MODEL SELECTION

- Our goal is to answer to the following questions:
  - Given a model, how can we conclude that it is "good enough" to use?
  - Having 2 or more "good enough" models, how to choose the "better" model?
- There is a lot of literature on model selection but we will confine ourselves to a general framework.
- What is "good enough"? The answer depends on the raised problem.
  - Models are always simplified representations of reality. A model as complex as reality is useless.
  - We want to capture the main of data using a model as simple as possible (parsimony principle or Occam's razor).
  - Technically speaking, there are two ways to analyze a model adequacy: Graphical representation and **testing procedures**.

REPRESENTATIONS OF THE DATA AND MODEL

- Basic idea: To compare the collected data with the proposed model(s)
- The proposed model can be represented by its distribution function or by its density (or probability) function
- The collected data can be represented by
  - The empirical cumulative distribution function
  - An histogram
  - Observed values (discrete data)
- **Important point**: Observed data can be censored and truncated. If so, we need to "correct" the distribution of the proposed model: $F(x) \rightarrow F^*(x)$.

  For instance, if $X \sim F(x)$ and the observed values are truncated at point t, the distribution function of the observed value is $F^*(x) = \begin{cases} 0 & x \le t \\ \dfrac{F(x)-F(t)}{1-F(t)} & x > t \end{cases}$. Remember that the density of the observed values is $f^*(x) = \dfrac{f(x)}{1-F(t)}$ for $x > t$ and that $F^*(x) = \int_t^x f^*(u)\,du$

GRAPHICAL COMPARISONS (see Loss Models)

- How to choose an adequate graphical procedure for observed data?
  - Discrete versus continuous variables
  - Grouped versus individual data
  - Sample size (namely for histograms)
- Graphical comparison
  - Plot the observed data and the proposed model on the same graph
  - Sometimes (namely when using the ecdf) it is more readable to plot $D(x) = F_n(x) - F^*(x \mid \theta)$
  - Use a P-P plot
- **P-P plot**: plot $F_n(x)$ against our parametric estimate of $F*(x)$. If the model fits well, the plotted points will be near the 45º line running from (0,0) to (1,1). Some authors refer that, as $E\left(F(X_{(j)}\right) = j/(n+1)$, where $X_{(j)}$ is the *j*-th order statistic in the sample, it is preferable to use $F_n^*(x) = \frac{n}{n+1} F_n(x) = \frac{j}{n+1}$ (assuming no ties) instead of $F_n(x)$.

  Challenging question: Can you prove that $E\left(F(X_{(j)}\right) = j/(n+1)$?

- **Example 16.1 and 16.2** (partially) – Consider Data Sets **B** and C. For this example and all that follow, in Data Set B replace the value 15743 with 3476 (to allow the graphs to fit comfortably on a page). Truncate Data Set B at 50 and Data Set C at 7500. Estimate the parameter of an exponential model for each data set. Plot the appropriate functions and comment on the quality of the fit of the model. Repeat this for Data Set B censored at 1000 (without any truncation). Example 16.2 → Plot $D(x)$

  Case 1 – Data Set B with truncation at *t*=50.

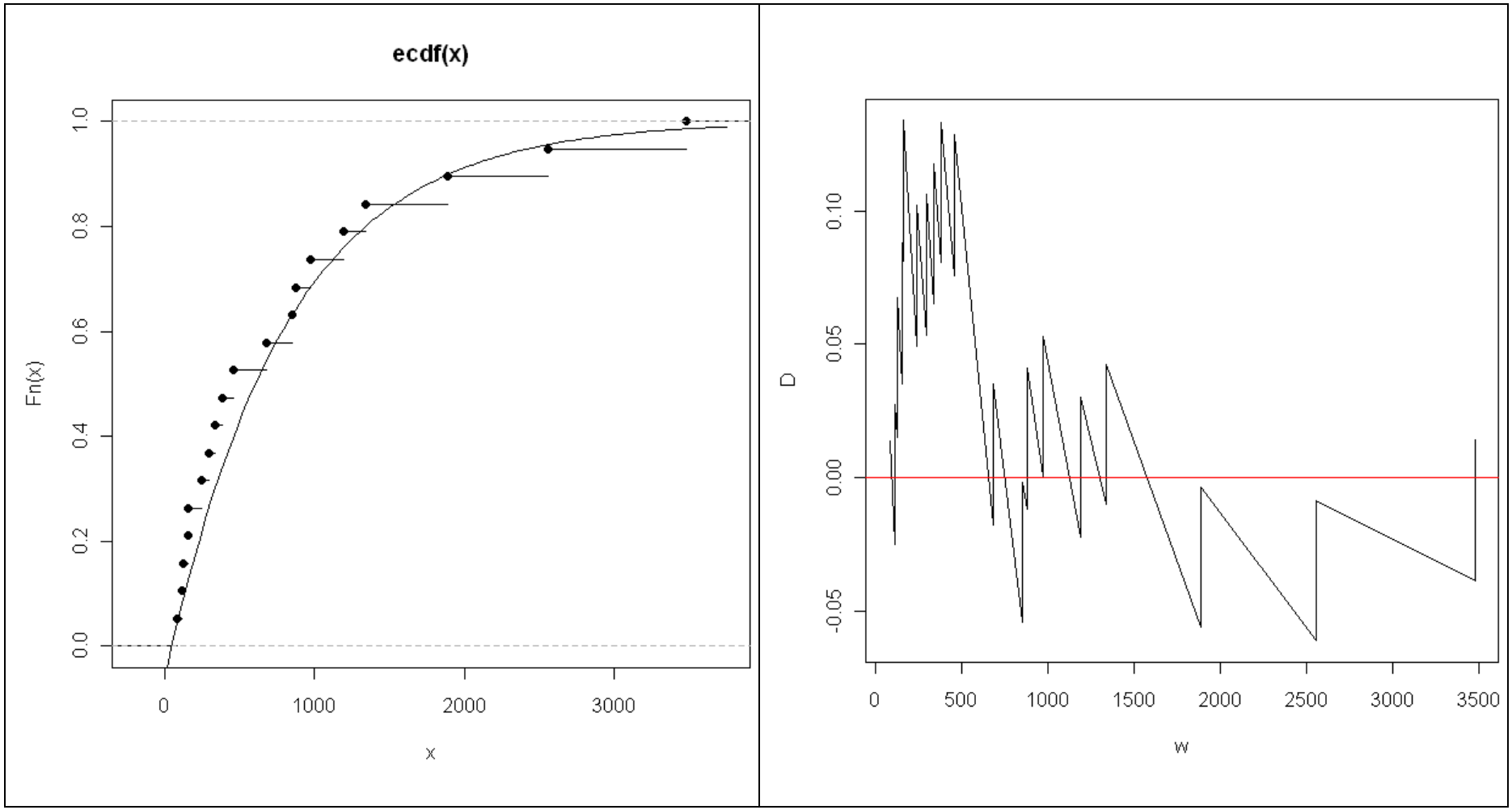
$$f^*(x\,|\,\theta) = f(x\,|\,x>t) = f(x)/(1-F(t)) = \theta^{-1}e^{-(x-t)/\theta},\ x>t$$

$$\ell(\theta\,|\,\mathbf{x}^*) = \sum_{i=1}^{19}\ln f^*(x_i\,|\,\theta) = \sum_{i=1}^{19}\left(\ln f(x_i\,|\,\theta) - \ln(1-F(t)\,|\,\theta)\right) = \sum_{i=1}^{19}\left(-\frac{x_i}{\theta} - \ln\theta + \frac{t}{\theta}\right)$$

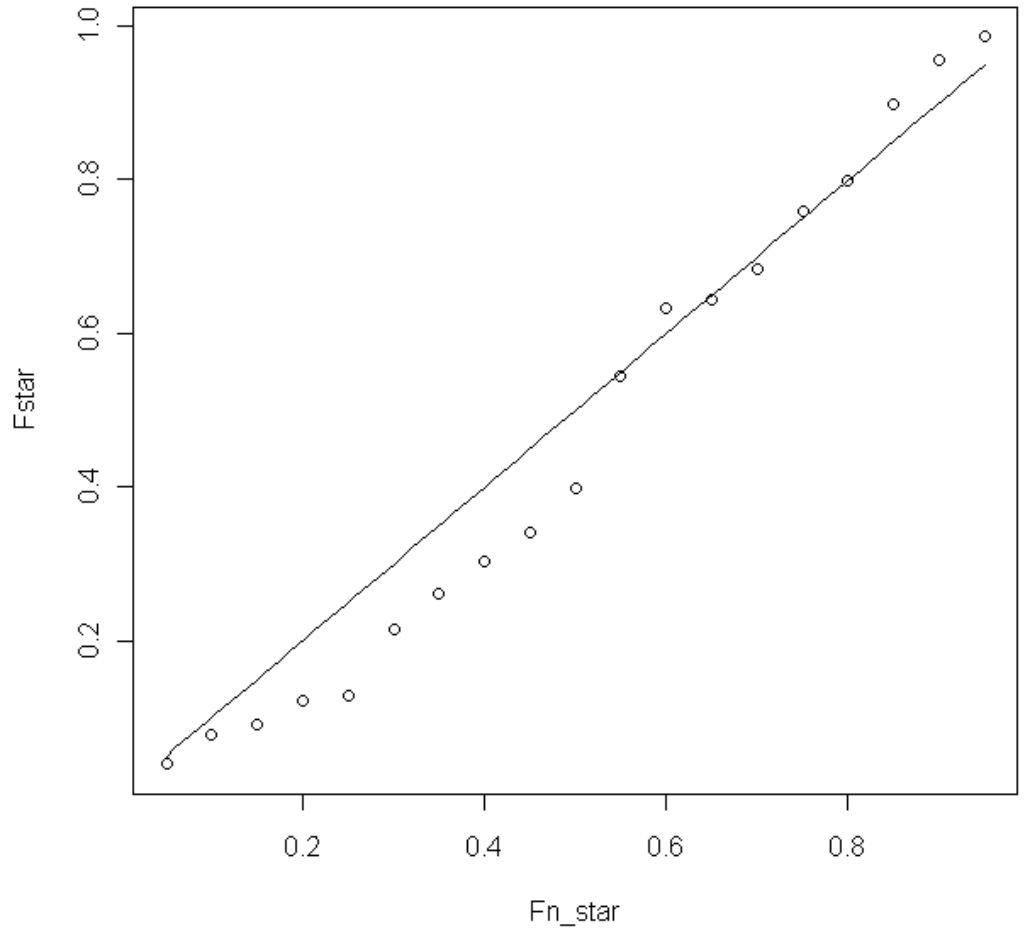
$$\ell'(\theta\,|\,\mathbf{x}^*) = \frac{n\bar{x}}{\theta^2} - \frac{n}{\theta} - \frac{nt}{\theta^2};\ \ell'(\theta\,|\,\mathbf{x}^*) = 0 \Leftrightarrow \frac{n\bar{x}}{\theta^2} - \frac{nt}{\theta^2} = \frac{n}{\theta} \qquad \text{then } \hat{\theta} = \bar{x} - t \text{, i.e. } \hat{\theta} = 802.3158$$

$$\hat{F}^*(x) = \frac{\hat{F}(x) - \hat{F}(t)}{1-\hat{F}(t)} = \frac{e^{-50/802.3158} - e^{-x/802.3158}}{e^{-50/802.3158}} = 1 - \exp\left(-\frac{x-50}{802.3158}\right) \qquad x>50$$

  As the sample size is small we use the ecdf.



The model seems to understate the distribution function at smaller values of x .



F^*- For each observed value x_j, calculate $F^*(x_j)$

F_n * - The empirical values are given by $F_n(x_j) = j/n,\ j = 1,2,\cdots,n$

Hypothesis Testing

- More accurately, one can test the hypothesis

  H0: The data came from a population with the stated model

  H1: The data did not come from such population

- The test statistic is usually a measure of how close are the data from the distribution specified in the null hypothesis

- H0 can be:
  - A simple hypothesis, i.e. the null completely specifies the distribution. This is the most adequate situation for many testing procedures (critical values for the tests can be deduced) but rarely happens in practical situations;
  - A composite hypothesis, i.e. the null specifies a family of distributions and some unknowns parameters remain (and have to be estimated before computing the test).
    When we use the same sample to estimate unknown parameters and to compute the test, the test statistics tends to be smaller than it would be using pre-specified parameters since the parameter estimates are “optimized” (they are chosen to fit as closely as possible the data). For a given level of the type I error probability, we will under reject H0 unless some correction is made.

- Sometimes when we have a large sample we can randomly split our sample in 2 sub-samples and use one of them to estimate the parameters and the other sub-sample to compute the test. However this is not a usual procedure.
- Among the many possible adjustment tests, we will discuss:
    - The **Kolmogorov-Smirnov** test;
    - The Anderson-Darling test (similar to KS approach but using a different test statistic – see Loss Models for a presentation of Anderson-Darling test)
    - The **chi-square goodness of fit** test.

Kolmogorov- Smirnov test

- The original goal of the Kolmogorov-Smirnov test is to test $H_0 : F(x) = F_0(x)$, $-\infty < x < \infty$ against $H_1 : F(x) \neq F_0(x)$ for some $x \in \Re$.
- $F_0(x)$ is the distribution function of a **continuous random variable** and all the distribution function parameters are specified (H0 is a simple hypothesis).
- The test statistic reflects the **maximum distance between the empirical cumulative distribution function** (which is a step function) **and the distribution function** (possibly corrected to allow for truncation and/or censoring), $F^*(x)$. This test can only be applied when the observed data are not grouped (we need to evaluate the empirical distribution function as well as possible).
- Let us define the test statistic as $D_n = \sup_{-\infty<x<\infty} \left| F_n(x) - F^*(x) \right|$. Note that in *Loss Models* the test statistic is referred as $D_n = \max_{-\infty<x<\infty} \left| F_n(x) - F^*(x) \right|$ which is not, strictly speaking, correct.
- The rejection region is obviously defined for D_n greater than a critical value.

- To get the distribution of the test statistic we use 2 theorems (proofs are omitted as they are complex)
  - **Theorem 1 –** The distribution of the test statistic D_n does not depend on $F_0(x)$.
  - **Theorem 2 –** For $z \geq 0$, $\lim_{n\to\infty} \Pr(\sqrt{n}\, D_n \leq z) = 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 z^2}$
  - Tables with the usual critical values for D_n are available for small values of *n* and with approximate values for larger n (for $\alpha = 0.10$, $1.22/\sqrt{n}$; for $\alpha = 0.05$, $1.36/\sqrt{n}$; for $\alpha = 0.01$, $1.63/\sqrt{n}$). We used a crude approximation given by $\sqrt{-(1/2)\times\ln(\alpha/2)}$

    Challenging question: Using theorem 2 obtain the approximation $\sqrt{-(1/2)\times\ln(\alpha/2)}$
- To get the value of the test statistic we must observe that $D_n = \sup_{-\infty<x<\infty} \left|F_n(x) - F^*(x)\right|$ has to be obtained at one of the observed values x_j. Then sort the observed value and compute

  $$D_n = \max_{i=1,2,\ldots n}\left[F^*(x_{(i)}^-) - F_n(x_{(i)}^-); F_n(x_{(i)}) - F^*(x_{(i)})\right] = \max_{i=1,2,\ldots n}\left[F^*(x_{(i)}) - (i-1)/n; (i/n) - F^*(x_{(i)})\right]$$

  Remember that we are assuming that $F^*(x)$ is continuous.
- Using theorem 2 we can approximate (for large samples) the *p*-value of D_n.

- **Example 16.4 and 16.5**– Calculate D for Example 16.1 (16.4) and complete the KS test (16.5)

  Since the data are grouped for case 2 we only can compute D_n for cases 1 and 3 (we only discuss case 1). For both situations, H0 is a composite hypothesis and we need to estimate the parameter of the exponential distribution. The results are

  Case 1

  From example 16.1 we get $\bar{x} = 852.31579$ $t = 50$ $\hat{\theta} = 802.31579$

  $D_n = 0.133952$ (details on next slide) critical value ($\alpha = 0.05$) $\approx 1.36 / \sqrt{19} = 0.312$

  Then we do not reject that the data came from an exponential distribution.

  3 comments:

  1) As we **estimated one parameter**, the test became conservative that is **we under reject H0**;
  2) When the ample size is small, it is difficult to get a rejection;
  3) Our critical value can be improved using a table of values of the KS test for small samples or using simulations techniques

i	x_i	F*(x_i)	F_n(x_i)=i/n	F_n(x_i-)=(i-1)/n	D+	D-
1	82	0.03910	0.05263	0.00000	0.01353	0.03910
2	115	0.07782	0.10526	0.05263	0.02744	0.02519
3	126	0.09038	0.15789	0.10526	0.06752	0.00000
4	155	0.12267	0.21053	0.15789	0.08786	0.00000
5	161	0.12921	0.26316	0.21053	0.13395	0.00000
6	243	0.21381	0.31579	0.26316	0.10198	0.00000
7	294	0.26223	0.36842	0.31579	0.10619	0.00000
8	340	0.30334	0.42105	0.36842	0.11772	0.00000
9	384	0.34051	0.47368	0.42105	0.13317	0.00000
10	457	0.39787	0.52632	0.47368	0.12845	0.00000
11	680	0.54398	0.57895	0.52632	0.03496	0.01767
12	855	0.63335	0.63158	0.57895	0.00000	0.05440
13	877	0.64327	0.68421	0.63158	0.04094	0.01169
14	974	0.68389	0.73684	0.68421	0.05295	0.00000
15	1193	0.75940	0.78947	0.73684	0.03007	0.02256
16	1340	0.79968	0.84211	0.78947	0.04242	0.01021
17	1884	0.89832	0.89474	0.84211	0.00000	0.05621
18	2558	0.95610	0.94737	0.89474	0.00000	0.06137
19	3476	0.98602	1.00000	0.94737	0.01398	0.03865

Using R

```
> x=c(82,115,126,155,161,243,294,340,384,457,680,855,877,974,1193,1340,1884,2558,3476)
>
> theta_hat=mean(x)-50
>
> trunc_expon_dist=function(x,theta,t) {
+   # x must be greater than or equal to t
+   (exp(-t/theta)-exp(-x/theta))/exp(-t/theta)
+   }
>
> ks.test(x,"trunc_expon_dist",theta=theta_hat,t=50)

        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.134, p-value = 0.841
alternative hypothesis: two-sided
```

Final comment to the Kolmogorov-Smirnov test:

- The test can be adapted to one sided H1 hypothesis like $H_1 : F(x) > F_0(x)$ or $H_1 : F(x) < F_0(x)$.
- When applied to discrete distributions (to avoid) the test is conservative and under reject H0.
- The test can be adapted to test if 2 different samples came from the same population. This generalization is called the 2 samples Kolmogorov-Smirnov test.

Anderson-Darling test

- Like the KS test, the goal of the Anderson-Darling test is to test $H_0 : F(x) = F_0(x)$, $-\infty < x < \infty$ against $H_1 : F(x) \neq F_0(x)$ for some $x \in \Re$.
- $F_0(x)$ is the distribution function of a continuous random variable where all parameters are known (H0 is a simple hypothesis).
- To take into account that the observed data could have been truncated and/or censored we replace $F_0(x)$ by $F^*(x)$ (as we did before) and $-\infty < x < \infty$ by $t < x < u$ where $t = -\infty$ (or 0 if the random variable is non-negative) if there is no left truncation and $u = +\infty$ if there is no right censoring.
- The main difference between the KS and the Anderson-Darling tests is the way by which we evaluate the discrepancies between the model and the observed data. Now our test statistic will be given by

$$A^2 = n \int_t^u \frac{\left(F_n(x) - F^*(x)\right)^2}{F^*(x)\left(1 - F^*(x)\right)} f^*(x)\, dx$$

We are evaluating the expected value of the weighted squared difference between $F_n(x)$ and $F^*(x)$

The critical values for the Anderson-Darling test are dependent on the specific distribution that is being tested. Tabulated values and formulas have been published for a few specific distributions (normal, lognormal, exponential, Weibull or logistic).

Chi-square goodness-of-fit test

- Unlike the KS and AD tests, the chi-square test can be used with discrete data. However it is an asymptotic test.
- **Basic idea**: To compute a chi-square goodness of fit test we define a partition of the support of the proposed distribution in k classes, $\{A_1, A_2, \cdots, A_k\}$ and we compare the expected number of observations in each class of the partition under the null hypothesis with the observed number in the sample.
- Formally, let $p_j = \Pr(X \in A_j \mid H_0)$ and let N_j be the number of observations that fall in class *j* (note that $p_j > 0$, $\sum_{j=1}^{k} p_j = 1$ and $\sum_{j=1}^{k} N_j = n$).
  - Expected number, under the null, $E_j = n\, p_j$. Observed number $O_j = N_j$
  - Test statistic: $\chi^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j}$

  In the early nineties, Karl Pearson showed that χ^2 follows asymptotically a chi-square distribution with $k-1$ degrees of freedom. This result has been obtained for a multinomial distribution with no unknown parameters.

  The result can be used with large samples when they are no unknown parameters and the expected number of observations in each class is greater than a given threshold.

There are 2 alternative rules that are commonly used to establish this threshold:

- **More conservative** – All the expected values are greater than or equal to 5, i.e. $E_j = n\,p_j \geq 5$, $j = 1,2,\cdots,k$. When the sample is quite large we can increase this value to 10.
- **Less conservative** – None of the expected values may be less than 1; No more than 20% of the expected values may be less than 5.

- How to compute the test when the original data came from a continuous random variable with no unknown parameters but possibly left truncated at *t* and right censored at *u*?
  - Choose a set of $k-1$ values, $t = c_0 < c_1 < \cdots < c_{k-1} < c_k = u$ and split the random variable's domain into k intervals $(c_{j-1}; c_j)$, $j = 1,2,\cdots,k$.
  - Compute the probability associated with each interval, $p_j = F^*(c_j) - F^*(c_{j-1})$, and calculate the expected number of observations in each interval, $E_j = n\,p_j$.
  - Compute the χ^2 statistic. If the observed value of the statistic is greater than the adequate percentile of a χ^2 distribution with $k-1$ d.f., reject the null.

Comments:

1. Note that we are testing H_0': $p_j = p_{0j}$ and not the initial hypothesis that observations, X, came from a population with distribution $F^*(x)$, that is, $H_0 : X \sim F^*(x)$. As H_0' can be deduced from H_0 (but the inverse is not true), the rejection of H_0' implies the rejection of H_0 but when we do not reject H_0' the conclusion that H_0 should not be rejected is less comfortable.
2. If the set of intervals does not allow us to comply with the rules about the expected number of observation in each interval we must aggregate intervals. As we generally want as many intervals as possible to get the hypothesis H_0' as closed as possible with H_0 we generally use intervals with the same approximate probability.
3. When the original data is a discrete random variable we choose the intervals as close as possible to the outcomes of the variable. Usually we need to aggregate values in the right tail of the distribution to meet the rules about the expected number of observations in each class.

- How to carry the test when they are *p* unknown parameters?
  - **Parameters estimation**: The more adequate procedure is to use maximum likelihood estimators based on grouped sample (called minimum chi-square estimation), that is, the likelihood function will be given by $L(\theta_1,\theta_2,\cdots,\theta_p \mid n_1,n_2,\cdots,n_k)=\prod_{j=1}^{k}\left(F^*(c_j \mid \theta_1,\cdots,\theta_p)-F^*(c_{j-1} \mid \theta_1,\cdots,\theta_p)\right)^{n_j}$.

    As this procedure is, most of the time, annoying one can follow the usual maximum likelihood based on individual data, $L(\theta_1,\theta_2,\cdots,\theta_p \mid x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n} f^*(x_i \mid \theta_1,\cdots,\theta_p)$.
  - Once the parameters have been estimated $\hat{p}_j = F^*(c_j \mid \hat{\theta}_1,\cdots,\hat{\theta}_p)-F^*(c_{j-1} \mid \hat{\theta}_1,\cdots,\hat{\theta}_p)$, and $E_j = n\,\hat{p}_j$. The test statistic is $\chi^2=\sum_{j=1}^{k}\frac{(O_j-E_j)^2}{E_j}$.
  - When the parameters have been estimated using minimum chi-square estimation, the test statistic follows asymptotically a chi-square distribution with $n-1-p$ degrees of freedom. If the alternative estimation has been used the distribution will lie somewhere between a chi-square distribution with $n-1-p$ and $n-1$ degrees of freedom. The usual procedure is to consider $n-1-p$ degrees of freedom.

Example 16.7 – Perform the chi-square goodness-of-fit test for the exponential distribution for the continuing example. We will use the same intervals as in *Loss Models*.

Case 1 – Data Set B truncated at 50 – The test is not adequate since the sample size is only 19. To follow the rules about the minimum expected number of observations in each class we must merge classes 1-3 and classes 5-6. Exercise: Do it and compute the p-value of the test

$$E_j = n\times\left(F^*(c_j \mid \hat{\theta}) - F^*(c_{j-1} \mid \hat{\theta})\right) = n\times\left(e^{-c_{j-1}/\hat{\theta}} - e^{-c_j/\hat{\theta}}\right)\times e^{t/\theta} = n\times\left(\exp\left((t-c_{j-1})/\hat{\theta}\right) - \exp\left((t-c_j)/\hat{\theta}\right)\right)$$

Using individual data → $\hat{\theta} = 802.32$ (see example 16.1) *p*-value=0.8436 4 df

Using grouped data → $\hat{\theta}_G = 743.2042$ (numerical maximization using EXCEL) *p*-value=0.8651 4 df

				Individual data		Grouped data	
j	c_{j-1}	c_j	n_j	E_j	χ^2	E_j	χ^2
1	50	150	3	2.2265	0.2687	2.3920	0.1546
2	150	250	3	1.9656	0.5444	2.0908	0.3953
3	250	500	4	3.9644	0.0003	4.1468	0.0052
4	500	1000	4	5.0289	0.2105	5.0784	0.2290
5	1000	2000	3	4.1427	0.3152	3.9139	0.2134
6	2000	Inf	2	1.6719	0.0644	1.3780	0.2807
		Sum			**1.4035**		**1.2782**

Example – Conduct an approximate goodness-of-fit test for the Poisson model assuming that we observed the first 2 columns of the following table

k	n_k	Expected	Q_j
0	85500	84279.2073	17.6833
1	13000	14414.6945	138.8417
2	1400	1232.7087	22.7032
3+	100	73.3896	9.6487

Parameter estimation using grouped data for the last interval:

$$\ell(\theta) = \sum_{k=0}^{2} n_k \ln f(k \mid \theta) + n_3 \ln\left(1 - F(2 \mid \theta)\right) \rightarrow \hat{\theta} = 0.171035$$

$$Q_{obs} = \sum Q_j = 188.8788 \qquad \text{p-value=0.0000}$$

We reject the hypothesis that the data are Poisson distributed

LIKELIHOOD RATIO TEST

- The likelihood ratio test provides an answer to the question "Is it more likely that the observations came from a population with distribution A than from a population with distribution B" **when the distributions are nested.**
- **Nested distributions:** Distribution A and B are nested when one of them is a special case of the other that can be obtained by means of a set of linear constraints. For instance, an Exponential distribution with parameter θ is nested in the Gamma family (the exponential distribution is obtained from the Gamma distribution when $\alpha = 1$) or a normal distribution with mean 3 is nested in the normal family of distributions.

Before applying the likelihood ratio test to model selection let us present the test.

- $X \sim f(x \mid \theta)$ Our purpose is to test $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ where $\{\Theta_0, \Theta_1\}$ is a partition of Θ. Note that θ can be a vector.
- Likelihood ratio – $\lambda(x_1, x_2, \cdots, x_n)$

  The likelihood ratio, called λ or $\lambda(x_1, x_2, \cdots, x_n)$, is defined by $\lambda(x_1, x_2, \cdots, x_n) = \dfrac{\sup_{\theta \in \Theta_0} L(\theta \mid x_1, x_2, \cdots, x_n)}{\sup_{\theta \in \Theta} L(\theta \mid x_1, x_2, \cdots, x_n)}$

  where $L(\theta \mid x_1, x_2, \cdots, x_n)$ is the likelihood function of θ given the observed sample.
- Comments:
  - The denominator is the likelihood evaluated at $\hat{\theta}$, i.e. $L(\hat{\theta})$, the likelihood evaluated at the maximum likelihood estimate;
  - The numerator follows the same approach but with a constraint: $\theta \in \Theta_o$;
  - As it is obvious we get $0 \leq \lambda \leq 1$ (usually $0 < \lambda \leq 1$);
  - If we consider a random sample (before observation) we get $\Lambda = \lambda(X_1, X_2, \cdots, X_n)$ instead of $\lambda(x_1, x_2, \cdots, x_n)$. As it is obvious, Λ is a statistic and, consequently, follows a sampling distribution.

- **Likelihood ratio test**

  A likelihood ratio test is any test with a rejection region $W = \{(x_1, x_2, \cdots, x_n) : \lambda(x_1, x_2, \cdots, x_n) < c\}$, with $0 \leq c \leq 1$.

- **Comments:**
  - The definition of the rejection region is intuitive;
  - As with the Neyman-Pearson procedure we define α and then we get c. To do so, it is necessary to know the sampling distribution of Λ given H_0 (or the distribution of an equivalent statistic).

- **Asymptotic distribution of the likelihood ratio**

  If some regularity conditions are fulfilled in the population, $-2\ln\Lambda \overset{\circ}{\sim} \chi^2_{(r)}$ where r is the difference between the number of free parameters specified by $\theta \in \Theta$ and the number of free parameters specified by $\theta \in \Theta_0$.

- **Example 16.9 –** You want to test the hypothesis that the population that produced Data Set B (using the original largest observation) has a mean that is other than 1200. Assume that the population has a gamma distribution and conduct the likelihood ratio test at a 5% significance level. Also, determine the *p*-value.

  $X \sim G(\alpha,\theta)$ $\quad \mu = \alpha\theta$ $\quad H_0 : \mu = 1200$ against $H_1 : \mu \neq 1200$

  Denominator:

  $$f(x \mid \alpha, \mu = 1200) = \frac{\theta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-x/\theta},\ x > 0,\ \alpha > 0$$

  $$L(\alpha \mid x_1, \cdots, x_n, \mu = 1200) = \prod_{i=1}^{n} \frac{\theta^{\alpha}}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-x_i/\theta};\ \ell(\alpha) = \sum_{i=1}^{n} \left(\alpha \ln\theta - \ln\Gamma(\alpha) + (\alpha-1)\ln x_i - x_i/\theta\right)$$

  Using numerical optimization we get $\hat{\alpha} = 0.556158$, $\hat{\theta} = 2561.138$ and $\ell(\hat{\alpha},\hat{\theta}) = -162.293$

Numerator:

$$\mu = 1200 \Leftrightarrow \alpha\theta = 1200 \Leftrightarrow \theta = 1200/\alpha$$

$$f(x \mid \alpha, \mu = 1200) = \frac{1200^{\alpha}\,\alpha^{-\alpha}}{\Gamma(\alpha)} x^{\alpha-1}\, e^{-\alpha x/1200},\ x > 0,\ \alpha > 0$$

$$\ell(\alpha) = \sum_{i=1}^{n} \left(\alpha \ln 1200 - \alpha \ln \alpha - \ln \Gamma(\alpha) + (\alpha - 1)\ln x_i - \alpha x_i / 1200\right)$$

Using numerical optimization we get $\hat{\alpha}_0 = 0.549549$ and $\ell(\hat{\alpha}_0) = -162.466$

The test

Although we have a small sample we will use the asymptotic version of the test. The result should be read very carefully.

$$-2\ln\lambda = -2(-162.466 + 162.293) = 0.344328$$

$p - value = 0.557342$ using a chi-square with 1 degree of freedom.

- **Likelihood ratio test and nested models**

  When 2 models are nested, one of them can be obtained from the other by mean of a set of linear constraints. We put in H_0 the constrained model and in H_1 the "free" model and then perform a likelihood ratio test.

  When the null is a limiting rather than a particular case of the alternative the test may still be used but with a more complex sampling distribution (mixture of chi-square distributions). However it is reasonable to still use the test provided that it is clearly understood that we are not performing a formal test.

- **Example –** Let us use data from Example 16.9 to decide if is reasonable to use an exponential distribution instead of a Gamma distribution.

  $X \sim G(\alpha,\theta)$ $\qquad$ $H_0 : \alpha = 1$ against $H_1 : \alpha \neq 1$

  Numerator:

  $f(x \mid \theta, \alpha = 1) = \theta^{-1} e^{-x/\theta},\ x > 0,\ \theta > 0$

  $L(\theta \mid x_1, \cdots, x_n, \alpha = 1200) = \prod_{i=1}^{n} \theta^{-1} e^{-x_i/\theta} = \theta^{-n} e^{-n\bar{x}/\theta}$

  $\ell(\theta) = -n \ln \theta - n\bar{x}/\theta$ $\qquad$ $\hat{\theta}_0 = \bar{x} = 1424.4$ $\qquad$ $\ell(\hat{\theta}_0) = -n \ln \bar{x} - n = -165.23$

Denominator (see Example 16.9)

$\hat{\alpha} = 0.556158$, $\hat{\theta} = 2561.138$ and $\ell(\hat{\alpha}, \hat{\theta}) = -162.293$

The test

Same comment about the sample size (see example 16.9)

$-2\ln\lambda = -2(-165.23 + 162.293) = 5.873432$

$p - value = 0.015371$ using a chi-square distribution with 1 degree of freedom.

SELECTING A MODEL

In selecting a model 2 ideas should be present:

- Parsimony – "the simpler the better" (for the same "quality")
- Restrict the set of possible models – if you try hundreds of models, some of them will fit the data by chance.

Keeping in mind these 2 ideas, model selection is always a judgment-based approach (from my point of view). Some points that deserve consideration:

- A clear understanding of the problem is necessary. For instance you have to be prepared to answer to questions like "It is more important to fit well the tail or to match the mode?"
- How (using which kind of models) has this problem be solved before? Have these models proved well in the past? If not, why?
- Some statistical procedures (namely those presented in this chapter) help to eliminate models.
- When we compare nested models we can use the likelihood ratio test.

- *Loss Models* presents 5 criteria in a section called score-based approach.
  - Lowest value of the Kolmogorov-Smirnov test statistic
  - Lowest value of the Anderson-Darling test statistic
  - Lowest value of the chi-square goodness of fit statistic
  - Highest *p*-value for the chi-square goodness of fit statistic
  - Highest value of the likelihood function at its maximum
- From my point of view it is not correct to use these criteria to select a model unless the *p*-values obtained are quite different (fourth criterion). Note that all these criteria, but the fourth, violate the parsimony principle as the number of estimated parameters is not considered.

- We can add 2 more criteria: the Aikaike and the Schwarz criteria (both are preferable to the 5th criterion)
  - **Aikaike criterion**: $AIC = -2\ln L + 2r$ where L stands for the value of the likelihood function at its maximum and r for the number of estimated parameters; the lesser the value of AIC the better.
  - **Schwarz criterion**: $SBC = \ln L - (r/2)\ln n$ where *n* is the sample size. Using this criterion presented in *Loss Models*, the greater the value of SBC, the better. To use a scale similar to that used with AIC, the Schwarz criterion is generally presented as $BIC = -2\ln L + r\ln n = -2 \times SBC$. Using the latter expression, the lesser the value of BIC the better.
- Comment:

  The main difference between these two criteria is the way used to penalize the number of parameters. Using AIC a new parameter is relevant if it increases the log-likelihood by more than 1 and using the BIC by more than $\ln n$.

- **Example 16.11** – For the continuing example in this chapter, choose between the exponential and Weibull models for the data.

  Case 1 – Sample truncated at 50 with no censoring (19 observations)

	K-S	A-D	χ^2 (p-value)	Loglik	SBC	Likelihood ratio test
Exponential	0.1340	0.4292	1.4034 (0.8436)	-146.06	-147.35	0.7585 (0.3838)
Weibull	0.0887	0.1631	0.3615 (0.9481)	-145.68	-148.63	

  Both distributions seem acceptable – K-S approximate critical value (5%)=0.3120 and A-D approximate critical value (5%)=2.492 – and then we use the **parsimony** principle or the likelihood ratio test to **choose the exponential**.