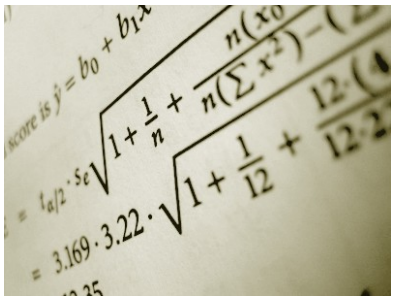
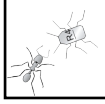
Crash course on statistical analysis



These slides are based on a course by Susan Stepney



Outline

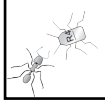
Why we need statistics

Collecting data

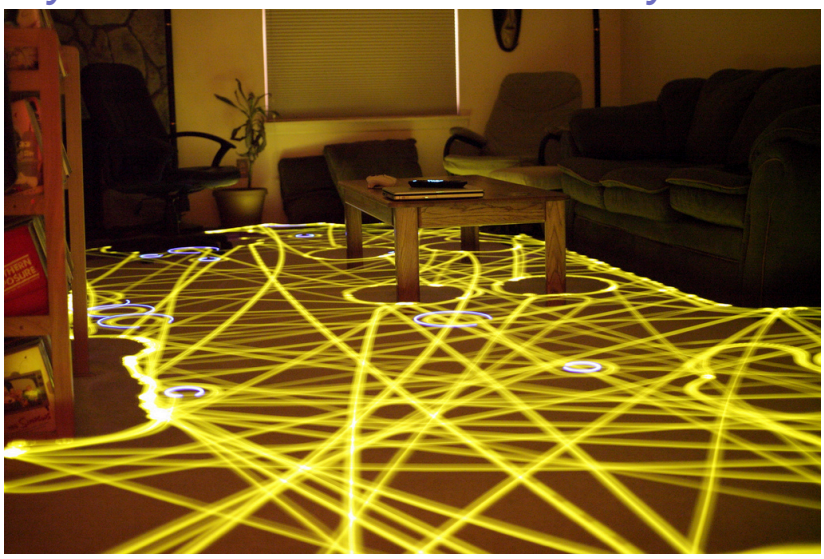
Describing data

Comparing data (statistical tests)

Plotting data



Why do we need statistical analysis?



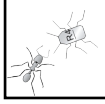
Why do we need statistical analysis?

	Roomba	Team 1	Team 2
Empty arena	81.50	91.97	87.01
Cluttered arena	51.02	66.80	63.53
Round arena	50.16	46.21	75.70

These data tell you **nothing** about which algorithm is better.

Statistics

- A statistic is a summary of the data, a value that captures some characteristic of the data
- The statistics you should calculate depend on what you are interested in saying about the data and the kind of data we have
- Data should be collected accordingly
- Statistics allow you to “tell a story” about the data



Outline

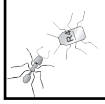
Why we need statistics

Collecting data

Describing data

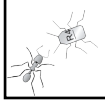
Comparing data (statistical tests)

Plotting data



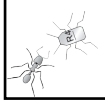
Collecting data

- How you collect data depends on the question you are asking in your experiments.
- The following decisions need to be made:
  1. What type of data needs to be collected?
  2. How much data do you need to collect?
  3. What experiments do you conduct?



Types of data

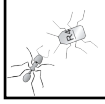
- Nominal / categorical:
  - Data can be grouped into categories (e.g., on / off; blue / brown / green eyes; time spent hitting obstacles / driving forward / spiraling / wall following)
- Ordinal / ranked:
  - Data can be ranked (a < b < c; e.g., very poor < poor < satisfactory < good < excellent …)
- Numerical:
  - Data can describe intervals (e.g., temperature, date) or ratios (e.g., length, mass, % of dust collected)



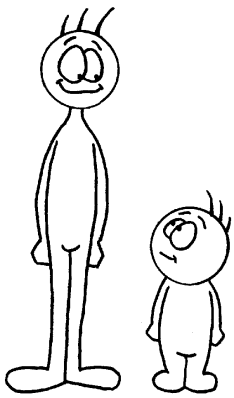
How much data?

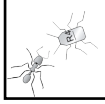
	Roomba	Team 1	Team 2
Cluttered arena	**51.02**	66.80	63.53

- Running an experiment once tells you nothing about it! The result may have been achieved through pure chance.



How much data?

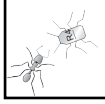




How much data?

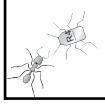
	Roomba	Team 1	Team 2
Cluttered arena	**51.02**	66.80	63.53

- Running an experiment once tells you nothing about it! The result may have been achieved through pure chance.
- Example: are Swiss people tall?
- Run an experiment under identical conditions at the very least 5 times. For scientific papers the minimum is closer to 20.



Experiments

- When comparing between experiments, make sure to change only one thing at a time! (e.g., increasing mutation rate and crossover rate at the same time is pointless)
- Often, baseline experiments are used to compare your experimental data. These generate data under some "default" condition, where no effect is expected. This data highlights the changes due to your experiments.



Outline

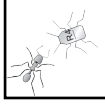
Why we need statistics

Collecting data

Describing data

Comparing data (statistical tests)

Plotting data



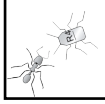
Mean and standard deviation (SD)

- Sample mean: $m = \frac{1}{n}\sum_{i=1}^{n} x_i$

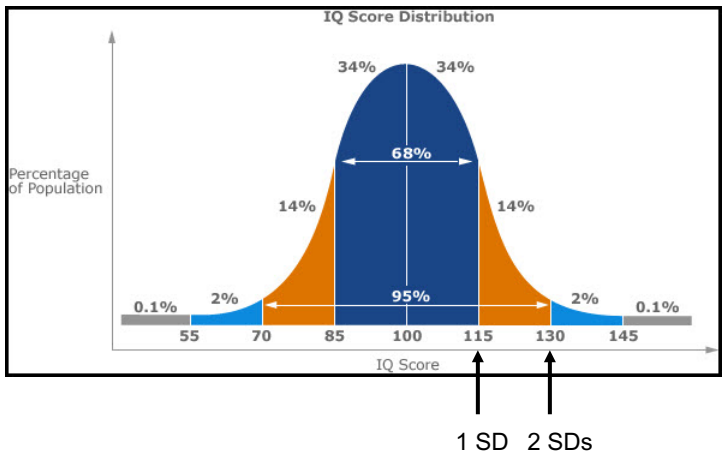
  Average value of data (measure of centrality)
  mean({-30, 1, 2, 3, 4}) = -4
  mean({0, 1, 2, 3, 4}) = 2

- Sample standard deviation $SD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(m - x_i)^2}$

  A measure of the spread of data values
  SD({-30, 1, 2, 3, 4}) = 14.6
  SD({0, 1, 2, 3, 4}) = 1.6

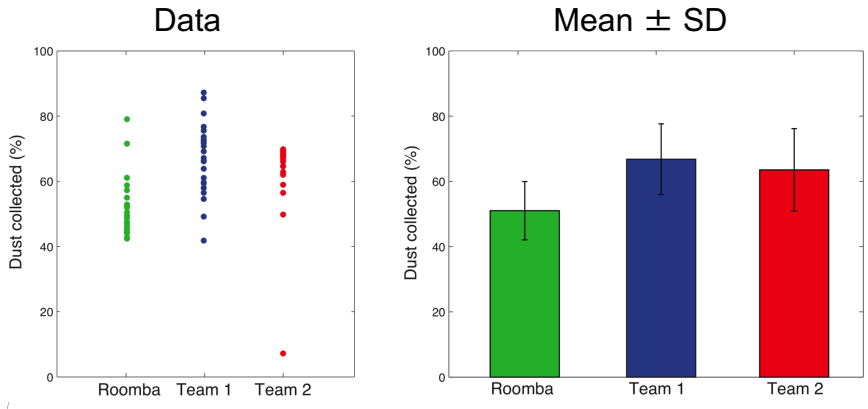


Mean and standard deviation (SD)



Mean and standard deviation (SD)

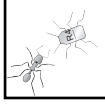
	Roomba	Team 1	Team 2
Cluttered arena	**51.02 ± 8.93**	66.80 ± 10.84	63.53 ± 12.67



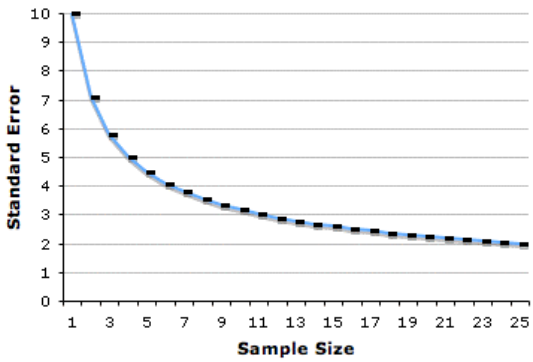
Standard deviation vs. standard error

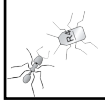
- Standard error of the mean $SE = \frac{SD}{\sqrt{n}}$

  A measure of the spread of data values
  SE({-30, 1, 2, 3, 4}) = 6.53
  SE({0, 1, 2, 3, 4}) = 0.72

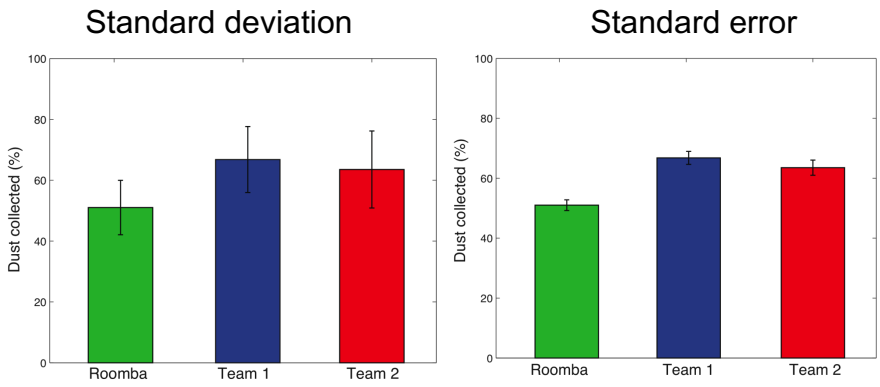


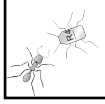
Standard deviation vs. standard error



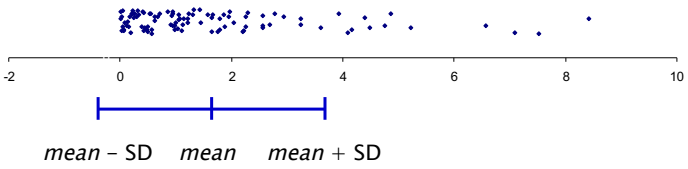


Standard deviation vs. standard error

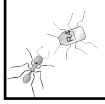




Problems with mean ± SD

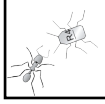


- No recognition of distribution (skew)
- *Mean* ± SD may be outside the possible range of the data!
- E.g., mean ± SD({-30, 1, 2, 3, 4}) = -4 ± 14.6

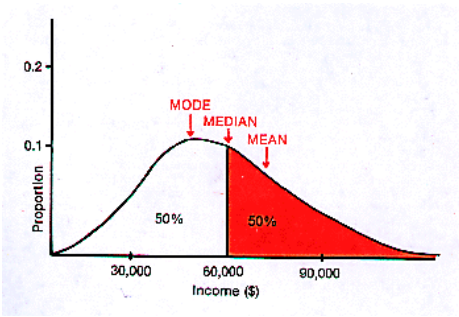


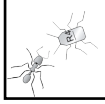
Median and quartiles

- Median: item with average rank
  Rank the items in order, and pick the middle one
  median({-30, 1, 2, 3, 4}) = 2
  median({0, 1, 2, 3, 4}) = 2
  The median is a nonparametric measure of centrality



Median and quartiles

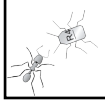




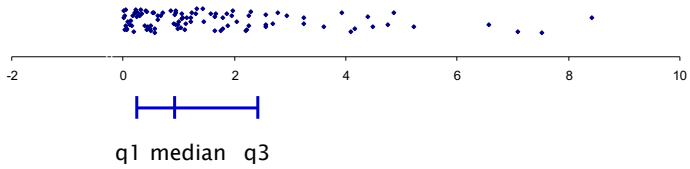
Median and quartiles

- Median: item with average rank
  Rank the items in order, and pick the middle one
  median({-30, 1, 2, 3, 4}) = 2
  median({0, 1, 2, 3, 4}) = 2
  The median is a nonparametric measure of centrality

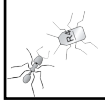
- Quartiles (Q1 = 25th; Q2 = 75th percentiles) are a nonparametric measure of spread
  The difference between quartiles can indicate skew
  Median = 50th percentile
  25th percentile({-30, 1, 2, 3, 4}) = 1
  75th percentile({-30, 1, 2, 3, 4}) = 3



Median and quartiles

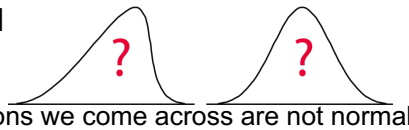


- half of the data is < median (by definition)
- quartiles here also indicate skew and quartiles are inside the data range

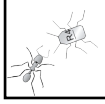


Normal distributions aren't normal

- It is (often incorrectly) assumed that the underlying distribution is normal



  - Most of the distributions we come across are not normal
  - Most well-known statistical tests require normal distribution for them to work and can give **very** wrong answers if not
  - Non-parametric must be used for non-normal distributions
- It often makes more sense to use median/quartiles than mean ± SD
- Non-normal distributions are affected by "outliers"
  - Long "tails" dominate the mean, and make the standard deviation misleading



Outline

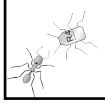
Why we need statistics

Collecting data

Describing data

Comparing data (statistical tests)

Plotting data

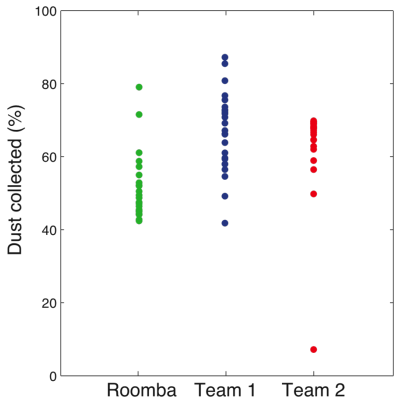


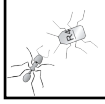
Comparing data

Do two sets of data differ?

For example:

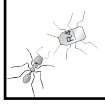
- Does my algorithm perform better than other algorithms?
- Does changing parameter *x* influence performance?





The null hypothesis

- A way of formulating the question
- Usually, H_0 is a hypothesis that the change has no effect
  E.g., the new algorithm is no different from the old
  E.g., the new parameter values give the same results as before
- The question is: can we reject the null hypothesis?
  Exhibiting statistically significant evidence against H_0 allows to reject it at a given confidence level
- Note: the point is not to prove that the null hypothesis true!
  If no effect is found, maybe because I haven't looked hard enough



The null hypothesis

- H_0 is assumed true unless data indicate otherwise

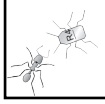
  We are measuring probability $p(\text{obs}|H_0)$, i.e., p("I see this sample" | "H_0 holds"), remember: H_0 = "two data sets were drawn from the same population"

- A small value of p means reject the null hypothesis

  If $p < \alpha$ (e.g., 5%), we can reject H_0 at a 1 - α (here 95%) confidence level

- Scientific conventions: $p < 0.05$, $p < 0.01$ or $p < 0.001$

  If $p > 0.05$: not significant



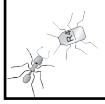
The null hypothesis

- Run baseline experiment and experiment where the baseline condition has changed

- By sampling from these two sets of data, we expect to find either:

  Change has no statistically significant effect: H_0 not rejected

  Change (in either direction) has statistically significant effect: H_0 rejected!



Important to remember!

- Do not compare apples and oranges!

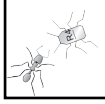
  E.g., highly optimized algorithm vs. prototype, comparing performance on different problems

- Collect enough data!

  Maybe you got lucky the one time your algorithm worked

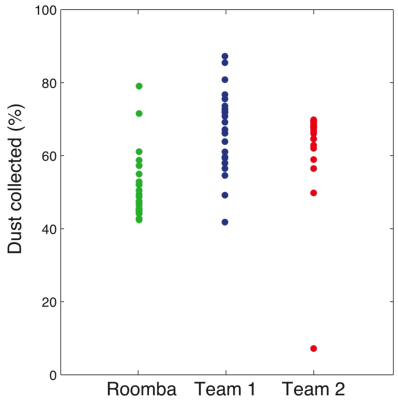
- Not rejecting H_0 does not imply proof of H_0!

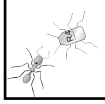
  We are measuring $p(\text{obs}|H_0)$, *not* $p(H_0|\text{obs})$



Using statistical tests

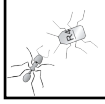
- Most statistical tests (e.g., various t-tests) require that the underlying distribution be normal!
- Use nonparametric tests. They make fewest assumptions.





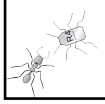
Are medians different?

- Use the nonparametric rank-sum test (a.k.a. Wilcoxon test, a.k.a. Mann-Whitney test) to test whether the medians of two samples are different
  - H_0: samples X and Y have the same medians
- In Matlab: `p = ranksum(X,Y,'alpha',0.05);`
- If $p < 0.05$, can reject H_0 at the 95% confidence level



Degrees of freedom (df)

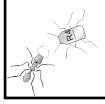
- The degree of freedom of a statistical test shows the number of values in the calculation of a statistic that are free to vary.
- *df* = number of samples – number of estimated parameters needed
- E.g., to compare medians, number of samples – 1 (only the median of one sample is needed)
- This conveys the accuracy of the statistical test.



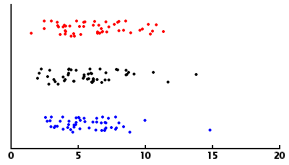
In a nutshell

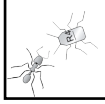
What to write in a paper:

"The performance of the Roomba is significantly worse if the wall sensor is added (without sensor, mean ± SD: 56 ± 12, with sensor: 48 ± 6, rank-sum test, $p < 0.01$)"



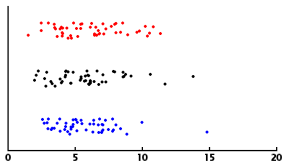
Are medians different?

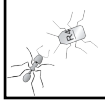




Are medians different?

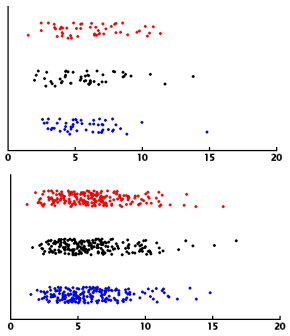
N = 50	rank sum *p*
blue/black	0.997
black/red	0.422
blue/red	0.368

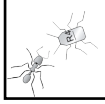




Are medians different?

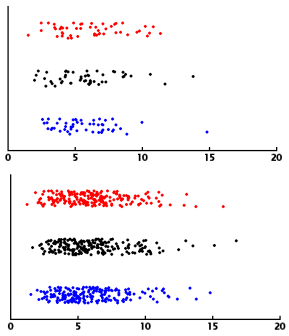
N = 50	rank sum *p*
blue/black	0.997
black/red	0.422
blue/red	0.368

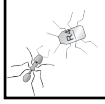




Are medians different?

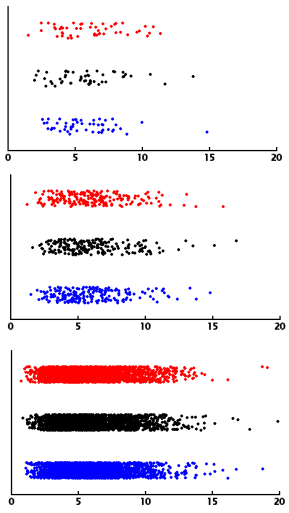
N = 50	rank sum *p*
blue/black	0.997
black/red	0.422
blue/red	0.368
N = 200	
blue/black	0.680
black/red	0.694
blue/red	0.413

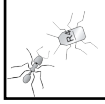




Are medians different?

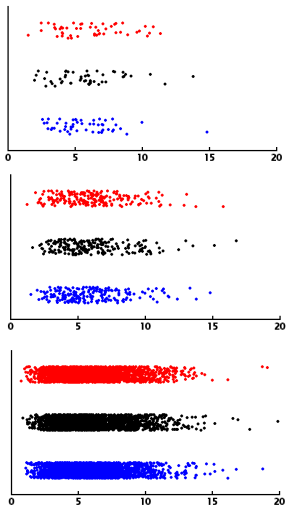
N = 50	rank sum *p*
blue/black	0.997
black/red	0.422
blue/red	0.368
N = 200	
blue/black	0.680
black/red	0.694
blue/red	0.413



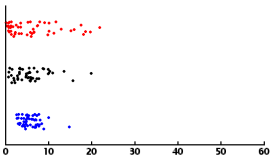


Are medians different?

N = 50	rank sum *p*
blue/black	0.997
black/red	0.422
blue/red	0.368
N = 200	
blue/black	0.680
black/red	0.694
blue/red	0.413
N = 2000	
blue/black	0.800
black/red	0.947
blue/red	0.826

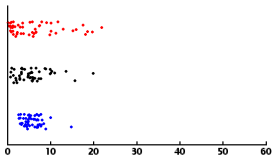


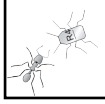
Are medians different?



Are medians different?

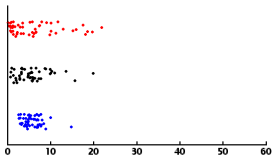
N = 50	rank sum *p*
blue/black	0.395
black/red	0.697
blue/red	0.316

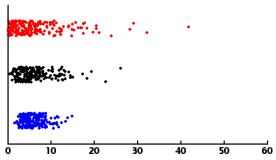


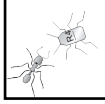


Are medians different?

N = 50	rank sum *p*
blue/black	0.395
black/red	0.697
blue/red	0.316

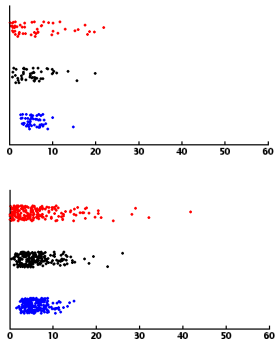


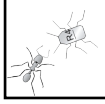




Are medians different?

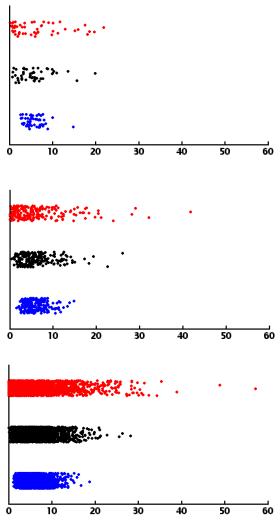
N = 50	rank sum *p*
blue/black	0.395
black/red	0.697
blue/red	0.316
N = 200	
blue/black	0.191
black/red	0.030
blue/red	0.002

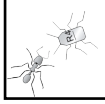




Are medians different?

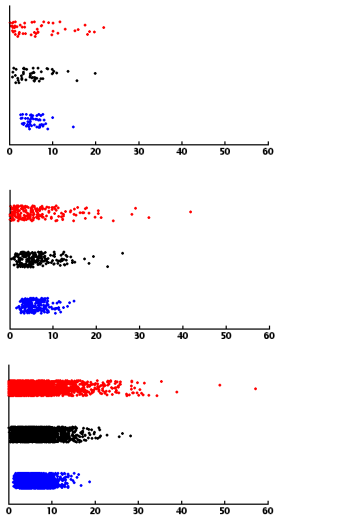
N = 50	rank sum *p*
blue/black	0.395
black/red	0.697
blue/red	0.316
N = 200	
blue/black	0.191
black/red	0.030
blue/red	0.002





Are medians different?

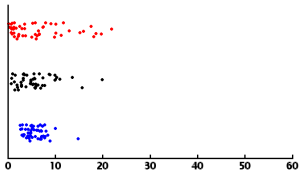
N = 50	rank sum *p*
blue/black	0.395
black/red	0.697
blue/red	0.316
N = 200	
blue/black	0.191
black/red	0.030
blue/red	0.002
N = 2000	
blue/black	0.000
black/red	0.000
blue/red	0.000

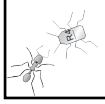


Kolmogorov-Smirnoff test

- If we can't demonstrate that the medians are different, use the nonparametric Kolmogorov-Smirnoff test (a.k.a. KS test) to test whether two samples are drawn from populations with the same distribution
  - H_0 : samples X and Y have the same distribution
- In Matlab: `[h,p] = kstest2(X,Y,'alpha',0.05);`
- If $p < 0.05$, can reject H_0 at the 95% confidence level
  - because the distributions are different in some way (maybe medians, maybe quartiles, maybe some other way, …)

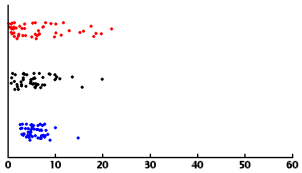
Kolmogorov-Smirnoff test

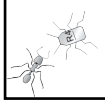




Kolmogorov-Smirnoff test

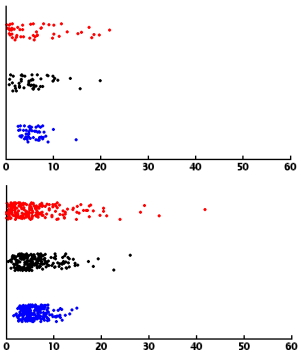
N = 50	rank sum	KS ***p***
blue/blk	0.395	0.241
blk/red	0.697	0.155
blue/red	0.316	**0.001**

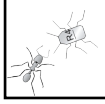




Kolmogorov-Smirnoff test

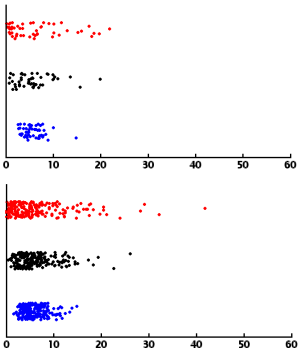
N = 50	rank sum	KS *p*
blue/blk	0.395	0.241
blk/red	0.697	0.155
blue/red	0.316	0.001

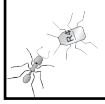




Kolmogorov-Smirnoff test

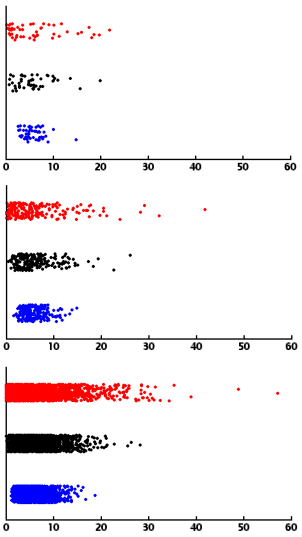
N = 50	rank sum	KS *p*
blue/blk	0.395	0.241
blk/red	0.697	0.155
blue/red	0.316	0.001
N = 200		
blue/blk	0.191	0.008
blk/red	0.030	0.001
blue/red	0.002	0.000





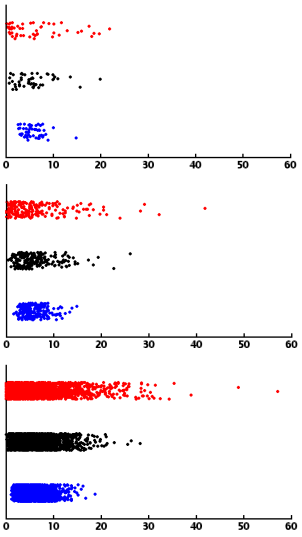
Kolmogorov-Smirnoff test

N = 50	rank sum	KS *p*
blue/blk	0.395	0.241
blk/red	0.697	0.155
blue/red	0.316	0.001
N = 200		
blue/blk	0.191	0.008
blk/red	0.030	0.001
blue/red	0.002	0.000



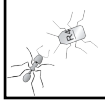
Kolmogorov-Smirnoff test

N = 50	rank sum	KS *p*
blue/blk	0.395	0.241
blk/red	0.697	0.155
blue/red	0.316	0.001
N = 200		
blue/blk	0.191	0.008
blk/red	0.030	0.001
blue/red	0.002	0.000
N= 2000		
blue/blk	0.000	0.000
blk/red	0.000	0.000
blue/red	0.000	0.000

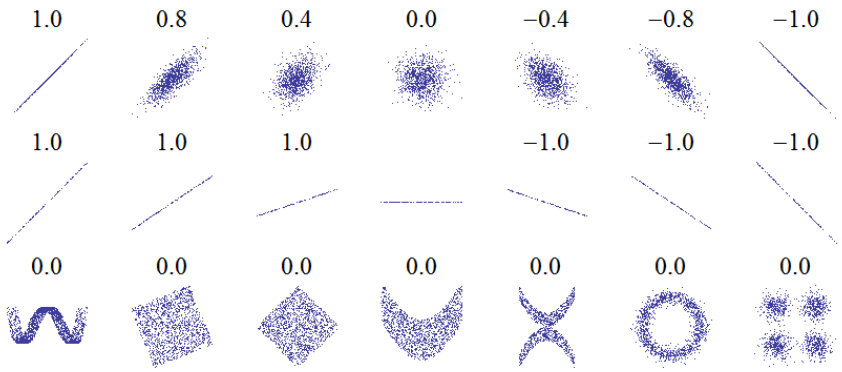


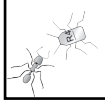
Other useful/common statistical tests

- Kruskal-Wallis test (Matlab: kruskalwallis)
  Ranksum for multiple datasets (null hypothesis: all are the same)
- Lilliefors test (Matlab: lillietest)
  Test whether a distribution is normal
- t-test (Matlab: ttest2)
  Parametric version of rank-sum (use if distribution is significantly normal)
- ANOVA (Matlab: anova1)
  Parametric version of Kruskal-Wallis (use if distributions are normal)
- Sign test (Matlab: signtest)
  Rank-sum comparing distribution x to a median of 0
- F-test (Matlab: vartest2)
  Test whether variances of two distributions are different
- Correlation tests (E.g., Matlab: corrcoef)

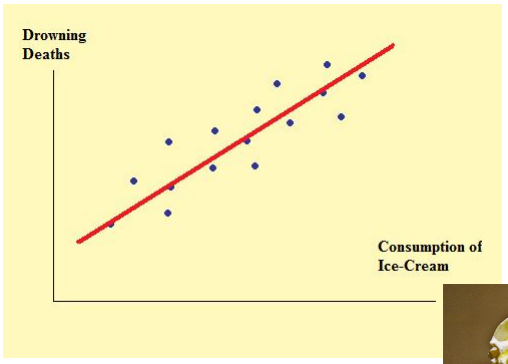


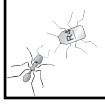
Correlations coefficients





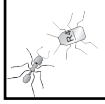
Correlation = Causation?



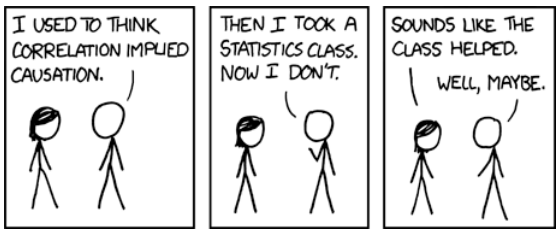


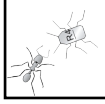
Correlation ≠ Causation!!!!!!!!!!!!!!!!

- Third factor causation:
  - As ice cream sales increase, the rate of drowning deaths increases sharply
    - ➢ Ice cream causes drowning?
- Reverse causation:
  - The more firemen fighting a fire, the bigger the fire
    - ➢ Fireman cause fire?
- Coincidence:
  - With a decrease in the number of pirates, there has been an increase in global warming
    - ➢ A lack of pirates causes global warming?



Correlation ≠ Causation!!!!!!!!!!!!!!!!!





Outline

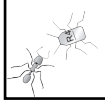
Why we need statistics

Collecting data

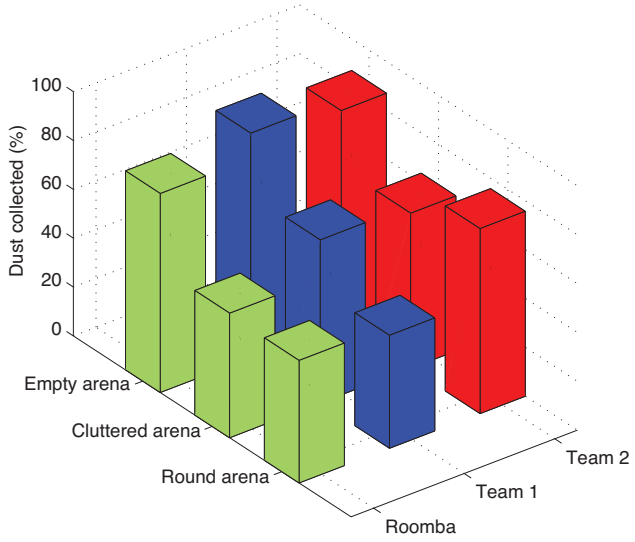
Describing data

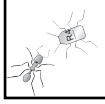
Comparing data (statistical tests)

Plotting data



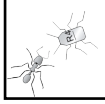
What is wrong with this plot?



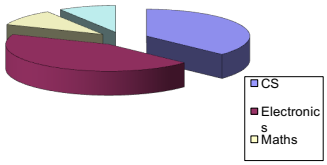


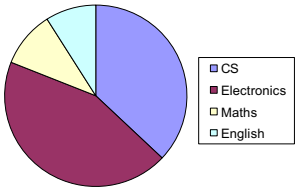
Common plotting mistakes

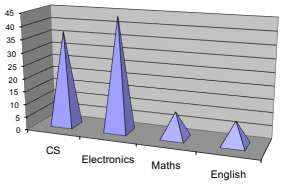
- Cluttered plots that obscure data (“chart junk”)
- Lack of information (SD, data distribution, etc.)
- More information than necessary
- Covering up strange data
- Using colours that are difficult to distinguish
- Using small font sizes
- Using plots that do not support an argument
- Forgetting to add labels to axes
- Using different scales when comparing data

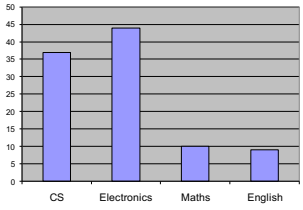


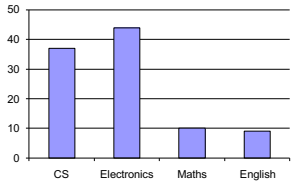
Common plotting mistakes

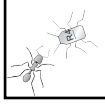




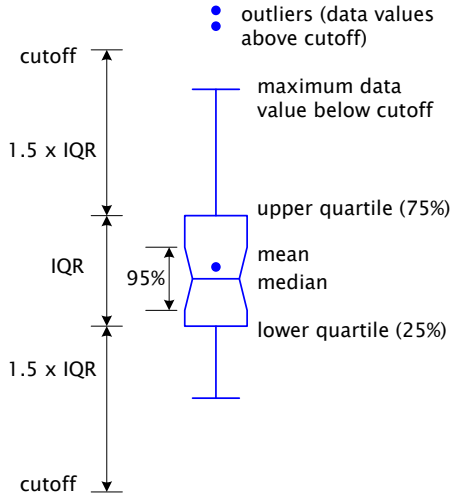


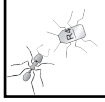




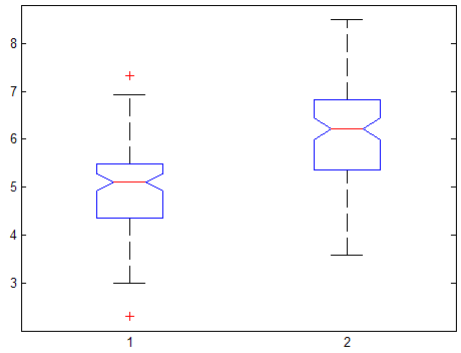


Boxplots

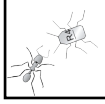




Boxplots



- If the notches in the box plot do not overlap, you can reject with 95% confidence that the true medians do not differ.

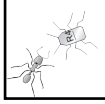


Finding anomalies in your data

If you have some "anomalous" data

Don't just discard it as an "outlier" – understand it!

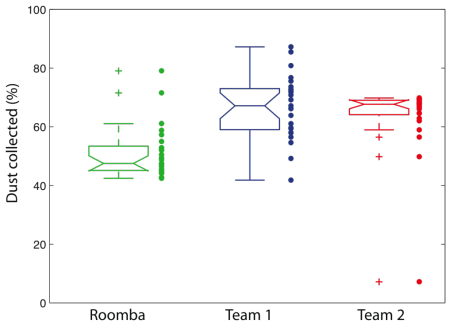
- Is it just a statistical fluctuation?
- Is it an error in the experimental design or implementation?
  If so, fix the problem, and rerun all the experiments
- Is it an interesting unexpected effect?
  Investigate it further! It might be the basis of a new discovery

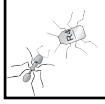


Boxplots - example

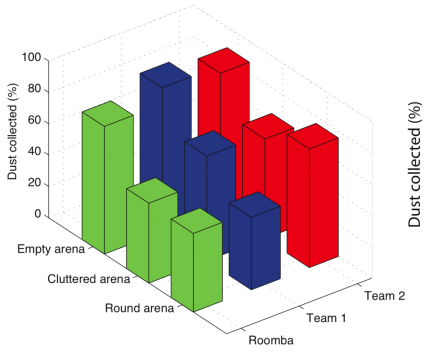
In Matlab: `p = boxplot([X,Y,Z],'Notches','on');`

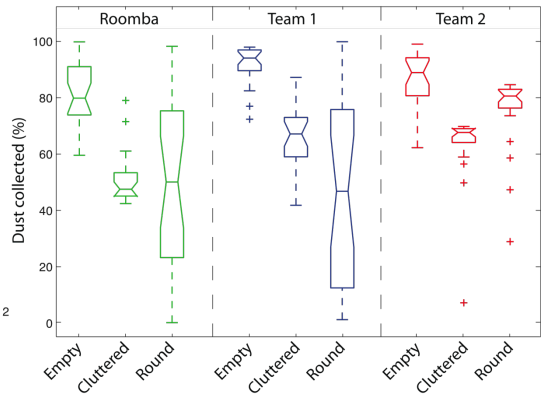
N = 25	rank-sum p	KS p
gr/blue	< 0.001	< 0.001
gr/red	< 0.001	< 0.001
blue/red	0.42	0.01



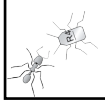


Boxplots - example

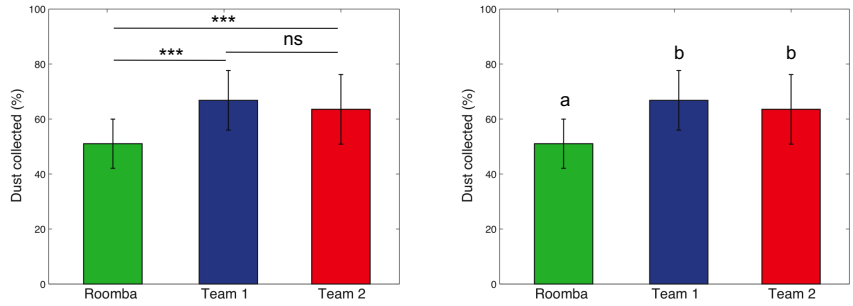


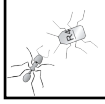


- Reduce clutter!
- Expose the important features of the data



Other plotting options





Summary

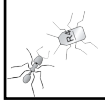
Why we need statistics

Collecting data

Describing data

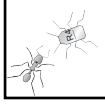
Comparing data (statistical tests)

Plotting data



Statistical analysis checklist

- Decide on data to measure, then run appropriate experiments to collect it
- Collect enough data
- Look at raw data
- Plot in simple, clear graphs that give enough information about the data (use boxplots)
- Run statistical (non-parametric!) tests to compare the data and "tell a story"
- Do not present more data than you need to make your point
- Be honest and do not hide anomalous data



If you want to learn more