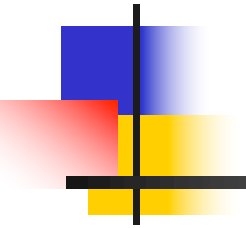


Input Modeling

Input models

- Input models have a fundamental role in simulation, since they determine the system behavior and outputs
- Examples:
  - Queueing systems: arrival process and service times
  - Processors: generation process for new jobs, jobs duration
  - Reliability analysis: times between failures, lifetime for components, repair times
- There are two approaches
  - Trace-driven simulations
  - Input models from general assumptions
  - Input models from representative data

Trace-Driven Simulation

- In some situations we may choose to substitute the stochastic input process with sets of real collected data
- Simulators based on the usage of real input data sets are called *trace-driven*
- A trace is a time ordered set of records on events, containing all the relevant data associated to the event
- Example: Queueing system
  - Observation of the arriving clients
  - For each client, we record the arrival time, the service duration and any other possible relevant information
  - In the simulation, we will use the collected arrival times A_i and service durations S_i to reproduce the system behavior

Trace-Driven Simulation

- Advantages
  - Input credibility, with respect to the hypothesis on the data distributions and their parameters
  - It is possible to validate the simulator by direct comparison of the outputs with the behavior of the real system
- Disadvantages
  - Collection and storage of several traces is expensive (in time and memory space)
  - The complexity of the simulator may increase due to the operations to read/manage the stored traces
  - Each trace may not be suitable to represent the system in situations different from the ones in which the data were collected
  - Trace duration may be too short to show rare system events (which often are what we are interested in)

Input models

1. Data collection from the real system
   - Costly, it takes time and resource commitments
   - Sometime the real system cannot be measured (e.g. it is not completely accessible to measurements)
2. Identification of a statistical distribution suitable to represent the observed data
3. Selection of the parameters for the distribution (possibly estimating them from the real data)
4. Goodness-of-fit test: if it fails, go back to step 2 and choose a new distribution

Identifying a distribution

- The aim is to identify a suitable distribution to represent the collected data
  - Build an empirical distribution from the data
  - Compare the empirical distribution with well known distributions
  - Choose the distribution better representing the empirical one

Possible families of distributions

Criteria to choose a distribution:

- **Binomial**: number of successes in *n* independent experiments, each one with success probability *p*
- **Geometric**: number of trials needed before a success in an experiment with success probability *p*
- **Poisson**: number of independent events occurring in a time interval with fixed duration

Possible families of distributions

- **Normal**: distribution for a process that could be described as the sum of independent components
- **Lognormal**: distribution for a process that could be described as the product of independent components
- **Exponential**: independent times between events, process without memory
- **Gamma**: extremely flexible distribution for non-negative random variables

Possible families of distributions

- **Beta**: extremely flexible distribution for random variables with limited support
- **Erlang**: distribution for a process that could be described as a sum of exponentials (a special case of the Gamma)
- **Weibull**: distribution for random variables with high variance (*heavy tail*)
- **Pareto**: distribution for random variables with high variance (*heavy tail*)

Possible families of distributions

- **Uniform (continuous and discrete)**: distribution representing elements with the same probability, it can be used when there is little information on the distribution for the input data
- **Empirical**: used when no known distribution provides a suitable fitting

Histograms

- They are used to observe the shape of a distribution

1. Divide the range of the data into intervals, usually all with the same length
2. Find the frequency of occurrences within each interval
3. Plot the graph that represents the probability density

Histograms

- The number of class intervals (and their width) depends on the number of observations and on the amount of dispersion or scatter in the data
- For continuous r.v., a rule of thumb is:
  Choose the number of class intervals equal to the square root of the sample size
- For discrete r.v., if the number of observations is large enough, choose a cell for each possible value. If there are few data points, adjacent cells should be joined to avoid the ragged appearance of the histogram

Quantile-quantile (q-q) Plot

- It provides a visual indication of how well a sample of data fits a distribution
- It is based on the distribution *quantiles*
- If X is a r.v. with CDF F(x), the *q*-quantile of X is that value *y* such that $F(y)=P(X \leq y)=q$, for $0<q<1$
- The *q*-quantile is therefore $y=F^{-1}(q)$

Quantile-quantile (q-q) Plot

- Given a r.v. X with CDF F(x) and given a sample of data from X

$$\{x_i, i=1, 2,\ldots, n\}$$

- The same data, ordered from the smallest to the largest, form the sequence

$$\{y_j, j=1, 2,\ldots, n\}$$

with $y_j \leq y_{j+1}$

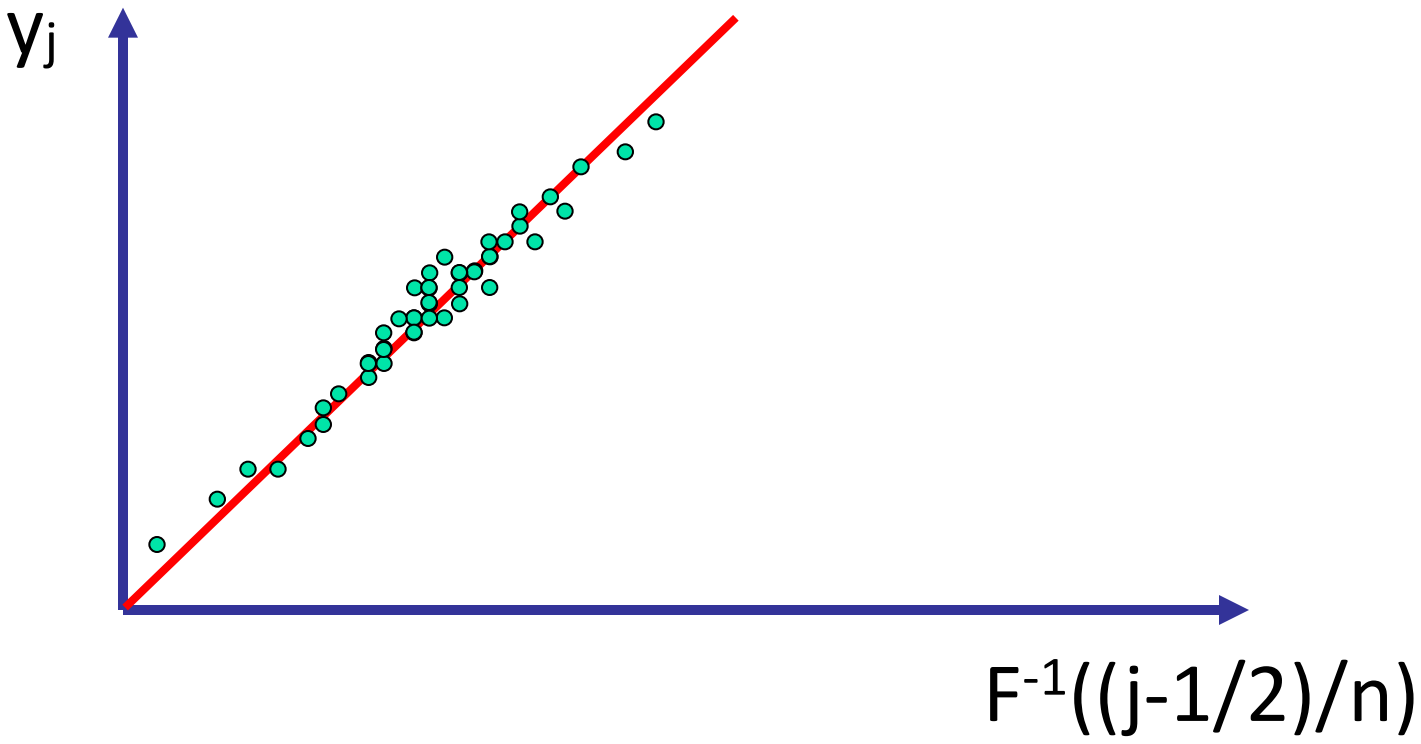
- y_j is an estimate of the (j-1/2)/n quantile of X

$$y_j \approx F^{-1}\left(\frac{j-1/2}{n}\right)$$

Quantile-quantile (q-q) Plot

- If we have chosen a distribution with CDF F(x) as a possible representation of a sample x_i of collected data, then a plot of the points defined by
  - Abscissa: $F^{-1}((j-1/2)/n)$
  - Ordinate: y_j
- is *approximately a straight line* if F(x) appropriate
- deviates in a systematic way from a straight line if F(x) is inappropriate

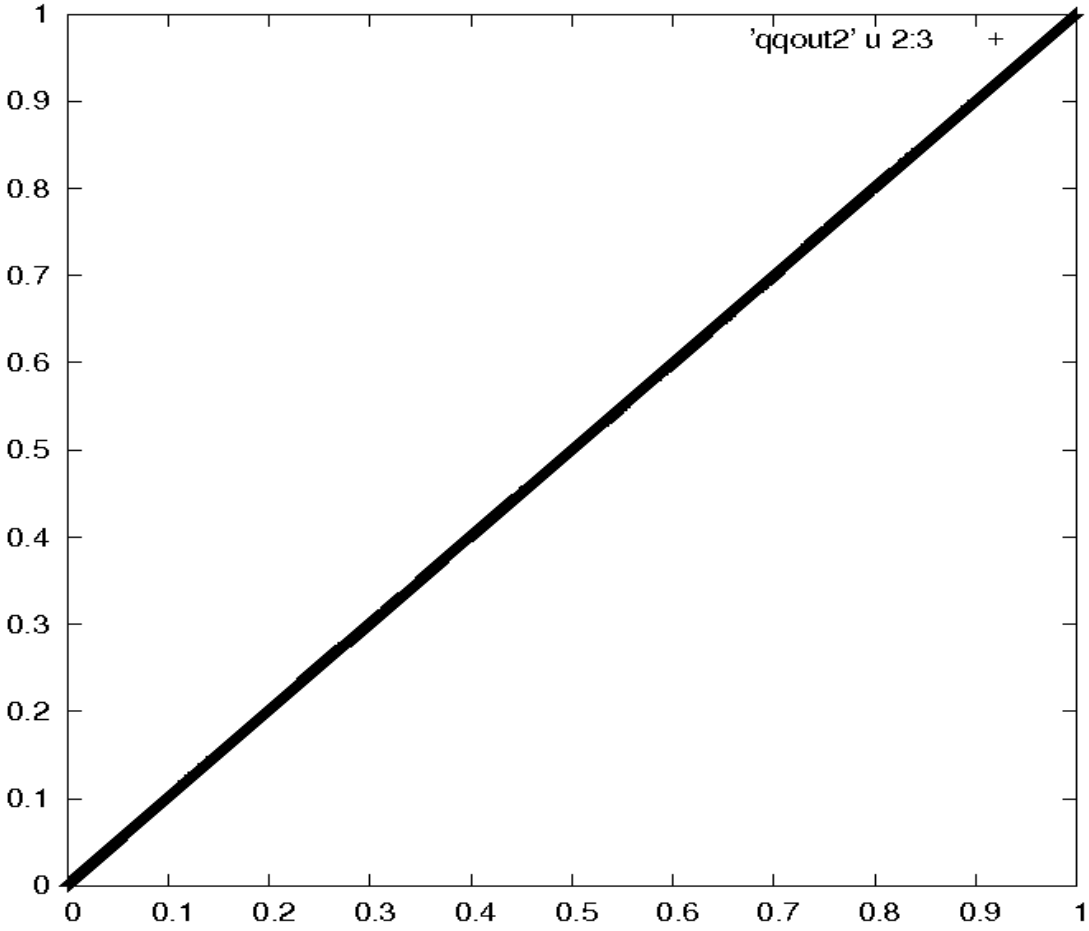
Quantile-quantile (q-q) Plot



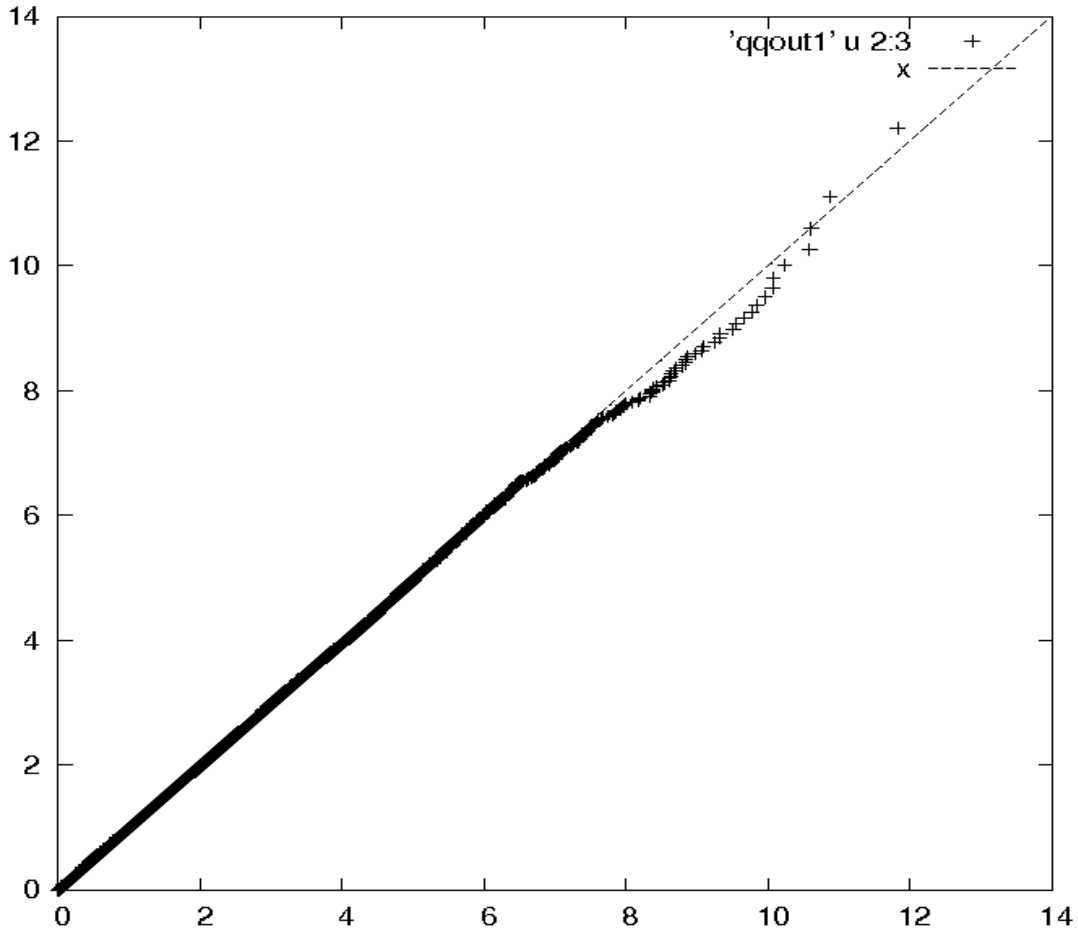
Quantile-quantile (q-q) Plot

- Values will never fall exactly on a straight line (even if X is exactly distributed as F(x), since it is a stochastic experiment)
- Adjacent values tend to stay consistently either above or below the ideal straight line
- Values at the extremes tend to dispersion (due to the variance), so they might diverge from the straight line

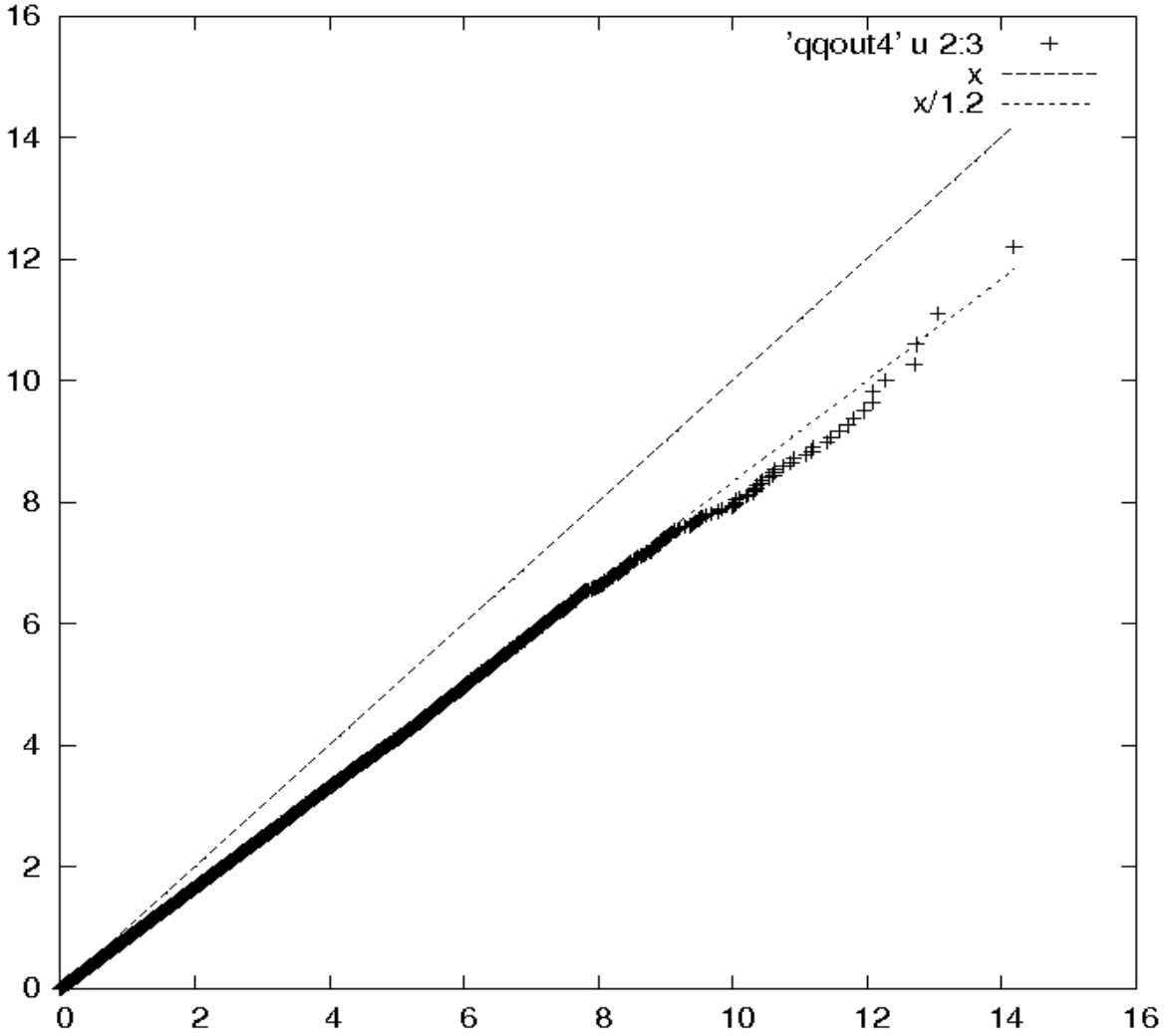
Example: Uniform



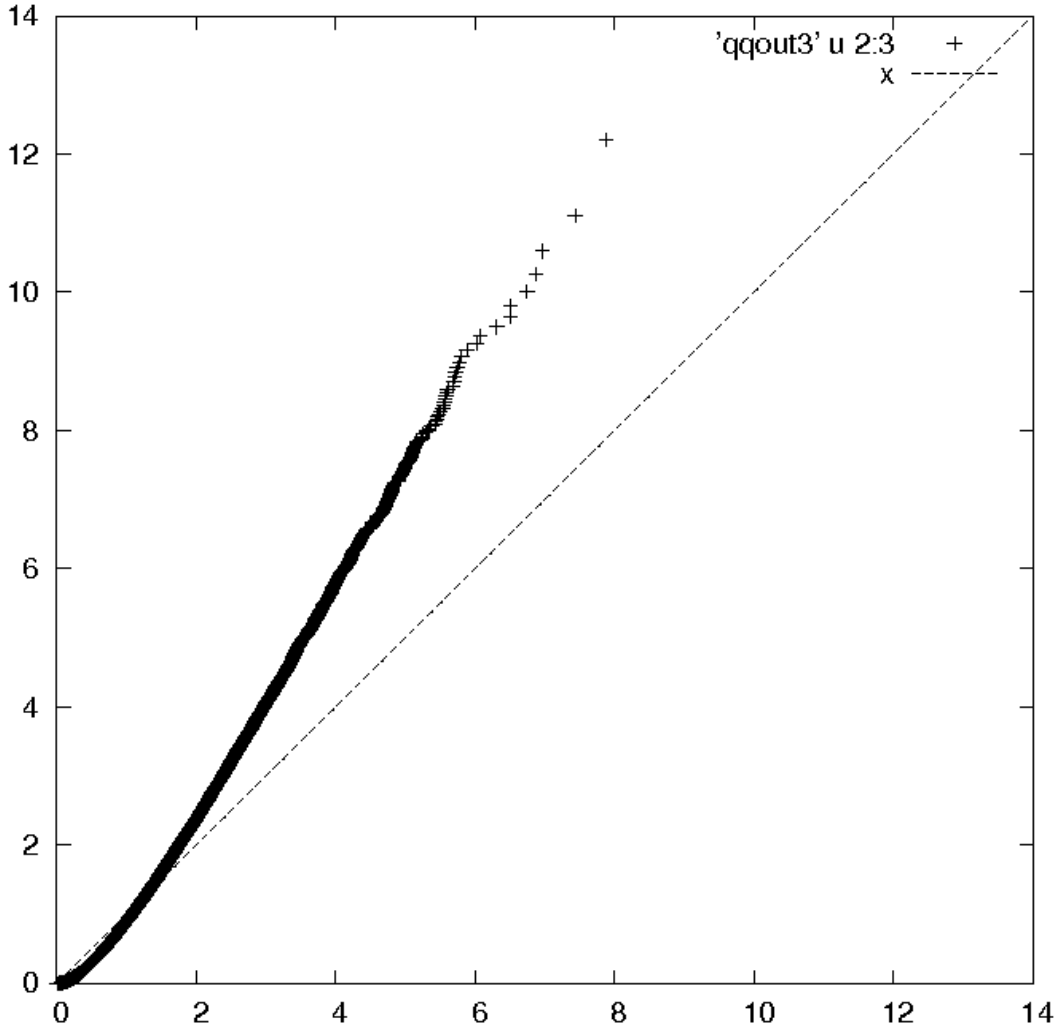
Example: Exponential



Setting parameters



Example: Exp vs Erlang



Parameter estimation

- After a family of distributions has been selected, the next step is to estimate the parameters of the distribution
- The most commonly used estimators for the parameters are the *sample mean* and, if needed, the *sample variance* (or any other higher order moment from the sample data)

Parameter estimation

- Given the set of observations

  $\{X_i, i=1, 2,\ldots, n\}$, estimators for the sample media and variance are defined as

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$S^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}{n-1}$$

Parameter estimation

- If the data are discrete and have been grouped in a frequency distribution with k distinct values, each one with frequency f_i, the estimators are defined as

$$\overline{X} = \frac{\sum_{i=1}^{k} f_i X_i}{n} \qquad S^2 = \frac{\sum_{i=1}^{k} f_i X_i^2 - n\overline{X}^2}{n-1}$$

where *n* is the total number of samples

Parameter estimation: examples

- **Poisson**: the parameter *a* is chosen equal to the measured mean,

$$a = \overline{X}$$

- **Exponential**: the parameter λ (rate) is selected as the inverse of the mean

$$\lambda = 1/\overline{X}$$

- **Normal**: we use the estimated mean and variance
- For more complex distributions (Weibull, Beta, Gamma), a maximum-likelihood estimator might be needed to determine the distribution parameters

Alternative approaches

- Maximum Likelihood Estimation (MLE) of θ:
  - Find the parameter(s) that maximize(s) the probability of generating such empirical data:

$$\theta^* = \arg\max(L(x_i \mid \theta))=$$

$$= \prod_i f_X(x_i|\theta)=\arg\max \sum_i \log(f_X(x_i|\theta))$$

  - Therefore under mild assumptions θ^* satisfies:

$$\sum_i \frac{d}{d\theta}\log(f_X(x_i|\theta))=0$$

Goodness-of-fit Test

- After selecting the distribution (with its parameters) representing the measured data, we can apply hypothesis tests to verify the goodness-of-fit
- The hypotheses to be verified is that the selected distribution provides a reasonable representation of the measured data
- There are tests, like the *chi-square test*, to check the goodness-of-fit

Chi-square test

- Given a hypothesis H_0 (e.g., data are sampled from a given distribution), we define a level of significance α of the test as

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ true})$$

  - H_0 is called *null hypothesis*
  - Frequently, α is set to 0.01, 0.05, or 0.10

- When H_0 is true, the test fails with probability α
- A statistical test should be repeated several times K (and would fail αK times)
- Positive test: *we have no evidence that H_0 is false*

Chi-square test

1. Define *n* intervals in the r.v. support, with interval *i* defined by (a_i, a_{i+1})

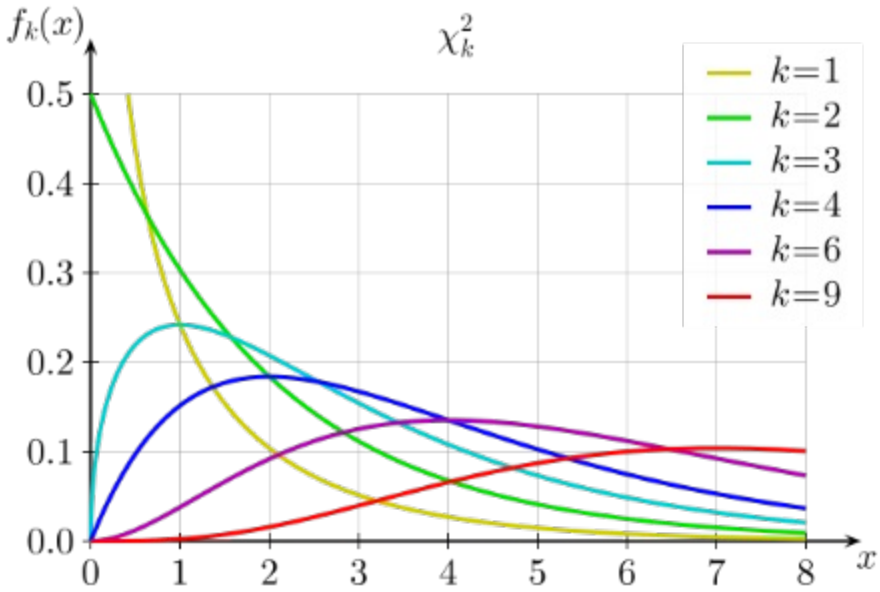
2. Generate *N* instances and compute the number O_i of instances falling in the *i*-th interval and compare it with the expected value, E_i,

$$E_i = N[F(a_{i+1}) - F(a_i)]$$

Chi-square test

3. Compute

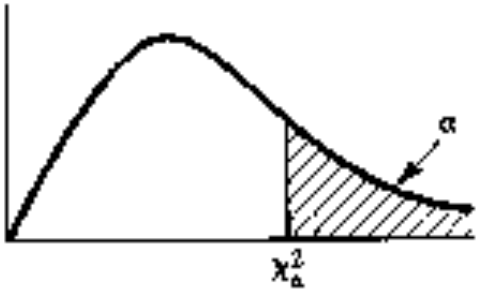
$$X = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$



4. If the hypothesis is true and N is large, *X* is distributed according to a chi-square distribution with *n-s-1* degrees of freedom, where *s* is the number of parameters estimated from the data

Chi-Square

Table A.6. Percentage Points of the Chi-Square Distribution with ν Degrees of Freedom



ν	$\chi^2_{0.995}$	$\chi^2_{0.01}$	$\chi^2_{0.025}$	$\chi^2_{0.05}$	$\chi^2_{0.10}$
1	7.88	6.63	5.02	3.84	2.71
2	10.60	9.21	7.38	5.99	4.61
3	12.84	11.34	9.35	7.81	6.25
4	14.96	13.28	11.14	9.49	7.78
5	16.7	15.1	12.8	11.1	9.2
6	18.5	16.8	14.4	12.6	10.6
7	20.3	18.5	16.0	14.1	12.0
8	22.0	20.1	17.5	15.5	13.4
9	23.6	21.7	19.0	16.9	14.7

Chi-square test

- Comparing X and the chi-square distribution with n-s-1 degrees of freedom, we decide if accepting or rejecting the null hypothesis H_0
- If $X>\chi^2_{n\text{-}s\text{-}1,\alpha}$

  reject the null hypothesis H_0 with level of significance α
- The values for $\chi^2_{n\text{-}1,\alpha}$ are known and usually tabulated

Chi-square χ^2

- The distribution χ^2 (chi-square) with *k* degrees of freedom is obtained when *k* independent variables with normal distribution N(0,1) are squared and summed
- The expression for the pdf is

$$f(x) = \frac{e^{-\frac{x}{2}} x^{\frac{k}{2}-1}}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)} \qquad x \geq 0$$

where $\Gamma(\cdot)$ is the Gamma function defined as

$$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt$$