Statistics Applied to Bioinformatics

The chi-squared (χ^2) test

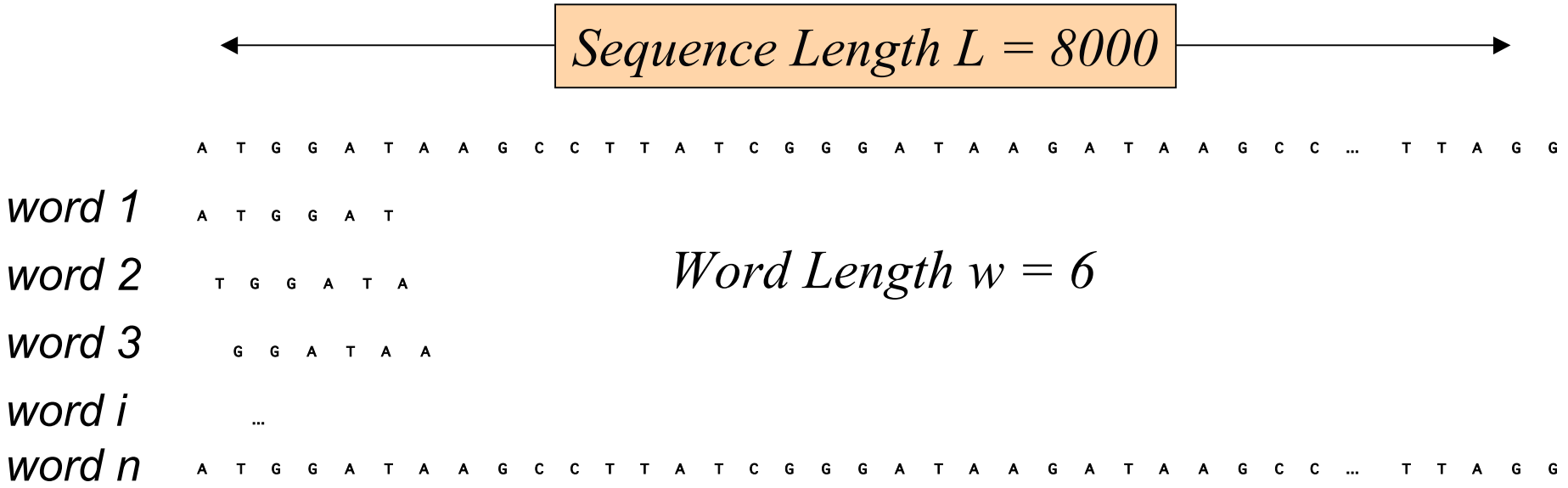
Jacques van Helden
Jacques.van.Helden@ulb.ac.be

Statistics Applied to Bioinformatics

Goodness of fit

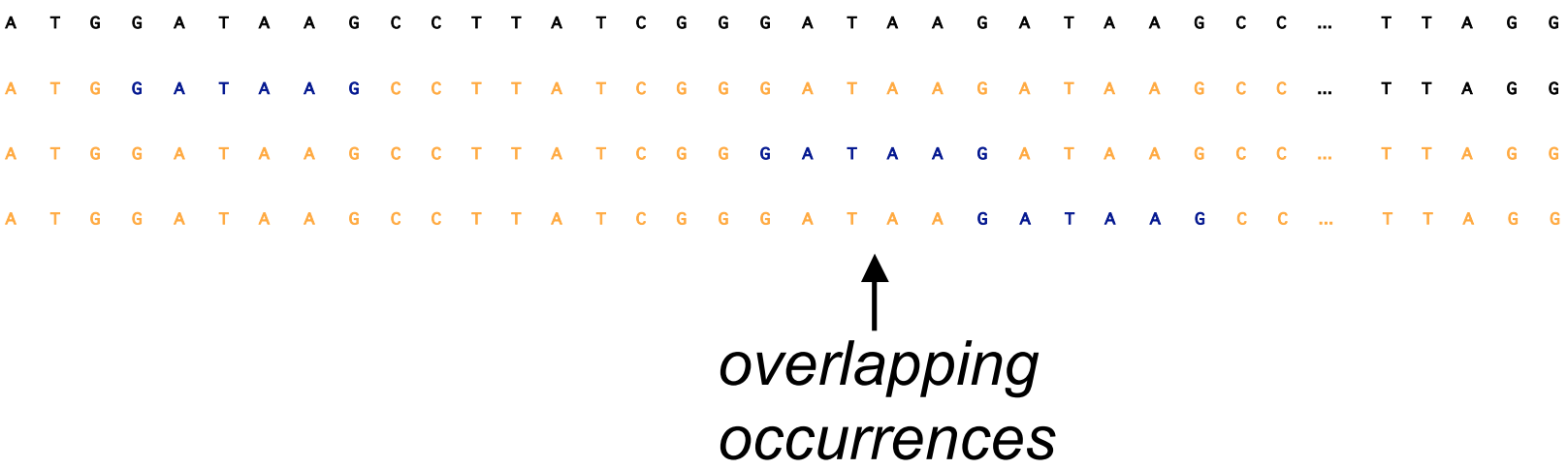
Jacques van Helden
Jacques.van.Helden@ulb.ac.be

Study case : word occurrences in DNA sequences



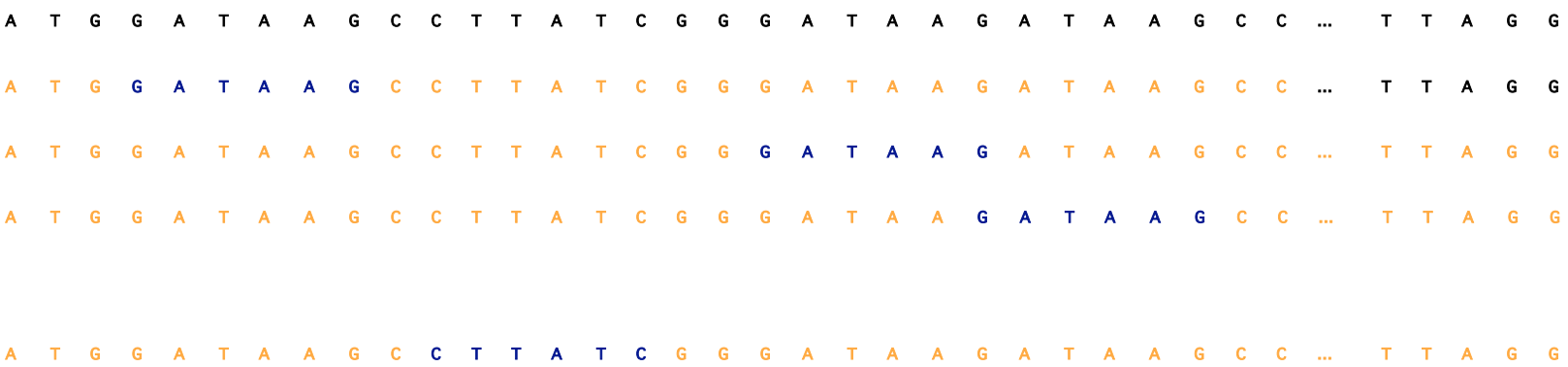
- Number of possible word positions
  n =
- If one also counts the reverse complementary strand, number of possible word positions
  n =

Observed occurrences – 1 strand



- Occurrences of GATAAG
  - obs = (with overlap)
  - obs = (without overlap)

Observed occurrences – 2 strands



- Occurrences of GATAAG counting on both strands
  - = (with overlap)
  - = (without overlap)

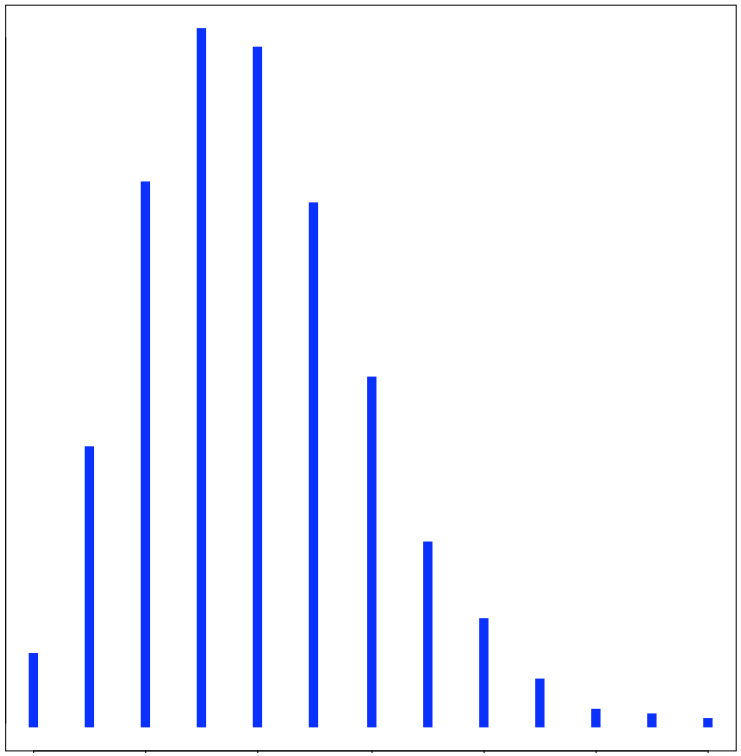
Word probabilities and expected occurrences

- Assuming all words are equiprobable
  - Number of possible words of length 6 =
  - Word proba =
- Expected occurrences for each word
  - Exp = (counting on 1 strand)
  - Exp = (counting on 2 strands)
- Warning, this is never the case in real conditions, and one has to define word-specific expected frequencies

Testing with Random sequence

- Generate 10,000 random sequences of length 8,000 each
- Count occurrences of GATAAG in each sequences (10,000 occurrences)

Observed occurrence distribution

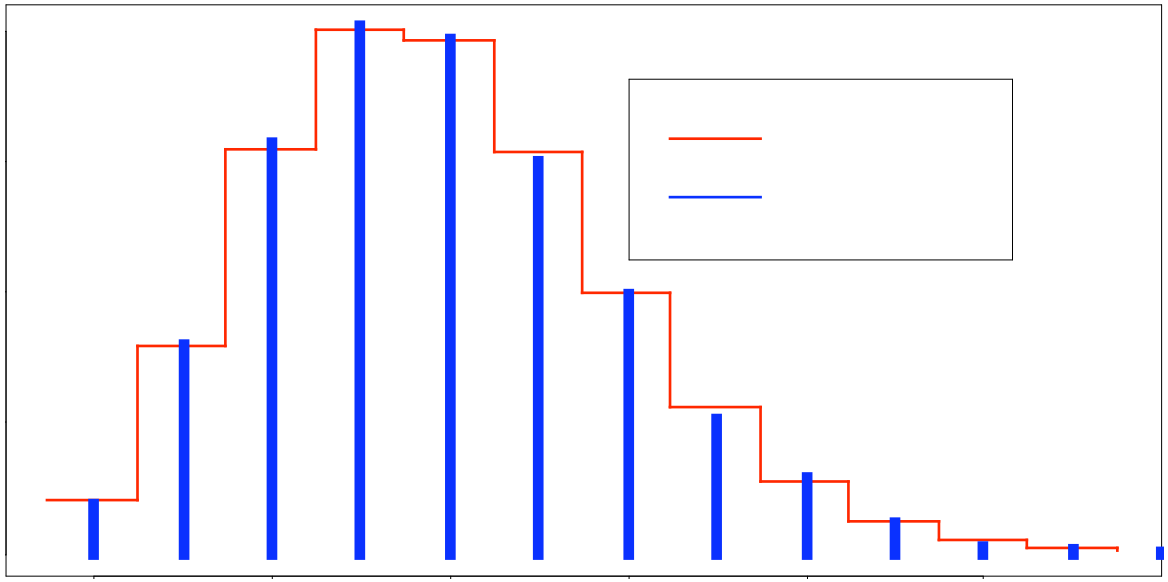


- Calculate distribution of occurrences : how many sequences contain
  - 0 occ obs[0]
  - 1 occ obs[1]
  - …
  - 12 occ obs[12]

Expected occurrence distribution

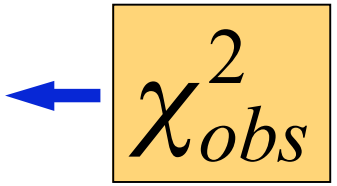
- Calculate expected occurrences with the binomial law
  - Trials =
  - Probability to find a word at a given position = 1/4096
  - For each occurrence between 0 and 12
    - Expected occurrence distribution
      = dbinom(occ,trials,proba)

Binomial fitting



Comparing observed and expected curves

occ	*obs*	*exp*	*d=obs-exp*	*d^2*	*d^2/exp*
0	191	201.6	10.6	111.4	0.55
1	793	787.0	-6.0	35.8	0.05
2	1566	1536.5	-29.5	872.4	0.57
3	2012	1999.6	-12.4	154.0	0.08
4	1962	1951.6	-10.4	107.8	0.06
5	1503	1523.7	20.7	430.1	0.28
6	994	991.3	-2.7	7.1	0.01
7	519	552.8	33.8	1141.0	2.06
8	293	269.7	-23.3	543.3	2.01
9	118	117.0	-1.0	1.1	0.01
10	31	45.6	14.6	214.3	4.70
11	15	16.2	1.2	1.4	0.09
12	3	5.3	2.3	5.1	0.97
SUM	**10000**	**9997.8**	**-2.2**	**3624.8**	**11.43**



The chi-squared (χ^2) test

- Hypothesis
  - observed distribution fits expected distribution
- Calculate the observed chi-squared

$$\chi^2_{obs} = \sum_{i=1}^{p} \frac{(n_{obs_i} - n_{\exp_i})^2}{n_{\exp_i}}$$

In our case $\chi^2_{obs} = 11.43$

- Calculate the degrees of freedom

k=number of classes - 1

In our case $k = 13 - 1 = 12$

- Calculate the theoretical chi-squared with alpha=0.05

$$\chi^2_{theor} = \chi^2_{1-\alpha}$$

In our case $\chi^2_{theor} = 21.0$

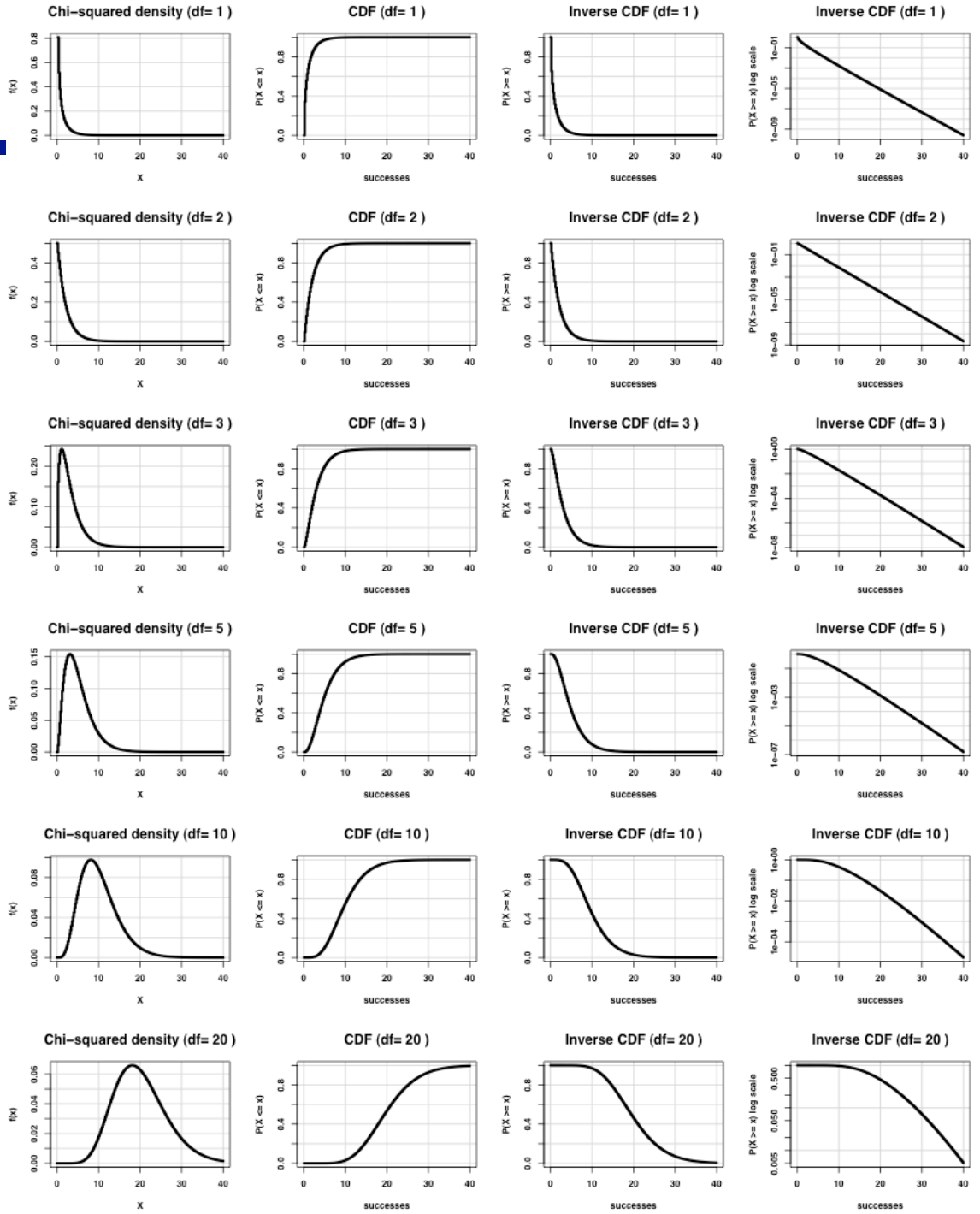
- Reject hypothesis if

$$\chi^2_{obs} \geq \chi^2_{theor}$$

In our case ***ACCEPT***

The theoretical chi2 distribution

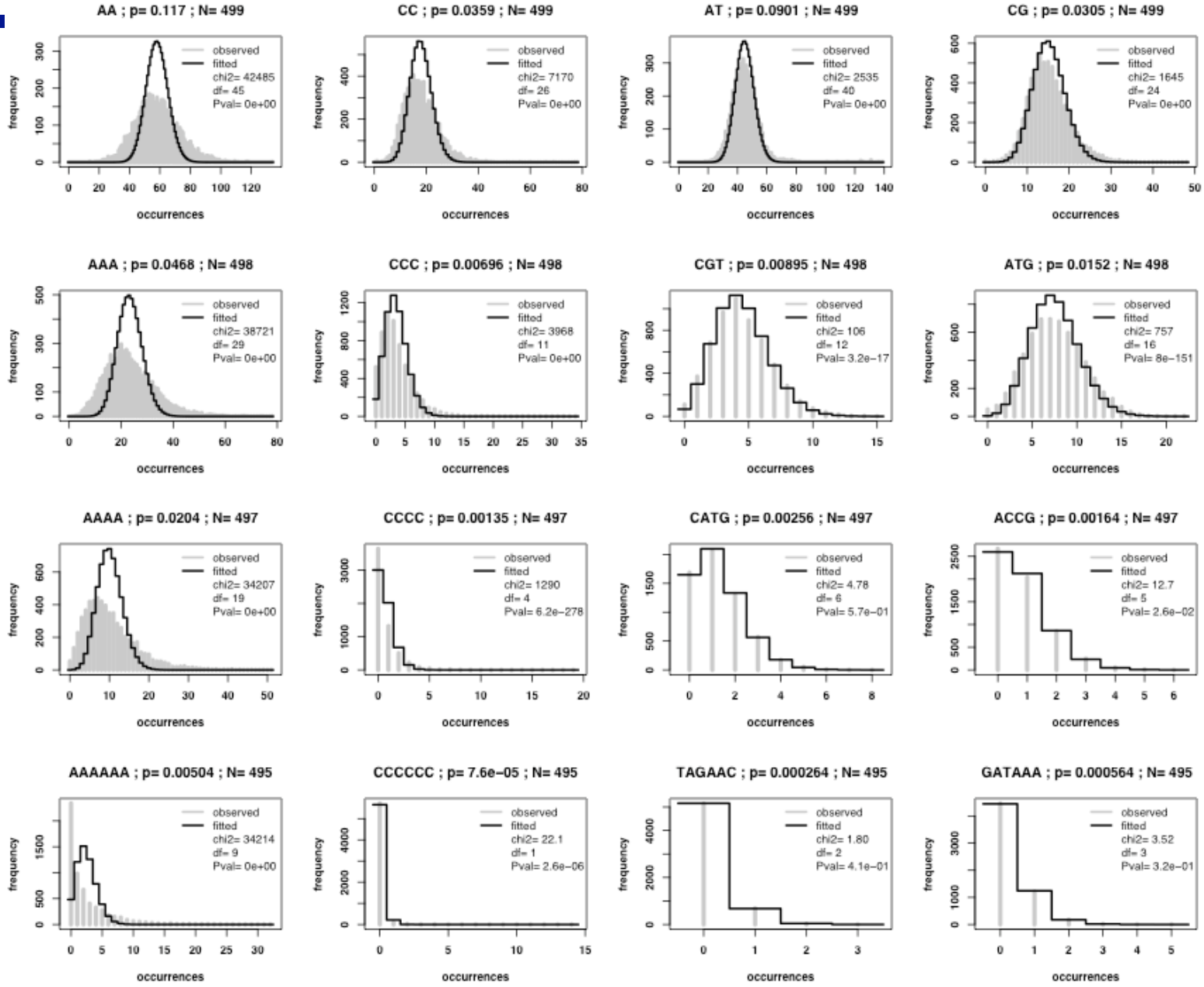
- The only parameter of the distribution is the number of degrees of freedom (df).
- When df increases, the chi2 tends towards a normal distribution.



Conditions of applicability

- Expected and observed values must be absolute frequencies
- Expected values must be > 1 for each class
- Less than 20% observed values < 5

Binomial fitting for word distributions in yeast promoters



Statistics Applied to Bioinformatics

The Kolmogorov-Smirnov test

Jacques van Helden
Jacques.van.Helden@ulb.ac.be

Kolmogorov-Smirnov - applicability

- One restriction to the χ^2 test is that the expected frequency of each class should be sufficient ($n_{exp.i} \geq 5$). When the sample is too small to fit this requirement, the Kolmogorov-Smirnov (KS) test can be applied.
- The KS test can only be applied for continuous and entirely defined distributions.
  - It is thus appropriate for normal, but not for binomial or Poisson fitting.
  - When parameters (e.g. mean, standard deviation) have to be estimated from the sample, the critical values must be adapted.
- The test consists in calculating the cumulative distribution of the sample, and comparing It to the theoretical cumulative distribution. For each class or value, the difference between observed and expected values is calculated, and the maximal difference is retained.
- This maximal difference is then compared with pre-defined tables.

Statistics Applied to Bioinformatics

Wilk-Shapiro normality test

Jacques van Helden
Jacques.van.Helden@ulb.ac.be

The Wilk-Shapiro test

- The Wilk-Shapiro test is specific for testing the normality of a sample. It has good power properties.
- It is based on a weighted sum of the sorted values (X'), where a specific weight (w_i) is associated to each value according to its rank.
- The weights are calculated on the basis of a standard normal distribution.
- The W statistics is then compared to critical values. Small values of W are evidence of departure from normality

$$W = \frac{\left(\sum_{i=1}^{n} w_i X_i^{'}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

$$w' = (w_1, w_2, \ldots, w_n) = MV^{-1}\left[\left(M'V^{-1}\right)\left(V^{-1}M\right)\right]^{-1/2}$$

W	Wilk statistics
X_i	values
X'	sorted values
w_i	weights
M	expected values of standard standard normal order statistics for a sample of size n

More info:
http://www.itl.nist.gov/div898/software/dataplot/refman1/auxillar/wilkshap.htm
http://www.itl.nist.gov/div898/handbook/prc/section2/prc213.htm