Lognormal Model for Call-Center Service Times and Hazard Rate Functions

Part 1. Service Times – lognormal?

- Review – Basics of Lognormal Distribution

Definition: X is a lognormal random variable if $\ln(X)$ is normally distributed with mean μ and variance σ^2.

Density: $f(x) = \dfrac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$.

Mean: $e^{\mu+\sigma^2/2}$.

Variance: $e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$.

CV: $\sqrt{e^{\sigma^2}-1}$.

Note that CV does not depend on μ.

For small σ $(\sigma < 0.5)$, one can use CV $\approx \sigma$.

Median: e^{μ}.

Mode: $e^{\mu-\sigma^2}$ (compare the mean, median and mode).

Hazard Rate: (*standard* lognormal random variable):

$$h(x) = \frac{f(x)}{S(x)} = \frac{\frac{1}{x}e^{-(\ln x)^2/2}}{\int_{\ln x}^{\infty} e^{-t^2/2}dt} \sim \frac{\ln x}{x} \quad (x \to \infty).$$

Part 1. Service Times – lognormal? (2)

- Service times data
  - November and December **service times** (64704 calls) for the four major service types: IN (5592), NE (7622), NW (5774) and PS (45716).
  - For every service type, we check if the lognormal distribution fits
    - Standard goodness-of-fit tests (chi-square, Kolmogorov-Smirnov) reject the lognormal hypothesis.
    - These tests are rarely applicable for large samples of real data b/c the test recognizes very small differences between real-data and theoretical distributions.
    - However, the fit can be good enough for applications.
  - Hence, we use two graphical tests, **histograms** and **Q-Q plots**, to compare the sample service-time and lognormal distributions, and check if the differences are really significant for our purposes.

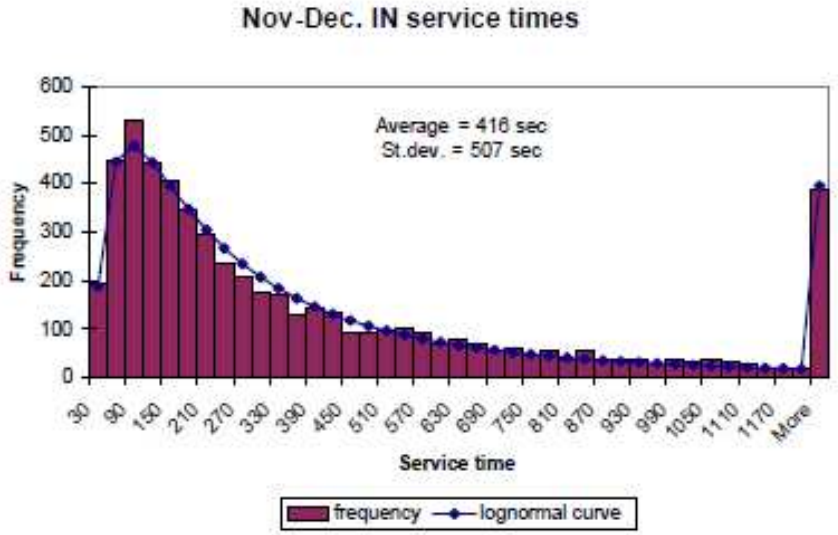
Part 1. Service Times – lognormal? (3)

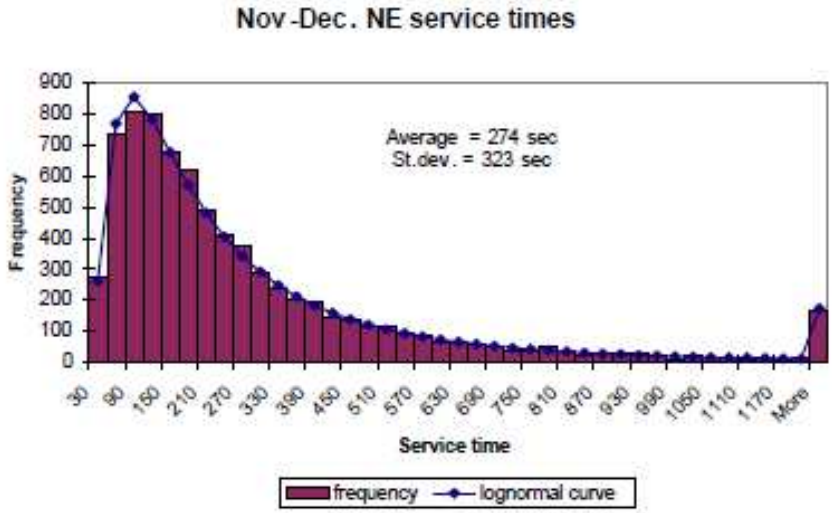
- Histograms of Service Times versus Lognormal Density

1. transform the sample of service times via Ln transformation (Ln(Service times))
2. estimate μ and σ and use the formulae on slide 3 to estimate the mean and standard deviation of the lognormal distribution and get the lognormal cdf F
3. define $T_0=0$, T_1, T_2, … according to a chosen histogram bin size, and compute the empirical frequency for each interval
4. fit lognormal distribution by calculating theoretical probabilities to fall into intervals $P_i=F(T_{i+1})-F(T_i)$ and getting theoretical frequencies by $N_i=N \times P_i$.
5. compare with the histogram

Part 1. Service Times – lognormal? (4)

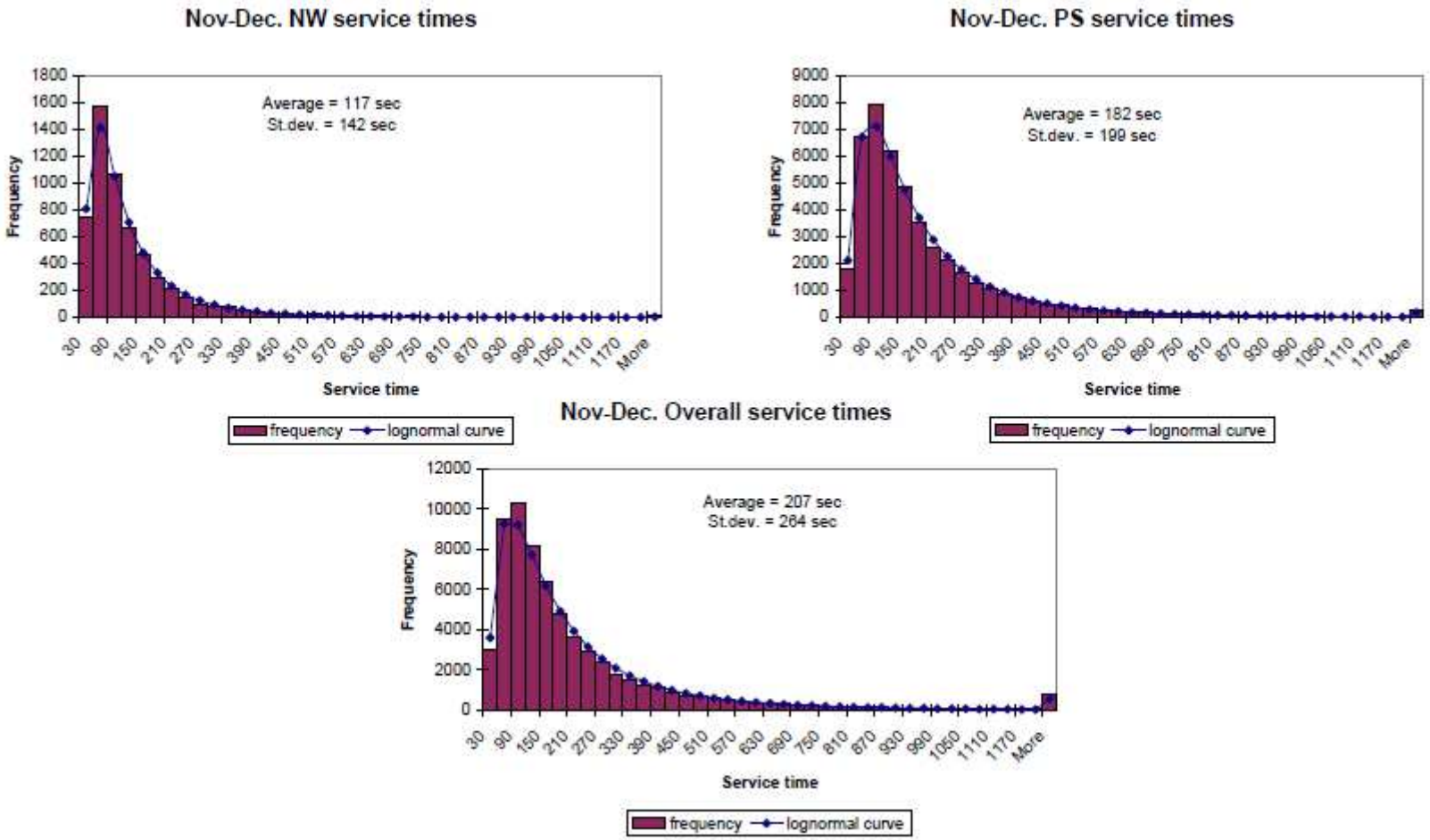
- **EX** Histograms of Service Times versus Lognormal Density
  - bin size: 30 seconds, chosen by trial-and-error
  - The fit seems good for all service types
    - IN: somewhat worse, but only in the “middle” of distribution.
    - PS and “overall” are similar, but PS seems slightly better.
    - good fit at the “tails” for IN and NE





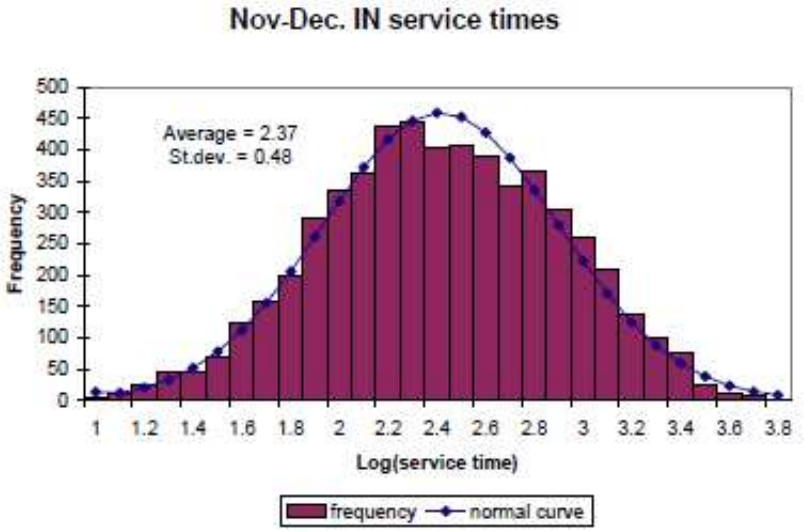
Part 1. Service Times – lognormal? (5)

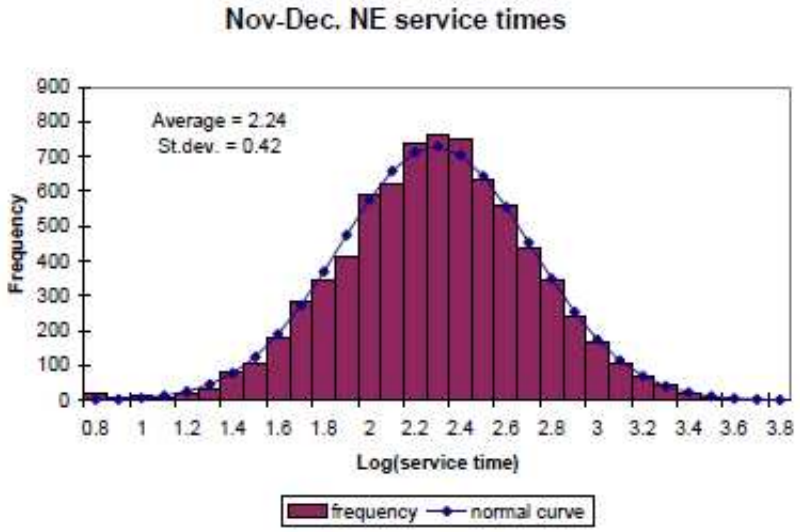
- **EX** Histograms of Service Times versus Lognormal Density



Part 1. Service Times – lognormal? (6)

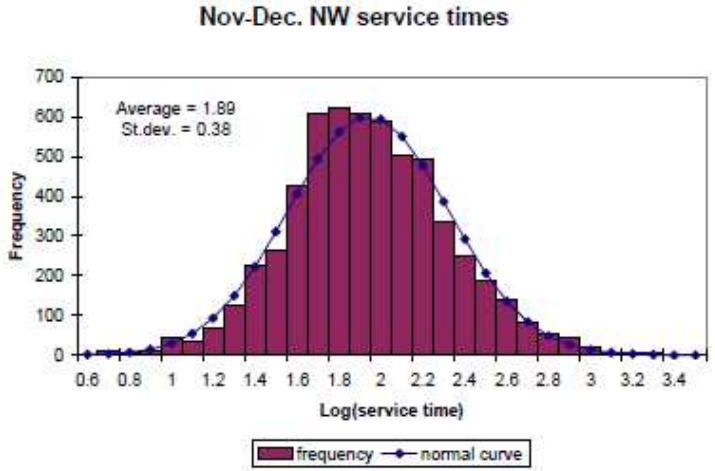
- **EX** Histograms of Log_{10} (Service Times) versus Normal Density
  - decimal logarithm: integers 1, 2 and 3 correspond to 10, 100 and 1000 seconds, respectively.
  - The fit for NE and PS service types is better than for the two other types. However, the normal curve seems a reasonable approximation for all service types.

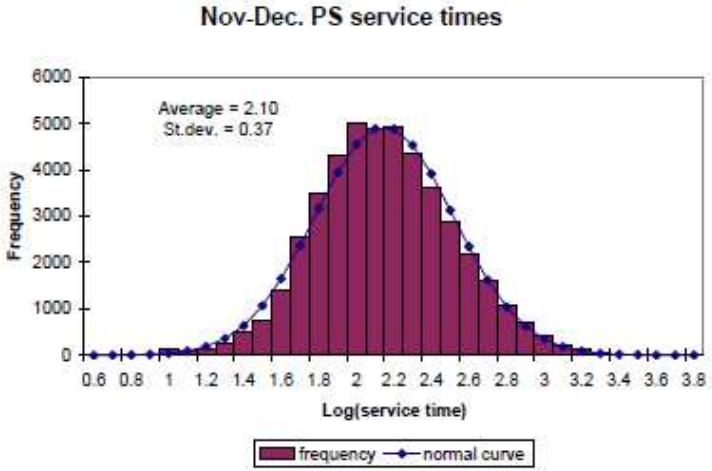


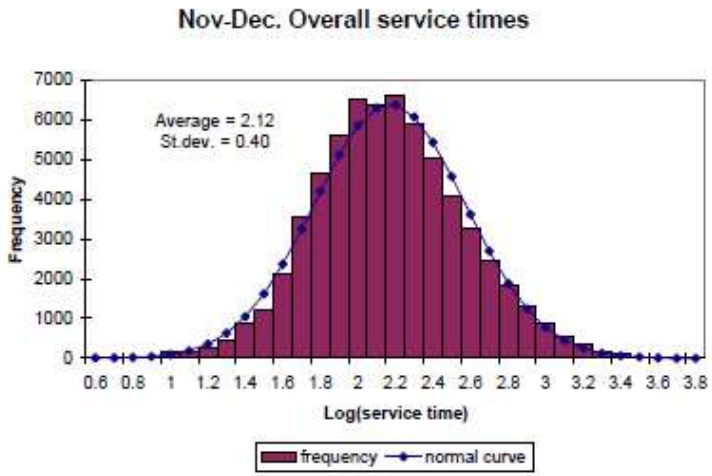


Part 1. Service Times – lognormal? (7)

- **EX** Histograms of Service Times versus Lognormal Density

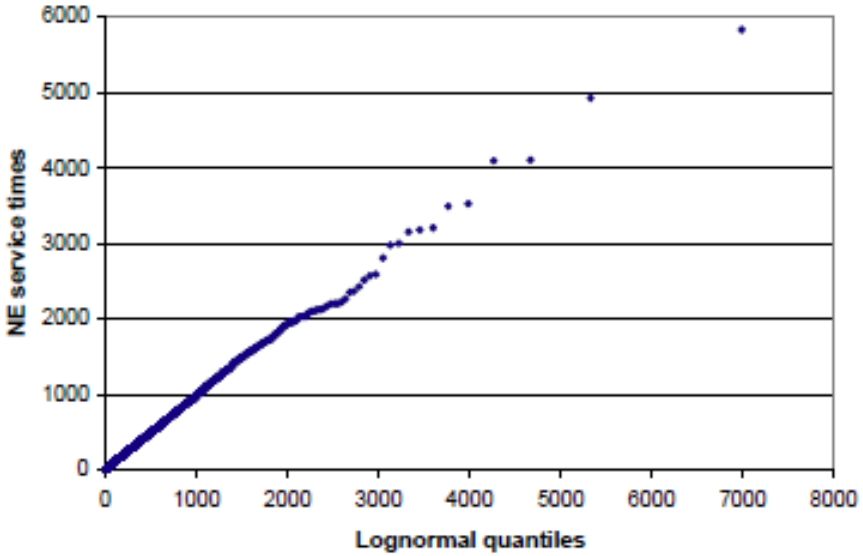


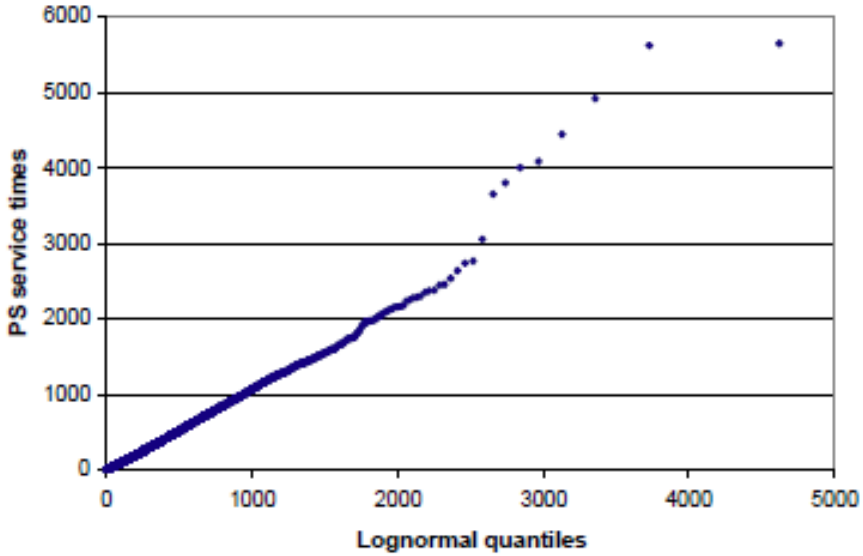




Part 1. Service Times – lognormal? (8)

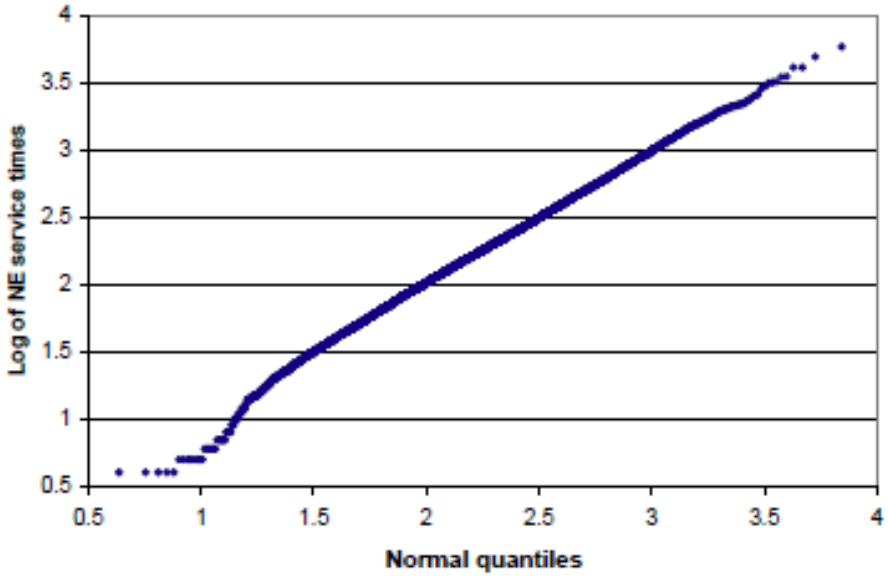
- **EX** Service Times versus lognormal-quantiles
  - A good fit to a straight line up to 30 minutes (1800 sec)
  - the center and the upper-right corner of both graphs include only a small number of large service times

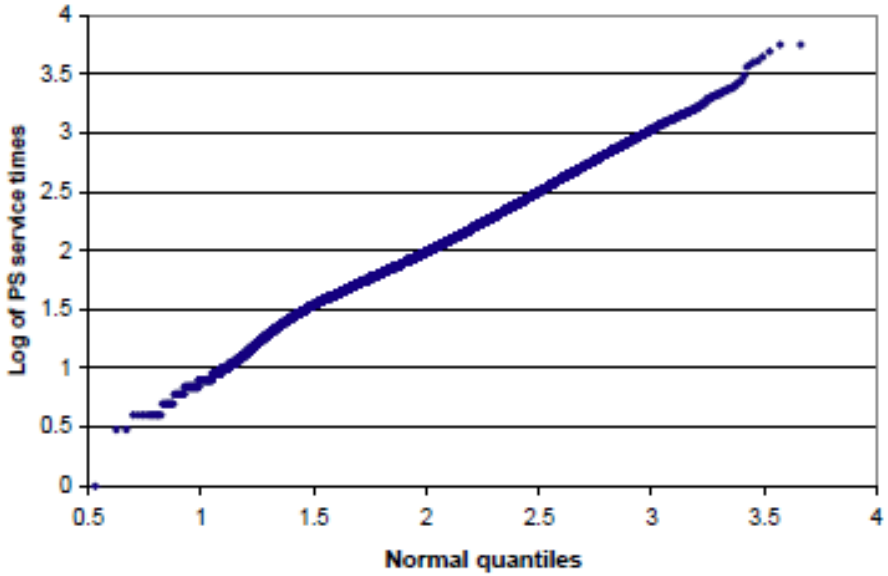




Part 1. Service Times – lognormal? (9)

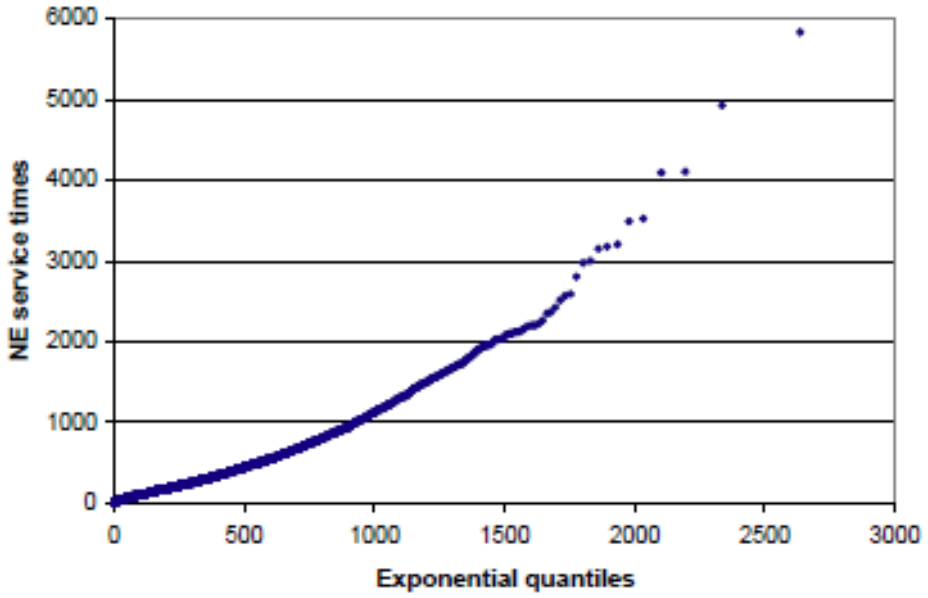
- **EX** Log_{10} (Service Times) versus normal-quantiles
  - a more *balanced plot* (main bulk of the data in the middle)
    - the normal probability plot is the most popular of QQ-plots
  - a straight line in the middle of the graph is observed with some *noise* at the edges.





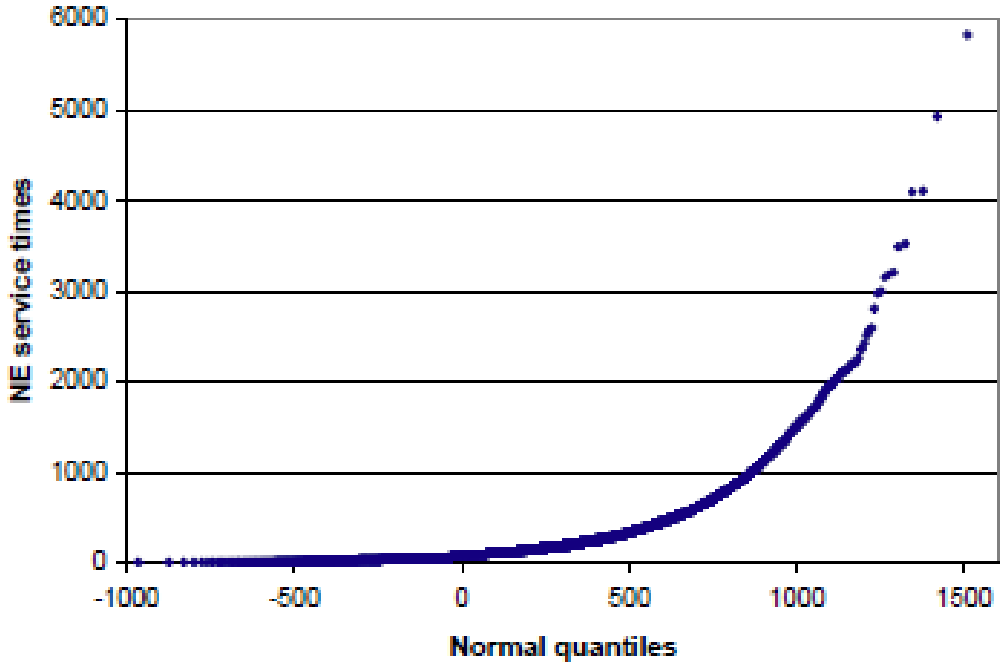
Part 1. Service Times – lognormal? (10)

- **EX** Service Times versus exponential-quantiles
  - moderately convex, below the 45° line
  - The behavior of the QQ-plot demonstrates that the sample data has a heavier tail than the theoretical distribution.
  - NOTE: QQ-plots are an excellent tool to compare tails of distributions; for example, a plot can show that the "exponential tail" is a good approximation even if the exponential hypothesis is strongly inconsistent for small values.



Part 1. Service Times – lognormal? (11)

- **EX** Service Times versus normal-quantiles
  - strongly convex, below the 45° line
  - the sample data has a heavier tail than the theoretical distribution.



Part 1. Service Times – lognormal? (12)

- **Conclusion**
  - The lognormal model provides a good approximation for the service time distribution of the four major service types.
  - The fit for NE and PS service type is better than for IN and NW.

- **Why Lognormal?**
  - Lognormal distribution arises frequently in applications.
  - We do not have a good “story” behind this distribution that can explain, even partially, its prevalence. It is not clear whether the lognormal distribution is so special.
  - Apparently, one can fit to “lognormal” data, as successsfully, also other rich enough families of distributions, for example Gamma.

Part 2. Hazard Rate Functions

- The hazard rate function **h(t)** *uniquely* determines the distribution of a non-negative random variable
  - $S(t) = 1 - F(t) = \begin{cases} \exp\left\{-\int_0^t h(t)\,dt\right\} & \text{continuous time} \\ \Pi_{i=0}^{t}\,[1 - h(i)] & \text{discrete time} \end{cases}$

- **Continuous** Case: For a continuous non-negative random variable T,

$$h(t) = \frac{f(t)}{S(t)}, \quad t \geq 0,$$

where $f(t)$ is the density of T and $S(t)$ is the survival function: $S(t) = \int_t^\infty f(u)du$.
Note that $\mathrm{P}\{T \leq t + \Delta \,|\, T > t\} \approx h(t) \cdot \Delta$.

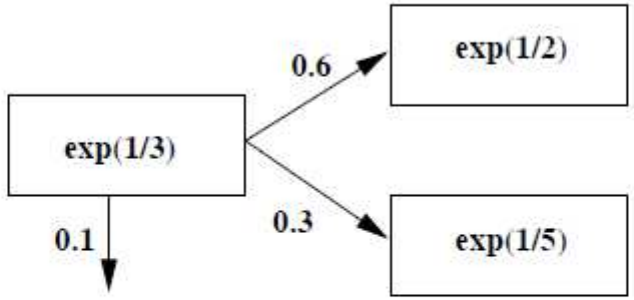
- **Discrete** Case: If T is a discrete non-negative random variables that takes values $t_1 < t_2 <$ … with corresponding probabilities $\{p_i, i \geq 1\}$, then its hazard-sequence, for i>0, is defined by

$$h(t_i) = \frac{p_i}{\sum_{j \geq i} p_j} = \frac{p_i}{S(t_i-)},$$

Part 2. Hazard Rate Functions (2)

- Theoretical Calculation

Example: consider the following service time distribution:



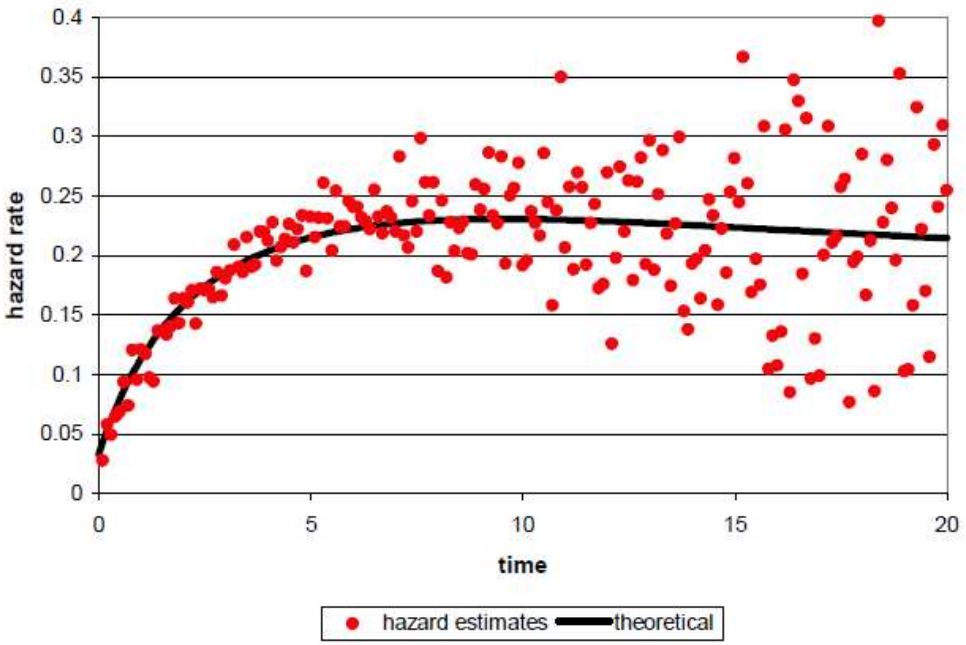
Its hazard rate can be calculated theoretically:

$$h(x) = \frac{0.15 \cdot e^{-x/5} + (29/60) \cdot e^{-x/3} - 0.6 \cdot e^{-x/2}}{0.75 \cdot e^{-x/5} + 1.45 \cdot e^{-x/3} - 1.2 \cdot e^{-x/2}}.$$

* $S(t) = 1 - F(t) = P(X > x) = 0.1P(X_1 > x) + 0.6P(X_1 + X_2 > x) + 0.3P(X_1 + X_3 > x)$
$= \ \ldots$

Part 2. Hazard Rate Functions (3)

- How do we estimate hazard rate from data?
  - A simulation experiment: 10,000 independent realizations of service times were simulated in Excel. The theoretical hazard rates were plotted and compared against estimates of the hazard rate, based on the simulation data.



- Comments:
  - The hazard-rate is neither increasing nor decreasing: *hump* pattern.
  - Value at $t = 0$: 1/3*0.1 - product of rate of the initial phase and exit probability.
  - Limit at $t = \infty$: 1/5 - rate of the longest final phase (exp(1/5))

Part 2. Hazard Rate Functions (4)

- Estimating the Hazard Rate ($\widehat{h}_j$, red dots on pg. 12):

The hazard rate is assumed to be constant on successive time intervals of length 0.1 between 0 and 20 (200 intervals overall). Formally, interval j is $\left(\frac{j-1}{10}, \frac{j}{10}\right]$, $j = 1, 2, \ldots, 200$.

The hazard estimate $\hat{h}_j$ for interval number j is calculated using the following formula:

$$\hat{h}_j = \frac{d_j}{b_j \left(r_{j-1} - \frac{1}{2} d_j\right)},$$

where

d_j = *number of events* (service terminations) in interval number j;

r_{j-1} = *number at risk* at the beginning of interval number j (number of services that have not terminated yet at time $\frac{j-1}{10}$);

b_j = length of interval number j (0.1 for all intervals, in our case).

The following provides some intuition for the above formula:
Let n denote the sample size. Then $\frac{d_j}{b_j \cdot n}$ is a reasonable estimate of the average density in interval number j and $\frac{r_{j-1} - 0.5 \cdot d_j}{n}$ is an approximation for the survival function in the center of this interval.

Remark. This estimation procedure is also valid for the **censored data**.

Remark. Handout that we install in "Related Materials" contains additional examples of phase-type distributions and their hazard rates.

Part 2. Hazard Rate Functions (5)

- Estimating the Hazard Rate ($\widehat{h}_j$, red dots on pg. 12): - continued

Part of Excel Table

Time	events	at risk	Hazard Estimate	Theoretical
0		10000		0.033
0.1	28	9972	0.028	0.044
0.2	58	9914	0.058	0.054
0.3	49	9865	0.050	0.063
0.4	64	9801	0.065	0.072
0.5	67	9734	0.069	0.080
0.6	91	9643	0.094	0.087
0.7	71	9572	0.074	0.095
0.8	115	9457	0.121	0.101
0.9	90	9367	0.096	0.108
1	113	9254	0.121	0.114
1.1	108	9146	0.117	0.119

- How do we use the estimates, $\widehat{h}_j$?
  - $\hat{S}(t) = \Pi_{i=0}^{t}\left[1 - \widehat{h}_i\right], \quad t = 0, 1, \ldots$
  - $\hat{E}[\tau] = \int_0^\infty \hat{S}(t)dt$
    - If $\tau \sim \exp(\theta)$, $E[\tau] = \frac{1}{\theta}$. Hence, $\hat{E}[\tau] = \frac{W_{total}}{\#\, abandoned}$