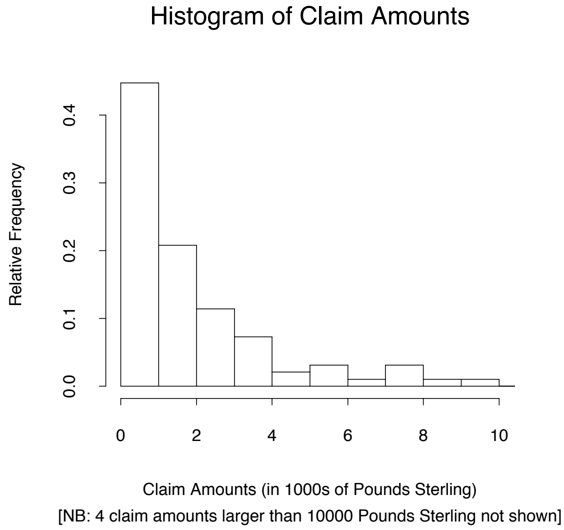
Example Claims Data

Claim Amounts (in £'s)

24	26	73	84	102	115
132	159	207	240	241	254
268	272	282	300	302	329
346	359	367	375	378	384
452	475	495	503	531	543
563	594	609	671	687	691
716	757	821	829	885	893
968	1053	1081	1083	1150	1205
1262	1270	1351	1385	1498	1546
1565	1635	1671	1706	1820	1829
1855	1873	1914	2030	2066	2240
2413	2421	2521	2586	2727	2797
2850	2989	3110	3166	3383	3443
3512	3515	3531	4068	4527	5006
5065	5481	6046	7003	7245	7477
8738	9197	16370	17605	25318	58524



Sample Statistics: $\overline{x} = 2989.83$, $s = 6856.11$, median $= 1233.5$

The Exponential Distribution

1. Characterisation

   - Probability Density Function (*pdf*):

$$f_X(x;\theta) = \frac{1}{\theta} e^{-x/\theta} \qquad \text{(mean parameterisation)}$$

or

$$f_X(x;\lambda) = \lambda e^{-\lambda x} \qquad \text{(rate parameterisation)}$$

   - Moments:

$$E_\theta(X) = \theta \qquad E_\theta(X^2) = 2\theta^2 \qquad Var_\theta(X) = \theta^2$$

or

$$E_\lambda(X) = \frac{1}{\lambda} \qquad E_\lambda(X^2) = \frac{2}{\lambda^2} \qquad Var_\lambda(X) = \frac{1}{\lambda^2}$$

   - Quantiles (Percentiles):

$$\begin{aligned}
\text{Median}_\theta(X) = x_{0.5} \text{ solves : } & \int_0^{x_{0.5}} \frac{1}{\theta} e^{-x/\theta} dx = 0.5 \\
\implies & \left[-e^{-x/\theta} \right]_{x=0}^{x_{0.5}} = 0.5 \\
\implies & -e^{-x_{0.5}/\theta} + 1 = 0.5 \\
\implies & x_{0.5} = \theta \ln(2) \\
\text{OR} \qquad x_p \text{ solves : } & \int_0^{x_p} \frac{1}{\theta} e^{-x/\theta} dx = p \\
\implies & x_p = \theta \ln\left(\frac{1}{1-p}\right)
\end{aligned}$$

2. Estimation of Parameters based on $x_1, \ldots, x_n$:
   - "Common Sense" Methods:

     θ is "Population Average", estimate by $\overline{x}$.
   - (Standard) Method of Moments (*MOM*):

     Solve system:

$$E_\theta(X) = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\vdots$$

$$E_\theta(X^k) = \overline{x^k} = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

     where k is number of parameters.
   - (Generalised) Method of Moments:

     Solve system:

$$E_\theta\{g_1(X)\} = \overline{g_1(x)} = \frac{1}{n}\sum_{i=1}^{n} g_1(x_i)$$

$$\vdots$$

$$E_\theta\{g_k(X)\} = \overline{g_k(x)} = \frac{1}{n}\sum_{i=1}^{n} g_k(x_i)$$

     e.g., $g_1(X) = \{X - E_\theta(X)\}^2$.

2. Estimation of Parameters based on $x_1, \ldots, x_n$ *(Continued)*:

   - Method of Percentiles (*MOP*):

     Solve system:

$$x_{p_1} = \hat{x}_{p_1}$$
$$\vdots$$
$$x_{p_k} = \hat{x}_{p_k}$$

     where $\hat{x}_p$ is observed p^{th} percentile of data, for some choice of $p_1, \ldots, p_k$.

     For example, for our data, *MOP* estimate based on median ($p_1 = 0.5$) solves:

$$\theta \ln(2) = 1233.5$$

     So

$$\hat{\theta}_{MOP} = 1233.5/\ln(2) = 1779.56$$

     (compared to

$$\hat{\theta}_{MOM} = \overline{x} = 2989.83$$

     using standard *MOM*).

2. Estimation of Parameters based on $x_1, \ldots, x_n$ *(Continued)*:

   - Maximum Likelihood Estimate (*MLE*)

     Likelihood Function:

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f_X(x_i; \theta) = \frac{1}{\theta^n} \exp\left(-\frac{1}{\theta}\sum_{i=1}^{n} x_i\right)$$

     Log-Likelihood Function: $l(\theta) = \ln\{L(\theta; x_1, \ldots, x_n)\}$

     Maximum Likelihood Estimate, $\hat{\theta}_{MLE}$, solves:

$$\frac{\partial l(\theta)}{\partial \theta_i} = 0, \qquad 1 \leq i \leq k$$

     For example, for exponential disribution:

$$l(\theta) = -n\ln(\theta) - \frac{n}{\theta}\overline{x}$$

     so that

$$\frac{dl(\theta)}{d\theta} = -\frac{n}{\theta} + \frac{n}{\theta^2}\overline{x} = 0 \quad \Longrightarrow \quad \hat{\theta}_{MLE} = \overline{x}$$

2. Estimation of Parameters based on $x_1, \ldots, x_n$ *(Continued)*:

- Maximum Likelihood Estimates *(Continued)*

  Maximum Likelihood Theorem: For large samples,

$$Pr_\theta \left\{ \frac{\hat{\theta}_{MLE} - \theta}{\sqrt{I^{-1}(\hat{\theta}_{MLE})}} \leq t \right\} \approx \Phi(t),$$

  where $\Phi(\cdot)$ is the standard normal *CDF* and

$$I(\theta) = -E_\theta\{l''(\theta)\}$$

  So, we can use $\hat{\theta}_{MLE} \pm 1.96\sqrt{I^{-1}(\hat{\theta}_{MLE})}$ as an approximate 95% confidence interval for θ.

  For example, for our data,

$$l''(\theta) = \frac{n}{\theta^2} - \frac{2n}{\theta^3}\overline{x},$$

  so that

$$I(\theta) = \frac{n}{\theta^2}$$

  Thus,

$$\sqrt{I^{-1}(\hat{\theta}_{MLE})} = \frac{\hat{\theta}_{MLE}}{\sqrt{n}} = \frac{2989.83}{\sqrt{96}} = 305.15$$

  and a 95% confidence interval is:

$$2989.83 \pm 1.96(305.15) = (2391.74,\ 3587.92)$$

3. Goodness-of-Fit Testing:

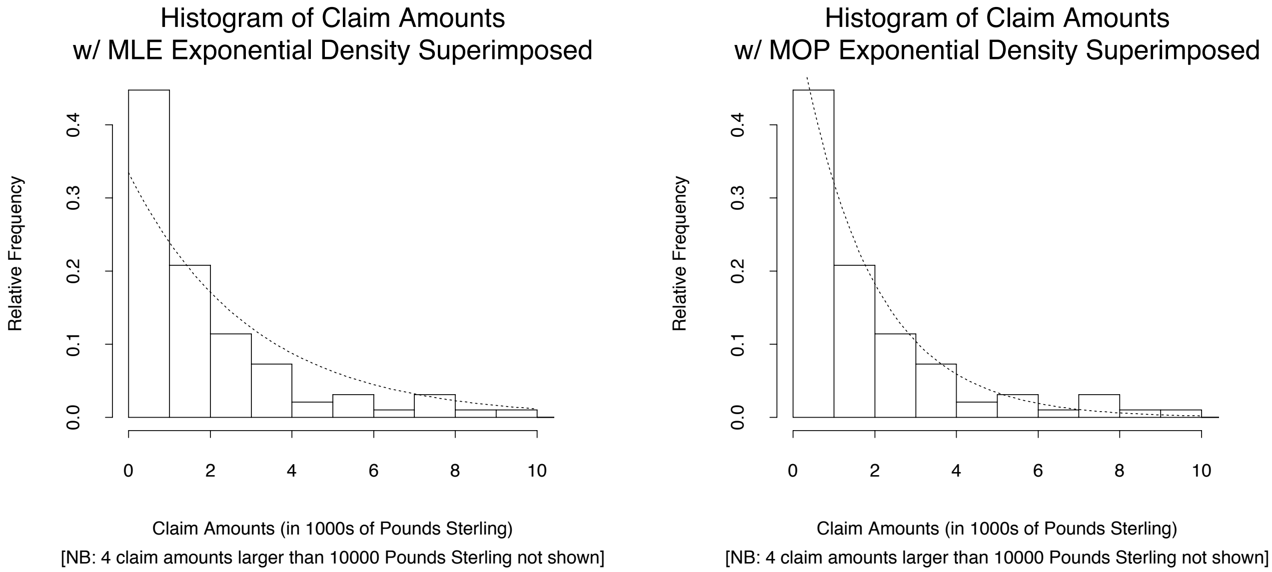
- "Common Sense" Methods:

  Compare parameter estimates based on different methods. For example, we had $\hat{\theta}_{MOP} = 1779.56$ versus $\hat{\theta}_{MOM} = \hat{\theta}_{MLE} = 2989.83$ (CAUTION: Only *MLE* has precision estimate so far)

  Use "structural facts" about the distribution. For example, we know that $E_\theta(X) = \theta = \sqrt{Var_\theta(X)}$ for exponential distributions. From data, we see that $\overline{x} = 2989.83 \neq s = 6856.11$ (Again, no precision estimates, though).

3. Goodness-of-Fit Testing *(Continued)*:

   - Histograms with Superimposed Density Curves:



3. Goodness-of-Fit Testing *(Continued)*:

- Pearson Chi-Squared Test:

  IDEA:

  · Data is n *iid* observations classified into k categories.

  · O_i = number of observations in category i.

  · "Theory": $p_i = Pr(\text{obs. in cat. } i)$

  · $E_i = np_i$ = expected # of obs. in ith category.

  · Measure discrepancy using *test statistic*:

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

  which has an approximate χ^2-distribution with a number of degrees of freedom equal to:

  $$df = k - 1 - (\#\, \textit{parameters estimated in determining } p_i).$$

3. Goodness-of-Fit Testing *(Continued)*:

   - Pearson Chi-Squared Test *(Continued)*:

     IMPLEMENTATION:

     · For continuous data, need to "discretise" using bins.

     · BIN CHOICE ISSUES:

       · Too few bins = oversmoothing

       · Too many bins = undersmoothing

       · Choose 5 to 15 bins.

     · Could use histogram bins (equal width):

Table 2.1: Observed and Expected Claim Amounts (in 1000 £'s) using the Exponential Distribution

	0-1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10	10+
O_i	43	20	11	7	2	3	1	3	1	1	4
$E_{i,MLE}$	27.3	19.5	14.0	10.0	7.2	5.1	3.7	2.6	1.9	1.3	3.4
$E_{i,MOP}$	41.3	23.5	13.4	7.6	4.4	2.5	1.4	0.8	0.5	0.3	0.3

$X^2 = 17.92$ (*MLE*) or $X^2 = 49.55$ (*MOP*)

with $df = 11 - 1 - 1 = 9$.

$Pr(\chi^2_9 > 17.92) = 0.0376$; $Pr(\chi^2_9 > 49.55) = 1.31 \times 10^{-7}$

Reject exponential distribution.

3. Goodness-of-Fit Testing *(Continued)*:

- Pearson Chi-Squared Test *(Continued)*:

  · Calculating E_i's:

$$\begin{aligned} Pr_\theta(a < X \leq b) &= \int_a^b \frac{1}{\theta} \exp\left(-\frac{x}{\theta}\right) dx \\ &= \exp\left(-\frac{a}{\theta}\right) - \exp\left(-\frac{b}{\theta}\right). \end{aligned}$$

  Thus, we estimate p_1 (using $\hat{\theta}_{MLE}$):

$$\begin{aligned} p_1 &= \exp(0) - \exp\left(-\frac{1000}{\hat{\theta}_{MLE}}\right) \\ &= 1 - \exp\left(-\frac{1000}{2989.83}\right) \\ &= 0.284. \end{aligned}$$

  So, $E_{1,MLE} = np_1 = 96(0.284) = 27.3$.

3. Goodness-of-Fit Testing *(Continued)*:

- Pearson Chi-Squared Test *(Continued)*:

  · Reliability of test requires $E_i > 5$.

  · Use equal-count bins. Try 12 bins, so that

$$E_i = 96/12 = 8 \text{ for all } i.$$

  · Equal-count bin construction:

    · Use $Pr_\theta(a < X \leq b)$ formula from above recursively.

    · For first bin, $(0, b)$, solve:

$$n\left\{1 - \exp\left(-\frac{b}{\theta}\right)\right\} = 8.$$

    Using $\hat{\theta}_{MLE}$ yields $b = 260.15$

    · For next bin, $(260.15, b)$, solve:

$$n\left\{\exp\left(-\frac{260.15}{\theta}\right) - \exp\left(-\frac{b}{\theta}\right)\right\} = 8.$$

    Using $\hat{\theta}_{MLE}$ yields $b = 545.11$. Continue.

Table 2.2: Observed and Expected Claim Amounts (in £'s) using the Exponential Distribution

Bin Range	O_i	E_i	Bin Range	O_i	E_i
0-260	12	8	2072-2618	5	8
260-545	18	8	2618-3285	6	8
545-860	10	8	3285-4145	6	8
860-1212	8	8	4145-5357	3	8
1212-1612	7	8	5357-7429	4	8
1612-2072	10	8	7429+	7	8

  · $X^2 = 23.00$, $df = 12 - 1 - 1 = 10$,

  · $Pr_\theta(\chi^2_{10} > 23.00) = 0.011$.