Introduction to statistics

Laurent Eyer
Maria Suveges, Marie Heim-Vögtlin SNSF grant

Recent history at the Observatory

- Request of “something on statistics” from PhD students, because of an impression of lack of knowledge
  - Daniel Schaerer
  - Amaury Triaud, Richard Anderson
  - Maria Suveges, Damien Segransan, Stéphane Paltani, Laurent Eyer
- Cafés statistiques in 2005-2006: http://obswww.unige.ch/~eyer/CAFSTAT/

Plan

- ~8 sessions
- Wednesday 15h - 17h
- 3 first lectures:
  - General Introduction, definitions, hypothesis testing, today
  - Chi2 statistics, maximum likelihood, wednesday 6 October 2010
  - Monte Carlo Markov chain, Wednesday 10 November 2010

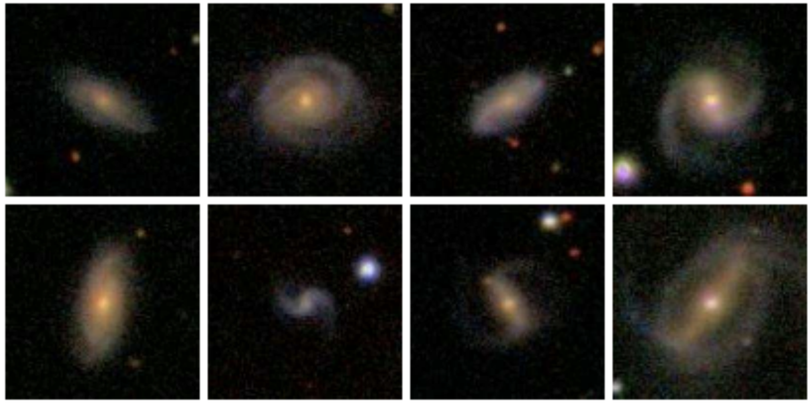
```
1) Statistical tests
  EYER, SUVEGES

2) Chi2 statistics, maximum likelyhood
   SEGRANSAN

3) Monte Carlo, Markov chain
   PALTANI

4) Robust, non-parametric statistics
   PALTANI, +SEGRANSAN, +EYER

5) Non-Gaussian statistics
   SUVEGES

6) Time series analysis
   EYER

7) Bayesian statistics
   EYER?,  SEGRANSAN?

8) Biases
   ??
```

Some statistical resources for this introduction

- Probabilités, analyse des données et statistique Gilbert Saporta
- Introduction à la statistique, Stephan Morgenthaler
- Advanced theory of statistics Maurice Kendall & Allan Stuart
- Statistics in theory and practice, Robert Lupton
- Introductory statistics with R, Peter Dalgaard
- wikipedia (it seems quite good, but always to take with care)
- ...

Random Variables

- Discrete:
  - Spectral type (G2V, KIII)
  - Galaxy type, galaxy zoo



Class	Button	Description
1		Elliptical galaxy
2		Clockwise/Z-wise spiral galaxy
3		Anti-clockwise/S-wise spiral galaxy
4		Spiral galaxy other (e.g. edge on, unsure)
5		Star or Don't Know (e.g. artefact)
6		Merger

- Continuous:
  - magnitude, flux, colour, radial velocity, parallax/distance, temperature, elemental abundances, magnetic field, age, etc...

Distribution

- Definition: density is a function $f(x)$ such that:

$$\Pr(a < X < b) = \int_a^b f(x)dx \qquad \Pr(X = k) = p(k)$$

- Distribution of one variable: univariate $f(x)$

- Distribution of several variables: multivariate $f(x, y, \ldots)$

- Marginalization: $u(x) = \int_{-\infty}^{\infty} f(x, y)dy$

- If and only if: $f(x, y) = u(x)v(y)$ the variables are independent

- Cumulative Distribution Function CDF:

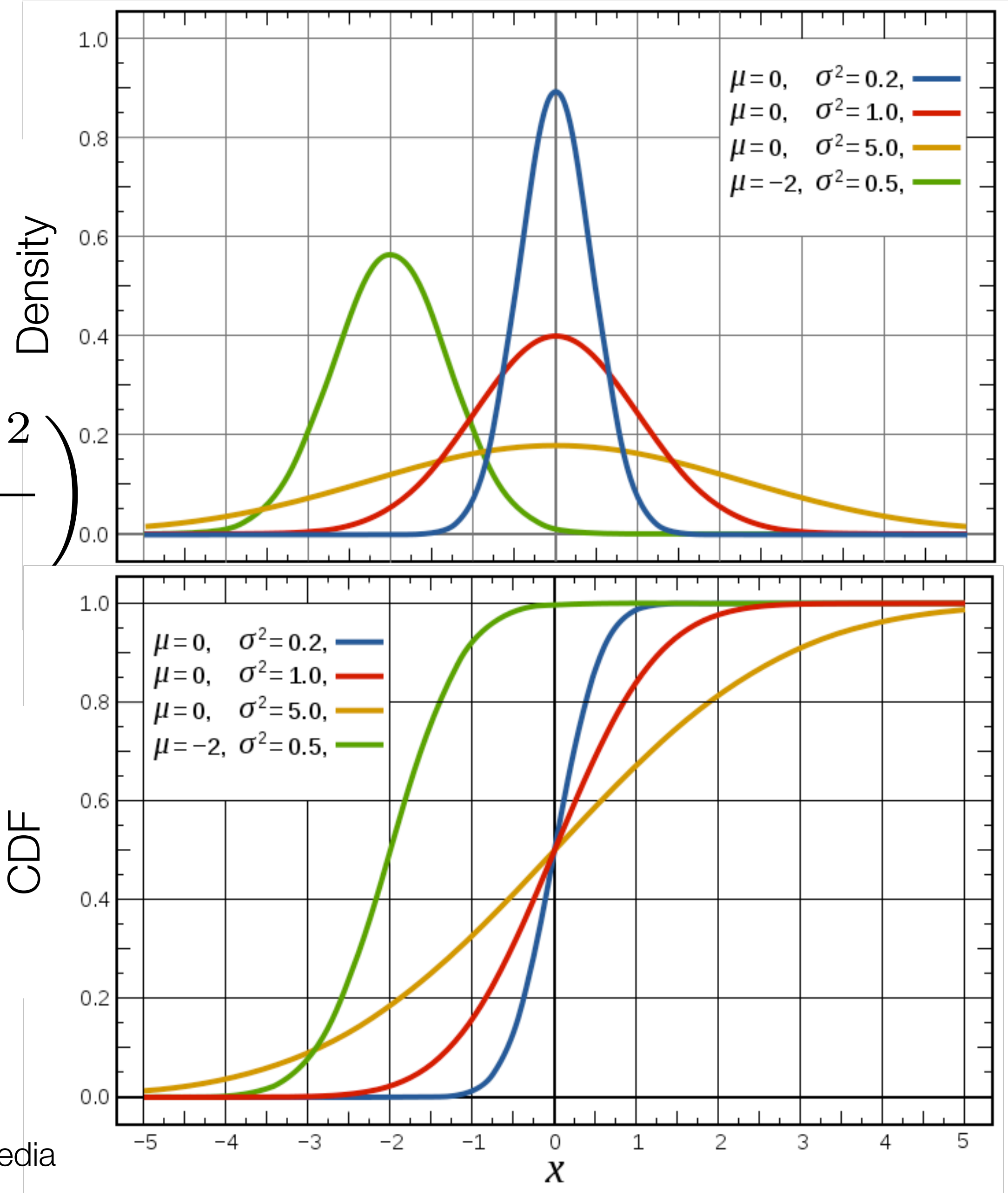
$$F(x) = \int_{-\infty}^{x} f(x')dx'$$

Digression on "LateXiT"

Example: Gaussian / Normal distribution

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

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



from wikipedia

Poisson distribution

Discrete probability distribution (no density)

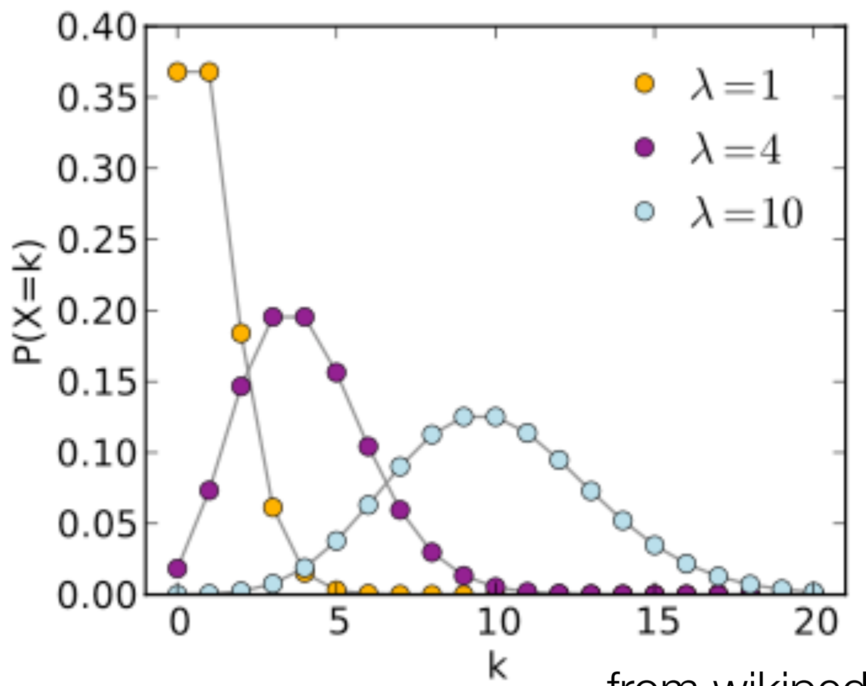
$$X \sim \text{Poisson}(\lambda)$$

Number of photons on a detector
Number of people in a shop

$$\Pr(X = k) = \exp(-\lambda)\frac{\lambda^k}{k!}$$

For large λ

$$\mathcal{N}(\lambda, \lambda)$$



from wikipedia

Moments of a distribution

- Information of location, the mean

$$E(X) = \mu = \int_{-\infty}^{\infty} x f(x) dx$$

Normal: μ
Poisson: λ

- Information of dispersion, the variance

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx$$

standard deviation

$$\sigma = \sqrt{\mathrm{Var}(X)}$$

Normal: σ^2
Poisson: λ

- Moment of order n about the mean:

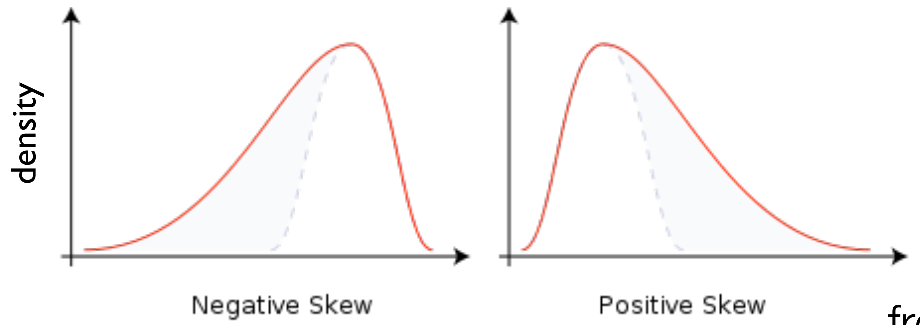
$$\mu_n = \int_{-\infty}^{\infty} (x-\mu)^n f(x) dx$$

3d and 4th moments of a distribution

- Skewness, asymmetry

$$\mu_3/\sigma^3 = \int_{-\infty}^{\infty} (x-\mu)^3 f(x) dx / \sigma^3$$

Normal: 0
Poisson: 1/√λ



from wikipedia

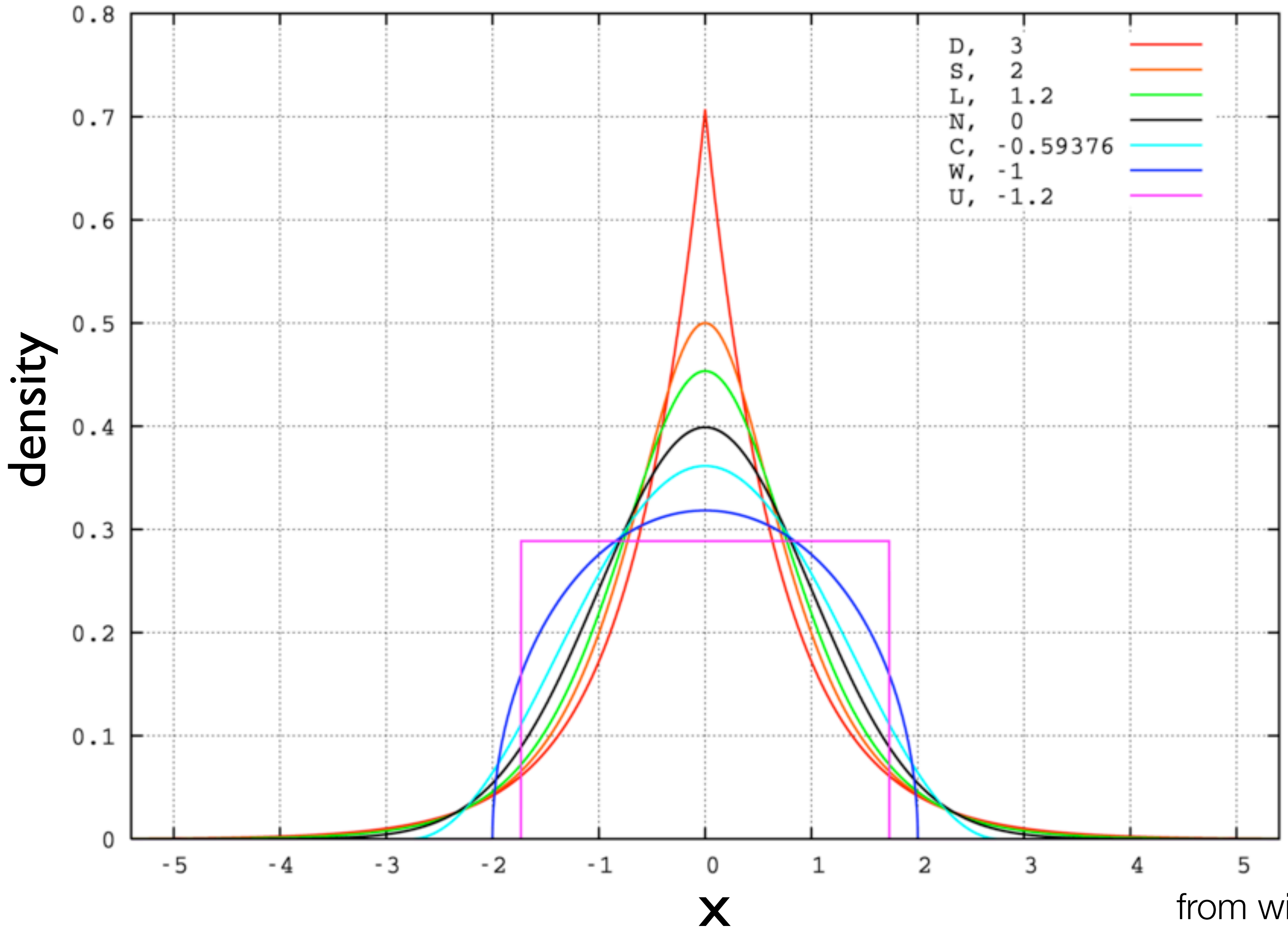
- Kurtosis

$$\mu_4/\sigma^4 = \int_{-\infty}^{\infty} (x-\mu)^4 f(x) dx / \sigma^4$$

$$\mu_4/\sigma^4 - 3$$

Normal: 0
Poisson: 1/λ

Example of different values of kurtosis: “boxiness” -- tail heaviness



from wikipedia

Covariance and correlation

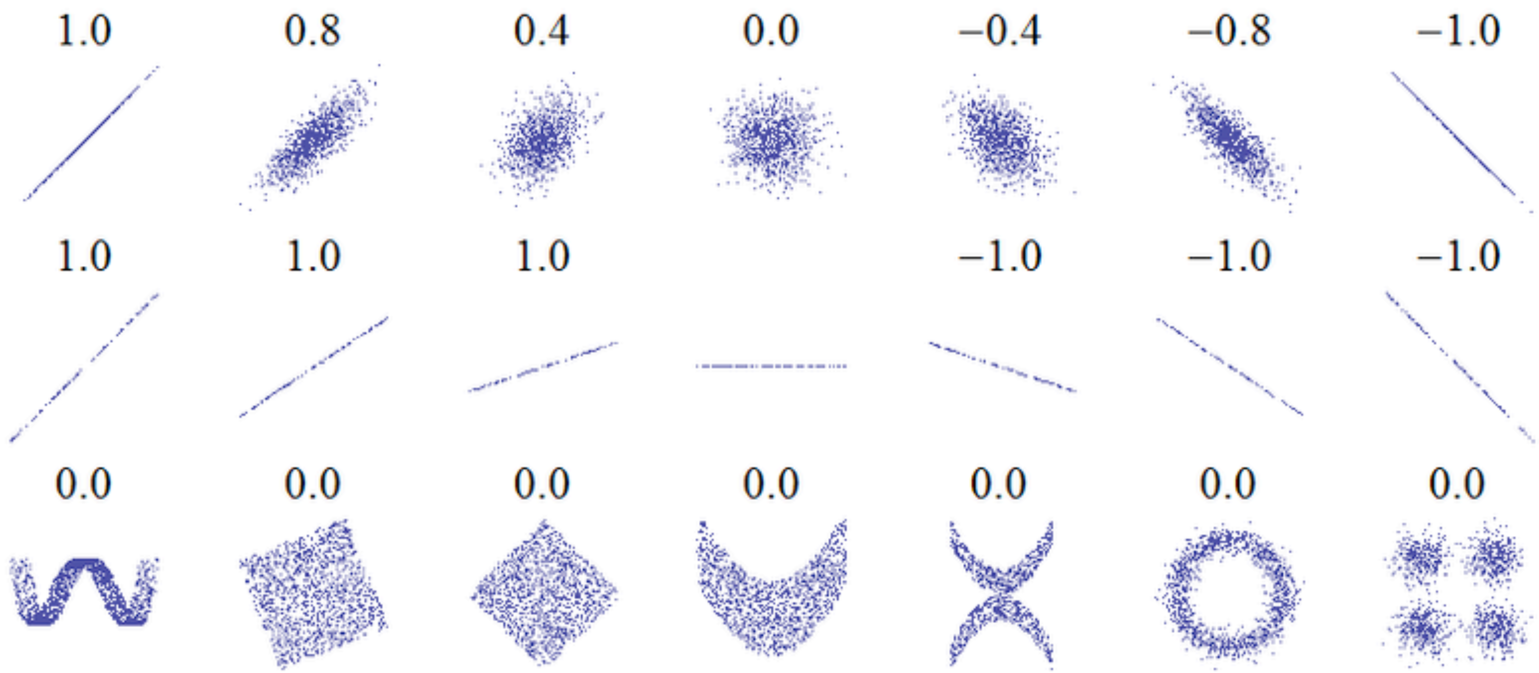
- Covariance

$$Cov(X,Y) = \int\int (x-\mu_x)(y-\mu_y) f(x,y)\,dx\,dy$$

- Correlation

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

Examples of correlation

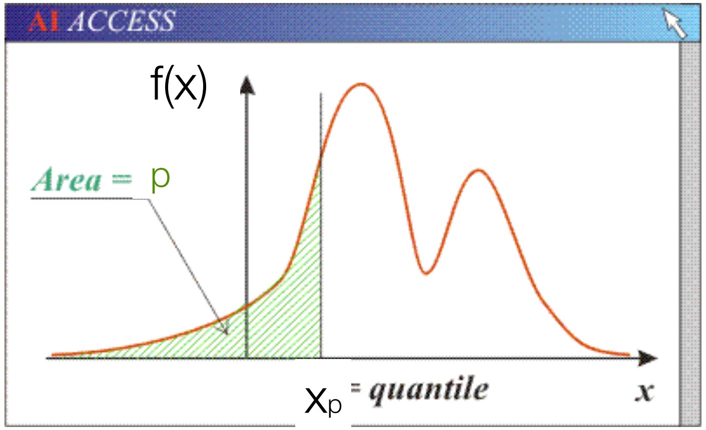


from wikipedia

Quantiles

- x_p : p-quantiles of f(x)

$$p = \int_{-\infty}^{x_p} f(x) dx$$



from www.aiacces.net

- Measure of location: Median

$$1/2 = \int_{-\infty}^{x_{1/2}} f(x) dx$$

- Measure of dispersion: Inter-quantile range

$$\mathrm{IQR} = x_{3/4} - x_{1/4}$$

Data, samples

- Usually we have observations, e.g. additive process

$$y_i = f(t_i) + \epsilon_i \qquad\qquad i = 1, \ldots, n$$

Deterministic random variable

- We want a characterisation of the deterministic and random parts

- Suppose something about the random variable, often normality: $\mathcal{N}(0, \sigma^2)$

Estimators

- Assumption of models
- Estimate the parameters of a distribution, moments
  - Exercise 1: Sample mean: $\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad E(\bar{X}) = \mu$
  - Exercise 2: Sample variance (bias):

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad E(\hat{\sigma^2}) = \frac{n}{n-1}\sigma^2$$

redefine ⇒

$$\hat{\sigma^2} = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- Sample quantiles are estimators of quantiles
  - Exercise 3: what is the sample median of {1 , 2, 3, 109812308}?

Central limit theorem

The distribution of the mean of a sufficiently large number of random variables can be approximated by a Gaussian distribution!

$$X_i, \ i = 1, \ldots, n \ \ \text{iid} \ \ \text{with} \ \ \mathrm{E}(\mathrm{X_i}) = \mu \ \ \mathrm{Var}(\mathrm{X_i}) = \sigma^2$$

iid= Independent identically distributed

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \quad \text{follows approximately} \quad \mathcal{N}(0, 1)$$

One reason why the Gaussian distribution is so important

Distribution derived from Normal distribution
1) Chi square distribution

If $X_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ ⇒ $\sum_{i=1}^{k} X_i^2 \sim \chi_k^2$

iid= Independent identically distributed

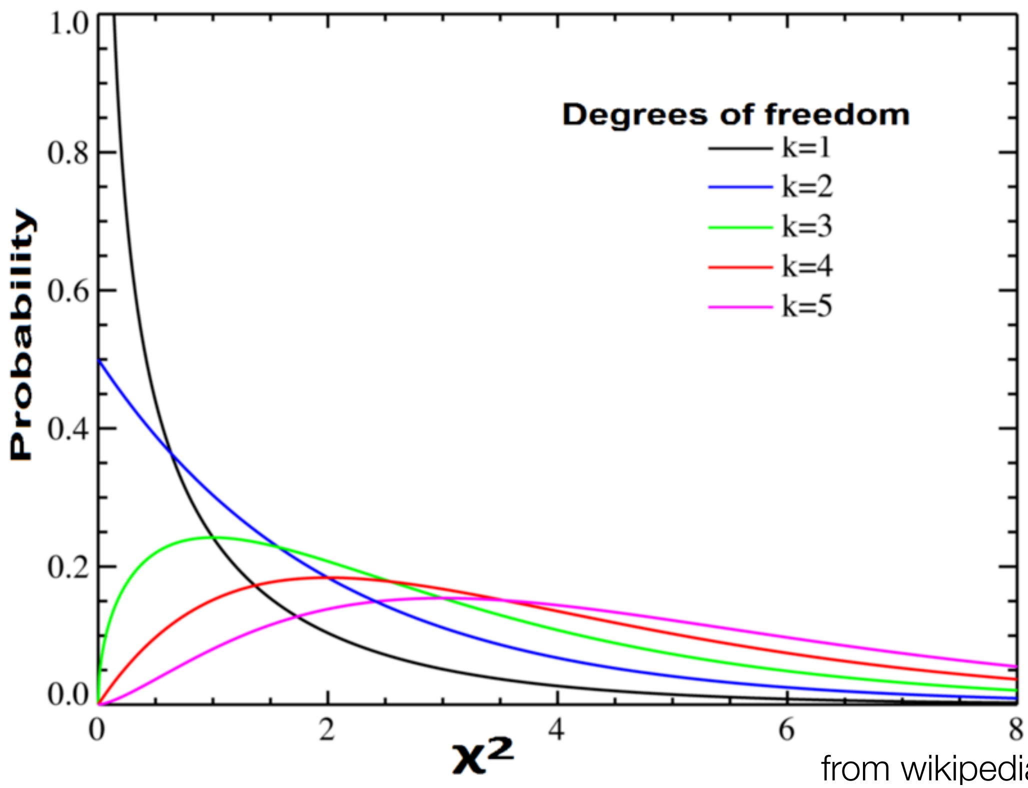
$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} \exp\left(-x/2\right)$$

mean: k
variance: 2 k
skewness: √(8/k)
kurtosis: 12/k

$$X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma)$$

$$\sum_{i=1}^{k} (X_i - \bar{X})^2/\sigma^2 \sim \chi_{k-1}^2$$

When k is large χ_k^2 approximates a $\mathcal{N}(k, 2k)$



from wikipedia

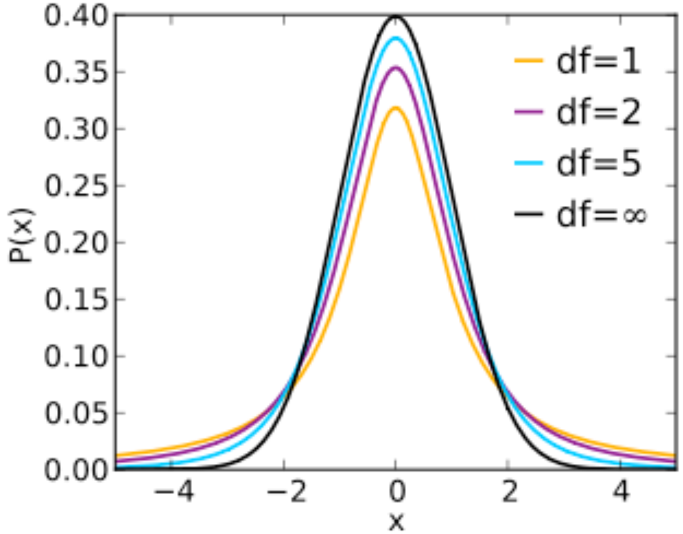
Distribution derived from Normal distribution
2) Student distribution

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

$$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\,\Gamma(n/2)}\left(1+\frac{x^2}{n}\right)^{-(n+1)/2}$$

Note $t_\infty = \mathcal{N}(0,1)$



mean: 0 n>1

NaN n=0,1

variance: n/(n-2) n>2

∞ 1<n≤2

otherwise NaN

skewness: 0 n>3

kurtosis: 6/(n-4) n>4

Estimators of Variance of different statistics

from Kendall & Stuart

Statistic	Variance multiplied by n	Notes
Mean, m'_1	$\mu_2 \; (= \sigma^2)$	True for any population with finite second moment μ_2.
Sample variance, m_2	$\mu_4 - \mu_2^2$	For normal parent, $= 2\sigma^4$.
Sample s.d., s	$(\mu_4 - \mu_2^2)/(4\mu_2)$	For normal parent, $= \sigma^2/2$.
Third moment, m_3	$\mu_6 - \mu_3^2 - 6\mu_4\,\mu_2 + 9\mu_2^3$	For normal parent, $= 6\sigma^6$.
Fourth moment, m_4	$\mu_8 - \mu_4^2 - 8\mu_5\,\mu_3 + 16\mu_2\mu_3^2$	For normal parent, $= 96\sigma^8$.
$\sqrt{b_1} = m_3/m_2^{3/2}$	6	For normal parent only. See **12.18** and Exercise 12.9.
$b_2 = m_4/m_2^2$	24	For normal parent only. See Exercise 12.10.
Coefficient of variation, V	See Example 10.5	For normal parent, $= V^2/2$ approx.
Pearson mode (cf. **6.3**)	See Yasukawa (1926) for formulae and tables	Distribution skew for moderate n.
Mean deviation	See **10.13**	For normal parent, $= \sigma^2(1-2/\pi)$.
Gini's mean difference	See **10.14**	For normal parent, $= (0{\cdot}8068)^2\,\sigma^2$.
Median	$1/(4y_0^2)$ when y_0 is ordinate at median	For normal parent and small samples see Hojo (1931) and K. Pearson (1931). For large normal samples equals $(1{\cdot}2533)^2\sigma^2$.
Quartile	$3/(16y^2)$ where y is the ordinate at the quartile	For normal parent, $= (1{\cdot}3626)^2\,\sigma^2$. See also Hojo (1931).
Deciles	See **10.10**	For normal parent, (deciles 4, 6) $= (1{\cdot}2680)^2\,\sigma^2$; (deciles 3, 7) $= (1{\cdot}3180)^2\,\sigma^2$; (deciles 2, 8) $= (1{\cdot}4288)^2\,\sigma^2$; (deciles 1, 9) $= (1{\cdot}7094)^2\,\sigma^2$.
Semi-interquartile range	$\frac{1}{4}\{3/16y_1^2 + 3/16y_2^2 - 1/8y_1\,y_2\}\,\sigma^2$, where y_1, y_2 are the quartile ordinates	For normal parent, $= (0{\cdot}7867)^2\,\sigma^2$.
Correlation coefficient r	See Example 10.6	For bivariate normal parent,

Graphical representation QQ Plots

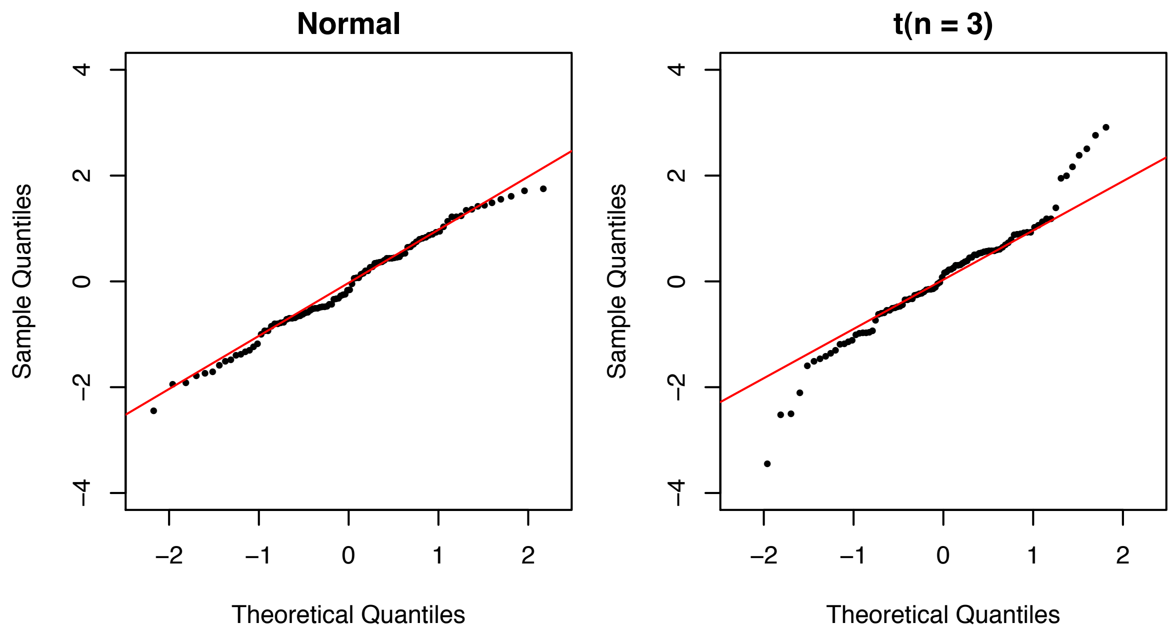
$$X_1, \ldots, X_n \qquad X_i \overset{\text{iid}}{\sim} F(x)$$

$$X_{(1)}, \ldots, X_{(n)}$$

$$\downarrow \qquad \downarrow$$

$$F^{-1}(\frac{1}{n+1}) \qquad F^{-1}(\frac{n}{n+1})$$ Theoretical

Comment on figures: label and numbers large enough, quantity and units



End of the Introduction