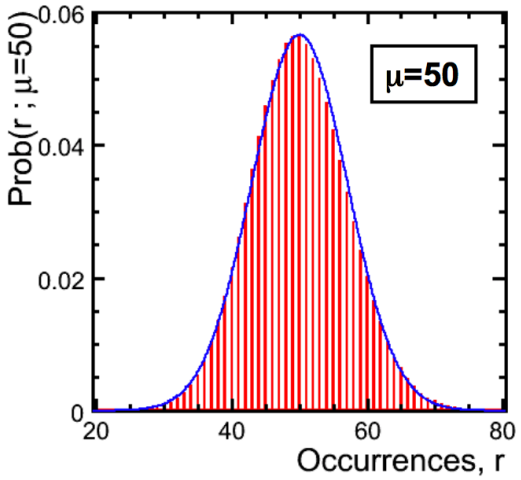
Poisson distribution → **Gaussian distribution**

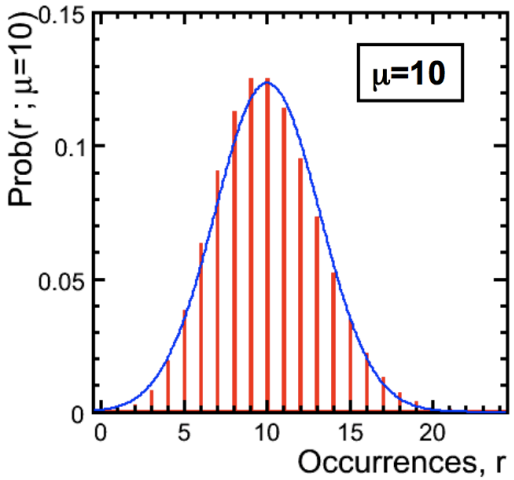
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\langle x \rangle = \mu$$

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

$$P(x;\mu) \approx k e^{-\frac{(x-\mu)^2}{2\mu}}$$





Gaussian distribution

- When the number of data (μ) is large,

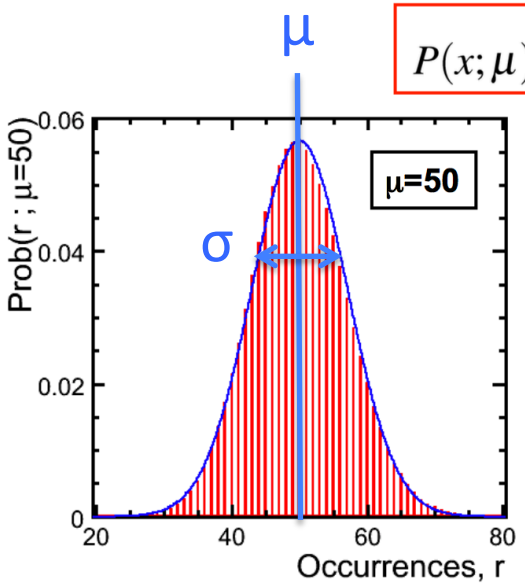
Poisson distribution → **Gaussian distribution**

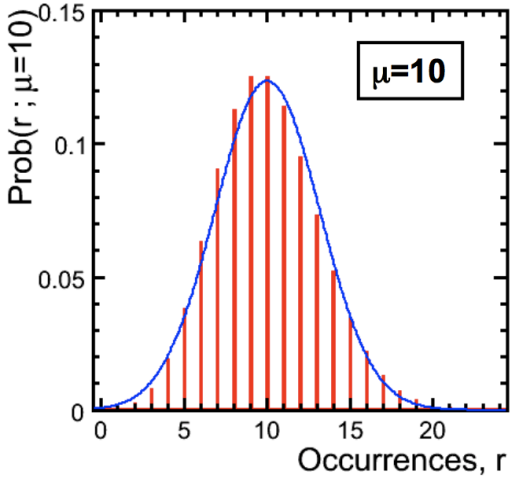
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\langle x \rangle = \mu$$

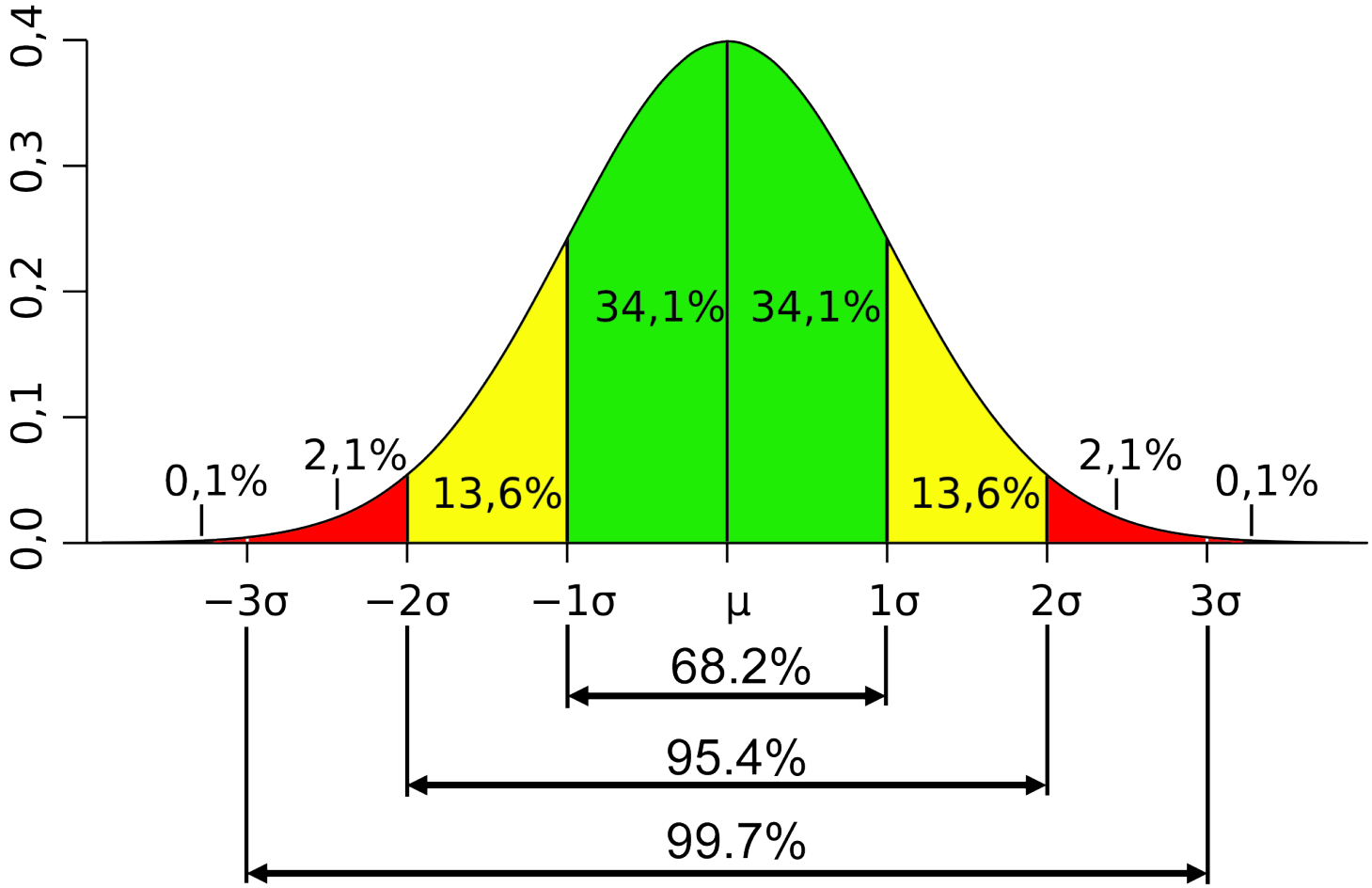
$$\langle (x-\mu)^2 \rangle = \sigma^2$$

$$P(x;\mu) \approx k e^{-\frac{(x-\mu)^2}{2\mu}}$$



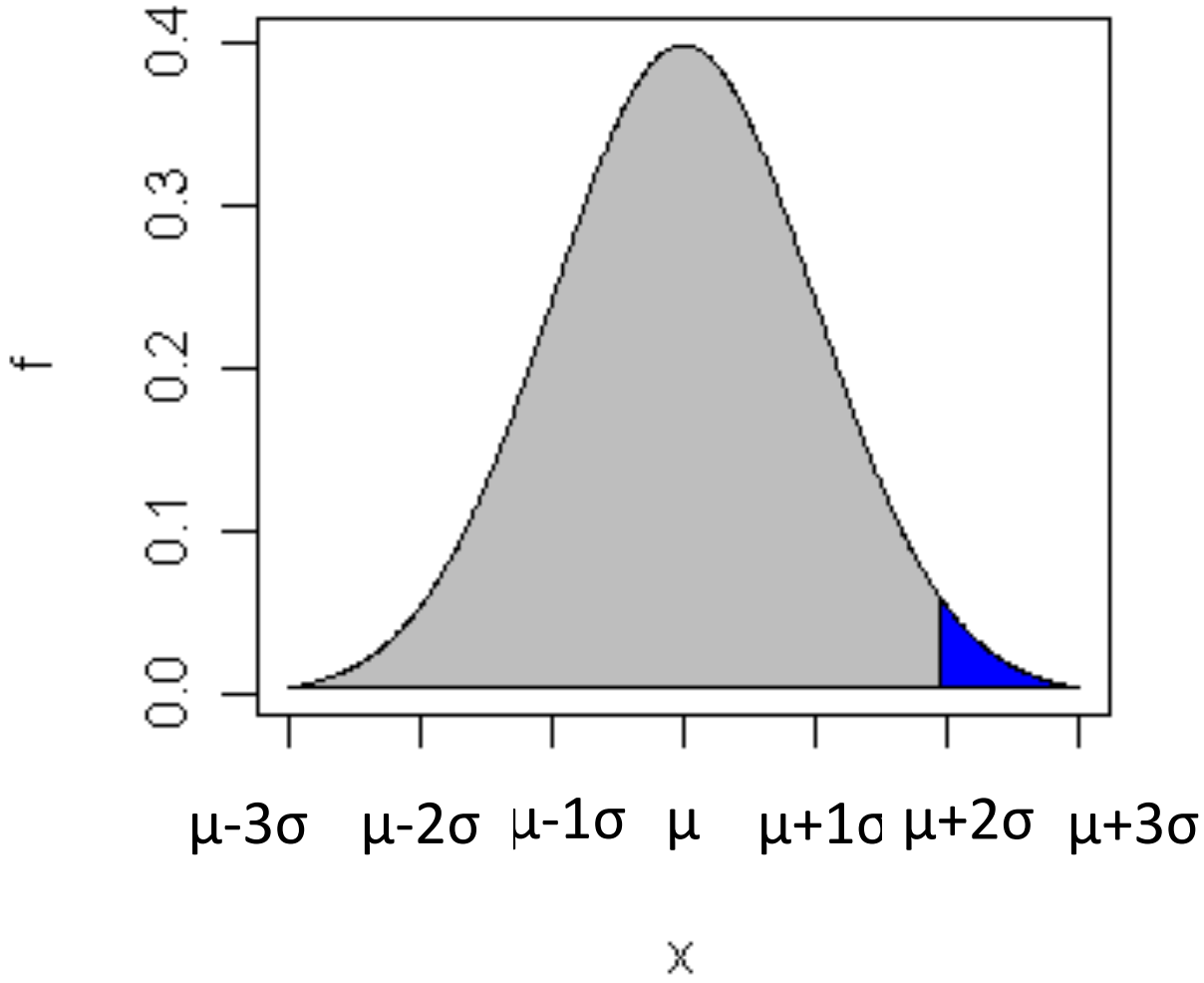


Detection significance vs false-detection rate



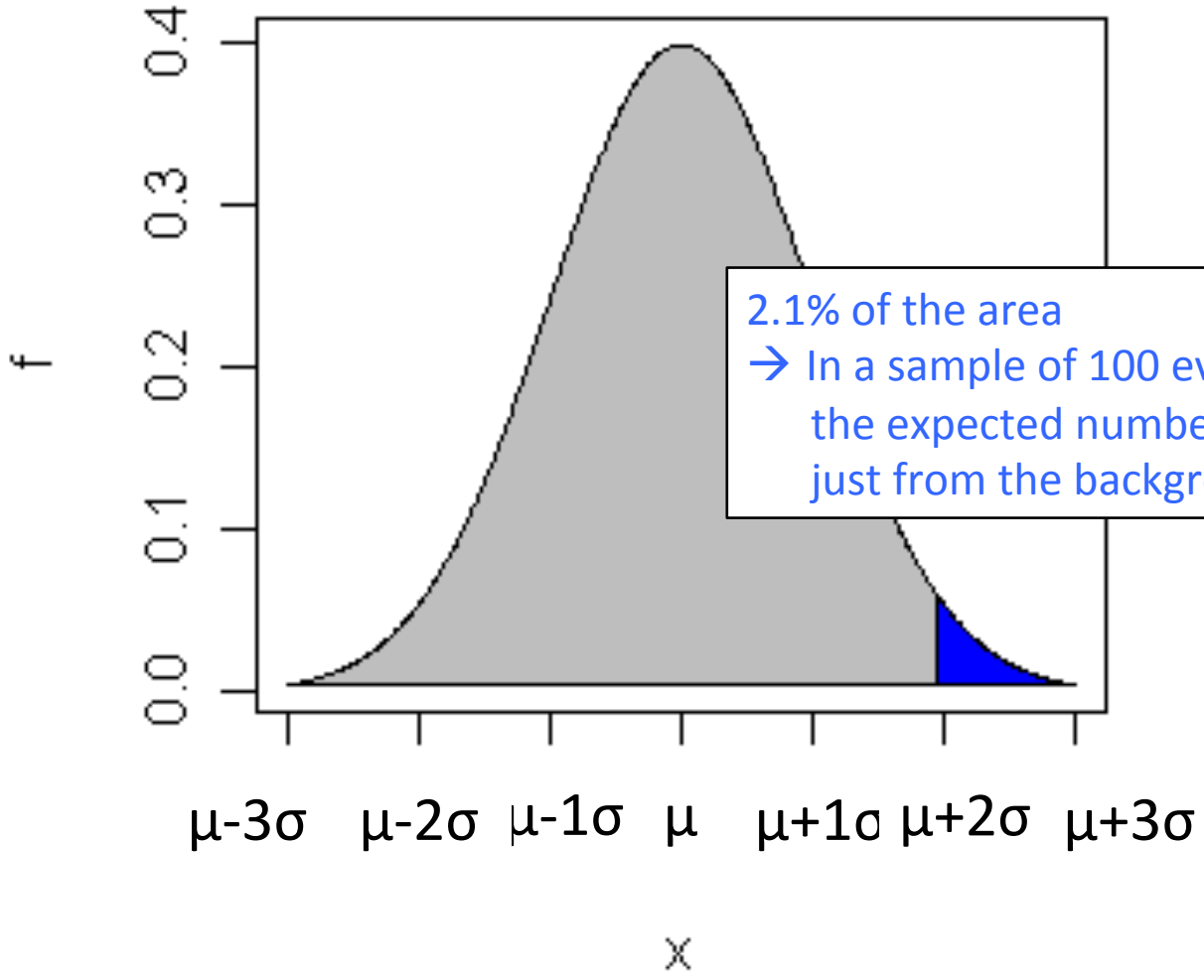
Source: https://kanbanize.com/blog/normal-gaussian-distribution-over-cycle-time/

Detection significance vs false-detection rate



Source: http://www.statsref.com/HTML/index.html?normal_distribution.html

Detection significance vs false-detection rate



2.1% of the area
→ In a sample of 100 events,
the expected number of 2σ detection
just from the background is 2.1 times.

Source: http://www.statsref.com/HTML/index.html?normal_distribution.html

Data fitting

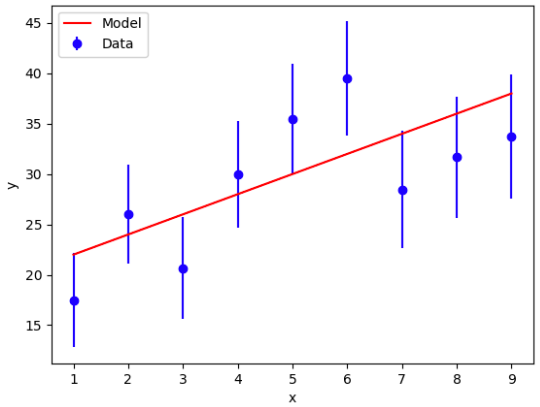
Maximum Likelihood as estimation for data with Gaussian distribution

The model parameters that have the best-chance to explain the observed data

- The likelihood function:

$$L(\theta) = \prod_{i=1}^{n} p(x_i; \theta)$$

The probability of measuring x_i, given a model θ



Data fitting

Maximum Likelihood as estimation for data with Gaussian distribution

The model parameters that have the best-chance to explain the observed data

- The likelihood function:

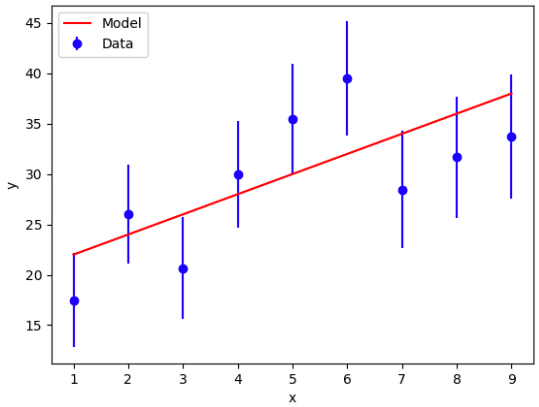
$$L(\theta) = \prod_{i=1}^{n} p(x_i; \theta)$$

The probability of measuring x_i, given a model θ

- If all your data have Gaussian distribution

$$L = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i}} e^{\frac{x_i - \mu_i}{2\sigma_i}}$$

$$ln(L) = -\frac{N}{2} ln(2\pi\sigma_i^2) - \sum_{i=1}^{N} \frac{(x_i - \mu_i)^2}{2\sigma_i^2}$$



Data fitting

Maximum Likelihood as estimation for data with Gaussian distribution

The model parameters that have the best-chance to explain the observed data

- The likelihood function:

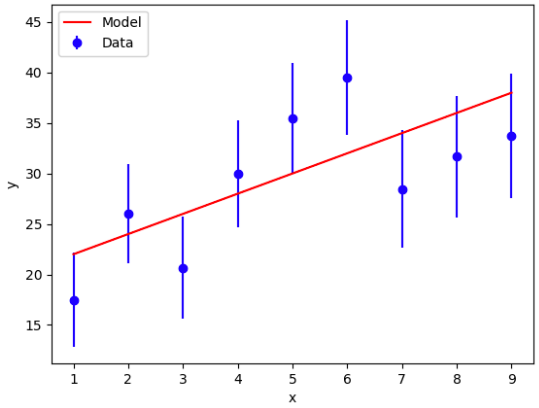
$$L(\theta) = \prod_{i=1}^{n} p(x_i; \theta)$$

The probability of measuring x_i, given a model θ

- If all your data have Gaussian distribution

Only related to data, not model
→ don't need for maximizing L when finding the right model

$$ln(L) = -\frac{N}{2} ln(2\pi\sigma_i^2) - \sum_{i=1}^{N} \frac{(x_i - \mu_i)^2}{2\sigma_i^2}$$



Data fitting

Maximum Likelihood as estimation for data with Gaussian distribution

The model parameters that have the best-chance to explain the observed data

- The likelihood function:

$$L(\theta) = \prod_{i=1}^{n} p(x_i; \theta)$$

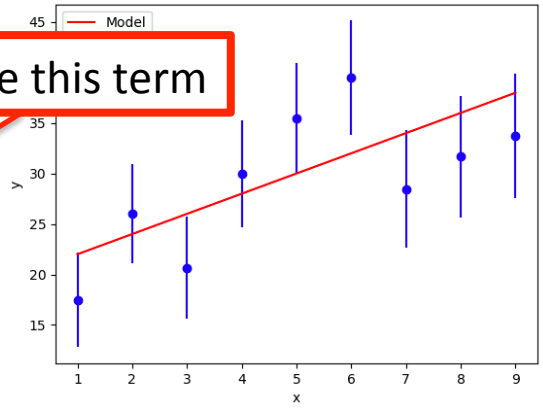
The probability of measuring x_i, given a model θ

- If all your data have Gaussian distribution

$$L = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i}} e^{\frac{x_i - \mu_i}{2\sigma_i}}$$

$$ln(L) = -\frac{N}{2} ln(2\pi\sigma_i^2) - \sum_{i=1}^{N} \frac{(x_i - \mu_i)^2}{2\sigma_i^2}$$

Maximize this term



Data fitting

Maximum Likelihood as estimation for data with Gaussian distribution

The model parameters that have the best-chance to explain the observed data

Maximize this term

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

Minimize this term

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

Data fitting

Maximum Likelihood as estimation for data with Gaussian distribution

The model parameters that have the best-chance to explain the observed data

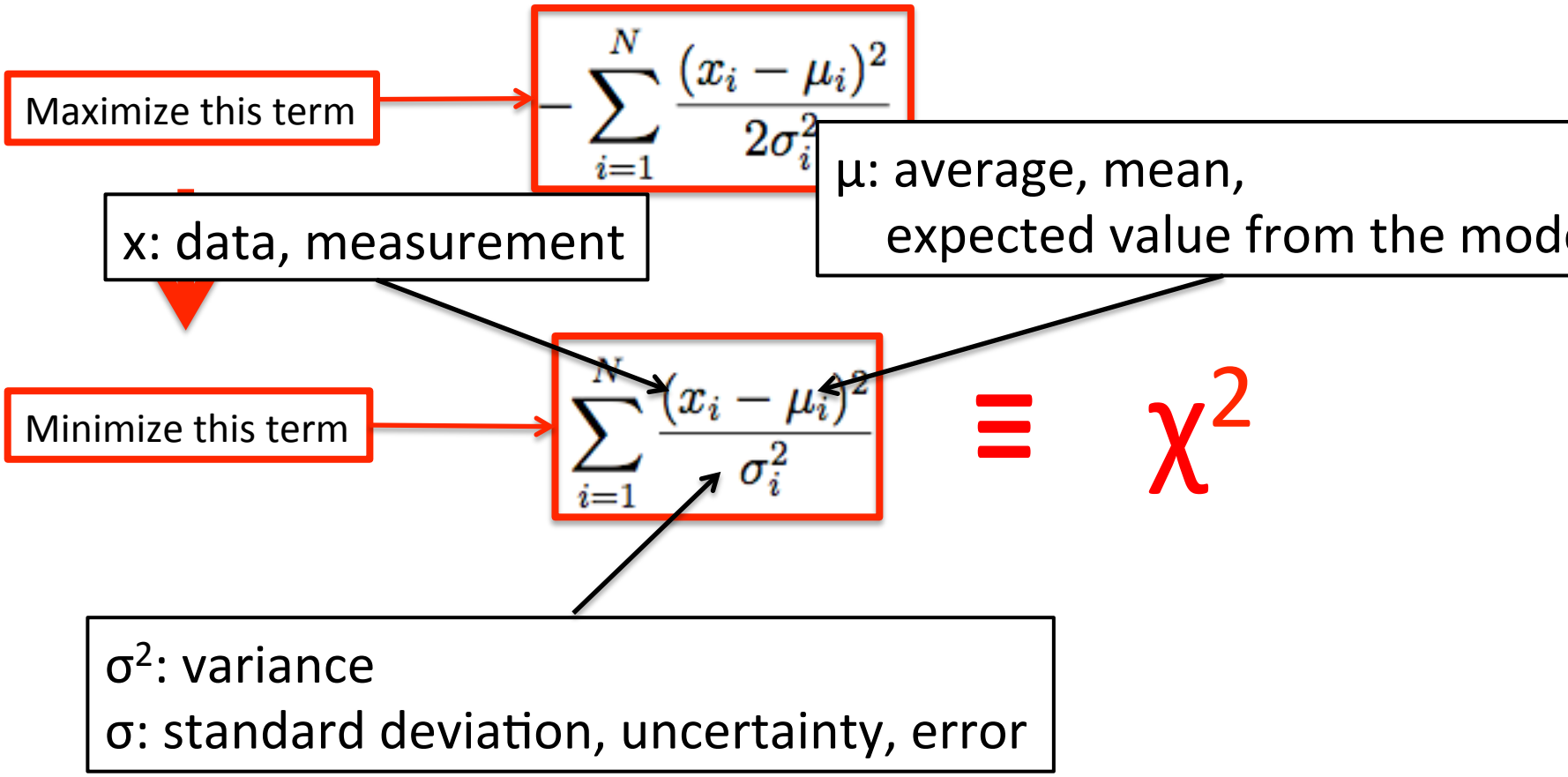
Maximize this term → $-\sum_{i=1}^{N} \frac{(x_i - \mu_i)^2}{2\sigma_i^2}$

Minimize this term → $\sum_{i=1}^{N} \frac{(x_i - \mu_i)^2}{\sigma_i^2} \equiv \chi^2$

Data fitting

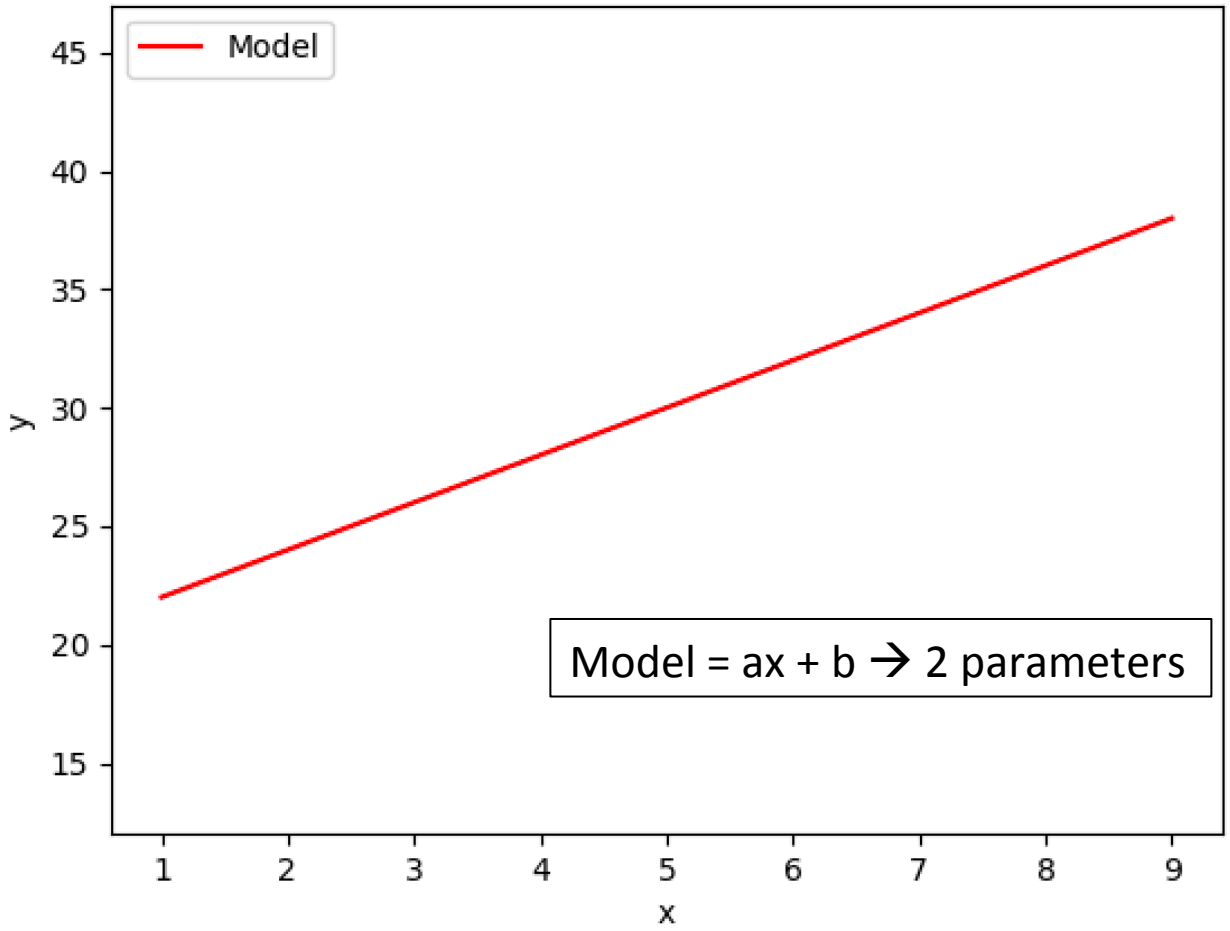
Maximum Likelihood as estimation for data with Gaussian distribution

The model parameters that have the best-chance to explain the observed data



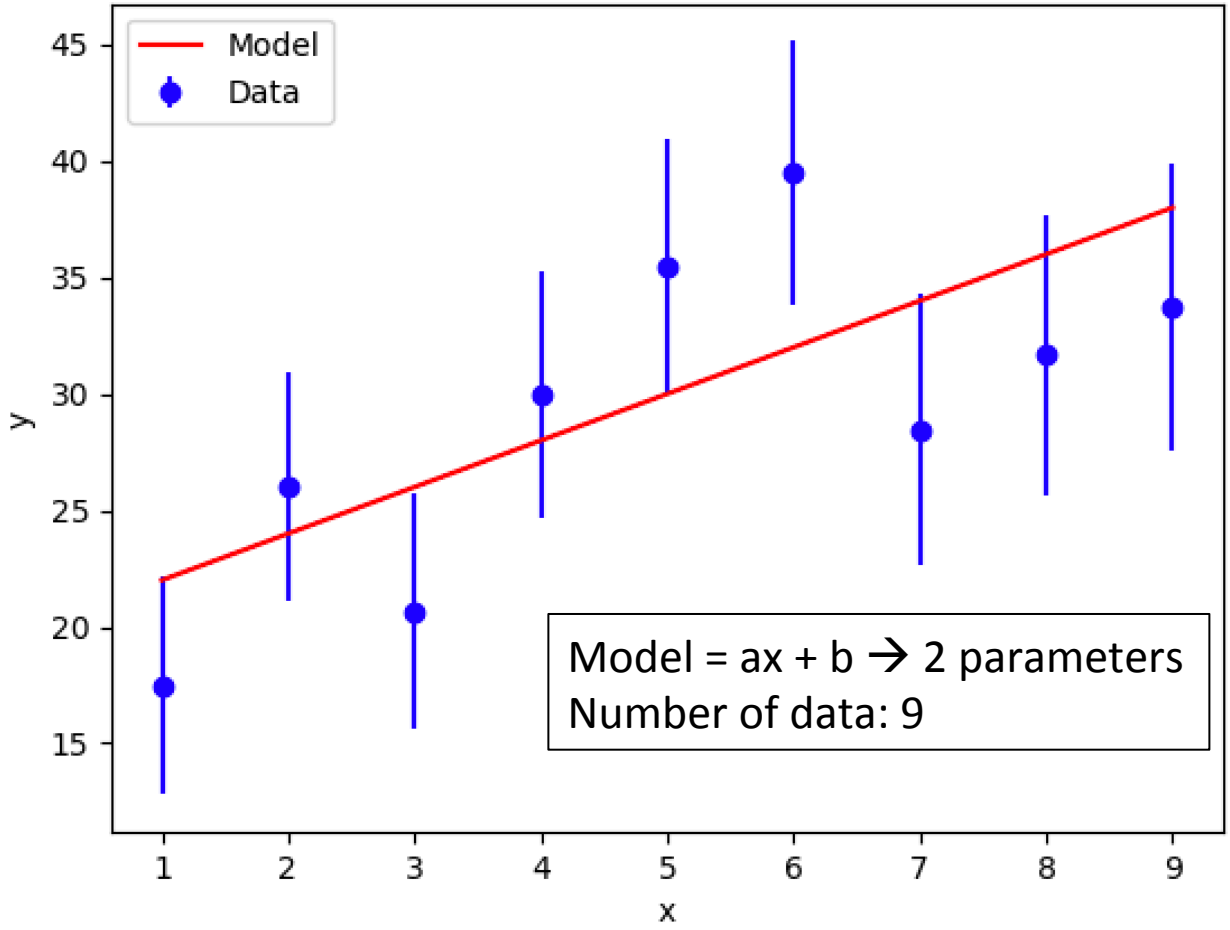
Data fitting

Maximum Likelihood as estimation for data with Gaussian distribution



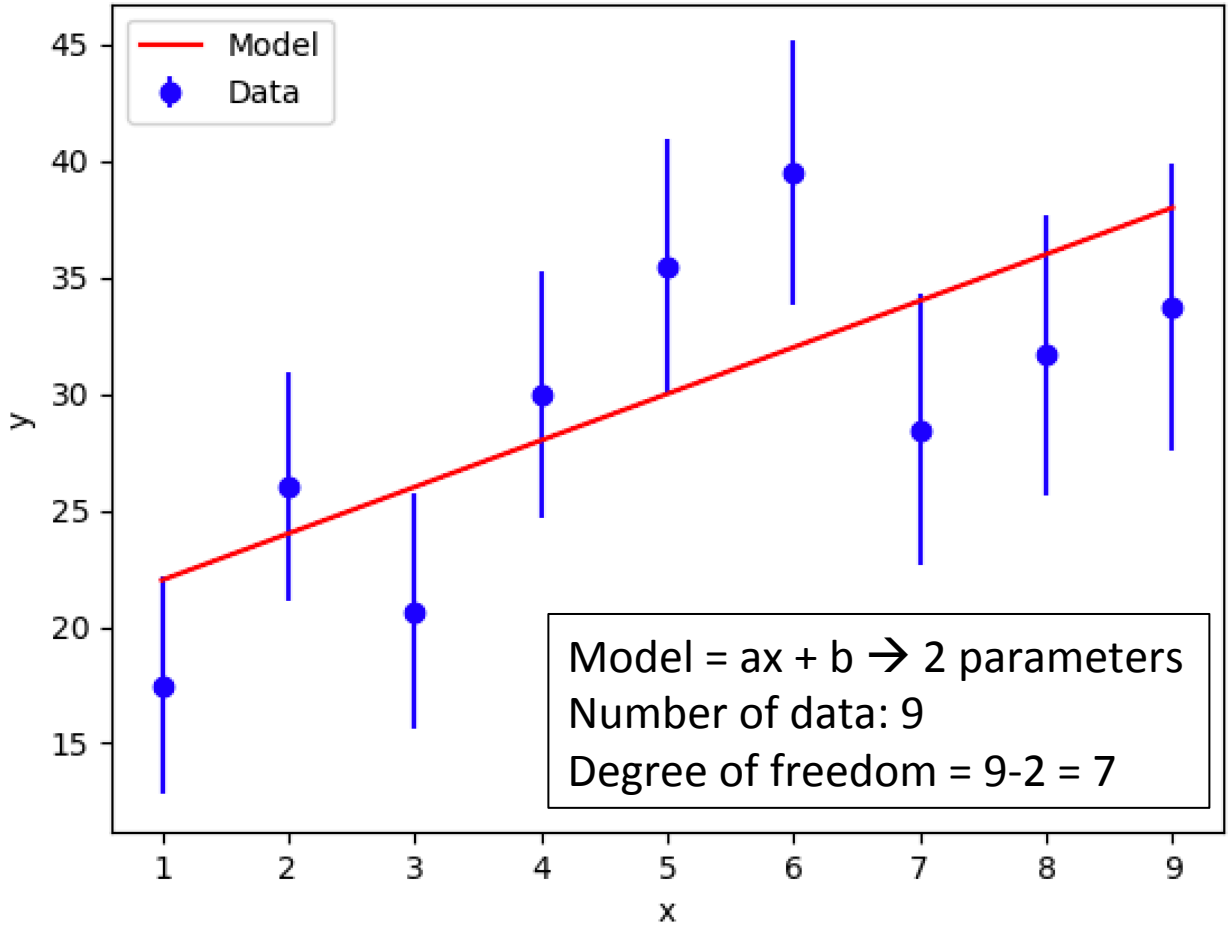
Data fitting

Maximum Likelihood as estimation for data with Gaussian distribution



Data fitting

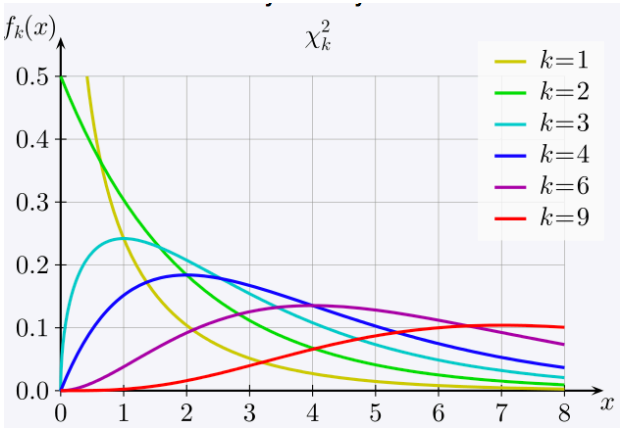
Maximum Likelihood as estimation for data with Gaussian distribution



Chi-square distribution

- If χ2 is a sum of the squares of independent normal/Gaussian random variables.

$$f(x;\,k) = \begin{cases} \dfrac{x^{\frac{k}{2}-1}e^{-\frac{x}{2}}}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)}, & x > 0; \\ 0, & \text{otherwise.} \end{cases}$$



x: value of χ^2

k: degree of freedom

$$\langle x \rangle = k$$

→ IF you have enough data points, the reduced $\chi^2 = \chi^2/k \sim 1$