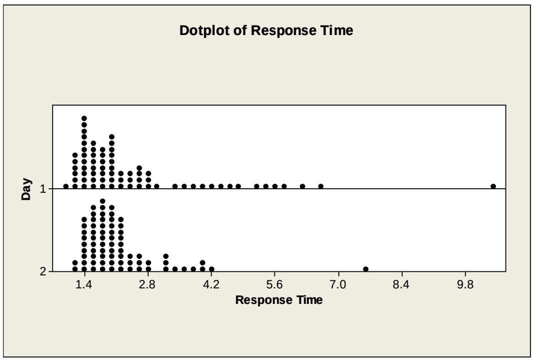
Response Times - Solution

(1) Descriptive statistics

We will in section (1)-(5) below disregard any variations in response times during the working hours, but will return to this from (6) on.

First consider the distribution of the observed response time before and after the system change. We see from the dotplot below that the distributions are both skewed with a long right tail, with a longer tail before the system change than after.



Here are some descriptive statistics

Descriptive Statistics: Response Time
```
Variable       Day  Count   Mean  SE Mean  StDev  Minimum     Q1  Median     Q3  Maximum
Response Time  1       72  2.523    0.193  1.641    1.080  1.475   1.955  2.785   10.390
               2       72  2.170    0.112  0.953    1.140  1.583   1.940  2.310    7.680
```

The mean response time is relevant for the total search time, while the median, quantiles and maximum are relevant for judging the response time as an immediate irritation. We see that both the mean and the median are reduced after the system change, the mean more so than the median. Whether this reduction is statistically significant will be investigated below.
We also see that the variation measured by the standard deviation (StDev) is reduced after the system change. The minimum may indicate a lower threshold on the response time of about one second, perhaps for a physical reason.

(2) - (4) One and Two-sample analyses

The sample mean and median may be taken as estimates of the respective long run mean (expectation) and long run median. Here follows these estimates with associated 95% confidence intervals.

One-Sample T: RespTime_1; RespTime_2
```
Variable     N   Mean  StDev  SE Mean       95% CI
RespTime_1  72  2.523  1.641    0.193  (2.138; 2.909)
RespTime_2  72  2.170  0.953    0.112  (1.946; 2.394)
```

The number of times the response time exceed 5 seconds are 7 among the 72 response times the first Wednesday and 1 among the 72 the next Wednesday. Using the fractions as estimates of the probabilities of exceeding 5 seconds before and after the system change, we can compute estimates and approximate confidence intervals by the standard textbook formula as follows:

$$\frac{7}{72} \pm 1.96 \cdot \sqrt{\frac{7/72 \cdot (1-7/72)}{72}} \text{ and } \frac{1}{72} \pm 1.96 \cdot \sqrt{\frac{1/72 \cdot (1-1/72)}{72}}$$

or using standard software

Confidence Interval for One Proportion: Exceed5_1; Exceed5_2

```
Variable   X   N  Sample p         95% CI
Exceed5_1  7  72  0.097222  (0.028791; 0.165654)
Exceed5_2  1  72  0.013889  (0.000000; 0.040921)

Using the normal approximation.
The normal approximation may be inaccurate for small samples.
```

The computation assumes normal approximation, which is best for large samples and p close to ½, but is far from being so here. Some software may provide exact confidence intervals (by more complicated formulas usually not given in textbooks). Here they are, and we see that the approximate intervals are too optimistic.

Confidence Interval for One Proportion: Exceed5_1; Exceed5_2

```
Variable   X   N  Sample p         95% CI
Exceed5_1  7  72  0.097222  (0.039990; 0.190110)
Exceed5_2  1  72  0.013889  (0.000352; 0.074971)
```

For estimates of the median, see below.

In the given context we may be willing to assume that the system change cannot worsen things, i.e. the situation is one-sided. We have therefore

Null hypothesis: No change in response times with the system change.
Alternative hypothesis: Response times improved with the system change

These hypotheses may be expressed formally in a variety of ways by relevant parameters. For the organisation the total search time is most important, and this may be expressed by the expected response. For the individual the felt nuisance by occasional long response times may be more important, and this may be expressed by probabilities of extremes.
One may also think of expressing hypothesis in terms of median change, but this seems less relevant in the given context.

The hypotheses are expressed in terms of expected response time may be tested by the two-sample t-test:

Two-Sample T-Test and CI: Response Time; Day

```
Two-sample T for Response Time

Day   N   Mean  StDev  SE Mean
1    72  2.523  1.641    0.193
2    72  2.170  0.953    0.112

Difference = mu (1) - mu (2)
Estimate for difference:  0.354
95% CI for difference:  (-0.089; 0.797)
T-Test of difference = 0 (vs > 0): T-Value = 1.58  P-Value = 0.058  DF = 113
```

We see that the difference in means is not statistically significant at the 5% significance level, but close to being so. We have done a two-sample analysis not assuming equal variances in the two groups. If we had taken the standard textbook assumption of equal variances, not justified here, we would have obtained the same p-value and same conclusion.

Exact calculation of confidence guarantees and p-values assumes normally distributed observations. This is clearly not justified here.

The common non-parametric alternative to the two-sample t-test is the Wilcoxon (Mann-Whitney) test, but note that this is in fact a test for equal medians. This test came out as follows, with no significant difference as well:

Mann-Whitney Test and CI: RespTime_1; RespTime_2

```
             N  Median
RespTime_1  72  1.9550
RespTime_2  72  1.9400

Point estimate for ETA1-ETA2 is 0.0100
95.0 Percent CI for ETA1-ETA2 is (-0.2000;0.2500)
W = 5237.5
Test of ETA1 = ETA2 vs ETA1 > ETA2 is significant at 0.4729
```

Software may also provide the following, which also provides the individual confidence intervals of the medians

Mood Median Test: Response Time versus Day

```
Mood median test for Response Time

Chi-Square = 0.00    DF = 1    P = 1.000

                             Individual 95.0% CIs
Day  N<=  N>  Median  Q3-Q1  ------+---------+---------+---------+
1     36  36   1.955  1.310  (------------------*---------------)
2     36  36   1.940  0.727        (-----------*---------)
                             ------+---------+---------+---------+
                                 1.80      1.92      2.04      2.16
Overall median = 1.945
```

The estimate of the difference between the probabilities of response times exceeding 5 seconds, and the testing of they being equal, came out as follows. We see that the difference is significant at the 5% level.

Test and Confidence Interval for Two Proportions: Exceed5_1; Exceed5_2

```
Variable   X   N  Sample p
Exceed5_1  7  72  0.097222
Exceed5_2  1  72  0.013889

Difference = p (Exceed5_1) - p (Exceed5_2)
Estimate for difference:  0.0833333
95% CI for difference:  (0.00975631; 0.156910)
Test for difference = 0 (vs > 0):  Z = 2.22  P-Value = 0.013

Fisher's exact test: P-Value = 0.031

* NOTE * The normal approximation may be inaccurate for small samples.
```

However, the computation leading to P=0.013 is based on the U-test with normal approximation, which may not be very good for skew cases (here small probabilities), even if the sample is moderate. Some software provides exact P-values as well. Here Fishers exact test says at that the difference in probabilities is still significant at 5% level, but if we had not assumed a one-sided situation at the outset, the exact p-value would have been doubled, and not significant at 5% level (while the standard textbook p-value would have made us to believe that it was!)

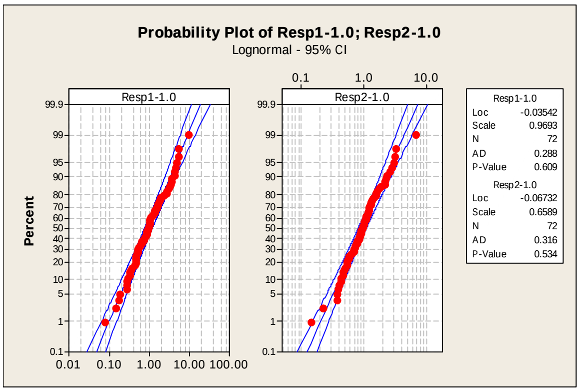
Note on assumptions for the tests:

For the t-test the normal assumption is clearly not satisfied. The W-test is often recommended as an alternative, in particular in situations where expectation and median coincide. This is not the case here, and the W-test has seemingly low power against the alternatives to the null hypothesis relevant in this context. The t-test is not significant at 5% level either. The U-test assumes independence and constant probability. This is questionable, but not a serious objection.

All estimation and tests are based on the assumption of constancy over the working hours. This is likely not to be the case, as we shall see in (8). If the lunch period has lower expected response times than otherwise, a possible way out is to omit this period from the analysis. However, we will overcome such an objection by more sophisticated analysis below.

(5) Distribution analysis

With specific distributional assumptions, we have better opportunity to study problems analytically, and at the same time observational randomness will be ”smoothed out”. We want a distribution which is both analytically tractable and fits available data well. By physical reasons the response times may be expected to never become smaller than a threshold a. For the excess a lognormal distribution may be worth a try, since it starts out at zero and has a long right tail, as the dotplots have shown. The choice of a=1.0 seems reasonable from the observed minimum. Two probability plots for the fit of observed (Response Time – 1.0) to lognormal before and after the system change follows. They show that the points are approximately following a straight line, with a high P-value. This supports a log-normal assumption, even if the one point outside the confidence limits at the right end may indicate a distribution with slightly heavier tail. Other distributions may be tried out as well, e.g. the gamma distribution, which does not give the same good fit.



In the box we get the corresponding estimates of location and scale for the corresponding normal distribution of $X = \ln(T-1)$. With these estimates as true values we compute (estimate) the probabilities of response time more than 5 seconds to be

$$P(T>5) = P(\ln(T-1) > \ln(5-1)) = P(X > 1.3863) \text{ where } X \sim \mathrm{N}(\mu,\sigma)$$

Before: $\mu = -0.03542 \quad \sigma = 0.9693 \quad P(X > 1.3863) = 0.0712$

After: $\mu = -0.06732 \quad \sigma = 0.6589 \quad P(X > 1.3863) = 0.0137$

We now want to find the maximal response time we can guarantee with 99.5% certainty after the system change. Note that $T = 1 + e^X$ and that $P(X \le 1.6299) = 0.995$ for the given $X \sim \mathrm{N}(\mu,\sigma)$. We then get

$$P(T \le 1 + e^{1.6299}) = P(T \le 6.1033)$$

We can therefore give a 99.5% guarantee that any individual response time is at most 6.1033.

(6) - (7) Matched pair analysis

We now take the difference between the response times before and after the system change for the same period of the day. We now allow the expected response times to vary over the working hours, but assume constant expected change, which is now the key parameter.
The matched pair t-test is based on normality assumptions on the differences. A histogram of the observed differences will show a distribution with long tail to the right as well, but the violation of normality is not as critical here as for the two sample test above.

Paired T-Test and Confidence Interval: RespTime_1; RespTime_2

```
              N   Mean  StDev  SE Mean
RespTime_1  72  2.523  1.641    0.193
RespTime_2  72  2.170  0.953    0.112
Difference  72  0.354  1.574    0.185

95% lower bound for mean difference: 0.045
T-Test of mean difference = 0 (vs > 0): T-Value = 1.91  P-Value = 0.030
```

We see that the change is statistically significant at the 5% level, while we did not get significance by the previous two-sample t-test.

For the corresponding test of the median change (Wilcoxon signed rank test) the change turned out not significant, but this test does not pick up what we are interested in anyway

Paired difference test tests do not provide the same opportunity to formulate hypotheses about unacceptable response times.

If the lunch period has lower expected response times than otherwise, a possible way out is to omit this period from the analysis.

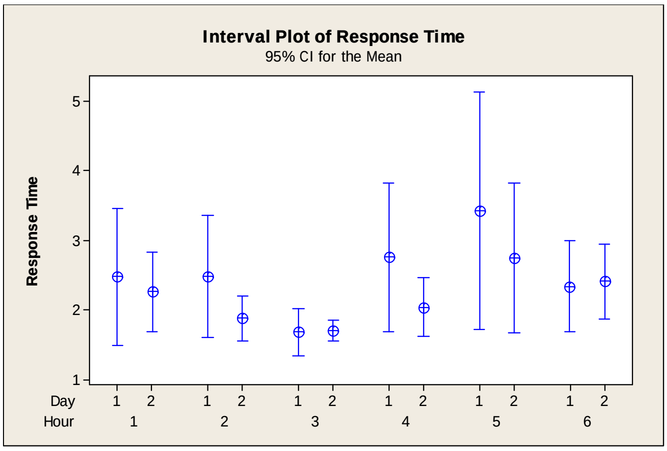
(8) ANOVA with variations over working hours

Descriptive statistics (mean, maximum,standard deviation) are given for each day and each hour of the day as follows:

Tabulated statistics: Day; Hour

```
Rows: Day   Columns: Hour

          1        2        3        4        5        6

1     2.473    2.483    1.677    2.753    3.424    2.330
      6.270    5.260    2.810    6.550   10.390    4.360
     1.5466   1.3809   0.5309   1.6781   2.6756   1.0291

2     2.258    1.874    1.705    2.037    2.739    2.405
      4.170    3.210    2.050    3.790    7.680    4.050
     0.8948   0.4996   0.2359   0.6669   1.6852   0.8358

Cell Contents:   Response Time   :  Mean
                 Response Time   :  Maximum
                 Response Time   :  Standard deviation
```

The result is more easily comprehended by the following plot, showing the means and the associated confidence intervals for the true expectation within each day and hour.



We see clearly that variation in the expected response times over the working hours, both before and after the system change. The lunch hour (hour 3) has short response times, while the next to last hour (hour 5) has long response times. The other hours have intermediate response times, with slight differences. We also see the improvement from Day 1 (before) and Day 2 (after), except for the last hour (which may be just bad luck). Note however that the confidence intervals are often fairly wide, due to the fact that now there are only 12 observations behind the estimation of each separate expectation. In fact doing a one-factor ANOVA for Day1 with Hour as factor, this factor does not come out as statistically significant (P=0.202).

One-way ANOVA: RespTime_1 versus Hour_1

```
Source  DF      SS    MS     F      P
Hour_1   5   19.47  3.89  1.50  0.203
Error   66  171.76  2.60
Total   71  191.24

S = 1.613   R-Sq = 10.18%   R-Sq(adj) = 3.38%

                              Individual 95% CIs For Mean Based on
                              Pooled StDev
Level   N   Mean  StDev   ---+---------+---------+---------+------
1      12  2.473  1.547           (---------*--------)
2      12  2.483  1.381            (--------*--------)
3      12  1.677  0.531   (---------*--------)
4      12  2.753  1.678              (---------*--------)
5      12  3.424  2.676                     (--------*---------)
6      12  2.330  1.029          (--------*---------)
                          ---+---------+---------+---------+------
                           1.0       2.0       3.0       4.0
```

A two-factor ANOVA with Hour and Day as factors comes out as follows:

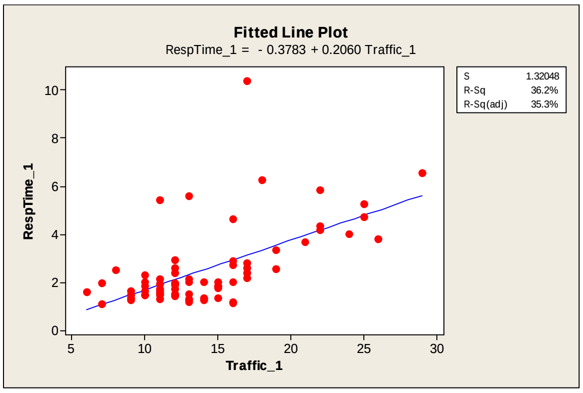
Two-way ANOVA: Response Time versus Hour; Day

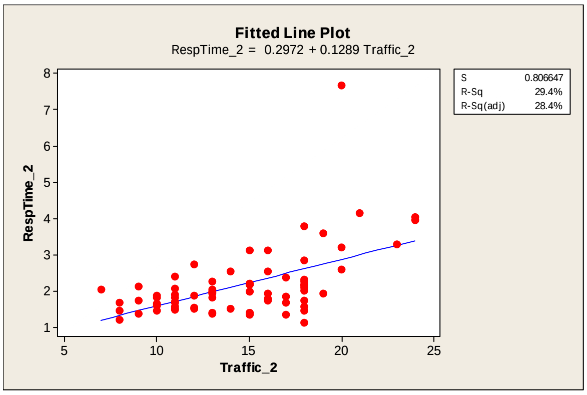
```
Source    DF       SS       MS     F      P
Hour       5   24.040  4.80797  2.84  0.018
Day        1    4.509  4.50854  2.67  0.105
Error    137  231.679  1.69109
Total    143  260.227

S = 1.300   R-Sq = 10.97%   R-Sq(adj) = 7.07%
```

We see that Hour comes out statistically significant, but Day does not. This may come as a surprise, but it is partly due to the large variations. Since Day has two categories and we have assumed one-sidedness, it is legitimate to divide its P-value in half, to obtain P=0.053, still not significant at 5% level.

(9) Regression analysis

Plotting Response Times against Traffic before and after the system change looks like this:





Here regression lines are fitted as well. We see that there is a tendency for response times to increase with traffic. The relationship has some linear features, except some outlying observations, and perhaps a tendency for a jump at some point (about 20).

A multiple regression analysis with Traffic and Day as explanatory variable gave the following:

Regression Analysis: Response Time versus Traffic; Day

```
The regression equation is
Response Time = 0.500 + 0.174 Traffic - 0.431 Day

Predictor     Coef  SE Coef      T      P
Constant    0.5003   0.4077   1.23  0.222
Traffic    0.17428  0.02096   8.31  0.000
Day        -0.4313   0.1841  -2.34  0.021

S = 1.10318   R-Sq = 34.1%   R-Sq(adj) = 33.1%
```

We see that both Traffic and Day are significant at the 5% level. The one-sided p-value for Day becomes 0.021/2=0.0105. The explanatory power measured by R-squared is 34.1% only. Expected increase in response time per addition of a request is 0.174 and the expected reduction in response time with the system change is 0.431.

The outlying observations of course ruin the basic assumption of constant variance and normality for making exact inferences. Nevertheless we feel comfortable with the general conclusion. Since the outliers are typically due to some special cause, it makes sense to remove them from the observations. The two most outlying observations are no. 53 and no.124, and the next two are no. 51 and no.52, so something special may have happened then. Removing these four observations the explanatory power measured by R-squared is increased to 51% with S reduced to 0.718. ThePp-value of Day changes only from 0.021/2 to 0.019/2. However, the regression coefficient is a bit smaller in absolute value, which tells us that the outlying observations may have caused the effect of the system change to appear larger than it really is.

Our observations are collected from one-minute periods five minutes apart. It is a possibility that problems mounts and that dependencies in response times may occur for subsequent periods. This may cause positive covariances that may have been taken into account in some analyses, but it is hardly a major problem, and could safely be neglected. Analysis of autocorrelation functions will indicate a time series of the MA(1) type for both Response Time and Traffic, i.e. some effects from one period to the next, but not longer.

(10) Explaining the probability of extremes

If we are mainly interested in the risk of large response times, say above 5 seconds, we may do a logistic regression as follows:

Binary Logistic Regression: Exceed5 versus Traffic; Day

```
Logistic Regression Table
                                                  Odds     95% CI
Predictor       Coef    SE Coef      Z      P  Ratio  Lower  Upper
Constant    -5.81408    1.44466  -4.02  0.000
Traffic     0.221325  0.0761255   2.91  0.004   1.25   1.07   1.45
Day         -2.04912    1.10416  -1.86  0.063   0.13   0.01   1.12

Log-Likelihood = -23.536
Test that all slopes are zero: G = 14.721, DF = 2, P-Value = 0.001
```

Again we see that Traffic and Day are significant at 5% level (p=0.063/2=0.0315)

(11) Poisson traffic?

The number of requests in given time period being Poisson distributed conforms with equal request tendency over time and that they come independent of each other. Independence may be reasonable, but the results above indicate differences during the day, e.g. lower frequency in the lunch period. The Poisson assumption may be tested with a chi-square test, where we compare the observed frequency in the periods with the expected according to the Poisson distribution. If the lunch period is taken out (and perhaps also the busiest hour (hour 5), it turns out that such a test provides no evidence for rejecting the Poisson assumption. The average per minute for the whole day is 14, while it is 16 in the busiest hour. Planning using the latter is reasonable.

Cumulative Distribution Function

```
Poisson with mean = 16
 x  P( X <= x )
25     0.986881
```

so that $P(X > 25) = 1 - P(X \le 25) = 0.0131$.

By hand with normal approximation (and continuity correction) we get instead

$$P(X > 25) = 1 - P(X \le 25) \approx 1 - G((25.5 - 16)/\sqrt{16}) = 1 - G(2.375) = 1 - 0.9912 = 0.0088$$