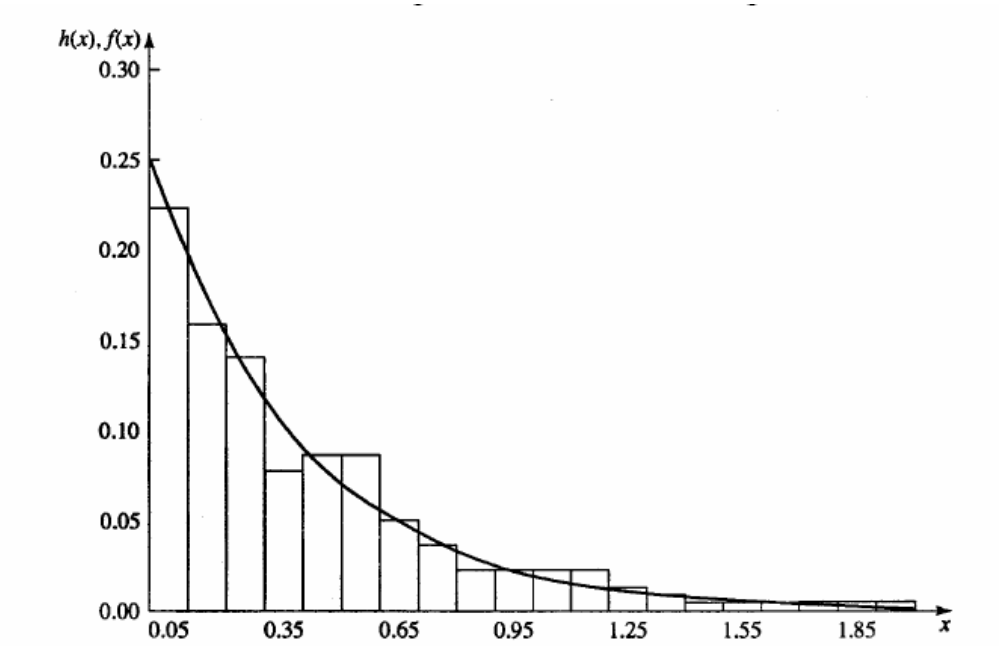
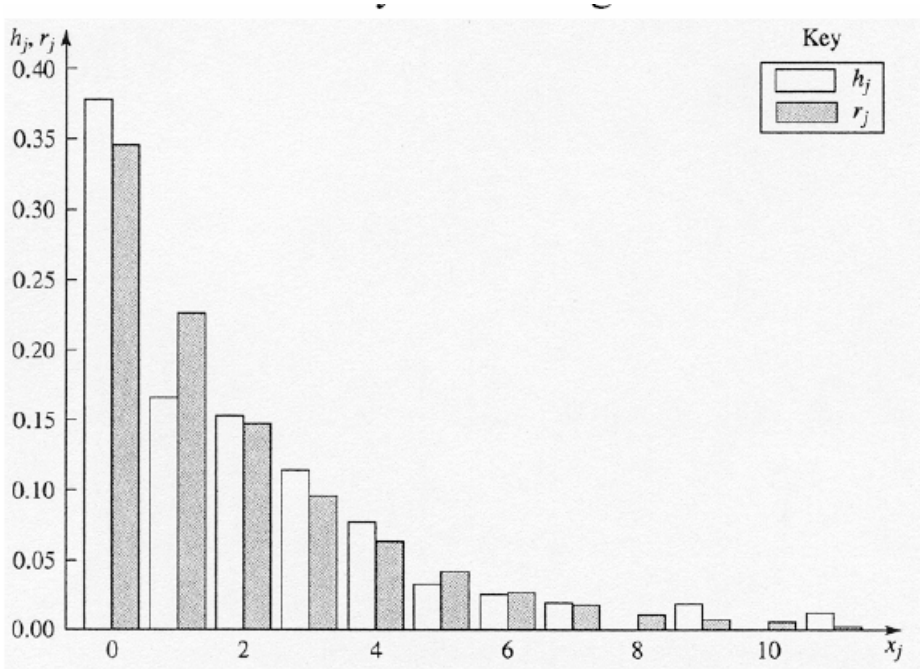
Selecting Input Probability Distributions 2
(Chapter 6, Law)

- **Activity III: Determining How Representative the Fitted Distributions Are**
  - Having hypothesized a family of distributions and estimated parameters, the final activity is to determine whether the hypothesized distribution is a good fit.
  - The main question here is: Does the fitted distribution agree with the observed data?
  - There two approaches to answer this question: Heuristic and formal statistical tests.

- **Heuristic approaches for goodness of fit**
  - For continuous data a *density/histogram* plot is used. The data histogram, with step Δb, is plotted on the same graph with $\Delta b\hat{f}(x)$, where $\hat{f}(x)$ is the hypothesized density.
  - This allows visual inspection of the goodness of fit.
  - Suppose data is arranged in intervals $[b_j, b_{j-1}), j = 1, \ldots, k$. Let h_j be the proportion of data in interval $[b_j, b_{j-1})$.
  - A *frequency comparison* works by graphically comparing h_j and $r_j = \int_{b_{j-1}}^{b_j} \hat{f}(x)dx$.



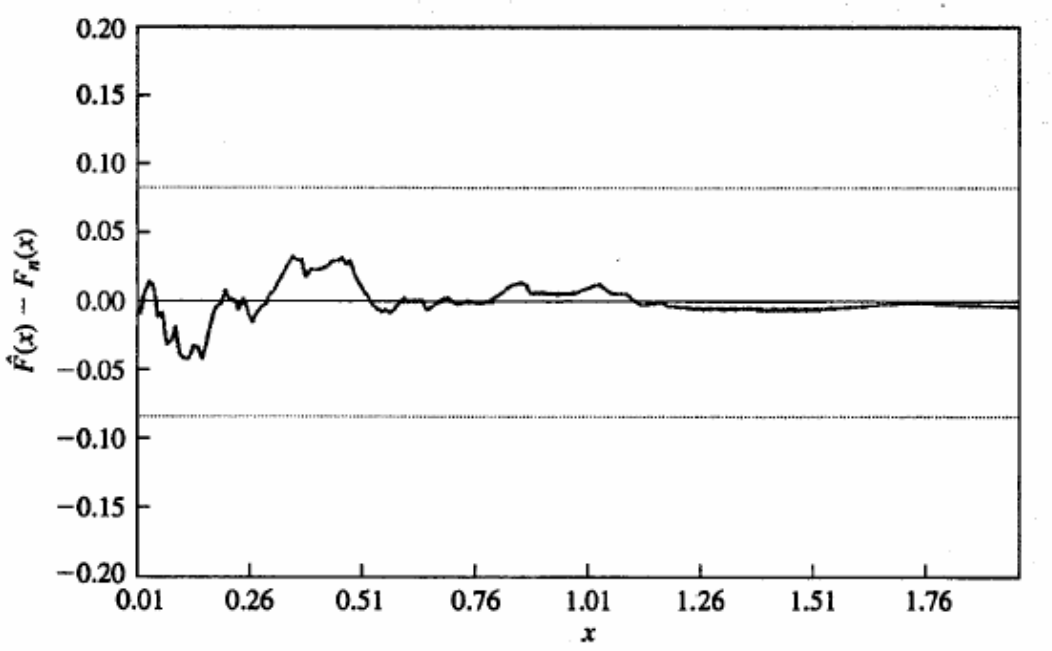
- For discrete distributions, a *frequency comparison* is done by comparing the number of data observations equal to x_j, h_j, with the fitted probability mass function $\hat{p}(x_j)$.



- The fitted distribution function, $\hat{F}(x)$, with an empirical distribution function based on *n* observations, $X_1, \ldots, X_n$, is

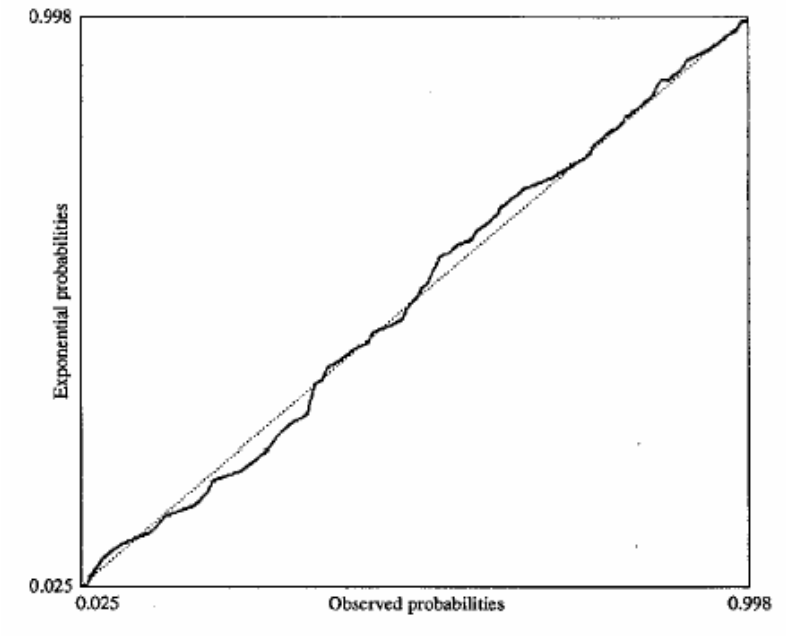
$$F_n(x) = \frac{\text{number of } X_i\text{'s} \le x}{n}.$$

- Since $\hat{F}(x)$ and $F_n(x)$ have similar shapes in most cases, a *distribution function difference plot* for $\hat{F}(x) - F_n(x)$ is used.

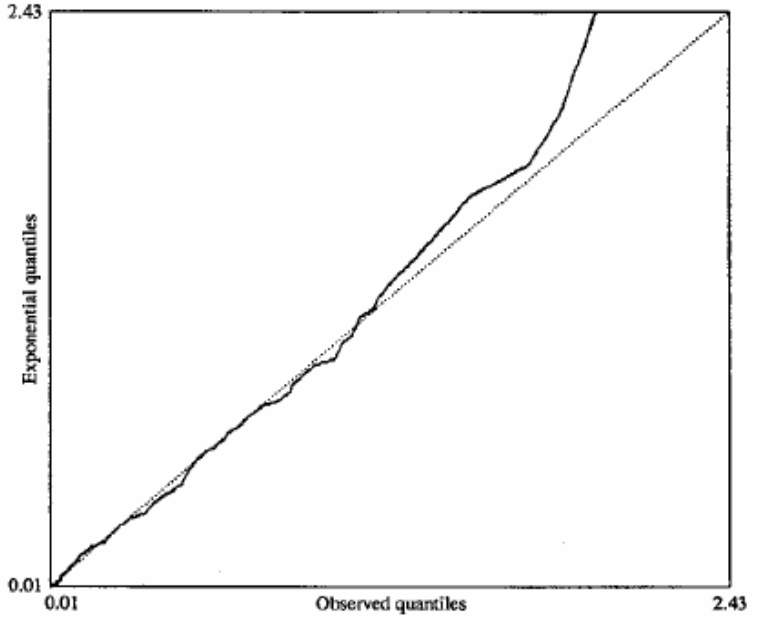


- Another way of comparing empirical and fitted distribution functions is through *probability plots*.
- To do a probability plot, the data is arranged such that $X_1 < X_2 < \ldots < X_n$. Then, an empirical distribution function is developed as $\tilde{F}_n(X_i) = (i - 0.5)/n$.

- A *P-P plot* is a plot of pairs $(\hat{F}(X_i), \tilde{F}_n(X_i))$. A straight line indicates a perfect fit.
- *P-P plots* are sensitive to misfits in the center.



- A *Q-Q plot* is a plot of quantile pairs $(\hat{F}^{-1}((i-0.5)/n), X_i)$
- *Q-Q plots* are sensitive to misfits in the tails.



- **The χ^2 goodness of fit test**
  - This is similar to the χ^2 test we saw for $U(0,1)$.
  - This test works as follows.
    - Given n data points, divide range of data into k intervals, $[a_0, a_1)$, $[a_1, a_2)$, ..., $[a_{k-1}, a_k)$.
    - Count the number of observations that fall in interval j, $N_j, j = 0, 1, \ldots, k-1$.
    - Find the expected number of observations in each interval, $E_j = np_j$, where $p_j = \hat{F}(a_j) - \hat{F}(a_{j-1})$.
  - This test is then performed as follows.
    - H_0: X_i's are iid with distribution function $\hat{F}(x)$
    - H_a: X_i's are not iid with distribution function $\hat{F}(x)$
    - Test Statistic

$$\chi^2 = \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j}$$

    - Rejection region: Reject H_0 if $\chi^2 > \chi^2_{k-1,\,1-\alpha}$.
  - As a guideline the intervals, $[a_{j-1}, a_j)$, are selected based on an *equiprobable approach*, i.e., $p_1 = p_2 = \ldots = p_k$, and such that $np_j \geq 5$.

- **Example (6.15, Law)**
  - An exponential distribution with $\hat{F}(x) = 1 - e^{-x/0.399}$ was fitted to 219 inter-arrival time observations (p. 322 Law).
  - To perform the χ^2 test $k = 20$ intervals are used with an *equiprobable approach* having $p_j = 1/20$.
  - Then, setting $a_0 = 0$, and $a_{20} = \infty$, $a_j, j = 1, 2, \ldots, 19$ are found such that $\hat{F}(a_j) = j/20$, which implies that $p_j = \hat{F}(a_j) - \hat{F}(a_{j-1}) = 1/20$. Then, the a_j's are found by inverting $\hat{F}(x)$ as

$$a_j = -0.399 \ln(1 - j/20).$$

  - See Example 6.15 Law for the details of the test.

- **The Kolmogorov-Smirnov goodness of fit test**
  - This can be seen as a formal comparison between the empirical and fitted distribution functions, $F_n(x)$ and $\hat{F}(x)$.
  - It has the advantage of not requiring grouping the data into intervals and being valid for any sample size over the χ^2 test.
  - However, it's not as general as χ^2.
  - H_0 and H_a are the same as for K-S are the same as for χ^2.

- Assume that data is arranged such that $X_1 < X_2 < \ldots < X_n$. Then, $F_n(X_i) = i/n$.
- The test statistic for KS is AD_n an adjustment of $D_n = \max(D_n^+, D_n^-)$, where

$$D_n^+ = \max_{i=1,\ldots,n}\left(i/n - \hat{F}(X_i)\right), \quad D_n^- = \max_{i=1,\ldots,n}\left(\hat{F}(X_i) - (i-1)/n\right).$$

- E.g., for $U(0,1)$, $AD_n = \left(\sqrt{n} + 0.12 + 0.11/\sqrt{n}\right)D_n$.
- H_0 is rejected (implying that there is not enough evidence of a good fit) if D_n is too large.
- Critical values for deciding how large is a large D_n are not available for all distributions.
- Table 6-15 (Law) lists critical values for "all parameter known" (which can be used for U(0,1)), normal and exponential distributions.

- **Example**
  - Use K-S test to check if the following data is iid distributed as U(0,1). Use $\alpha = 0.05$.

    0.05, 0.14, 0.44, 0.81, 0.93

  - In this cases, $\hat{F}(X_i) = X_i$. The TS is found as follows.

i	1	2	3	4	5
X_i	0.05	0.14	0.44	0.81	0.93
i/n	0.2	0.4	0.6	0.8	1
$i/n - X_i$	0.15	0.26	0.16	−0.01	0.07
$X_i - (i-1)/n$	0.05	−0.06	0.04	0.21	0.13

- Then, $D_n^+ = 0.26$, $D_n^- = 0.21$, $D_n = 0.26$, $AD_n = 0.63 < c_{1-\alpha} = 1.358$ (Table 6.15, Law). Do not reject H_0.

- **Selecting a distribution in the absence of data**
  - If no data is available (e.g. system under study is not functional yet or access to explicit data is difficult), a distribution can be fitted based on subjective estimates.
  - One can ask "experts" for min, max, most likely (mode), and median values.
  - Depending on what subjective estimates are available, the following distributions can be used.

Available information	Distribution
min, max	Uniform
Mean (unbounded)	Exponential
min, max, mode	Triangular
min, max, mean, mode	Beta

  - See Law (Section 6.11) for details on how to select parameter for these distributions.