School of Computer Science

# EVALUATION METHODS AND STATISTICS
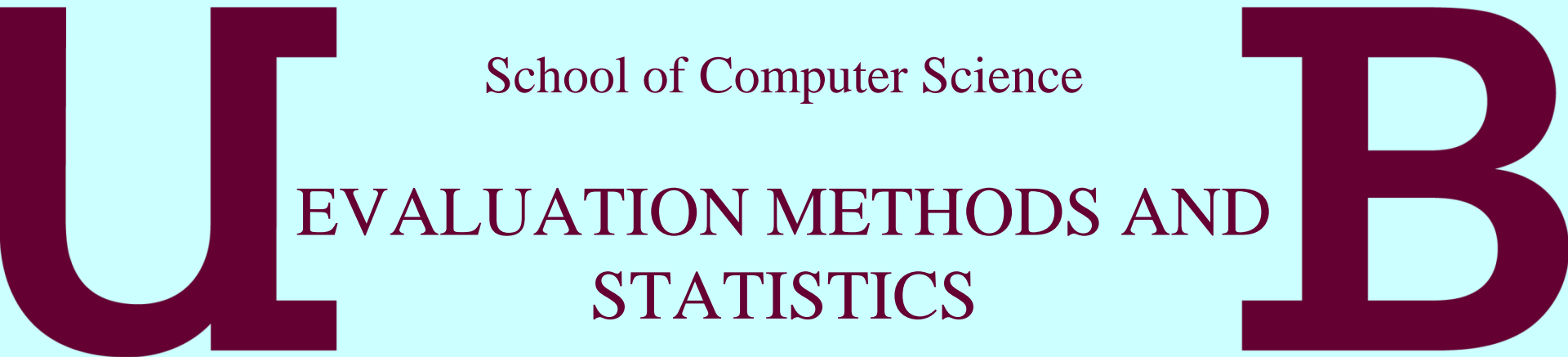
Prof. Chris Baber

Chair of Pervasive and Ubiquitous Computing

# The Drones Game



☐ The initial experimental design:

| Red Drones only | | | Red + Yellow Drones | | |
|:---:|:---:|:---:|:---:|:---:|:---:|
| Auto | Self | Both | Auto | Self | Both |

UNIVERSITY OF
BIRMINGHAM

# Participants

- ☐ Students on the MSc module 'Evaluation Methods and Statistics
  - N = c.140 students registered on EMS
  - n = 93 games completed.  But, 11 files which could not be used
  - n = 82.  But, 13 participants with errors in config for at least one trial.
  - n = 69 participants completed all three trials with correct config.

# Dependent Variables

- ☐ **Time per Target**
  - – Total time spent playing the game divided by the total 'drone Neutralised'

- ☐ **Beacon activation**
  - – Total 'beaconOnOff'

- ☐ **Sensitivity (d')**
  - – Applied only to 'drone Neutralised'

  d'= $ln\left(\frac{\left(H*(1-FA)\right)}{\left((1-H)*FA\right)}\right)$

# Hypotheses

☐ H1: When participants receive decision support (in the form of the polygon display), their sensitivity will be higher than when they do not have such support.

☐ H2: when participants can see the drones moving, their sensitivity will be higher than when they cannot see the drones.

UNIVERSITY OF BIRMINGHAM

# Data Preparation

- Data for dependent variable were tested for normality (using Shapiro Wilk test in R)
- If data were not shown to be normally distributed, a Box-Cox transformation was applied
- If data (following transformation) were not normally distributed, a non-parametric analysis was applied

# Time per Target
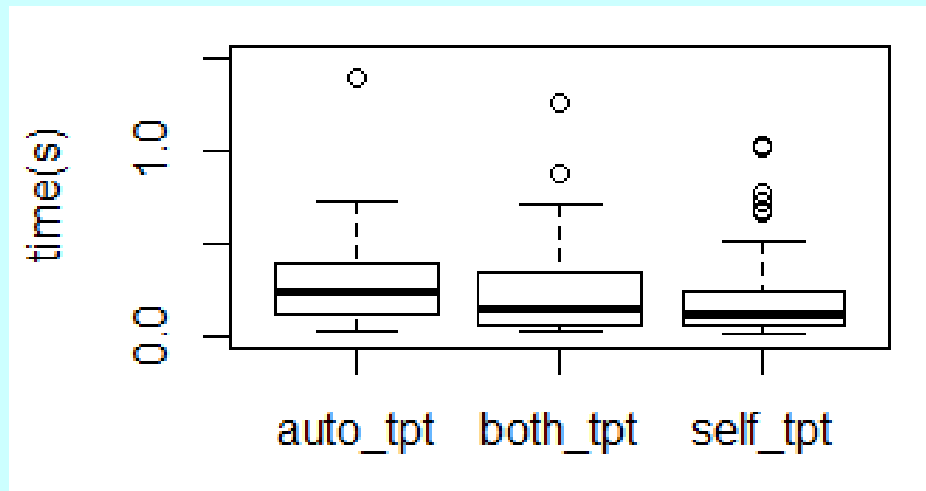


- Shapiro-Wilk: w= 0.251, p<2.2e-16.  Box-Cox transformation did not normalise data, so non-parametric tests applied.
- Friedman Analysis of Variance indicated a main effect of condition [$x^2(2)$ = 8.85293, p<0.05].
- Post-hoc, pairwise comparison showed significant difference between the 'auto' and 'self' conditions [z = 2.95, p<0.005, power (d) = 0.358].

# Friedman Analysis of Variance

- Non-parametric tests make no assumption about the underlying distribution of data

- Developed by Milton Friedman (economics Nobel laureate)

- Ranks the set of data and then asks whether the ranks are more likely to be binned in the independent variable categories

# Simple Example…

| Participant | Task 1 | Rank | Task 2 | Rank | Task 3 | Rank |
|---|---|---|---|---|---|---|
| 1 | 18 | | 18.5 | | 19 | |
| 2 | 17.9 | | 18 | | 18.5 | |
| 3 | 16.5 | | 16 | | 17 | |
| 4 | 21 | | 19 | | 20 | |
| 5 | 21 | | 22 | | 22.5 | |
| 6 | 17.5 | | 17.5 | | 19 | |
| 7 | 18 | | 19 | | 20 | |
| 8 | 15 | | 14 | | 15.5 | |
| 9 | 16 | | 17 | | 17.5 | |
| 10 | 18 | | 19.5 | | 20 | |
| 11 | 19 | | 18 | | 19.5 | |
| 12 | 17 | | 18 | | 19 | |

# Simple Example…

| Participant | Task 1 | Rank | Task 2 | Rank | Task 3 | Rank |
|---|---|---|---|---|---|---|
| 1 | 18 | 1 | 18.5 | 2 | 19 | 3 |
| 2 | 17.9 | 1 | 18 | 2 | 18.5 | 3 |
| 3 | 16.5 | 2 | 16 | 1 | 17 | 3 |
| 4 | 21 | 3 | 19 | 1 | 20 | 2 |
| 5 | 21 | 1 | 22 | 2 | 22.5 | 3 |
| 6 | 17.5 | 1.5 | 17.5 | 1.5 | 19 | 3 |
| 7 | 18 | 1 | 19 | 2 | 20 | 3 |
| 8 | 15 | 2 | 14 | 1 | 15.5 | 3 |
| 9 | 16 | 1 | 17 | 2 | 17.5 | 3 |
| 10 | 18 | 1 | 19.5 | 2 | 20 | 3 |
| 11 | 19 | 2 | 18 | 1 | 19.5 | 3 |
| 12 | 17 | 1 | 18 | 2 | 19 | 3 |
| Sum of ranks | | 17.5 | | 19.5 | | 35 |

# Friedman Statistic

$$Q = \frac{12}{mk(k+1)} \sum_{j=1}^{k} R_j^2 - 3m(k+1)$$

- k = levels of independent variable (columns)
- m = discrete measures (rows)
- R = rank across each row

# Simple Example…

k = 3 (one for each task)
m = 12 (one for each participant)
$\sum_{j=1}^{k} R_j^2$ = (17.5^2 + 18.5^2 + 35^2)

Q = 15.29

# Table of critical values

$$M = \frac{12}{nk(k+1)} \sum R_j^2 - 3n(k+1)$$

| n | k=3 α=5% | k=3 α=1% | k=4 α=5% | k=4 α=1% | k=5 α=5% | k=5 α=1% | k=6 α=5% | k=6 α=1% |
|---|---|---|---|---|---|---|---|---|
| 2 | — | — | 6.000 | — | 7.600 | 8.000 | 9.143 | 9.714 |
| 3 | 6.000 | — | 7.400 | 9.000 | 8.533 | 10.130 | 9.857 | 11.760 |
| 4 | 6.500 | 8.000 | 7.800 | 9.600 | 8.800 | 11.200 | 10.290 | 12.710 |
| 5 | 6.400 | 8.400 | 7.800 | 9.960 | 8.960 | 11.680 | 10.490 | 13.230 |
| 6 | 7.000 | 9.000 | 7.600 | 10.200 | 9.067 | 11.870 | 10.570 | 13.620 |
| 7 | 7.143 | 8.857 | 7.800 | 10.540 | 9.143 | 12.110 | 10.670 | 13.860 |
| 8 | 6.250 | 9.000 | 7.650 | 10.500 | 9.200 | 13.200 | 10.710 | 14.000 |
| 9 | 6.222 | 9.556 | 7.667 | 10.730 | 9.244 | 12.440 | 10.780 | 14.140 |
| 10 | 6.200 | 9.600 | 7.680 | 10.680 | 9.280 | 12.480 | 10.800 | 14.230 |
| 11 | 6.545 | 9.455 | 7.691 | 10.750 | 9.309 | 12.580 | 10.840 | 14.320 |
| 12 | 6.500 | 9.500 | 7.700 | 10.800 | 9.333 | 12.600 | 10.860 | 14.380 |
| 13 | 6.615 | 9.385 | 7.800 | 10.850 | 9.354 | 12.680 | 10.890 | 14.450 |
| 14 | 6.143 | 9.143 | 7.714 | 10.890 | 9.371 | 12.740 | 10.900 | 14.490 |
| 15 | 6.400 | 8.933 | 7.720 | 10.920 | 9.387 | 12.800 | 10.920 | 14.540 |
| 16 | 6.500 | 9.375 | 7.800 | 10.950 | 9.400 | 12.800 | 10.960 | 14.570 |
| 17 | 6.118 | 9.294 | 7.800 | 10.050 | 9.412 | 12.850 | 10.950 | 14.610 |
| 18 | 6.333 | 9.000 | 7.733 | 10.930 | 9.422 | 12.890 | 10.950 | 14.630 |
| 19 | 6.421 | 9.579 | 7.863 | 11.020 | 9.432 | 12.880 | 11.000 | 14.670 |
| 20 | 6.300 | 9.300 | 7.800 | 11.100 | 9.400 | 12.920 | 11.000 | 14.660 |
| ∞ | 5.991 | 9.210 | 7.815 | 11.340 | 9.488 | 13.280 | 11.070 | 15.090 |

For values of $n$ greater than 20 and/or values of $k$ greater than 6, use $\chi^2$ tables with $k-1$ degrees of freedom
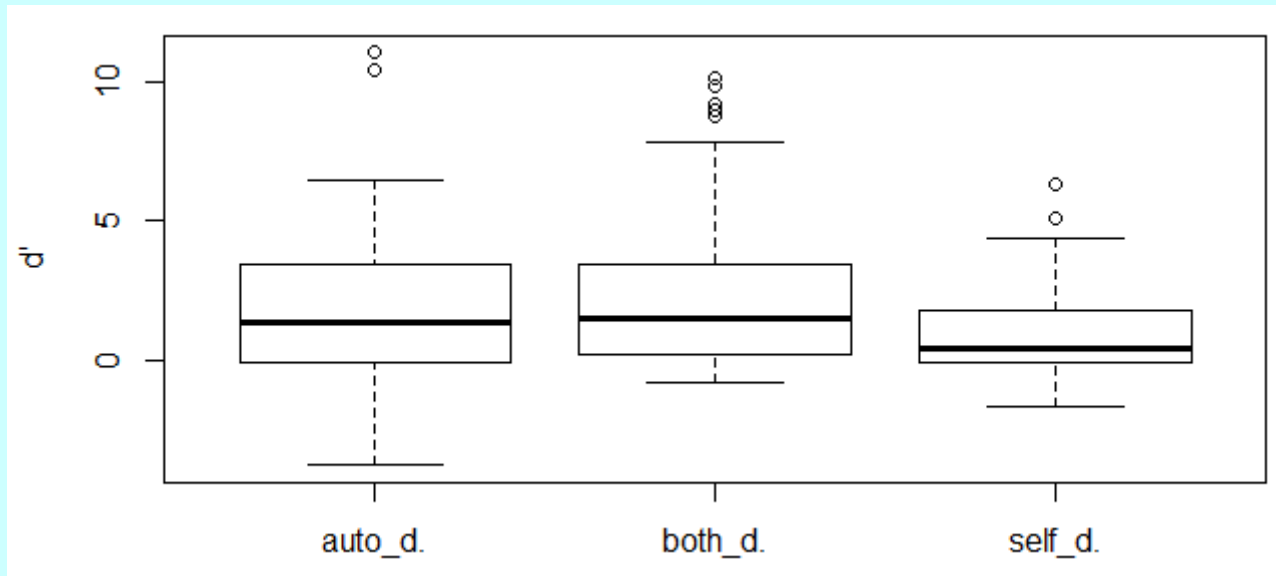
For n = 12, k = 3:  m = 6.5
Calculated Q > m, so reject null hypothesis at p<0.05

UNIVERSITY OF
BIRMINGHAM

# Beacon Activation
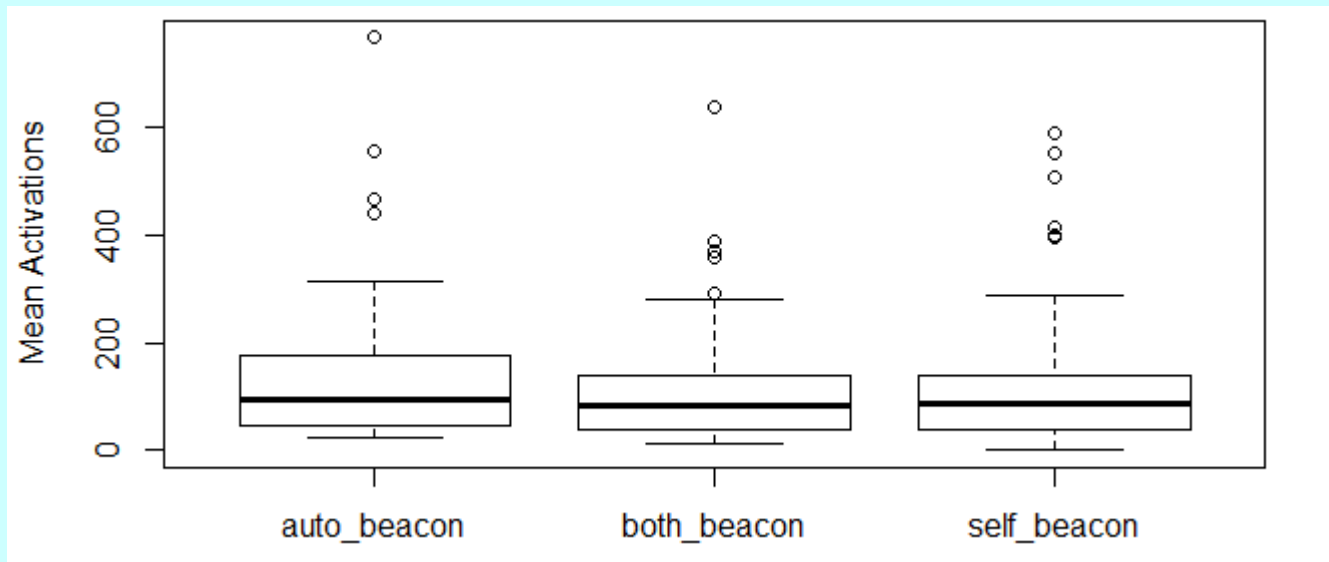


- Shapiro-Wilk: w = 0.776, p<2.2e-16. Box-Cox transformation did not normalise data, so non-parametric tests applied.

- Not significant effect of condition [$x^2$(68) = 4.067, p=0.131]

# Sensitivity



- Shapiro-Wilk…
- A significant main effect of condition [$x2(68) = 7.373$, $p<0.05$].
- Post-hoc, pairwise comparison using Wilcoxon Signed Rank Test showed significant differences between 'auto' and 'self' ($z= 2.188$, $p<0.05$, power (d) = 0.27) and between 'both' and 'self' ($z = 3.318$, $p<0.001$, d=0.40)

# Conclusions

☐ We can accept H1 and conclude that providing participants with decision support (in the form of polygon displays) improves sensitivity and can reduce time per target, relative to a display which does not provide this.

☐ We can reject H2 and conclude the seeing the drones does not improve sensitivity.