

Testing the Goodness-of-fit of a Proposed Probability Distribution



This Hypercard stack was prepared by:
Dennis L. Bricker,
Dept. of Industrial Engineering,
University of Iowa,
Iowa City, Iowa 52242
e-mail: dbricker@icaen.uiowa.edu

Chi-square Distribution

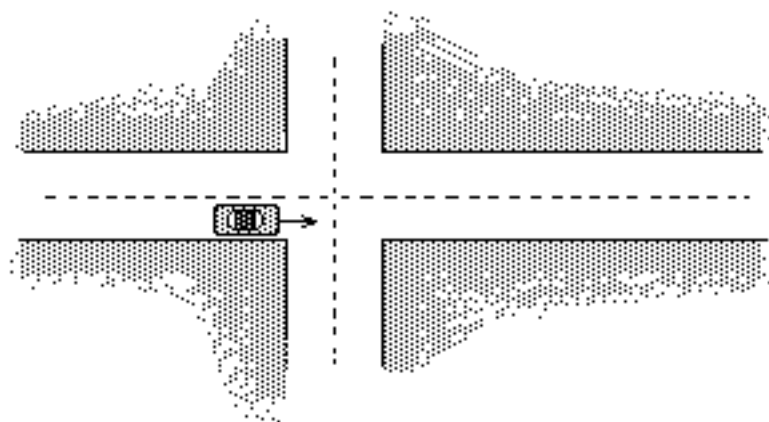
The *sum* of the *squares* of ν independent $N(0,1)$ random variables has the *chi-square* (χ^2) distribution.

$$f_{\nu}(y) = \frac{1}{2} \frac{(y/2)^{(\nu/2)-1} e^{-y/2}}{\Gamma(\nu/2)}, \quad y \geq 0 \quad \mu_Y = \nu, \quad \sigma_Y^2 = 2\nu$$

The parameter ν is referred to as "degrees of freedom"

Example

The number of eastbound vehicles arriving at an intersection in a 5-minute period between 7:00am & 7:05am was monitored for 5 workdays over a 20-week period (100 observations).

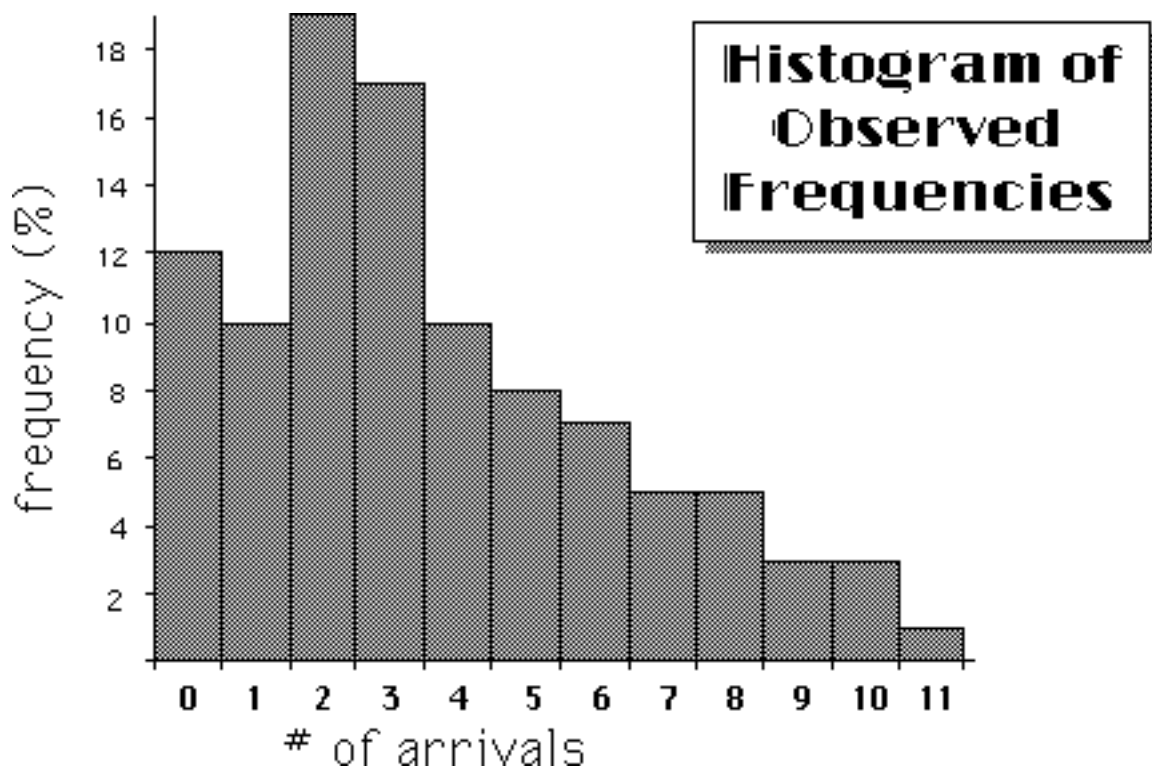


©D.L.Bricker, U. of IA, 1999

| #arrivals | frequency |
|-----------|-----------|
| 0 | 12 |
| 1 | 10 |
| 2 | 19 |
| 3 | 17 |
| 4 | 10 |
| 5 | 8 |
| 6 | 7 |
| 7 | 5 |
| 8 | 5 |
| 9 | 3 |
| 10 | 3 |
| 11 | 1 |

On 12 of the 100 days, no arrivals were noted; on 10 of the days, 1 arrival was noted; etc.

©D.L.Bricker, U. of IA, 1999



©D.L.Bricker, U. of IA, 1999

Suppose that p_i = probability of i arrivals between
7:00am and 7:05am each day

Then in a sample of size $N=100$, the number of days
that we observe i arrivals will have the *Binomial*
Distribution with parameters N and p_i :

$$P\{X = x\} = \frac{N!}{x!(N-x)!} p_i^x (1-p_i)^{N-x}$$

= probability that i arrivals are observed on
 x out of N days

©D.L.Bricker, U. of IA, 1999

It was guessed that the arrival process is Poisson, i.e., the time between arrivals has an *Exponential* distribution, and the number of arrivals in the 5-minute interval has a *Poisson* distribution:

$$P\{N_t = x\} = \frac{e^{-\lambda t}(\lambda t)^x}{x!} \quad \text{with mean } \lambda t$$

The average of the observed values is 3.64, so we estimate the arrival rate to be

$$\lambda = 3.64/5 \text{ min.} = 0.728/\text{min.}$$

©D.L.Bricker, U. of IA, 1999

$$P\{N_t = x\} = \frac{e^{-\lambda t}(\lambda t)^x}{x!}$$

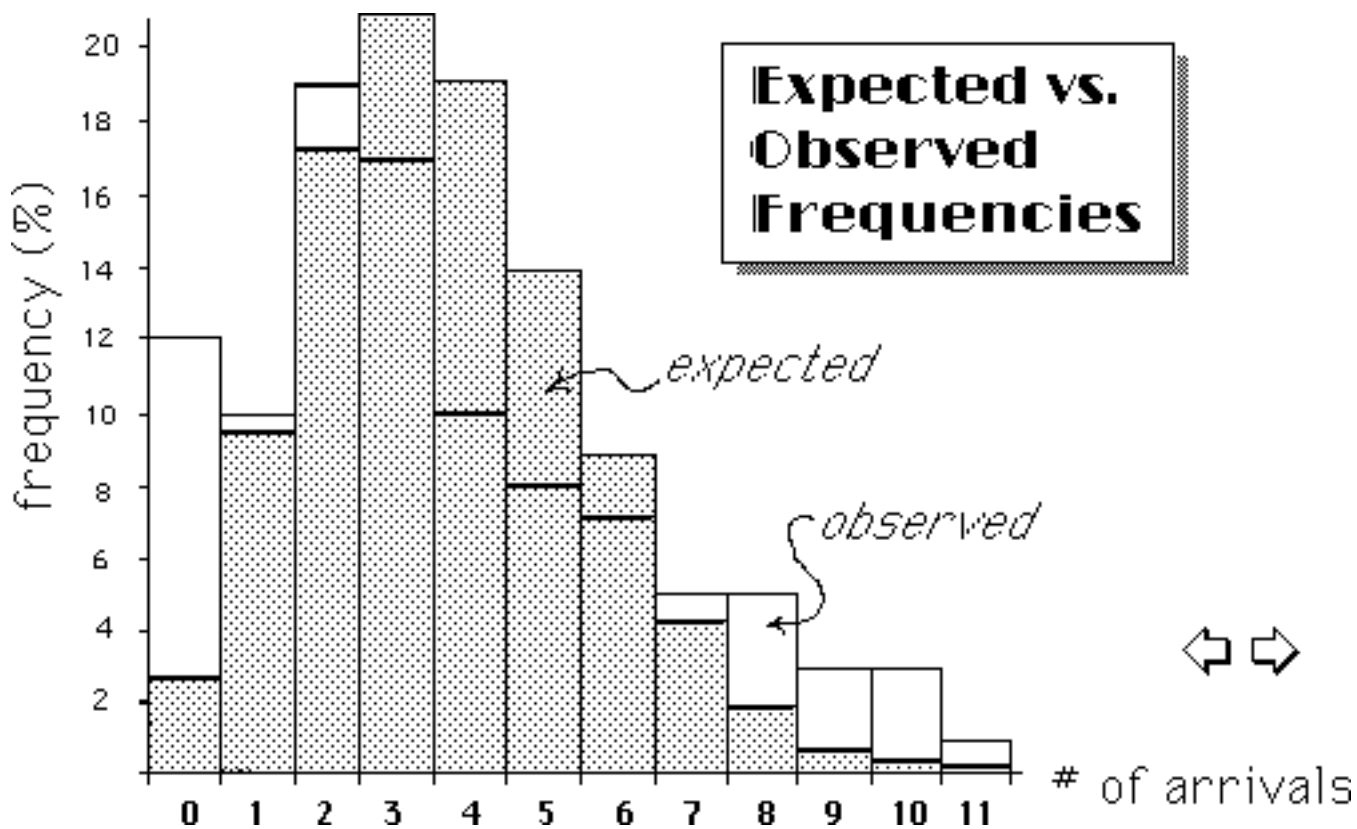
$$P\{N_5 = x\} = \frac{e^{-3.64}3.64^x}{x!}$$

For example, the probability of observing 0 arrivals is, according to this probability model,

$$P\{N_5 = 0\} = 0.026,$$

so that the expected number of observations of 0 arrivals during the 100 days is $0.026 \times 100 = 2.6$.

©D.L.Bricker, U. of IA, 1999



Goodness-of-Fit Test

Suppose that we have N observations of a discrete random variable

X , with $O_i = \#$ of observations of the i^{th} possible value.

Let $p_i =$ probability of observing this i^{th} value.

Then

O_i has the *binomial* distribution,
 with *expected value* $E_i = Np_i$
 and *variance* $= Np_i(1-p_i)$.

By the Central Limit Theorem, $\frac{O_i - Np_i}{\sqrt{Np_i(1-p_i)}}$ has approximately a $N(0,1)$ distribution, so that the sum of the squares

$$D' = \sum_{i=1}^k \frac{(O_i - Np_i)^2}{Np_i(1-p_i)}$$

has approximately a χ^2 distribution.

©D.L.Bricker, U. of IA, 1999

Owing to the lack of independence among the O_i ,
i.e., $O_i = N - \sum_{j \neq i} O_j$

the number of "degrees of freedom" of the variable is not k , but $k-1$, and each term is decreased by a factor of $(1-p_i)$:

$$\begin{aligned} D &= \sum_{i=1}^k \frac{(O_i - Np_i)^2}{Np_i(1-p_i)} (1-p_i) = \sum_{i=1}^k \frac{(O_i - Np_i)^2}{Np_i} \\ &= \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \end{aligned}$$

©D.L.Bricker, U. of IA, 1999

| i | O_i | E_i |
|-----|-------|-------|
| 0 | 12 | 2.6 |
| 1 | 10 | 9.6 |
| 2 | 19 | 17.4 |
| 3 | 17 | 21.1 |
| 4 | 10 | 19.2 |
| 5 | 8 | 14.0 |
| 6 | 7 | 8.5 |
| 7 | 5 | 4.4 |
| 8 | 5 | 2.0 |
| 9 | 3 | 0.8 |
| 10 | 3 | 0.3 |
| 11 | 1 | 0.1 |

The observed frequencies of the number of arrivals differs from the expected number for every i .

Are the differences between O_i and E_i large enough to justify rejecting the Poisson distribution with

$$\lambda = 0.728/\text{min.} \quad ?$$

©D.L.Bricker, U. of IA, 1999

| i | O_i | E_i | $(O_i - E_i)^2 / E_i$ |
|-----|-------|-------|-----------------------|
| 0 | 12 | 2.62 | 33.477 |
| 1 | 10 | 9.55 | 0.020 |
| 2 | 19 | 17.39 | 0.148 |
| 3 | 17 | 21.10 | 0.797 |
| 4 | 10 | 19.20 | 4.410 |
| 5 | 8 | 13.97 | 2.557 |
| 6 | 7 | 8.48 | 0.258 |
| 7 | 5 | 4.41 | 0.078 |
| 8 | 5 | 2.00 | 4.465 |
| 9 | 3 | 0.81 | 5.901 |
| 10 | 3 | 0.29 | 24.762 |
| 11 | 1 | 0.09 | 8.327 |

sum: 85.206

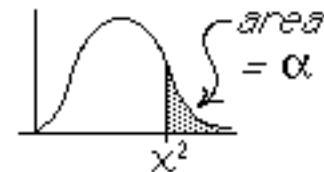
©D.L.Bricker, U. of IA, 1999

Chi-square Tables

From tables of the chi-square distribution, we can find the

value $\chi^2_{\alpha, \nu}$ such that

$$P\{D \geq \chi^2_{\alpha, \nu}\} = \alpha$$



After observing a sample of the random variable X , we can compute the observed value of D and compare it to the tabulated value of $\chi^2_{\alpha, \nu}$.

(ν = degrees of freedom)



©D.L.Bricker, U. of IA, 1999

| degrees of freedom | P{D > χ^2 } | | |
|--------------------|------------------|--------|--------|
| | 10% | 5% | 1% |
| 1 | 2.706 | 3.841 | 6.635 |
| 2 | 4.605 | 5.991 | 9.210 |
| 3 | 6.251 | 7.815 | 11.345 |
| 4 | 7.779 | 9.488 | 13.277 |
| 5 | 9.236 | 11.070 | 15.086 |
| 6 | 10.645 | 12.592 | 16.812 |
| 7 | 12.017 | 14.067 | 18.475 |
| 8 | 13.362 | 15.507 | 20.090 |
| 9 | 14.684 | 16.919 | 21.666 |
| 10 | 15.987 | 18.307 | 23.209 |
| 11 | 17.275 | 19.675 | 24.725 |
| 12 | 18.549 | 21.026 | 26.217 |
| 13 | 19.812 | 22.362 | 27.688 |
| 14 | 21.064 | 23.685 | 29.141 |
| 15 | 22.307 | 24.996 | 30.578 |
| 16 | 23.542 | 26.296 | 32.000 |

Chi-square Table

©D.L.Bricker, U. of IA, 1999

For our example, the number of cells in the histogram is 12, and the number of parameters estimated from the data is 1, so that the number of "degrees of freedom" is $12 - 1 - 1 = 10$.

For $\alpha = 5\%$, $\chi_{\alpha,10}^2 = 18.307$ from the table.

That is, the probability that the observed value of D exceeds 18.307, given the Poisson model with mean value 3.64 arrivals, is only 5%.

Since we calculated the value $D=85.2$, this would lead us to believe that the assumed model is *not* valid!

©D.L.Bricker, U. of IA, 1999

It has been recommended, however, that the observations be "lumped" together so that each cell contains at least 5 to 10 observations.

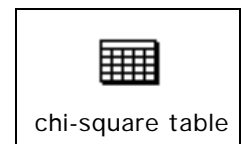
| i | O_i | E_i |
|-----|-------|-------|
| 0 | 12 | 2.62 |
| 1 | 10 | 9.55 |
| 2 | 19 | 17.39 |
| 3 | 17 | 21.10 |
| 4 | 10 | 19.20 |
| 5 | 8 | 13.97 |
| 6 | 7 | 8.48 |
| 7 | 5 | 4.41 |
| 8 | 5 | 2.00 |
| 9 | 3 | 0.81 |
| 10 | 3 | 0.29 |
| 11 | 1 | 0.09 |

The table is annotated with curly braces to group rows:

- Rows 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, and 11 are grouped together with a brace on the right labeled "22".
- Rows 7, 8, 9, 10, and 11 are grouped together with a brace on the right labeled "17".
- Rows 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, and 11 are grouped together with a brace on the right labeled "12.2".
- Rows 7, 8, 9, 10, and 11 are grouped together with a brace on the right labeled "7.6".

©D.L.Bricker, U. of IA, 1999

| i | O_i | E_i | $(O_i - E_i)^2 / E_i$ |
|----------|-------|--------|-----------------------|
| ≤ 1 | 22 | 12.181 | 7.914 |
| 2 | 19 | 17.391 | 0.148 |
| 3 | 17 | 21.101 | 0.797 |
| 4 | 10 | 19.202 | 4.410 |
| 5 | 8 | 13.979 | 2.557 |
| 6 | 7 | 8.480 | 0.258 |
| ≥ 7 | 17 | 7.621 | 11.541 |



Sums: 100 99.959 27.628

The degrees of freedom are now $7 - 1 - 1 = 5$, and for $\alpha = 5\%$, the table gives

$$\chi_{\alpha,5}^2 = 11.070$$

Again, since $D > \chi_{\alpha,5}^2$, we would reject the Poisson model with mean 3.64.

©D.L.Bricker, U. of IA, 1999

| Gap (sec.) | # gaps | Gap (sec.) | # gaps | Gap (sec.) | # gaps |
|------------|--------|------------|--------|------------|--------|
| 0-1 | 18 | 11-12 | 12 | 22-23 | 1 |
| 1-2 | 25 | 12-13 | 6 | 23-24 | 0 |
| 2-3 | 21 | 13-14 | 3 | 24-25 | 1 |
| 3-4 | 13 | 14-15 | 3 | 25-26 | 0 |
| 4-5 | 11 | 15-16 | 3 | 26-27 | 1 |
| 5-6 | 15 | 16-17 | 6 | 27-28 | 1 |
| 6-7 | 16 | 17-18 | 4 | 28-29 | 1 |
| 7-8 | 12 | 18-19 | 3 | 29-30 | 2 |
| 8-9 | 11 | 19-20 | 3 | 30-31 | 1 |
| 9-10 | 11 | 20-21 | 1 | | |
| 10-11 | 8 | 21-22 | 1 | | |

Example

Observed gaps in traffic

total observations=214

©D.L.Bricker, U. of IA, 1999

The average of the observed values of the gaps is

$$\frac{0.5 \times 18 + 1.5 \times 25 + 2.5 \times 21 + \dots + 29.5 \times 2 + 30.5 \times 1}{214}$$


= 7.66354 seconds. If the process is Poisson, i.e., if the gaps have an exponential distribution, the estimated arrival rate is

$$\lambda = \frac{1}{7.66354 \text{ sec.}} = 0.130488 / \text{sec.}$$

(midpoint of each cell was used in computing average.)

©D.L.Bricker, U. of IA, 1999

| Gap | O_j | p_j | E_j | $ O_j - E_j $ | $(O_j - E_j)^2 / E_j$ |
|-------|-------|-----------|---------|---------------|-----------------------|
| 0-1 | 18 | 0.122333 | 26.1793 | 8.17926 | 2.55547 |
| 1-2 | 25 | 0.107368 | 22.9767 | 2.02333 | 0.178175 |
| 2-3 | 21 | 0.094233 | 20.1659 | 0.834134 | 0.0345028 |
| 3-4 | 13 | 0.0827052 | 17.6989 | 4.69892 | 1.24752 |
| 4-5 | 11 | 0.0725876 | 15.5338 | 4.53375 | 1.32324 |
| 5-6 | 15 | 0.0637078 | 13.6335 | 1.36654 | 0.136973 |
| 6-7 | 16 | 0.0559142 | 11.9656 | 4.03436 | 1.36023 |
| 7-8 | 12 | 0.0490741 | 10.5018 | 1.49815 | 0.21372 |
| 8-9 | 11 | 0.0430707 | 9.21713 | 1.78287 | 0.344862 |
| 9-10 | 11 | 0.0378017 | 8.08957 | 2.91043 | 1.0471 |
| 10-12 | 20 | 0.062296 | 13.3313 | 6.66866 | 3.33583 |
| 12-13 | 6 | 0.0255565 | 5.46909 | 0.530915 | 0.0515389 |
| 13-16 | 9 | 0.0593941 | 12.7103 | 3.71033 | 1.0831 |
| 16-19 | 13 | 0.0401543 | 8.59303 | 4.40697 | 2.26013 |
| 19-24 | 6 | 0.0401612 | 8.5945 | 2.5945 | 0.783223 |
| 24-∞ | 7 | 0.043643 | 9.33961 | 2.33961 | 0.58608 |

$214 \times p_j$  $D = 16.5417$

©D.L.Bricker, U. of IA, 1999

$k=16$ intervals, and we have used the data to estimate 1 parameter (λ), so the number of "degrees of freedom" is $16-1-1=14$.

According to the table, for $\alpha = 5\%$, $\chi^2_{0.05,14} = 23.685$
i.e.,

$$P\{D \geq 23.685\} = 5\%$$

and since the observed value of D is 16.54, we do not reject the hypothesis that the traffic gap has an exponential distribution with $\lambda = 0.130/\text{sec}$.



chi-square table

©D.L.Bricker, U. of IA, 1999

In fact, $P\{D \geq 16.5\} \doteq 0.284$

That is, there is about a 28% probability that the value of D would exceed the observed value.

©D.L.Bricker, U. of IA, 1999

Example*"Incredibly good" data*

In a statistics laboratory, students were asked to draw samples of 4 from a bowl containing red and black balls in equal proportions, with the drawn balls being returned into the bowl after each test. 160 samples were to be drawn, and the results reported. Group A reported the following:

| | | | | | |
|----------------|----|----|----|----|----|
| # red balls | 0 | 1 | 2 | 3 | 4 |
| # observations | 9 | 40 | 59 | 41 | 11 |
| # expected | 10 | 40 | 60 | 40 | 10 |

©D.L.Bricker, U. of IA, 1999

| | | | | | |
|-----------------------|----------------|----|----------------|----------------|----------------|
| i | 0 | 1 | 2 | 3 | 4 |
| O_i | 9 | 40 | 59 | 41 | 11 |
| E_i | 10 | 40 | 60 | 40 | 10 |
| $O_i - E_i$ | 1 | 0 | 1 | 1 | 1 |
| $(O_i - E_i)^2 / E_i$ | $\frac{1}{10}$ | 0 | $\frac{1}{60}$ | $\frac{1}{40}$ | $\frac{1}{10}$ |

Sum:
0.24

The number of degrees of freedom is $5 - 1 = 4$, and the table gives, for $\alpha = 99\%$, $\chi_{0.99,4}^2 = 0.297$. Therefore, the probability that the students actually observed the results reported is less than 1%, leading the prof. to suspect that the lab report was faked!

| Time between arrivals (min.) | Frequency |
|------------------------------|-----------|
| 1 | 17 |
| 2 | 15 |
| 3 | 12 |
| 4 | 10 |
| 5 | 9 |
| 6 | 8 |
| 7 | 6 |
| 8 | 6 |
| 9 | 4 |
| 10 | 3 |
| 11 | 3 |
| 12 | 2 |
| 13 | 2 |
| 14 | 1 |
| 15 | 1 |
| 16 | 1 |

How good a fit to the observed data is the exponential probability distribution with mean interarrival time equal to 5 minutes, i.e.,

$$\lambda = \frac{1}{5\text{min.}} = 0.2/\text{min.} ?$$



©D.L.Bricker, U. of IA, 1999

The probability that T_1 , the next arrival time, falls in each interval is found using the *Exponential CDF*:

$$P\{T_1 \leq t\} = F(t) = 1 - e^{-\lambda t} = 1 - e^{-0.2t}$$

$$P\{t_0 \leq T_1 \leq t_1\} = F(t_1) - F(t_0) = e^{-\lambda t_0} - e^{-\lambda t_1}$$

For example,

$$\begin{aligned} P\{0 \leq T_1 \leq 1\} &= F(1) - F(0) = e^0 - e^{-0.2} \\ &= 1 - 0.8187307 \\ &= 0.181269 \end{aligned}$$

$$\begin{aligned} P\{1 \leq T_1 \leq 2\} &= F(2) - F(1) = e^{-0.2} - e^{-0.4} \\ &= 0.8187307 - 0.67032 \\ &= 0.148411 \end{aligned}$$

©D.L.Bricker, U. of IA, 1999

| i | O_i | p_i | E | $(E_i - O_i)^2 / E_i$ |
|----|-------|------------|-------------|-----------------------|
| 1 | 17 | 0.18126925 | 18.12692469 | 0.07005928 |
| 2 | 15 | 0.14841071 | 14.84107070 | 0.00170193 |
| 3 | 12 | 0.12150841 | 12.15084099 | 0.00187255 |
| 4 | 10 | 0.09948267 | 9.94826720 | 0.00026902 |
| 5 | 9 | 0.08144952 | 8.14495229 | 0.08976192 |
| 6 | 8 | 0.06668523 | 6.66852293 | 0.26585066 |
| 7 | 6 | 0.05459725 | 5.45972480 | 0.05346374 |
| 8 | 6 | 0.04470045 | 4.47004459 | 0.52365552 |
| 9 | 4 | 0.03659763 | 3.65976298 | 0.03163080 |
| 10 | 3 | 0.02996360 | 2.99636050 | 0.00000442 |
| 11 | 3 | 0.02453212 | 2.45321249 | 0.12187146 |
| 12 | 2 | 0.02008521 | 2.00852051 | 0.00003615 |
| 13 | 2 | 0.01644438 | 1.64443751 | 0.07688020 |
| 14 | 1 | 0.01346352 | 1.34635156 | 0.08909961 |
| 15 | 1 | 0.01102299 | 1.10229943 | 0.00949395 |
| 16 | 1 | 0.00902486 | 0.90248644 | 0.01053633 |

$E_i = 100p_i$

Sum: $D = 1.34619$

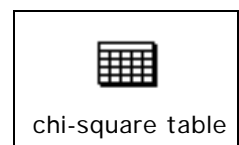
©D.L.Bricker, U. of IA, 1999

From a χ^2 -probability table, it appears that

$$P\{D \leq 1.4\} = 0.001\%$$

That is, the fit appears "too good"!

(One should be suspicious that the recorded data was fabricated or "doctored"!)



©D.L.Bricker, U. of IA, 1999

| Service Time (min.) | Frequency |
|---------------------|-----------|
| 1 | 51 |
| 2 | 23 |
| 3 | 12 |
| 4 | 7 |
| 5 | 4 |
| 6 | 2 |
| 7 | 1 |

How good a fit to the observations is the exponential distribution with mean service time of 2 minutes, i.e.,

$$\mu = 1/2 \text{ min.} = 0.5/\text{min.}?$$

©D.L.Bricker, U. of IA, 1999

| i | O_i | p_i | E_i | $(E_i - O_i)^2/E_i$ |
|---|-------|-----------|---------|---------------------|
| 1 | 51 | 0.393469 | 39.3469 | 3.4512 |
| 2 | 23 | 0.238651 | 23.8651 | 0.0313611 |
| 3 | 12 | 0.144749 | 14.4749 | 0.423164 |
| 4 | 7 | 0.0877949 | 8.77949 | 0.360679 |
| 5 | 4 | 0.0532503 | 5.32503 | 0.329707 |
| 6 | 2 | 0.0322979 | 3.22979 | 0.468262 |
| 7 | 1 | 0.0195897 | 1.95897 | 0.469441 |
| 8 | 0 | 0.0301974 | 3.01974 | 3.01974 |

Sum: $D=8.55355$

degrees of freedom = $8 - 1 - 1 = 6$.

What is χ^2 such that $P\{D > \chi^2\} = 5\%$?

©D.L.Bricker, U. of IA, 1999

From a χ^2 -probability table,

$$P\{D > \chi^2\} = 5\% \quad \text{for } \chi^2 = 12.59$$

The observed value of $D=8.55$ is, therefore, not so large as to indicate a poor fit of the model to the observed data.



©D.L.Bricker, U. of IA, 1999

| n | 99% | 95% | 90% | ... | 10% | 5% | 2% | 1% |
|----|--------|--------|--------|-------------------------|--------|--------|--------|--------|
| 1 | 0.0002 | 0.004 | 0.0158 | | 2.706 | 3.841 | 5.412 | 6.635 |
| 2 | 0.0201 | 0.103 | 0.211 | | 4.605 | 5.991 | 7.824 | 9.210 |
| 3 | 0.115 | 0.352 | 0.584 | | 6.251 | 7.815 | 9.837 | 11.341 |
| 4 | 0.297 | 0.711 | 1.064 | | 7.779 | 9.488 | 11.668 | 13.277 |
| 5 | 0.554 | 1.145 | 1.610 | | 9.236 | 11.070 | 13.388 | 15.086 |
| 6 | 0.872 | 1.635 | 2.204 | | 10.645 | 12.592 | 15.033 | 16.812 |
| 7 | 1.239 | 2.167 | 2.833 | | 12.017 | 14.067 | 16.622 | 18.475 |
| 8 | 1.646 | 2.733 | 3.490 | | 13.362 | 15.507 | 18.168 | 20.090 |
| 9 | 2.088 | 3.325 | 4.168 | | 14.684 | 16.919 | 19.679 | 21.666 |
| 10 | 2.558 | 3.940 | 4.865 | | 15.987 | 18.307 | 21.161 | 23.209 |
| 11 | 3.053 | 4.575 | 5.578 | <p>Chi-square Table</p> | 17.275 | 19.675 | 22.618 | 24.725 |
| 12 | 3.571 | 5.226 | 6.304 | | 18.549 | 21.026 | 24.054 | 26.217 |
| 13 | 4.107 | 5.892 | 7.042 | | 19.812 | 22.362 | 25.472 | 27.688 |
| 14 | 4.660 | 6.571 | 7.790 | | 21.064 | 23.685 | 26.873 | 29.141 |
| 15 | 5.229 | 7.261 | 8.547 | | 22.307 | 24.996 | 28.259 | 30.578 |
| 16 | 5.812 | 7.962 | 9.312 | | 23.542 | 26.296 | 29.633 | 32.000 |
| 17 | 6.408 | 8.672 | 10.085 | | 24.769 | 27.57 | 30.995 | 33.409 |
| 18 | 7.015 | 9.390 | 10.865 | | 25.989 | 28.869 | 32.346 | 34.805 |
| 19 | 7.633 | 10.117 | 11.651 | | 27.204 | 30.144 | 33.687 | 36.191 |
| 20 | 8.260 | 10.851 | 12.443 | | 28.412 | 31.410 | 35.020 | 37.566 |

degrees of freedom

©D.L.Bricker