Probability and Statistical Distributions

Machine Learning Course 160125 | knut.mora@fysik.su.se

Main themes, Chapter 3

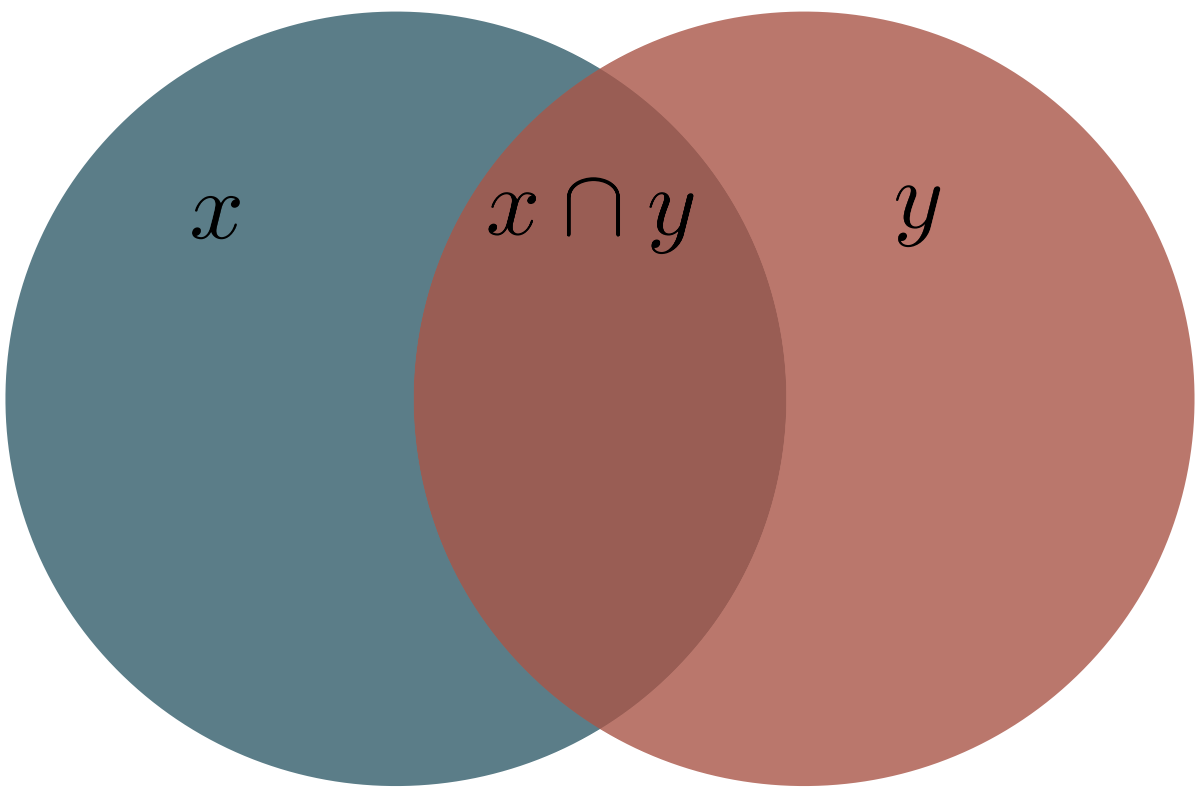
- Axioms and rules for probability, notation
- Conditional probability
- Probability distributions
- Random numbers
- Correlation

Notation and commonly used formulæ

$p(x) \quad p(x,y)$ Probabilities- must add to one

$$p(x \cap b) = p(x|y)p(y) = p(y|x)p(x)$$ < Bayes rule

$$p(x \cup y) = p(x) + p(y) - P(x \cap y)$$



transform variables:

$$p(x)dx = p(y(x))dy$$

Describing a distribution:

- FWHM
- moments:

$$m(k) = \int h(x) x^k dx$$

- Arithmetic mean (also known as the expectation value),

$$\mu = E(x) = \int_{-\infty}^{\infty} x h(x)\, dx \tag{3.22}$$

- Variance,

$$V = \int_{-\infty}^{\infty} (x-\mu)^2 h(x)\, dx \tag{3.23}$$

- Standard deviation,

$$\sigma = \sqrt{V} \tag{3.24}$$

- Skewness,

$$\Sigma = \int_{-\infty}^{\infty} \left(\frac{x-\mu}{\sigma}\right)^3 h(x)\, dx \tag{3.25}$$

- Kurtosis,

$$K = \int_{-\infty}^{\infty} \left(\frac{x-\mu}{\sigma}\right)^4 h(x)\, dx - 3 \tag{3.26}$$

- Absolute deviation about d,

$$\delta = \int_{-\infty}^{\infty} |x-d| h(x)\, dx \tag{3.27}$$

- Mode (or the most probable value in case of unimodal functions), x_m,

$$\left(\frac{dh(x)}{dx}\right)_{x_m} = 0 \tag{3.28}$$

- p% quantiles (p is called a percentile), q_p,

$$\frac{p}{100} = \int_{-\infty}^{q_p} h(x)\, dx \tag{3.29}$$

Standard estimators:

mean, $\overline{x}$, and the *sample standard deviation*, s, can be computed via standard formulas,

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{3.31}$$

and

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}. \tag{3.32}$$

Unbiased estimators- the expectation value is the true value
Consistent- bias and variance approaches zero with large sample size

Conditional probability

- If you have a multivariate pfd, you may be interested in the distribution of only one variable. Integrate away the rest, and get the marginal distribution.

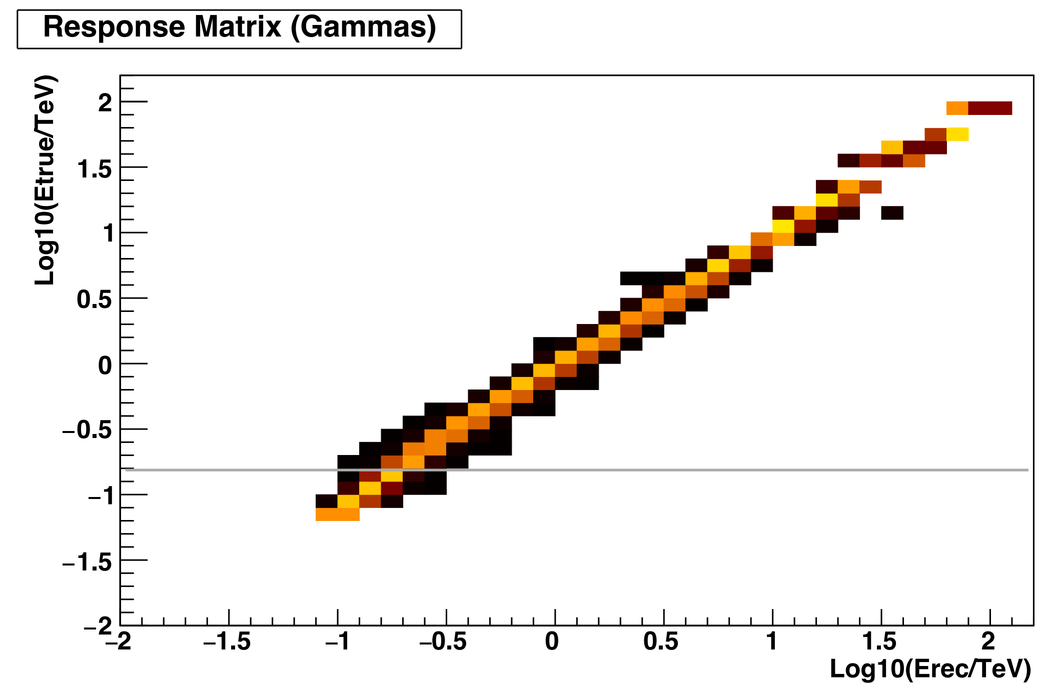
- You may also be interested in a slice of the multivariate pdf- the example to the right shows the distribution of reconstructed energy given one true energy.

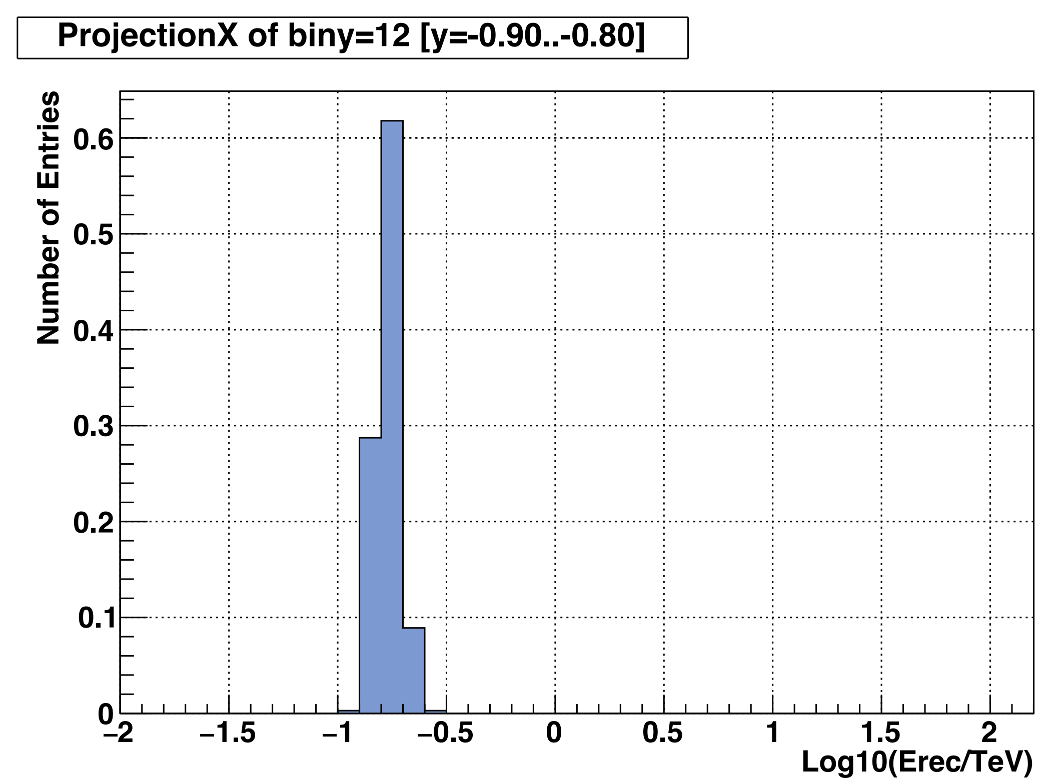
- In the latter case, you must renormalize using Bayes rule

- If they do not depend-

$$p(x,y) = p(x|y)p(y) = p(x)p(y)$$

$$h(x) = \int h(x,y)dy$$

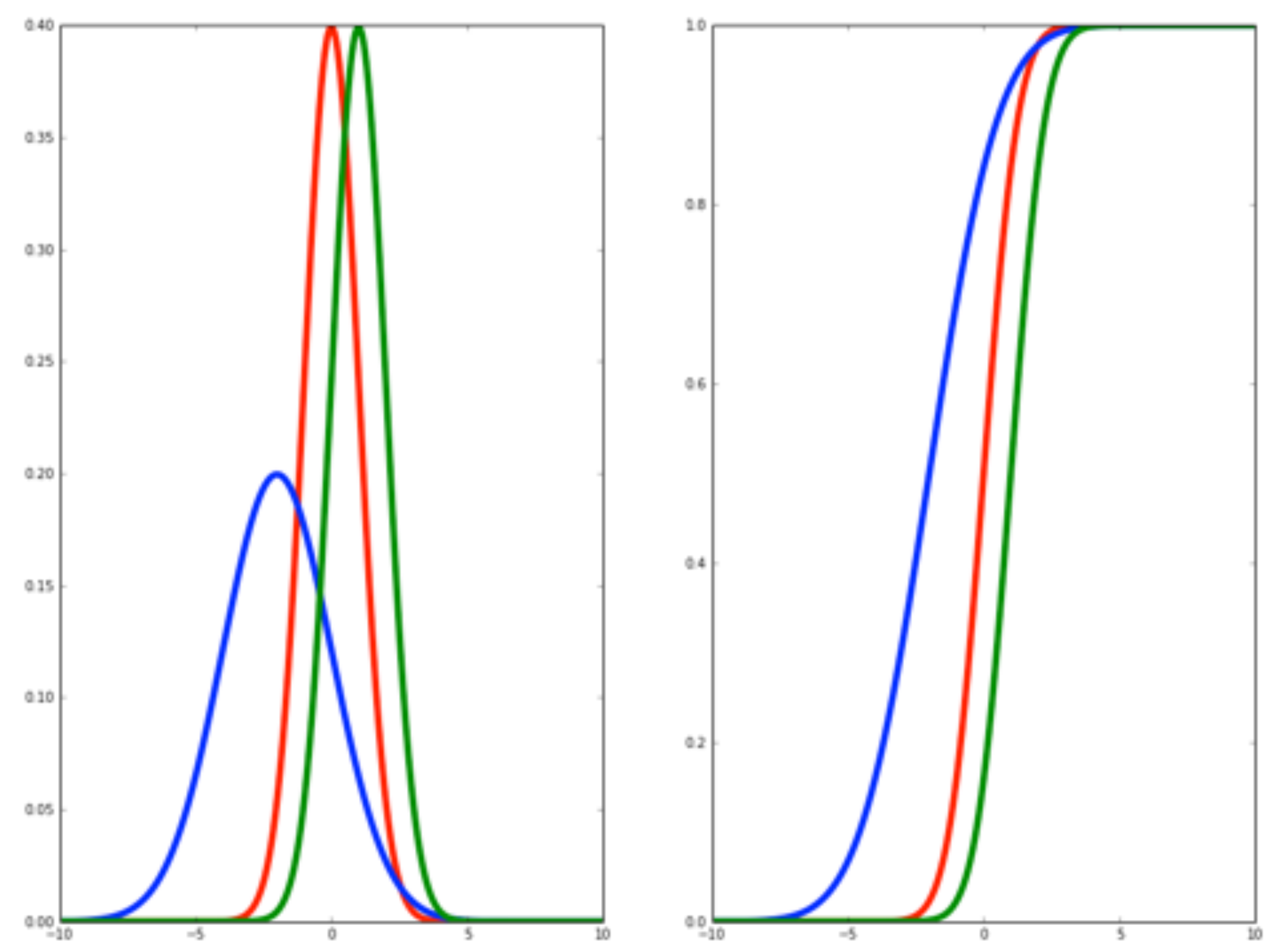


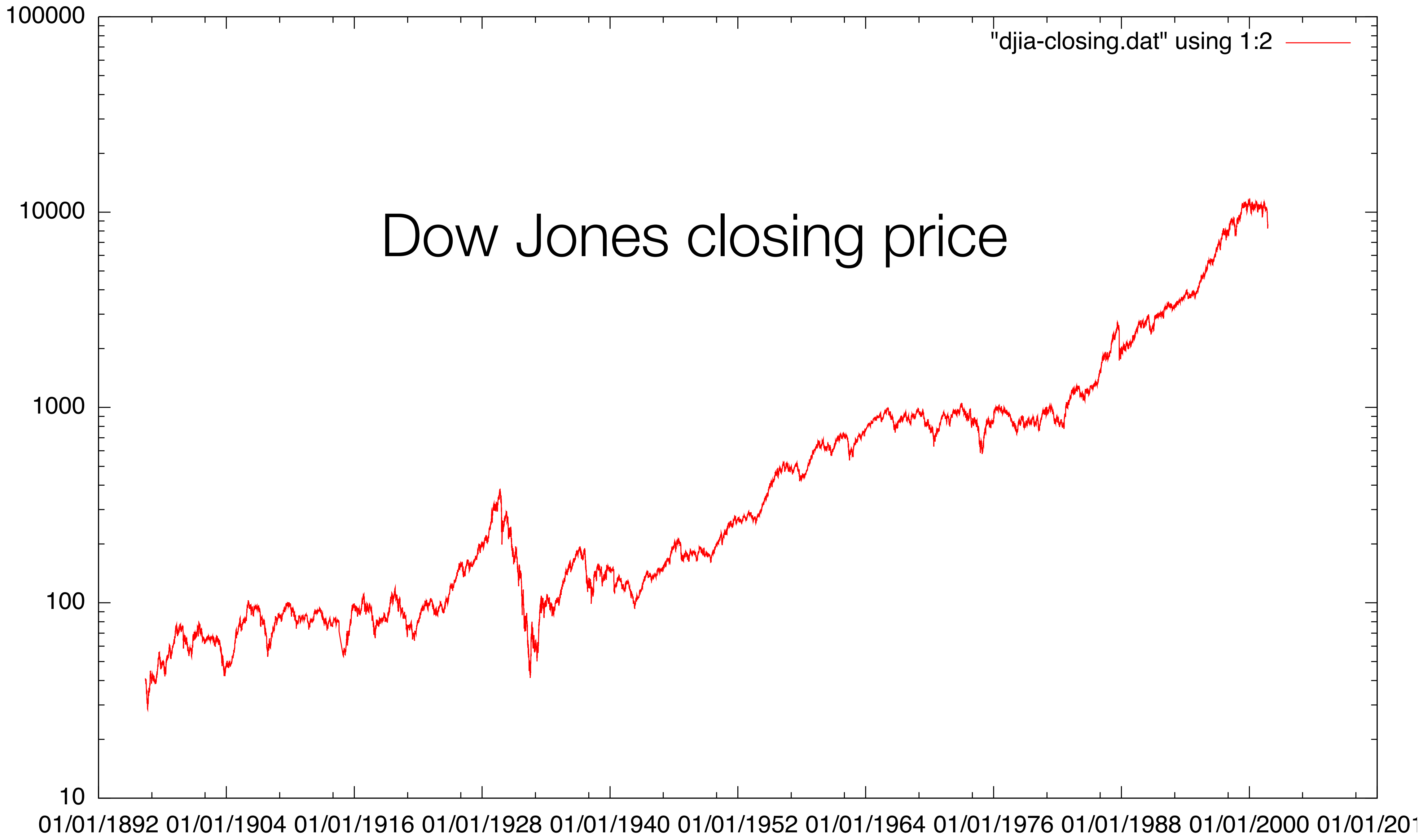


Gaussian Distribution

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

- AKA bell curve, normal
- Typical quantiles in units of sigma: 0.68 in +-1sigma, 0.95 in +-2sigma (approx)





The Central Limit Theorem

cross your fingers and hope for normality

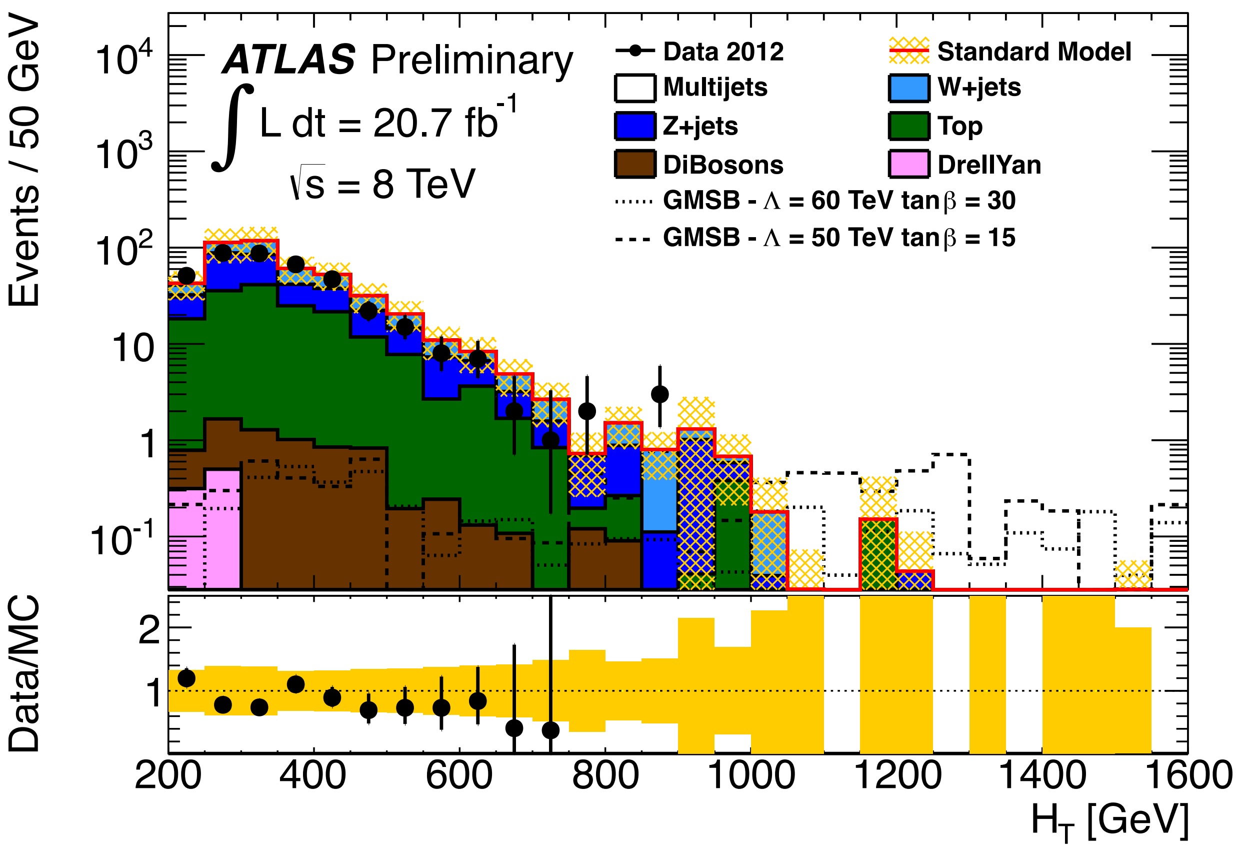
Poisson Distribution

$$h(n|\mu) = \frac{\mu^n}{n!} e^{-n}$$

- If your process has the same probability to happen each infinitesimal time interval, the number of events in a certain time is Poisson distributed

- Classic example: number of events in histogram or a counting experiment.

- In the limit of a high expected number, you get a gaussian with σ=√n (next page)



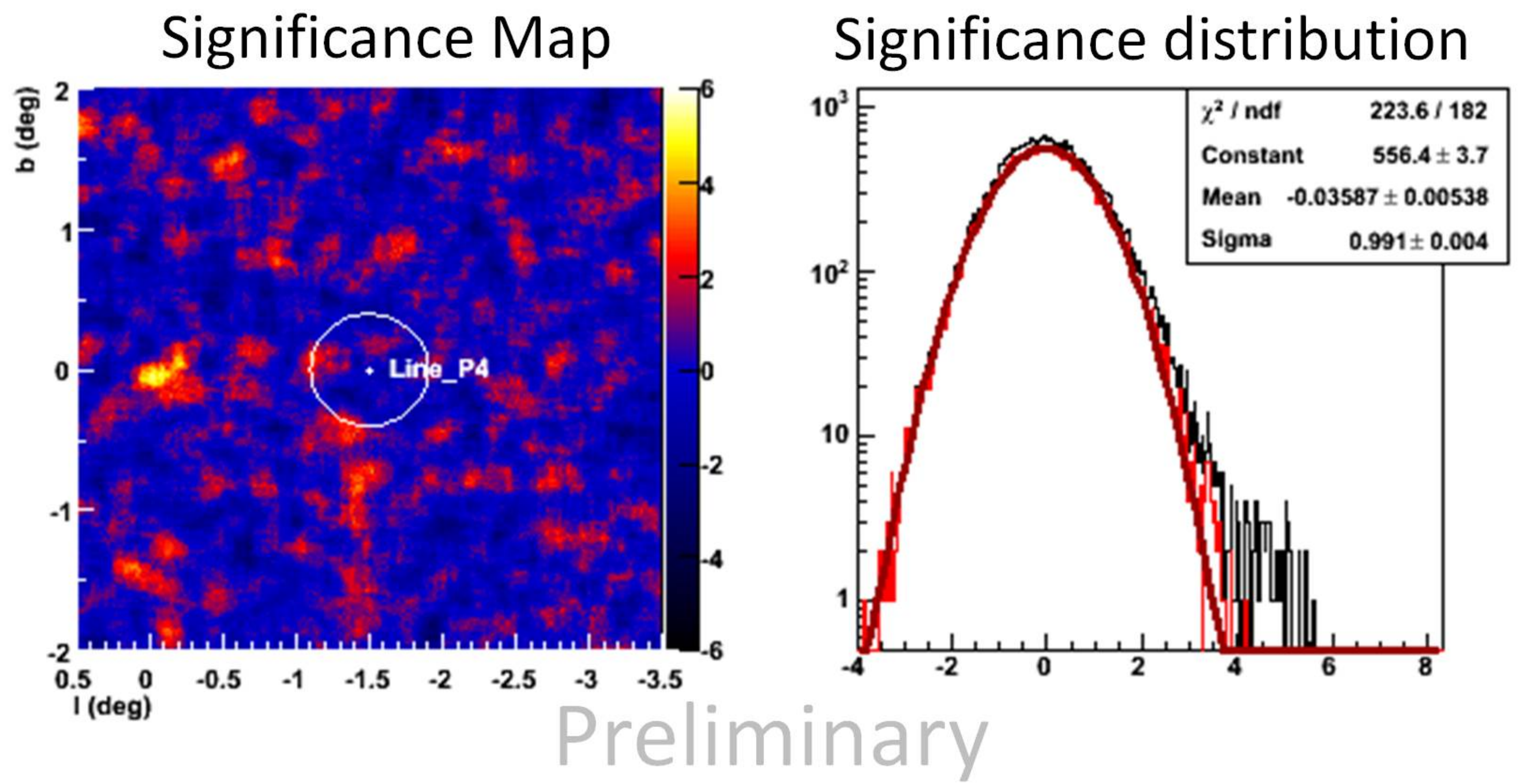


Figure 5: Significance map (left) and significance distribution (right) in the FOV. The ROI is represented with a white circle centred on the Fermi hotspot ($-1.5°$, $0°$). The known source HESS J1745-290 is clearly detected.

Significance distribution for photon counts

Large photon counts converge to a gaussian

Distributions and random numbers in Python

- scipy.stats contains a large number of statistical distributions
- you can generate random numbers, find pdf(pmf)s for continuous(discrete) distributions, as well as cumulative distributions
- Also inverses- useful if you want a limit, or need to compute a p-value.

```
In [1]: import scipy.stats as sps

In [2]: a = sps.norm(0,1)

In [3]: print a.rvs(3) ,"3 random #s"
[ 1.04566757  0.28175318  1.34379906] 3 random #s

In [4]: print a.pdf(1) ,"pdf(1)"
0.241970724519 pdf(1)

In [5]: print a.cdf(1) ,"cumulative function(1)"
0.841344746069 cumulative function(1)

In [6]: print a.ppf(0.5) ,"inverse cumulative function(0.5)"
0.0 inverse cumulative function(0.5)

In [7]: a.
a.args      a.entropy   a.kwds      a.moment    a.ppf
a.cdf       a.interval  a.mean      a.pdf       a.rvs
a.dist      a.isf       a.median    a.pmf       a.sf
```

Correlation

Upper right: variance and covariance in 2-D. Note that if x and y are independent, the correlation is 0. This implication does not work the other way!

Lower right: addition law for covariance

$$V_x = \int\int (x - \mu_x)^2 h(x, y) dy$$

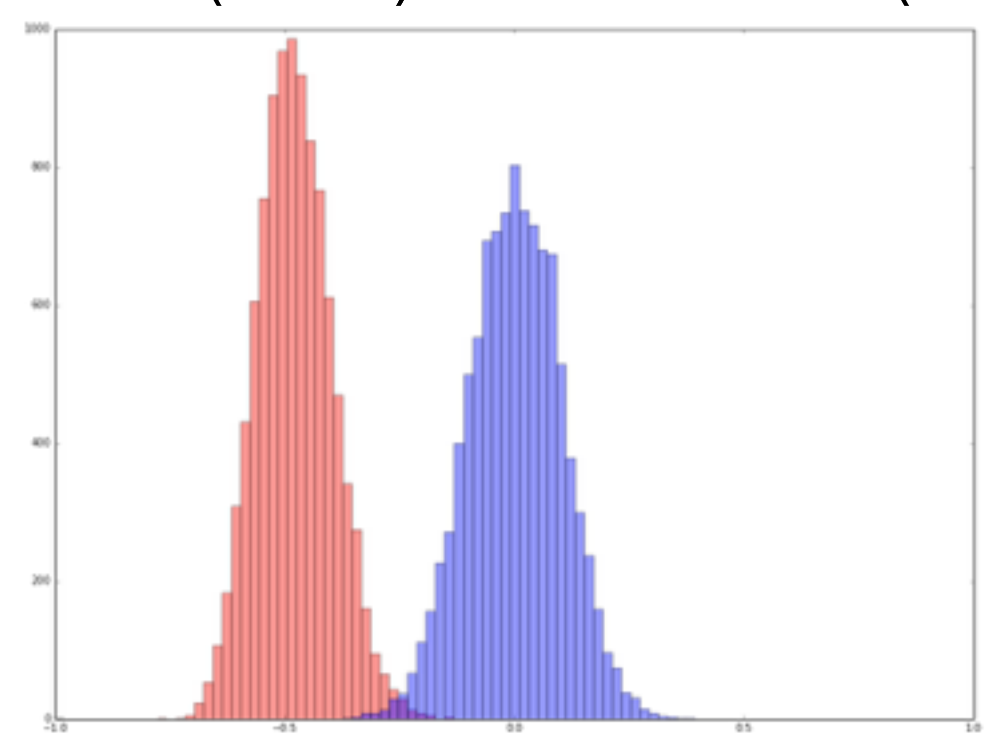
$$V_{xy} = \int\int (x - \mu_x)(x - \mu_y) h(x, y) dy$$

$$\mathrm{V}_{x+y} = \mathrm{V}_x + \mathrm{V}_y - 2\mathrm{V}_{xy}$$

2-D gaussian

- scipy.stats.multivariate_normal
- Normal gaussian, but replace variance and mean by matrixes
- “robust” estimate fcns provided by book
- Fun: I did not know about spearmans or kendall- they are quite neat.

Spearman rho for r=0(blue) and r=-0.5(red)



Main items for discussion

- What estimators and distributions do you use in your work?
- Are they biased?
- robust against outliers?