Statistics with Confidence

Tutorial at ICARIS 2009

Susan Stepney
Department of Computer Science, and
York Centre for Complex Systems Analysis

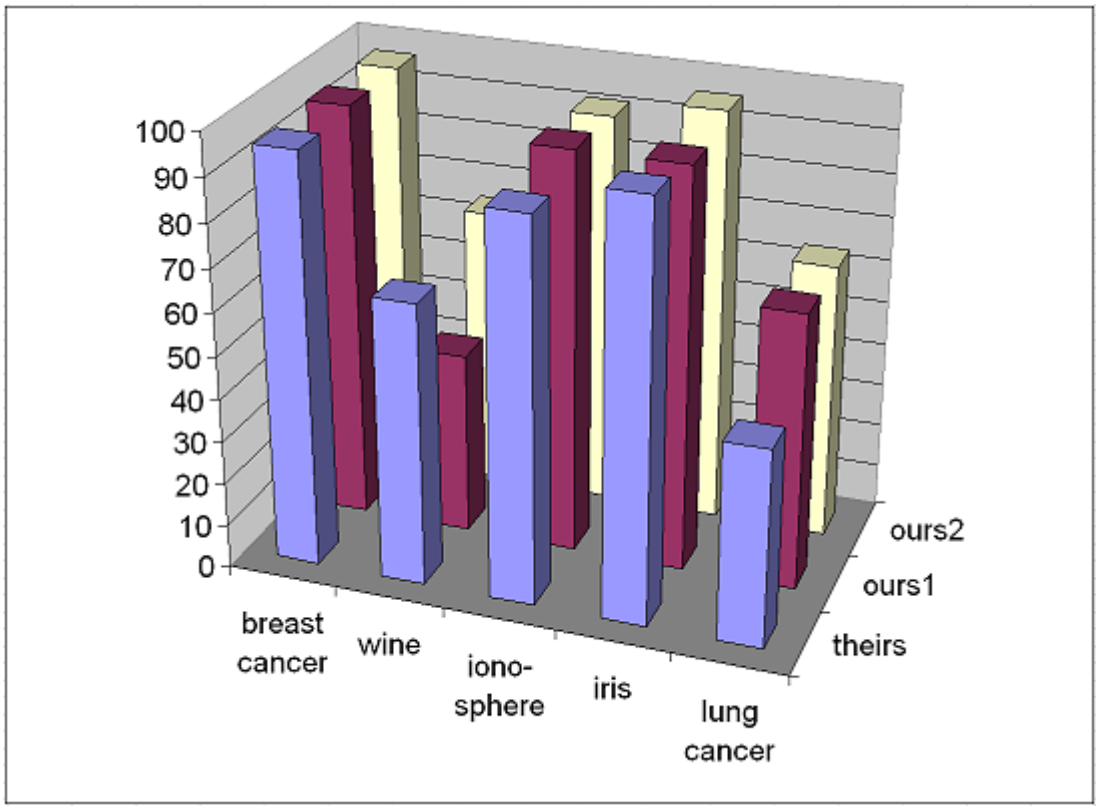
THE UNIVERSITY of York

how good is our algorithm?

data set	theirs (mean)	ours1 (mean)	ours2 (mean)
breast cancer	95.51	96.41	96.67
wine	65.49	42.37	65.23
ionosphere	88.72	93.18	91.67
iris	96.01	93.11	96.02
lung cancer	45.98	64.29	64.28

- we've all seen tables like this...
  - or worse, with *no* comparison with other results
- ***nothing*** can be concluded about which is "better"
  - is it a ***statistically significant*** difference?
  - is it an ***important*** (scientifically significant) difference?

how good is it? (in pictures)



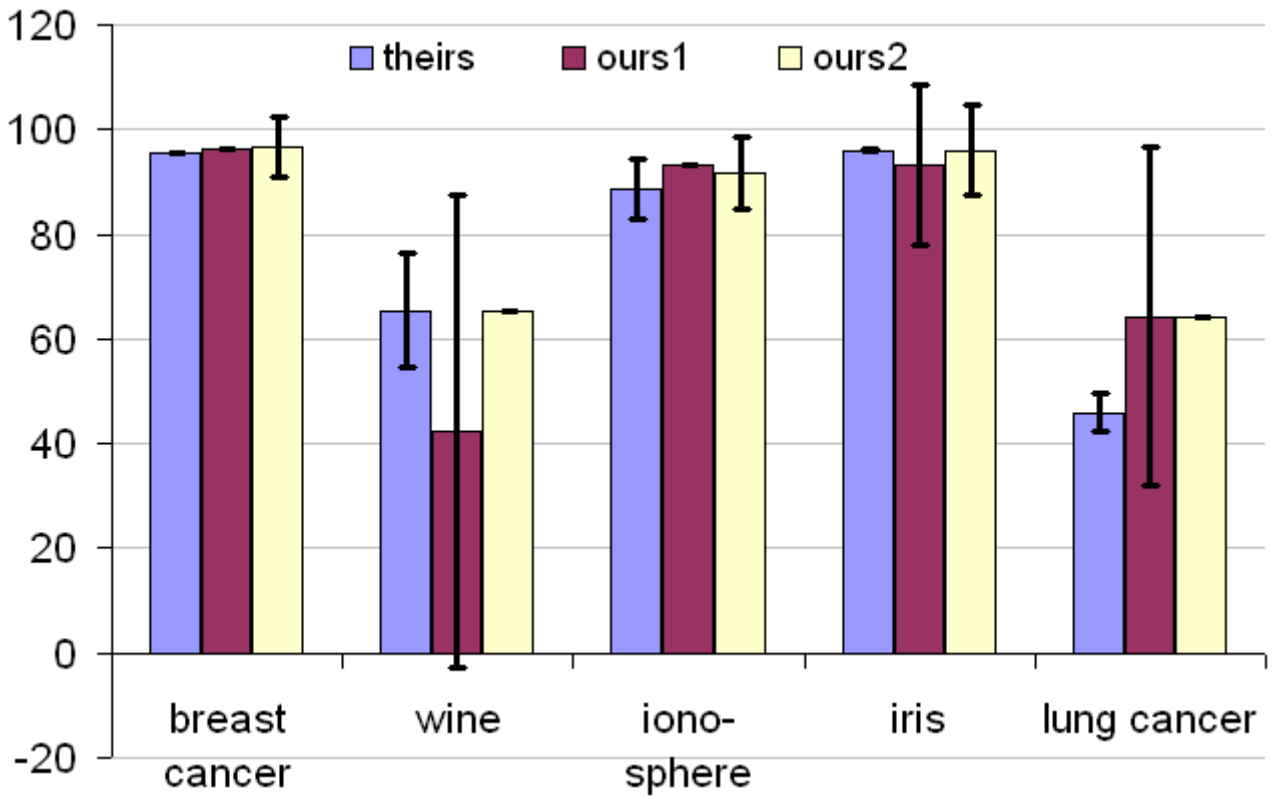
- we've all seen charts like this ...
  - or *worse* (using pyramids, ...)
- it's just *"chartjunk"*
  - it **obscures** the data (can't compare column heights)
    - which here isn't even very useful data to begin with!

how good is it, really?

data set	theirs (m ± 1sd)	ours1	ours2
breast cancer	95.51 ± 0.01	96.41 ± 0.02	96.67 ± 5.67
wine	65.49 ± 10.91	42.37 ± 45.01	65.23 ± 0.04
ionosphere	88.72 ± 5.67	93.18 ± 0.03	91.67 ± 6.98
iris	96.01 ± 0.11	93.11 ± 15.32	96.02 ± 8.54
lung cancer	45.98 ± 3.45	64.29 ± 32.32	64.28 ± 0.05

- this is more informative
- a (minimal!) basis for meaningful comparison
  - but still need to test for *statistical significance*
  - and should also test for *effect size* ("scientific significance")

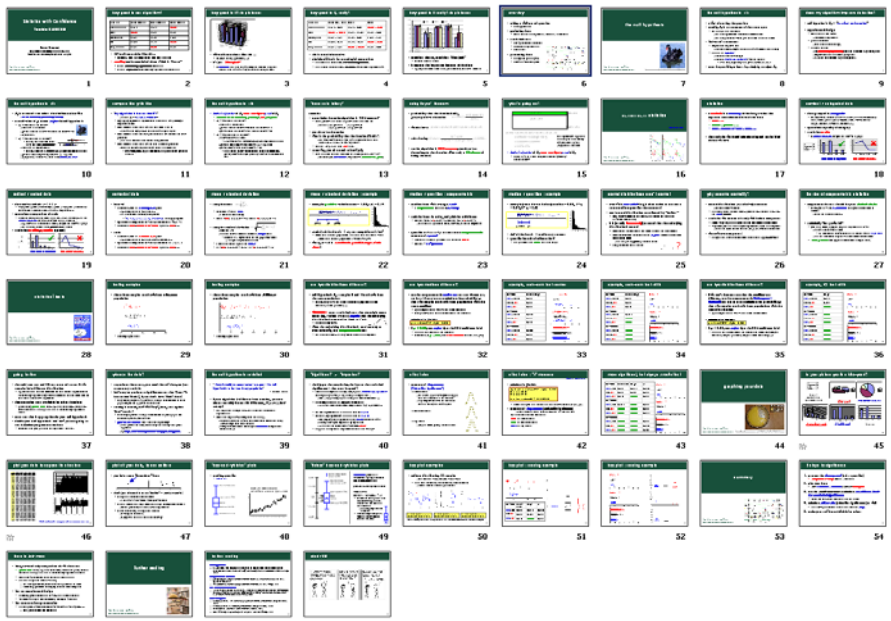
how good is it really? (in pictures)



- much less clutter, much less “chartjunk”
  - it could still be better
- it exposes the important features of the data
  - eg, why do some error bars go below 0%, or above 100% ??

overview

- asking a well-posed question
  - null hypothesis
- gathering data
  - kinds of data : categorical / ordinal / numerical
- statistical tests
  - non-parametric statistics
  - statistical significance
  - effect size
- presenting data
  - chartjunk: just say no!
  - box-and-whisker plots



the null hypothesis

http://www.wwt.org.uk/research/monitoring/images/Black_Swan.jpg

the null hypothesis (1)

- a way of casting the question
- usually, H_0 is a statement of the status quo
  - the change has no effect
    - the new algorithm is no different from the old
    - the new parameter values give the same results as before
- "universal" statements
  - impossible to prove true
    - no effect found, maybe because I haven't looked hard enough
  - but can be ***rejected***
    - exhibit a counter-example
    - exhibit *statistically significant* evidence against
      - even then, reject only at a given *confidence level*
- must be possible, at least *in principle*, to reject H_0

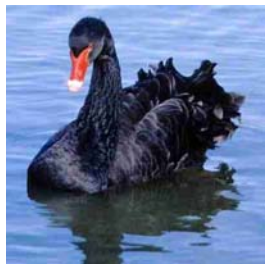
does my algorithm improve detection?

- null hypothesis H_0 = "*no effect* on detection"

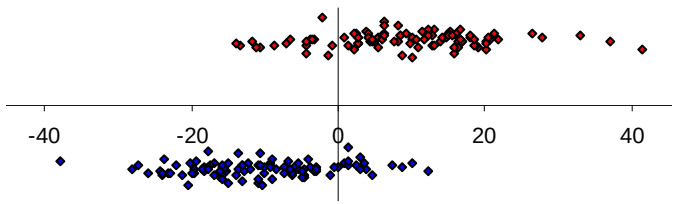
- experimental design
  - trial new and old styles
    - controlled experiment
  - gather statistics
    - this needs careful design!
  - analyse results
    - no *statistically significantly* effect on detection : H_0 not rejected
    - improvement in detection : H_0 rejected
    - *decrease* in detection : H_0 also rejected!

the null hypothesis (2)

- H_0 is assumed true unless data indicate otherwise
  - we are measuring probability $p(\text{obs}|H_0)$
- a small value of p means *reject* the null hypothesis
  - H_0 = "all swans are white"
    - I observe a black swan
    - p("this swan is black" | "all swans are white") = 0
    - H_0 rejected!



  - H_0 = "no effect on detection"
    - that is, "the detection rates have the same *distribution*"



    - I observe an improvement
      - that is, I observe a *statistically significantly* different distbn
    - p("I see this sample" | "H_0 holds") $< \alpha$ (eg, 5%)
    - H_0 rejected, at the $1-\alpha$ (here 95%) *confidence level*

compare like with like

- "my algorithm is better than X's"
  - cast as H_0 = "*no different* from X's"
  - my highly optimised algorithm is better than X's prototype implementation
  - mine worked better than X's the one time I ran it
    - I didn't dare try again, in case it didn't happen again
  - mine worked better on *this* problem than X's worked on *that* problem
    - and I had to search hard to find this problem
  - mine worked better than X's on this artificial problem
    - which is highly unrepresentative of the real world use
      - over-simplified, inputs too small, unrepresentative synthetic data, …

the null hypothesis (3)

- *lack of rejection* of H_0 DOES NOT IMPLY *proof* of H_0
  - because we are measuring $p(\text{obs}|H_0)$, *not* $p(H_0|\text{obs})$
  - H_0 = "all swans are white"
    - I observe a white swan
    - p("this swan is white" | "all swans are white") = 1
    - p("all swans are white" | "this swan is white") = ???
  - H_0 = "no effect on detection"
    - I don't observe an effect
      - I *don't* observe a *statistically significantly* different distbn
    - p("I see this sample" | "H_0 holds") $> \alpha'$ (eg, 99%)
    - p("H_0 holds" | "I see this sample") = ???
      - need to use Bayes' Theorem (and further information, p("H_0 holds"), that you probably don't have) to work this out

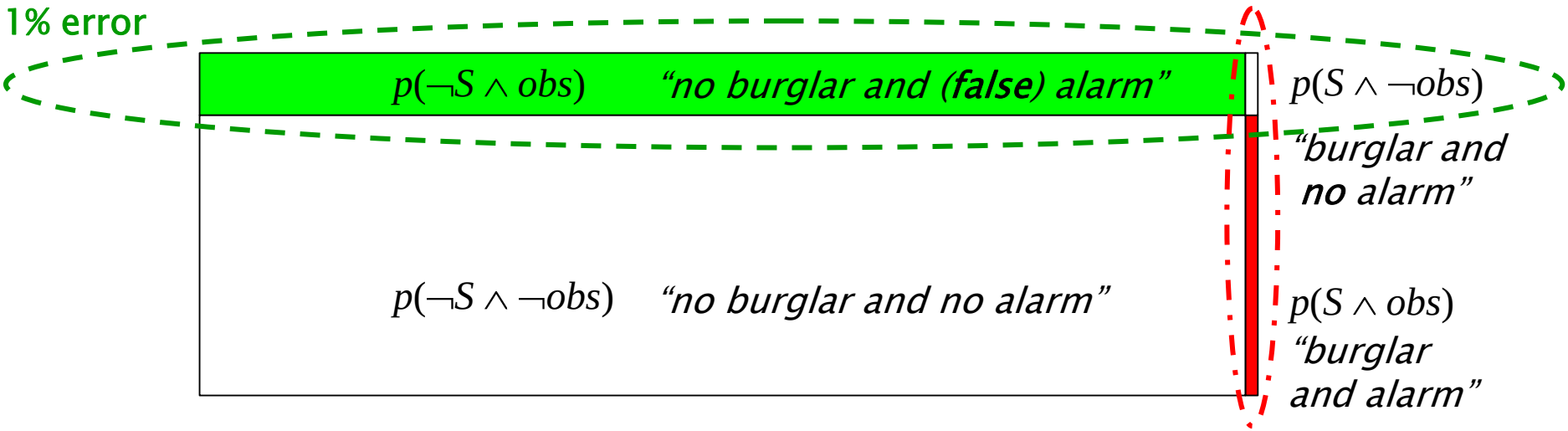
"base-rate fallacy"

- consider a "situation" S such as "you have an illness"; "there is an intruder", ... here, S is that "H_0 holds"
- a particular detection algorithm is "99% accurate"
  - false positive and false negative rates are both 1%
    - that is : $p(obs \mid S) = 0.99, \quad p(\neg obs \mid S) = 0.01$
      $p(\neg obs \mid \neg S) = 0.99, \; p(obs \mid \neg S) = 0.01$
      - where obs = result of a medical test, of the algorithm, ...
- you observe a detection
- what is the probability that the situation S holds?
  - what is $p(S \mid obs)$? it is *not* 99% !
- to answer, you also need to know $p(S)$
  - let's assume that the situation is *unlikely*, that $p(S) = 10^{-4}$
    - of course, we *don't know* p("H_0 holds") – *but that's the point*...

using Bayes' theorem

- probability that the situation holds, *given* a positive detection $p(S \mid obs) = \dfrac{p(S)\ p(obs \mid S)}{p(obs)}$
- we also have $p(obs) = p(obs \mid S)\ p(S) + p(obs \mid \neg S)\ p(\neg S)$
- substituting $p(S \mid obs) = \dfrac{10^{-4} \times 0.99}{0.99 \times 10^{-4} + 0.01 \times \left(1 - 10^{-4}\right)} = 0.00980\ldots$
- so: the algorithm is 99% accurate; you observe a detection; yet the situation ***S*** has only a 1% chance of being the case!

what's going on?



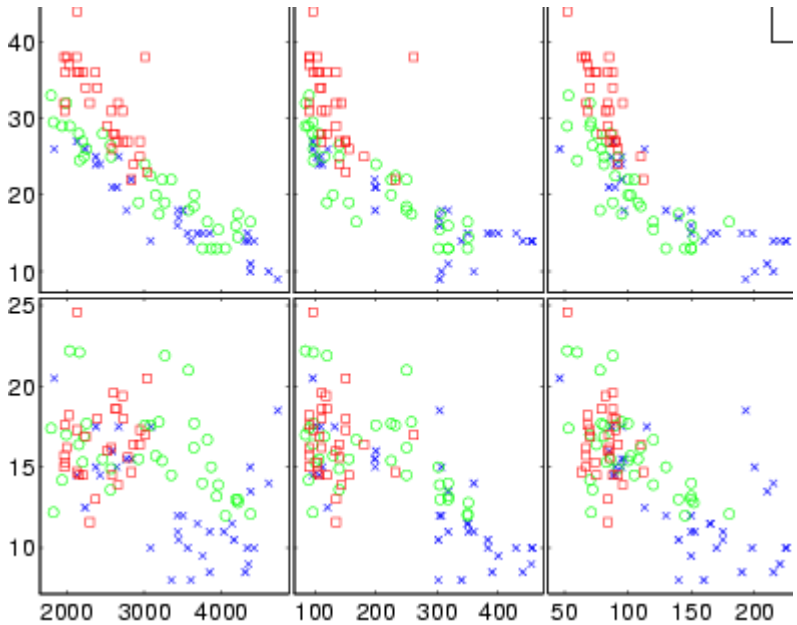
$$p(obs) = p(obs \mid S)\ p(S) + p(obs \mid \neg S)\ p(\neg S)$$
$$= 0.99 \times 10^{-4} + 0.01 \times \left(1 - 10^{-4}\right)$$
$$= 0.000099 + 0.009999$$
$$\approx 0.01\% + 1\%$$

the algorithm's apparent accuracy is here totally overwhelmed by the *rarity* of the situation

- *lack of rejection* of H_0 DOES NOT IMPLY *proof* of H_0
  - all we have is a kind of (probabilistic) "proof by contradiction"

lies, damned lies, and statistics

[Disraeli, attrib]



statistics

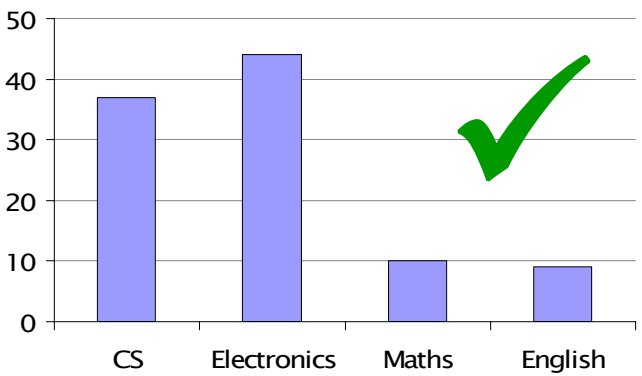
- a **statistic** is a **summary** of the data, a value that captures some characteristic of the data
  - **mode / ...**
  - **median / quartiles / ...**
  - **mean / standard deviation / skew / ...**

- the statistics we should calculate depend on the kind of data we have

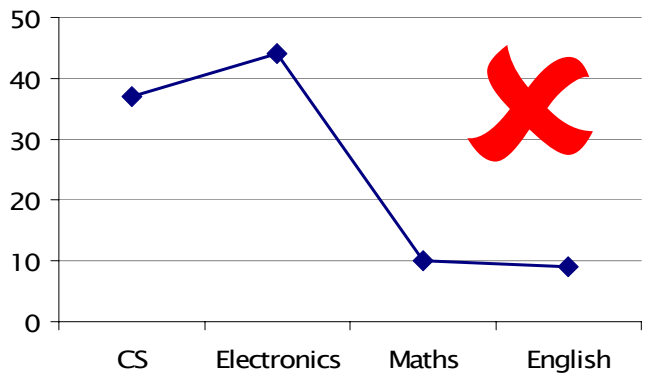
nominal / categorical data

- data grouped in ***categories***
  - on/off ; male/female ; butcher/baker/candlestickmaker ; blue/brown/green eyes ; Toyota/Ford/Volvo/other ; single/married/divorced/widowed ; …
- operations: equality of category
- statistics: **mode**
  - it ***makes no sense*** to "join the dots" in a graph

CS	37
Electronics	44
Maths	10
English	9



axis labels in any order

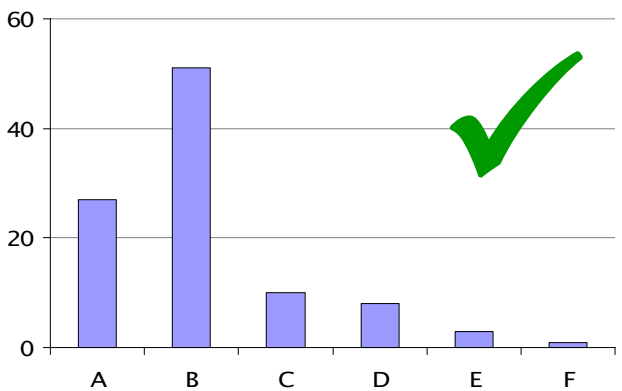


this ***makes no sense*** !!

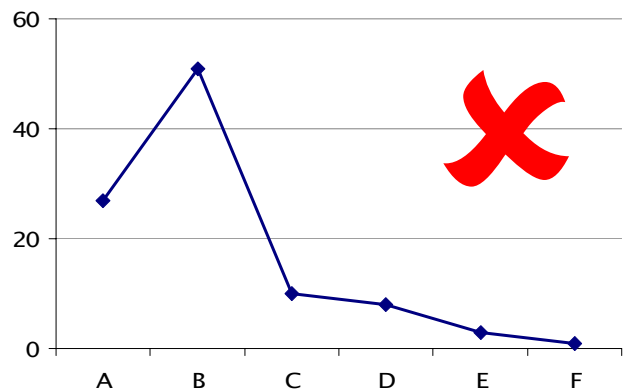
ordinal / ranked data

- data can be ranked: $a < b < c$
  - very poor < poor < satisfactory < good < very good < excellent ; Unclassified < Confidential < Secret < Top Secret ; …
- operations: comparison of ranks
  - numbers (integers) are often used to *encode* rankings, but it *makes no sense* to "calculate" with these numbers
    - eg, don't add them to find a mean, or "join the dots" in a graph
    - the encoding chosen could *just as well* be letters
- statistics: median, quartiles (spread)

A	27
B	51
C	10
D	8
E	3
F	1



axis labels in rank order



this *makes no sense* !!

numerical data

- interval
  - numerical, with an *arbitrary zero* point
    - eg, temperature in °C ; dates ; …
    - be careful with arbitrary zero points:
      - "the temperature rose by 1 °C (33 °F)" [newspaper report]
  - operations: comparison of values; arithmetic: a + b, a − b
  - statistics: **mean**, **standard deviation** (spread), **skew**, **...**

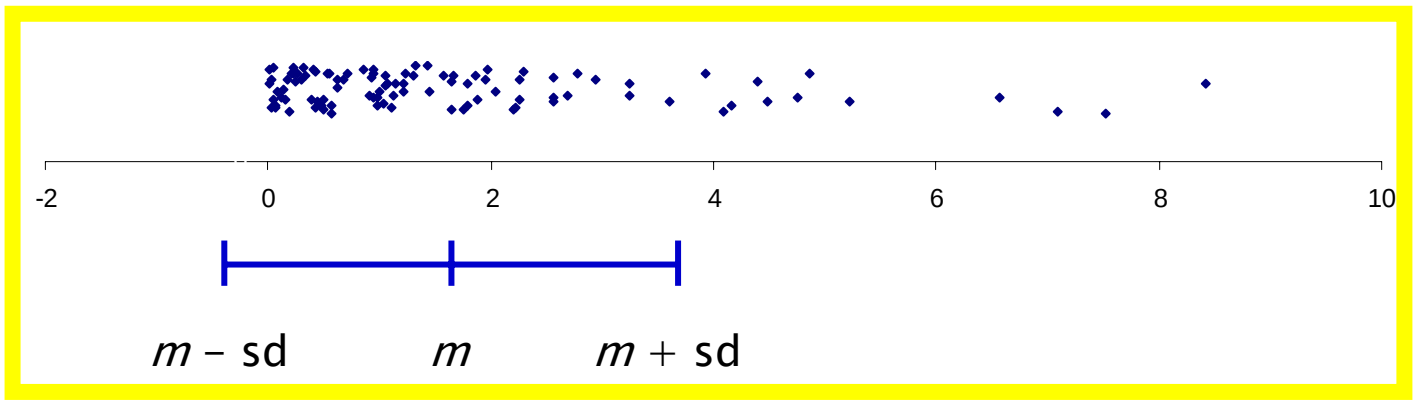
- ratio
  - numerical, with an *absolute zero* point
    - eg, temperature in K ; length ; mass ; lifetime ; …
  - operations: comparison of values; arithmetic : a ± b, a / b
  - statistics: **mean**, **standard deviation** (spread), **skew**, **...**

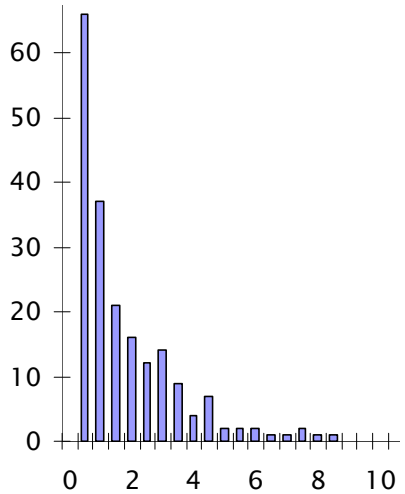
mean / standard deviation

- sample mean : $m = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - item with average ***value***
    - a measure of centrality
  - mean {–30, 1, 2, 3, 4} = –4 ; mean {0, 1, 2, 3, 4} = 2
- sample standard deviation $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(m - x_i)^2}$
    - note the $n-1$
    - because derived from the estimated ***sample*** mean
      - it would be n if you could use the *population* mean
        - » but you rarely know what that is
  - a measure of the spread of ***values***
  - sd {–30, 1, 2, 3, 4} = 14.6 ; sd {0, 1, 2, 3, 4} = 1.6

mean / standard deviation : example

- example: positive variable: mean = 1.69; s.d. = 2.14





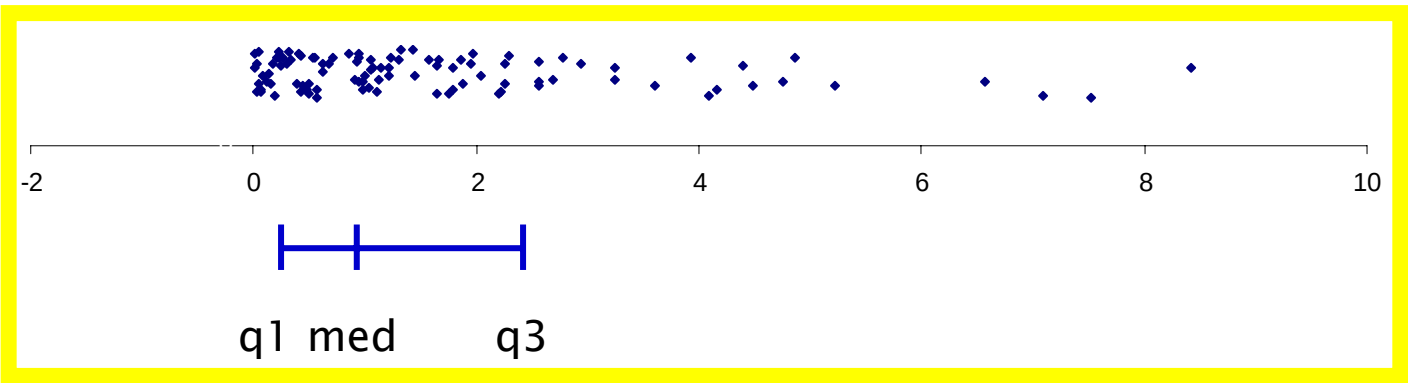
- *majority* of the data is $< m$; no recognition of skew
  - more than half the population has less than the mean value
    - more than half the population is "below average"!
- worse, m – sd is ***outside the possible range of the data*** !!

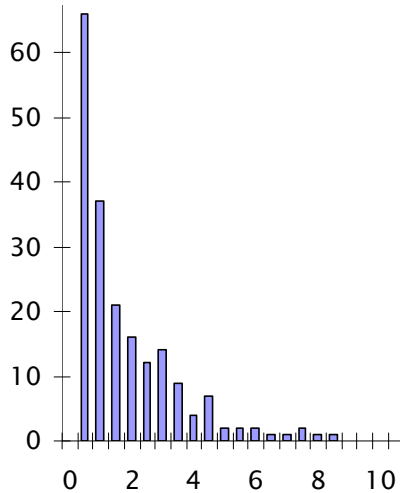
median / quartiles : nonparametric

- median : item with average *rank*
  - a nonparametric measure of *centrality*
- rank the items in order, and pick the middle one
  - median {−30, 1, 2, 3, 4} = 2 ; median {0, 1, 2, 3, 4} = 2
    - less affected by outliers (rank, not value, is what's important)
- quartiles (25^{th} and 75^{th} percentiles) are a nonparametric measure of *spread*
  - difference between quartiles can indicate skew
  - median = 50^{th} percentile

median / quartiles : example

- example (same data as before): median = 0.98, 1st q = 0.34, 3rd q = 2.32

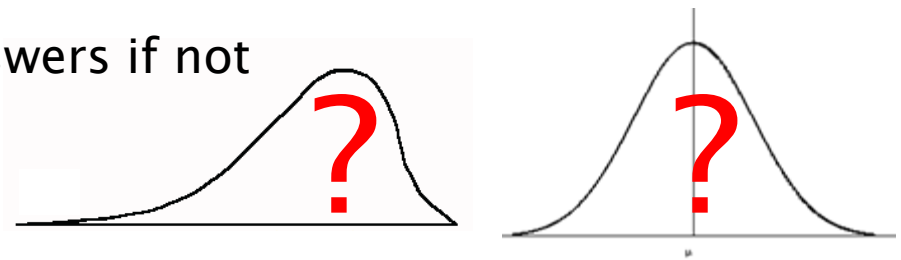




- *half* of the data is $<$ median (by definition)
- quartiles here also indicate skew
  - and quartiles are *inside* the data range

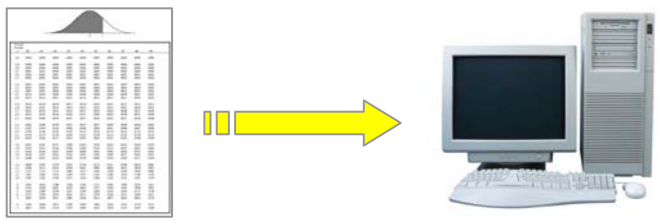
normal distributions aren't normal

- even with *numerical* data, 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
- it is (usually 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
  - nonparametric ones do not



why assume normality?

- normal distributions *are* relatively common
  - Central Limit Theorem
  - but : long tailed distributions, power laws, skewed data, etc, are ***also*** common
- statistics was an area of study well before computers
  - some very elegant analytic results for normal distributions
    - very little you can do analytically without *some* assumptions
      - there are *some* results that are *independent* of the distribution
  - calculate with tabulated results
    - table depends on distribution

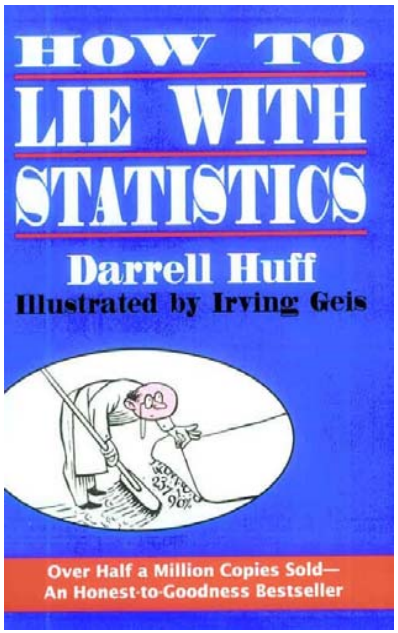


- we now have computers!
  - analyse the ***actual*** **distributions** rather than *approximations*

the rise of nonparametric statistics

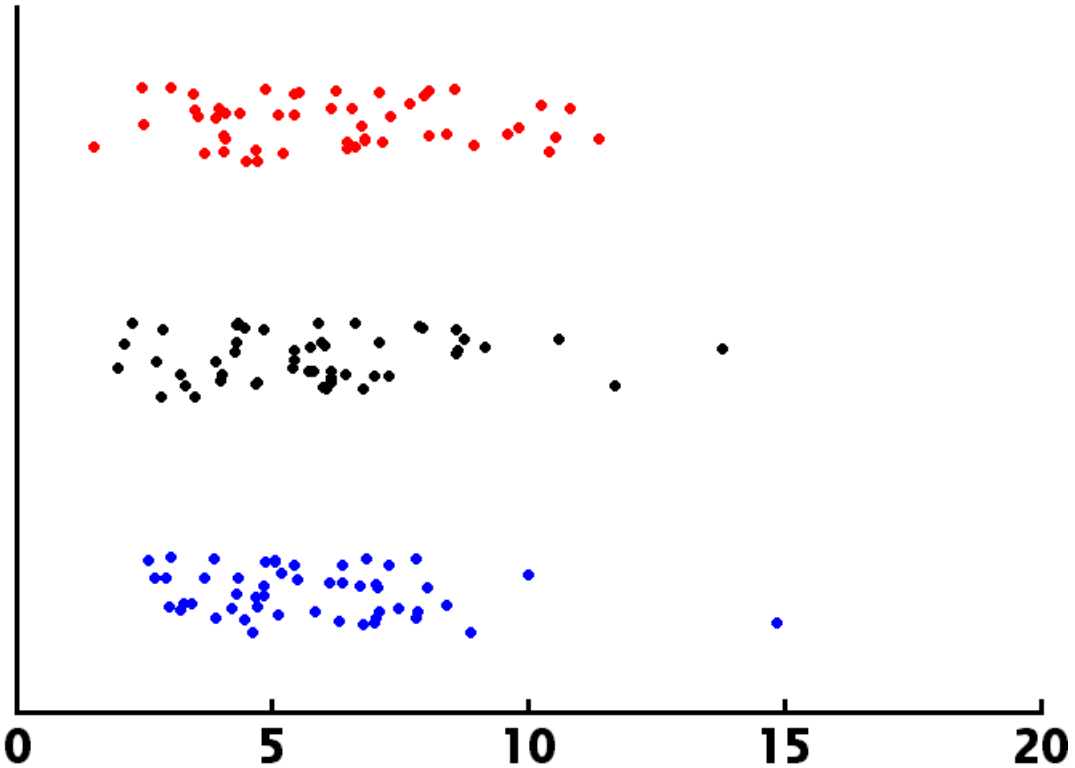
- nonparametric tests should be your *default choice*
  - if not, you have to *demonstrate* that your distribution is normal
    - there are statistical tests...
- technically "less powerful"
  - you need more samples to get a significance level
    - but it's not *that* much more
      - eg 100 for the rank-sum test *v* 95 for the *t*-test [Siegel 1988]
    - and the criticism is only relevant for normal distributions
  - not a problem with typical simulation sample sizes!

statistical tests



testing samples

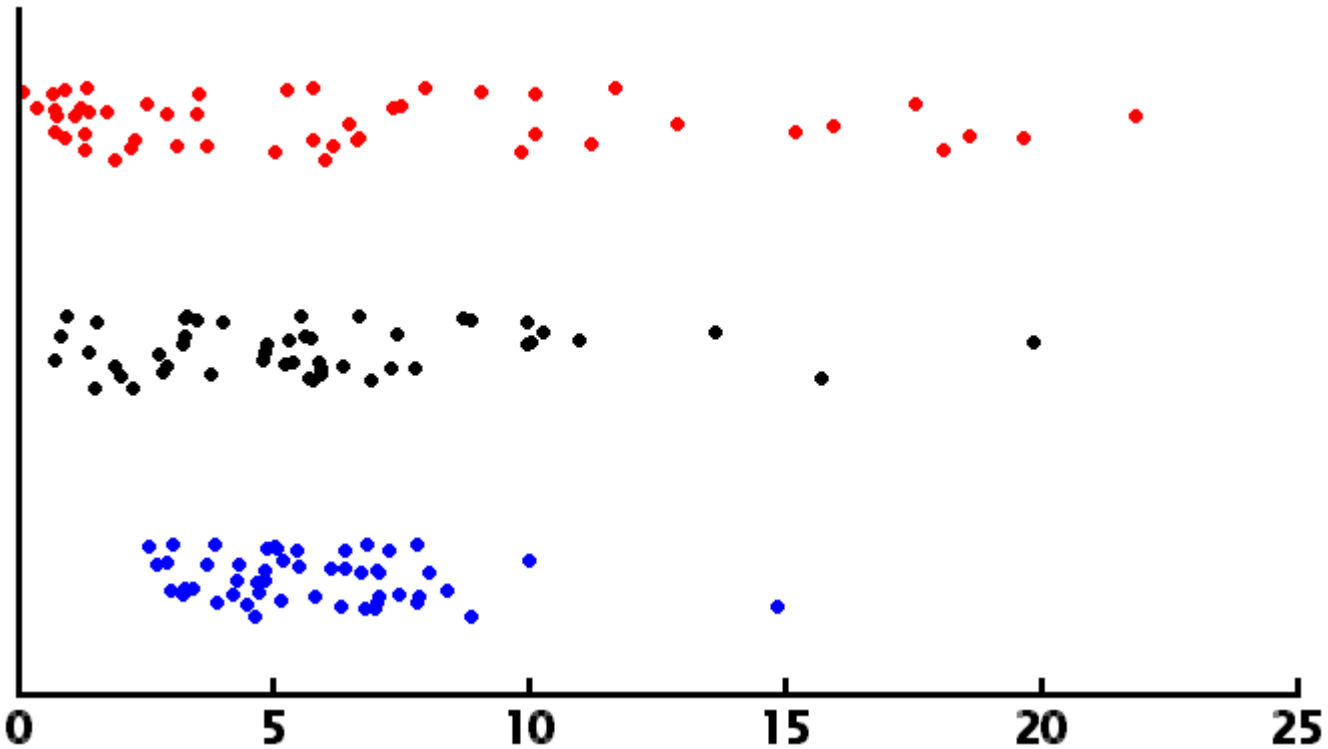
- these three samples are drawn from *the same* population



  - how might you tell?

testing samples

- these three samples are drawn from *different* populations

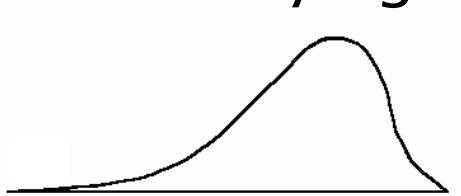


  - how might you tell?

are two distributions different?

- null Hypothesis H_0 : samples X and Y are drawn from the *same* population
  - that is, they have the same statistical properties: same medians, same quartiles, …

- **WARNING!** most statistical tests that you might come across (eg, various *t*-tests) ***require*** that the underlying distribution be normal for them to work
  - but it usually isn't !



- when the underlying distribution is non-normal, or even unknown, use **nonparametric** tests
  - the default choice, as they make fewest assumptions

are two medians different?

- use the nonparametric **rank-sum** test (aka Wilcoxon test, aka Mann-Whitney U test) to calculate the probability *p* that two samples are drawn from populations with the same *medians*
  - H_0 : samples X and Y have the same medians
  - if they don't, then their distributions are different

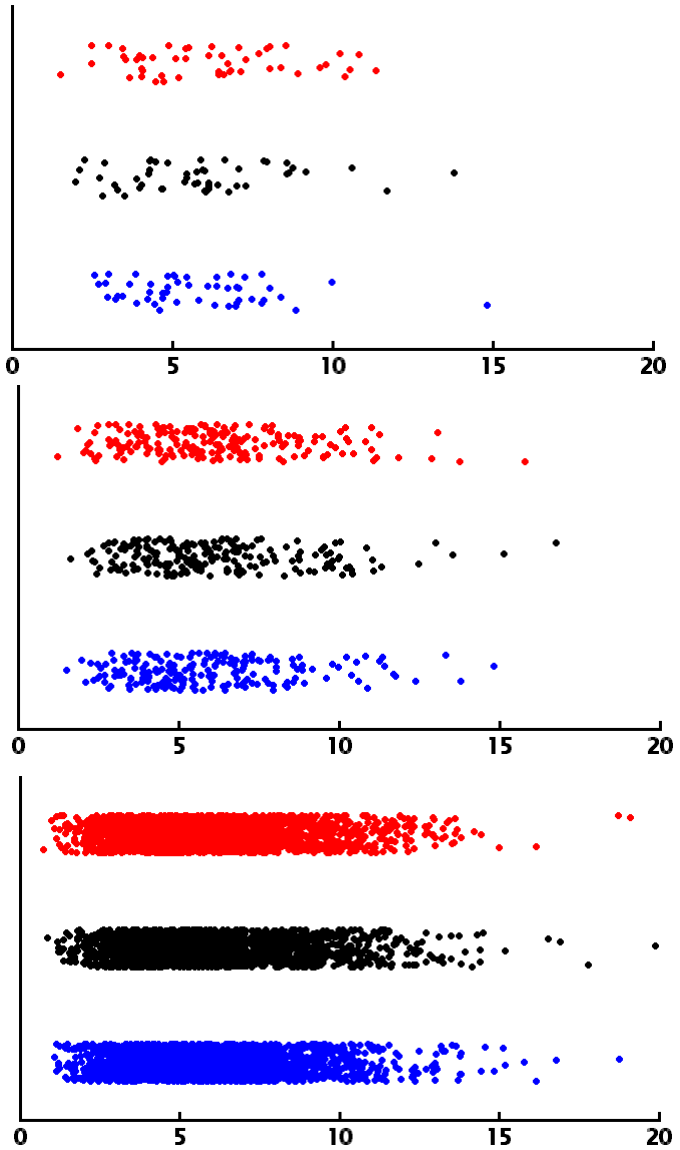
- calculate in Matlab:

```
p = ranksum(X,Y,'alpha',0.05);
```

- if $p < 0.05$, can *reject* H_0 at the 95% confidence level
  - because the medians are different
  - remember, if $p > 0.05$, this **does not mean** that we *accept* H_0

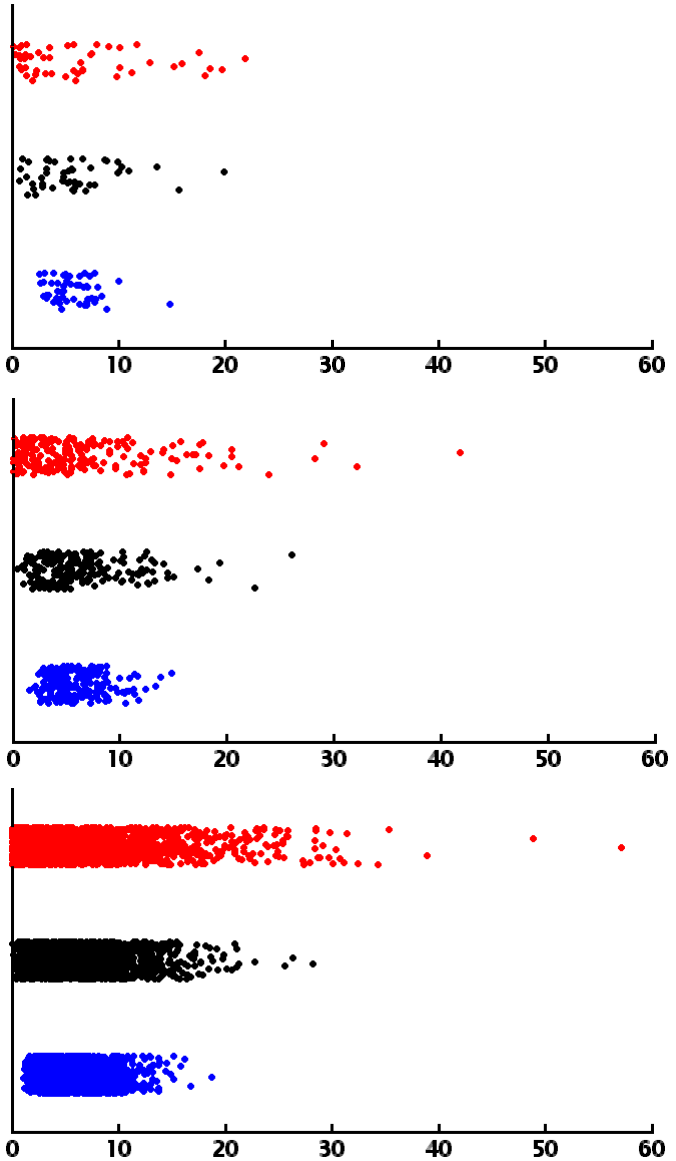
example, rank-sum test (same)

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



example, rank-sum test (diff)

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



are two distributions different?

- if we can't demonstrate that the *medians* are different, use the nonparametric **Kolmogorov–Smirnoff** test (aka KS test) to calculate the probability p that two samples are drawn from populations with the same distributions
  - H_0 : samples X and Y have the same distributions

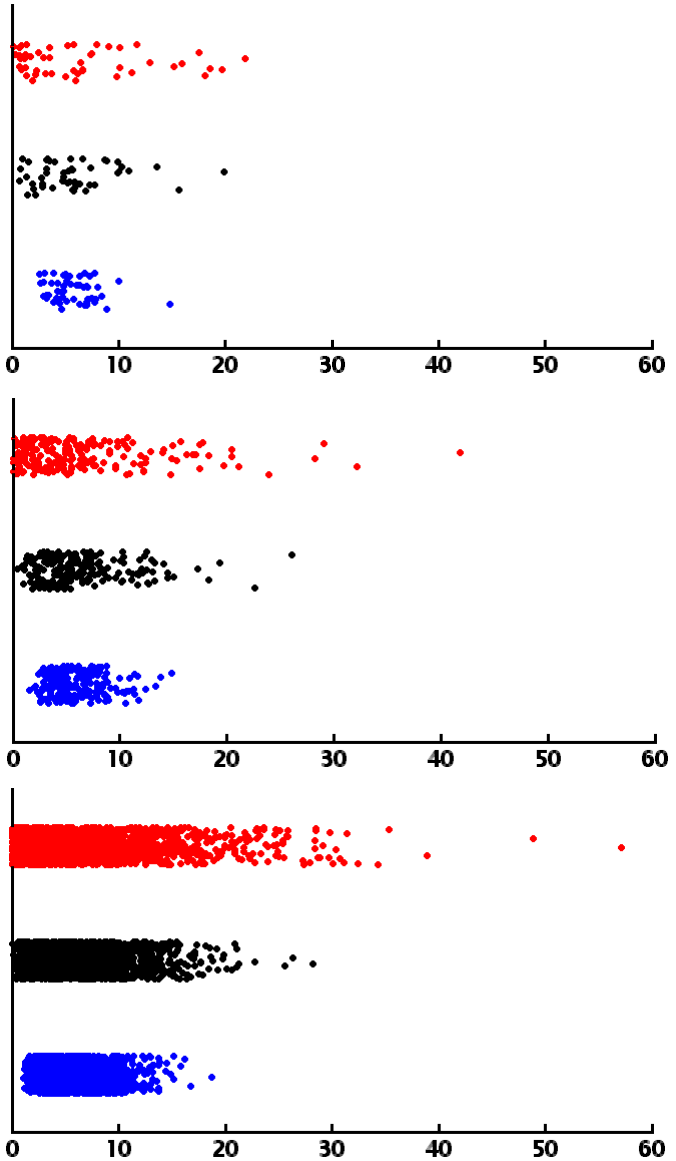
- calculate in Matlab:

```
p = ktest2(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, …)

example, KS test (diff)

N = 50	KS p
blue/black	0.241
black/red	0.155
blue/red	0.001
N = 200	
blue/black	0.008
black/red	0.001
blue/red	0.000
N = 2000	
blue/black	0.000
black/red	0.000
blue/red	0.000



going further

- the rank-sum test and KS test are used to test if two samples have different distributions
  - eg, whether the results from lots of runs of one algorithm are statistically significantly different from the results from lots of runs of an alternative algorithm
- there are other tests available for other situations
  - particularly paired data tests, eg, sets of (before, after) data, to test whether a change has had a statistically significant effect
- use a test that is appropriate for your null hypothesis
- decide your null hypothesis and how you are going to test it *before* you generate the data
  - to make sure you generate the right kind of data!

whence the data?

- to perform these tests, you need the raw data, not just a summary statistic
- how can you perform a significance test that "yours" is better than "theirs", if you don't have "their" data?
  - they will (surely?) have provided enough information in their paper for you to *replicate* their results
- turning it around, how will "they", later, test against "your" results?
  - (naturally!) provide enough information in your paper for *them* to replicate *your* results
  - provide the raw data (on a website repository)
    - even provide the code, so that they can exactly rerun your approach
    - eg: http://www-users.cs.york.ac.uk/~drw/papers/eurogp2009/

the null hypothesis revisited

- "there is really no good reason to expect the null hypothesis to be true in any population"

  – [Bakan 1967]

- if your algorithm *is* different from another, you can almost certainly detect this difference, *if you try hard enough*
  - the bigger the sample size, the better the statistical significance
  - mice are expensive; computers are cheap!
    - statistical tests are devised to demonstrate differences using *small* sample sizes (tens or less)
    - computer simulations routinely use enormous sample sizes (thousands...)

"significant" ≠ "important"

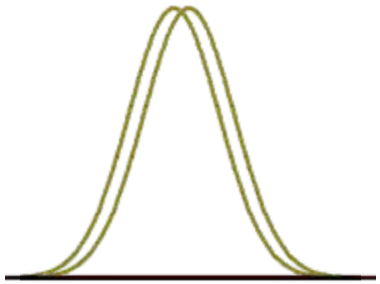
- the bigger the sample size, the better the statistical significance – that must be good?
  - not necessarily: can nearly *always* get a *statistically significant* result just by having a big enough sample size !
    - with *enough* samples, can distinguish two distributions …
  - but it might not be an *important* difference
    - … but the two distributions might still be very *very* similar …
  - the old algorithm has a success rate of 52.38%
  - whereas *my* algorithm's success rate is 52.41%
    - with improvement significant at the 99.9% confidence level
    - so why aren't you impressed by my result?
    - … because it's (probably) a very small *effect*
  - happens easily when experimental runs are "cheap"

effect size

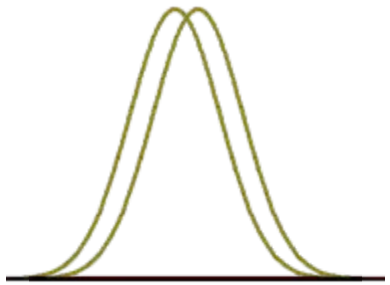
- measure of ***importance***
  ("scientific significance")

  a small effect

  - the data's *spread* is very much bigger than the difference in the medians
  - any "improvement" in the median is washed out by the data's spread

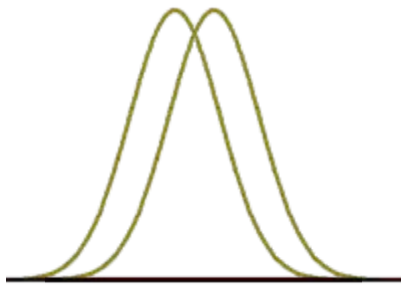


  a medium effect



  a big effect

  - *may* be worth getting excited about



effect size : "*A*" measure

[Vargha & Delaney 2000]

- calculate in Matlab:

```
[p,h,st] = ranksum(X,Y,'alpha',0.05);
N = size(X,1);
M = size(Y,1);

A = (st.ranksum/N - (N+1)/2)/M;
```

  - notice that you can do this and the rank-sum test *in one go* !
- measure of ***importance*** ("***scientific*** significance")
    - A lies between 0 and 1; if $A < 0.5$, use $1 - A$ in the test

0.5 : no effect (same medians)

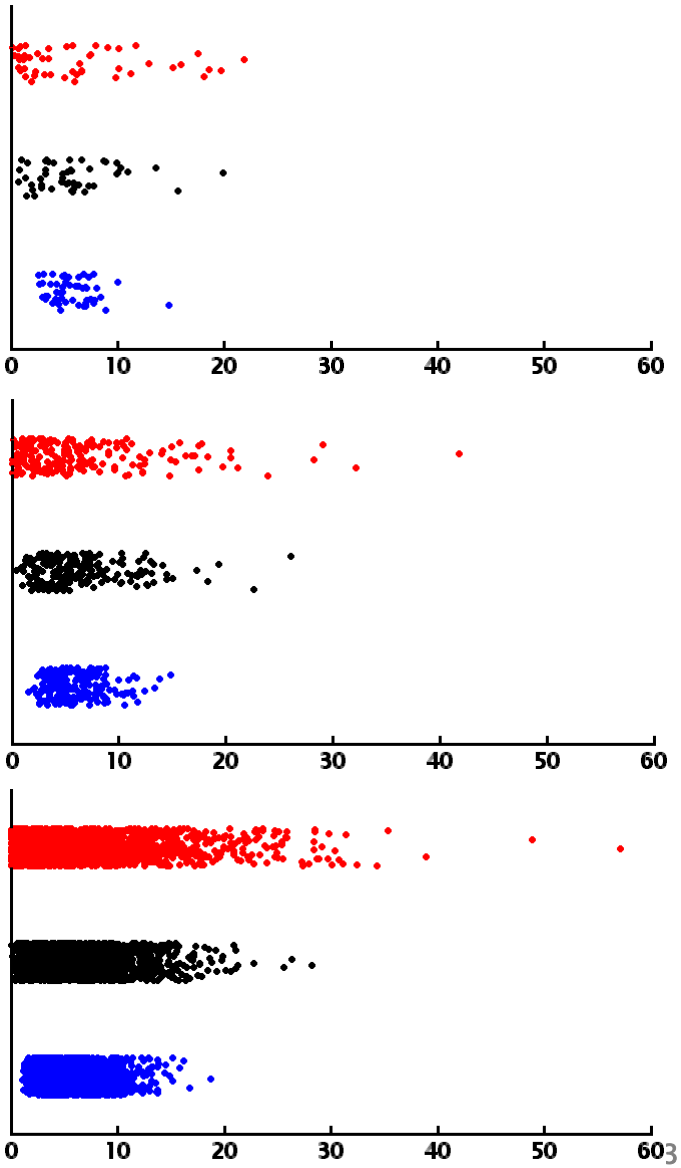
0.56 : a small effect

0.64 : a medium effect

0.71 : a big effect

more significant, but always *small* effect

N = 50	rank sum	KS p	A
blue/blk	0.395	0.241	0.550
blk/red	0.697	0.155	0.523
blue/red	0.316	0.001	0.558
N = 200			
blue/blk	0.191	0.008	0.538
blk/red	0.030	0.001	0.563
blue/red	0.002	0.000	0.591
N= 2000			
blue/blk	0.000	0.000	0.546
blk/red	0.000	0.000	0.570
blue/red	0.000	0.000	0.607



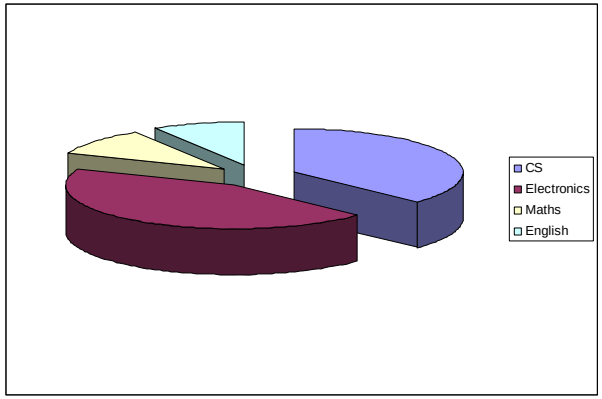
graphing your data

http://flowingdata.com/2008/09/19/pie-i-have-eaten-and-pie-i-have-not-eaten/

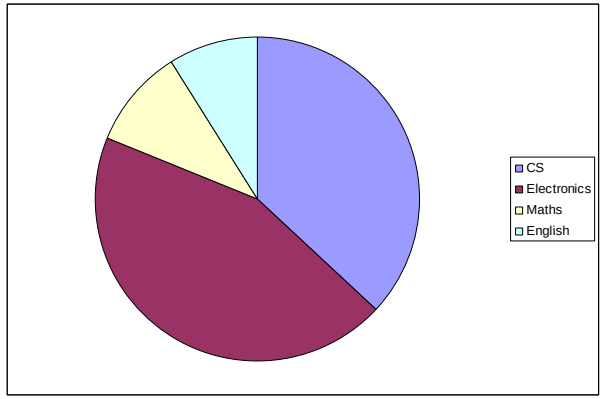
is your picture worth a kilo-word?

CS	37
Electronics	44
Maths	10
English	9

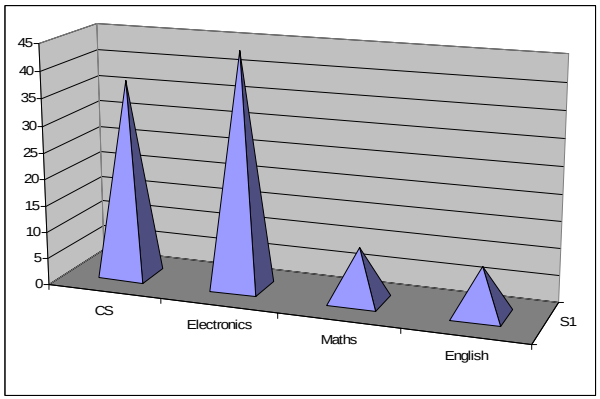
best !
for *this* data
(get more data?)



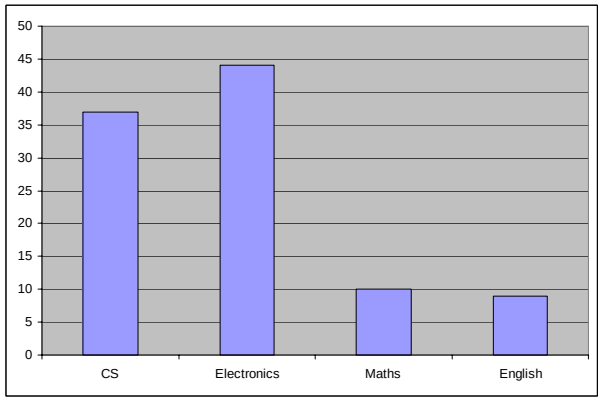
Chartjunk



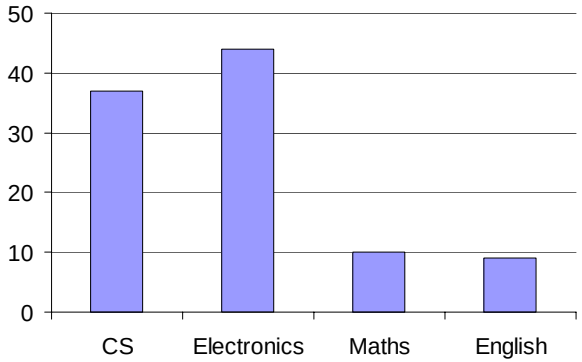
small distinctions obscured



More Chartjunk



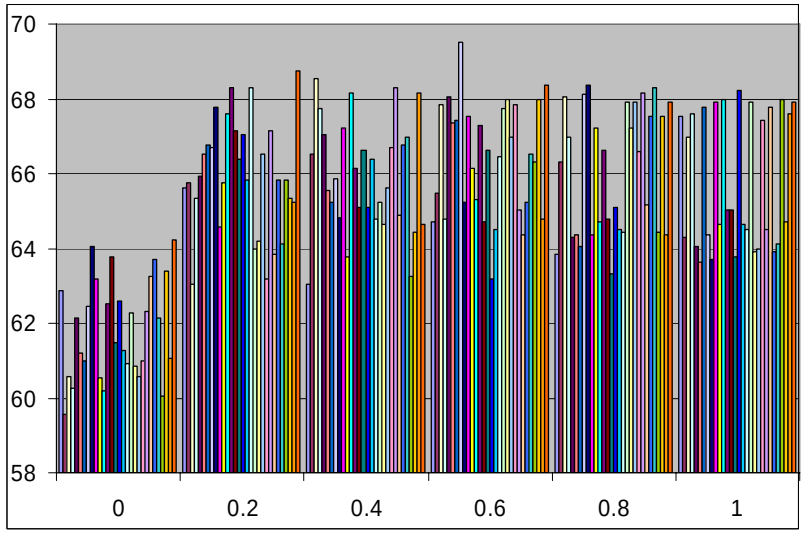
Cluttered

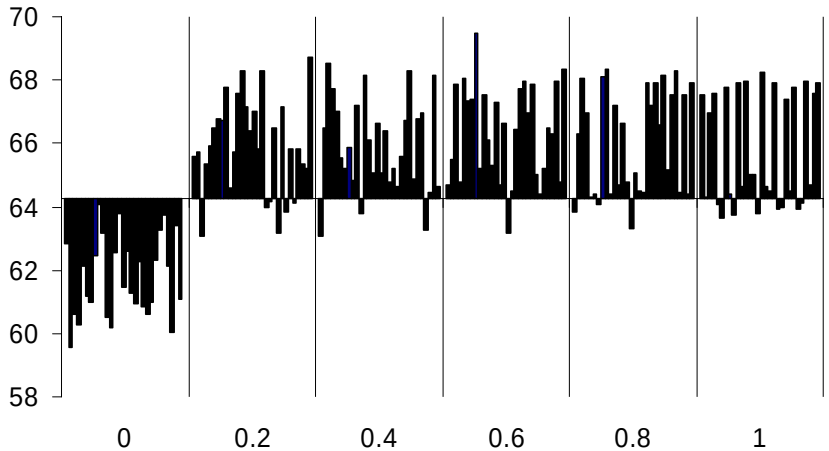


better

plot your data to expose its structure

0.00	0.20	0.40	0.60	0.80	1.00
62.86	65.61	63.06	64.70	63.85	67.54
59.58	65.74	66.51	65.48	66.32	64.31
60.59	63.06	68.55	67.86	68.05	66.96
60.26	65.35	67.73	64.77	66.96	67.60
62.13	65.93	67.03	68.05	64.31	64.05
61.20	66.51	65.55	67.35	64.38	63.65
61.00	66.77	65.22	67.41	64.05	67.79
62.46	66.71	65.87	69.50	68.11	64.38
64.05	67.79	64.83	65.22	68.36	63.72
63.19	64.57	67.22	67.54	64.38	67.92
60.53	65.74	63.79	66.13	67.22	64.64
60.19	67.60	68.17	65.29	64.70	67.98
62.53	68.30	66.13	67.28	66.64	65.03
63.79	67.16	65.09	64.70	64.77	65.03
61.47	66.39	66.64	66.64	63.32	63.79
62.60	67.03	65.09	63.19	65.09	68.24
61.26	65.81	66.39	64.51	64.51	64.64
60.93	68.30	64.77	66.45	64.44	64.51
62.27	63.98	65.22	67.73	67.92	67.92
60.86	64.18	64.64	67.98	67.22	63.92
60.59	66.51	65.61	66.96	67.92	63.98
61.00	63.19	66.71	67.86	66.58	67.41
62.33	67.16	68.30	65.03	68.17	64.51
63.26	63.85	64.90	64.38	65.16	67.79
63.72	65.81	66.77	65.22	67.54	63.92
62.13	64.11	66.96	66.51	68.30	64.11
60.05	65.81	63.26	66.32	64.44	67.98
63.39	65.35	64.44	67.98	67.54	64.70
61.06	65.22	68.17	64.77	64.38	67.60
64.24	68.74	64.64	68.36	67.92	67.92

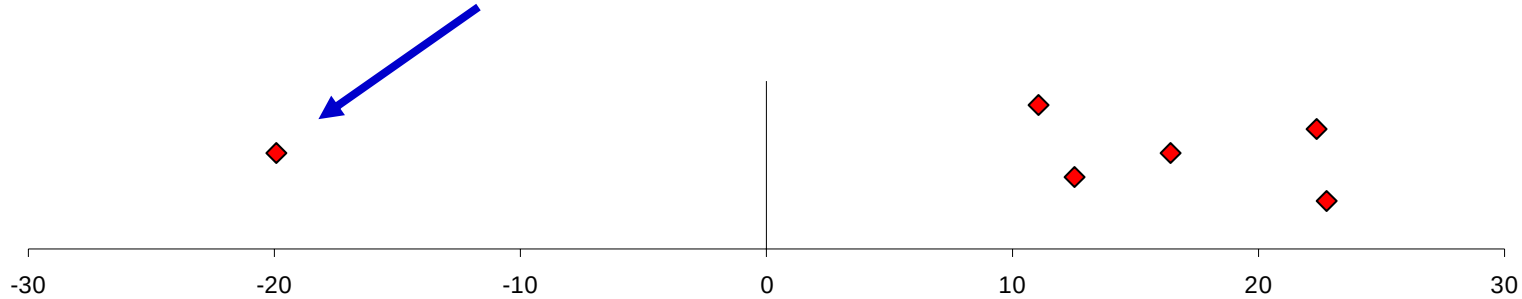




[Clark, Jacob, Stepney. Secret Agents Leave Big Footprints. *GECCO 2003*]

plot all your data, to see outliers

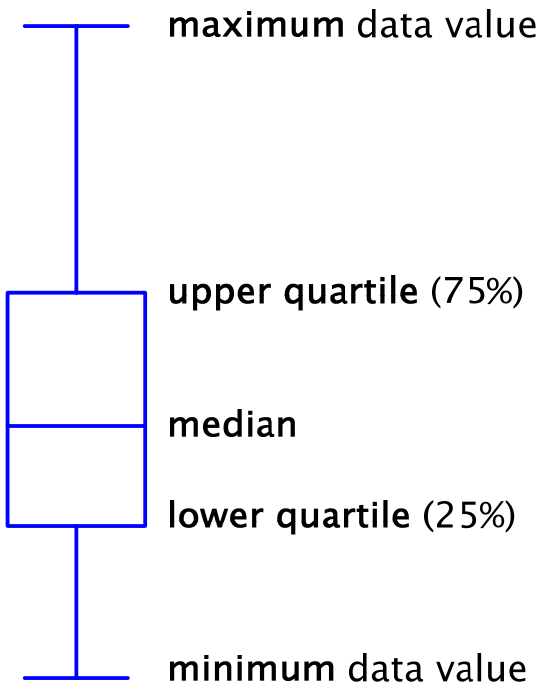
- you have some "anomalous" data



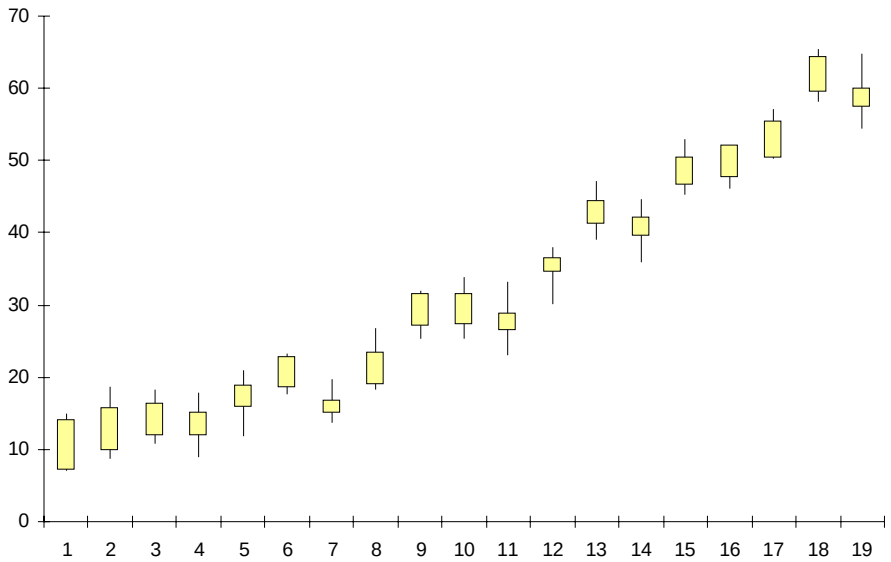
- don't just discard it as an "outlier" – *understand* it!
  - is it just a statistical fluctuation?
    - a once-in-a-blue-moon "six sigma" outlier?
  - is it an error in the experimental design or implementation?
    - fix the problem, and rerun *all* the experiments
  - is it in interesting unexpected effect?
    - investigate it further!
    - it *might* be the basis of a new discovery

“box-and-whisker” plots

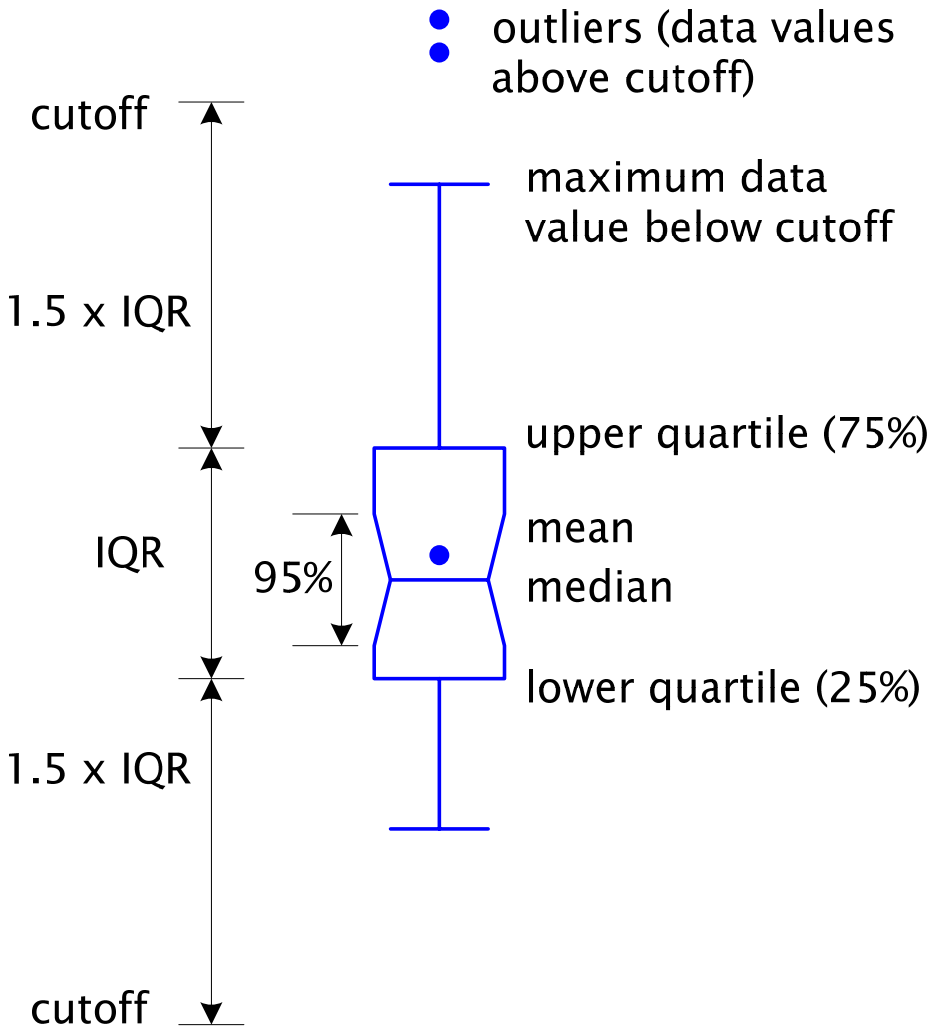
- median, quartiles
  - [Tukey, 1977]



if you *have* to use Excel, use
Chart Wizard > Standard Types > Stock:



“deluxe” box-and-whisker plots



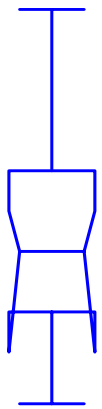
schematic plot: use a “cutoff” to highlight outliers (for *numerical* data)

plot *mean*, to highlight skew

“notches” at

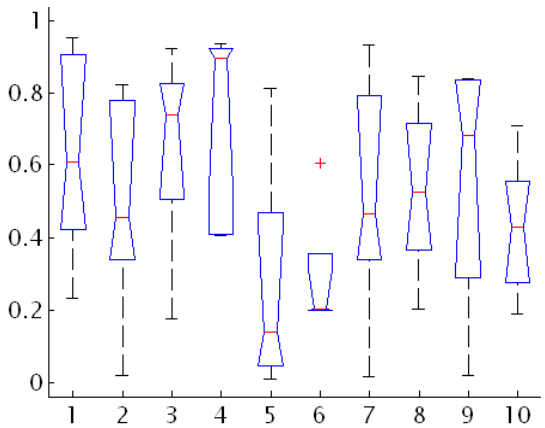
median $\pm$ 1.58(IQR / $\sqrt{n}$)

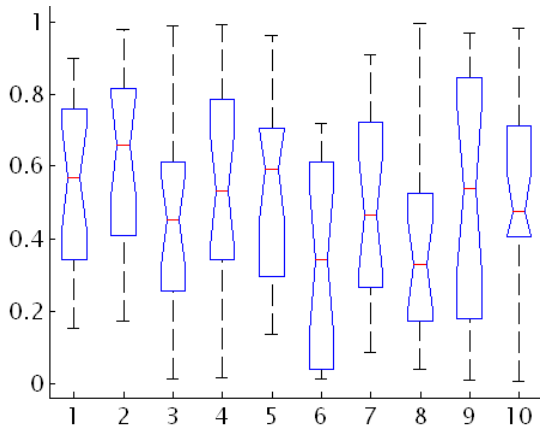
- if notches on separate bars do not overlap, ~ 95% confidence the medians are different
  - but do a proper test to check this
- small samples may have “folded back” notches outside the IQR:

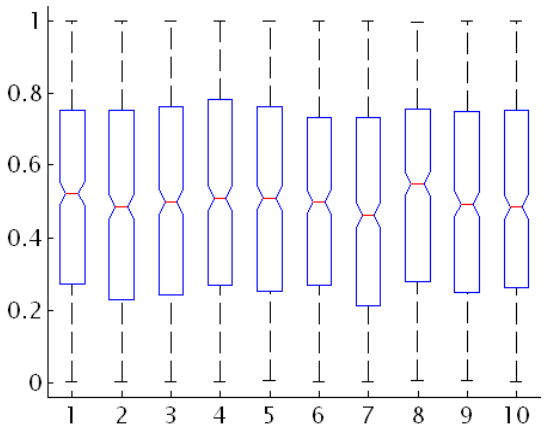


box plot examples

- uniform distribution, 10 samples
  - true median = 0.5, true IQR = 0.25–0.75
  - each of 5 elements, 20 elements, 500 elements
    - notice how the notches get smaller







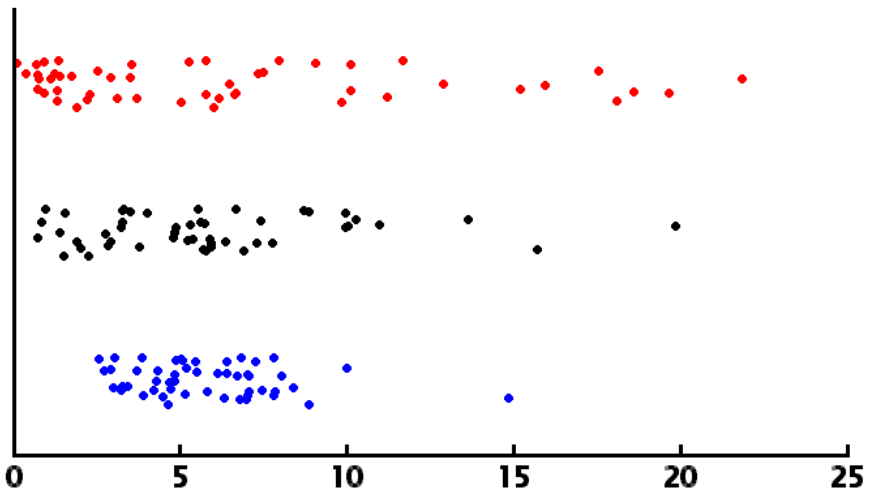

```
x = rand(5,10);
boxplot(x,'notch','on');
```

```
x = rand(20,10);
boxplot(x,'notch','on');
```

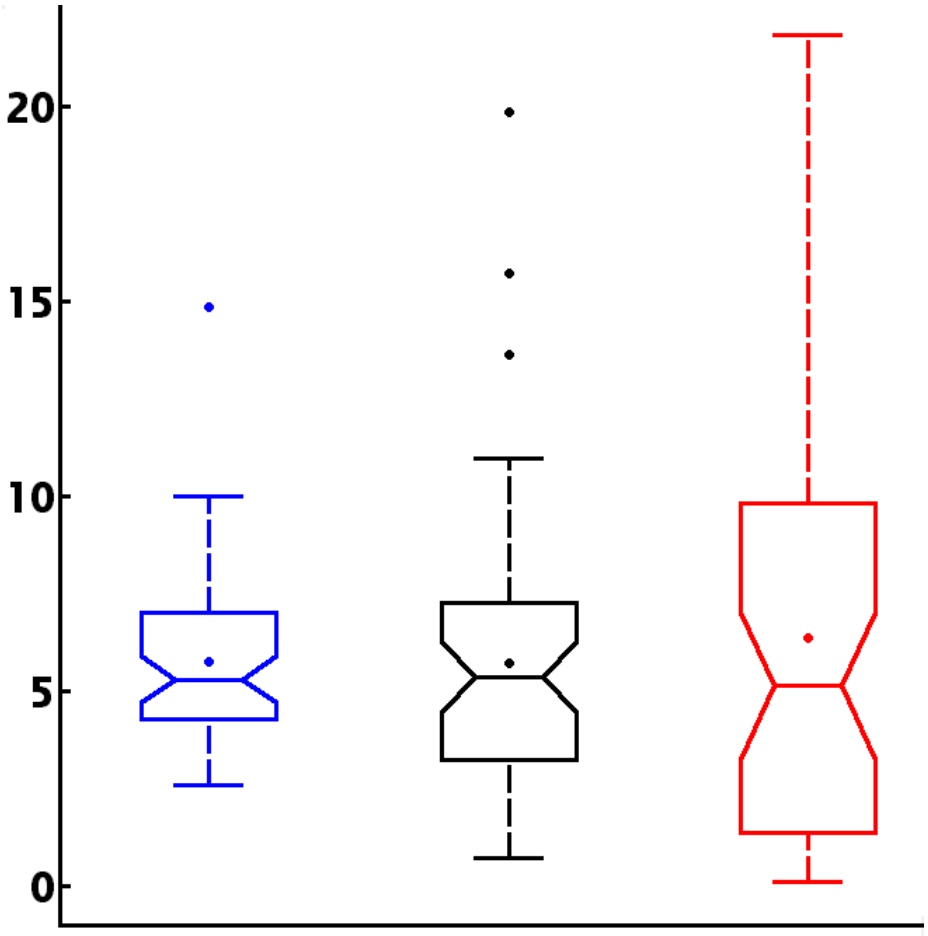
```
x = rand(500,10);
boxplot(x,'notch','on');
```

(the *entire* Matlab code that generated the random numbers, and drew the boxplot)

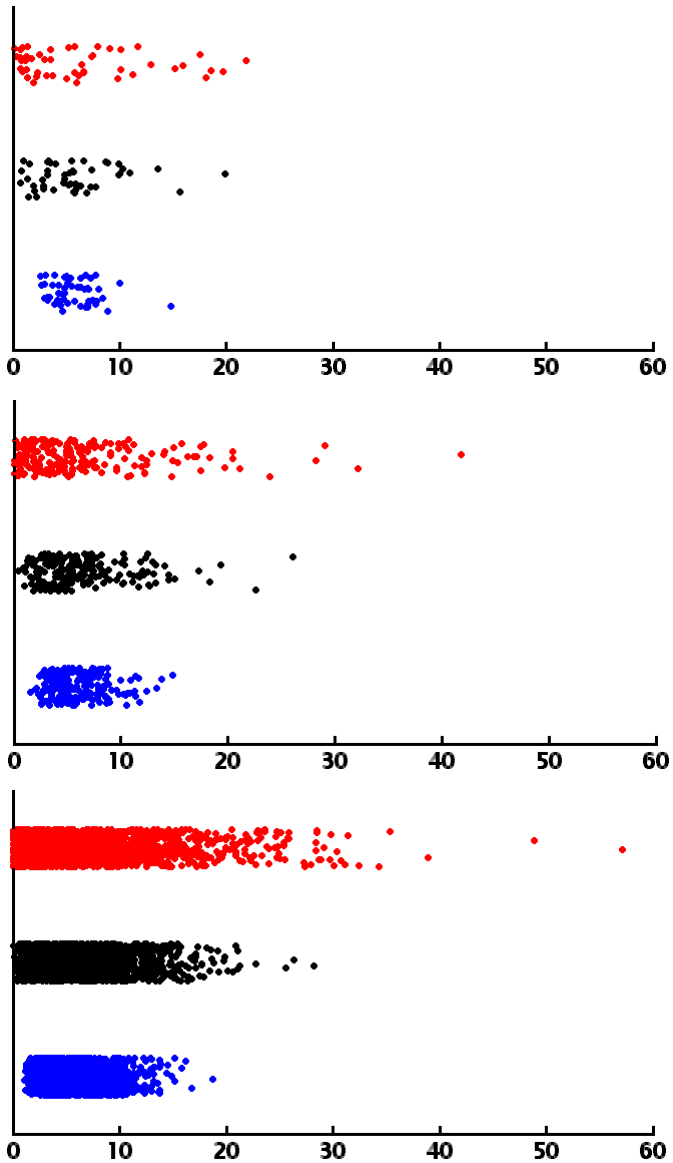
box plot : running example

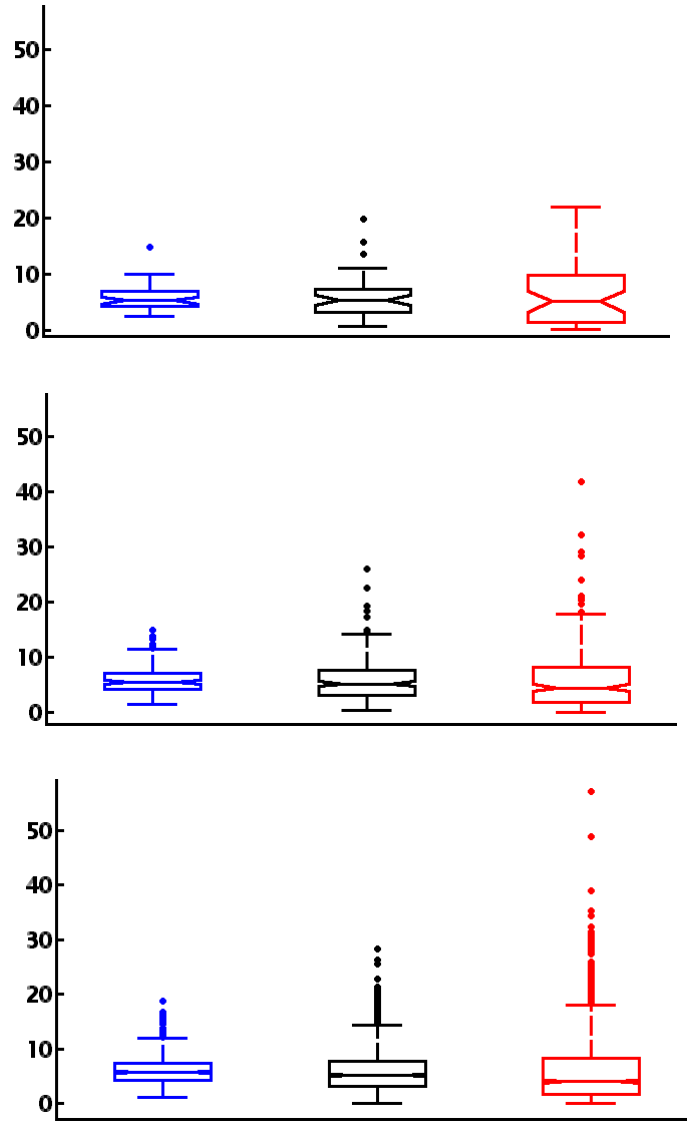


N = 50	rank-sum	KS p	A
blue/blk	0.395	0.241	0.550
blk/red	0.697	0.155	0.523
blue/red	0.316	0.001	0.558

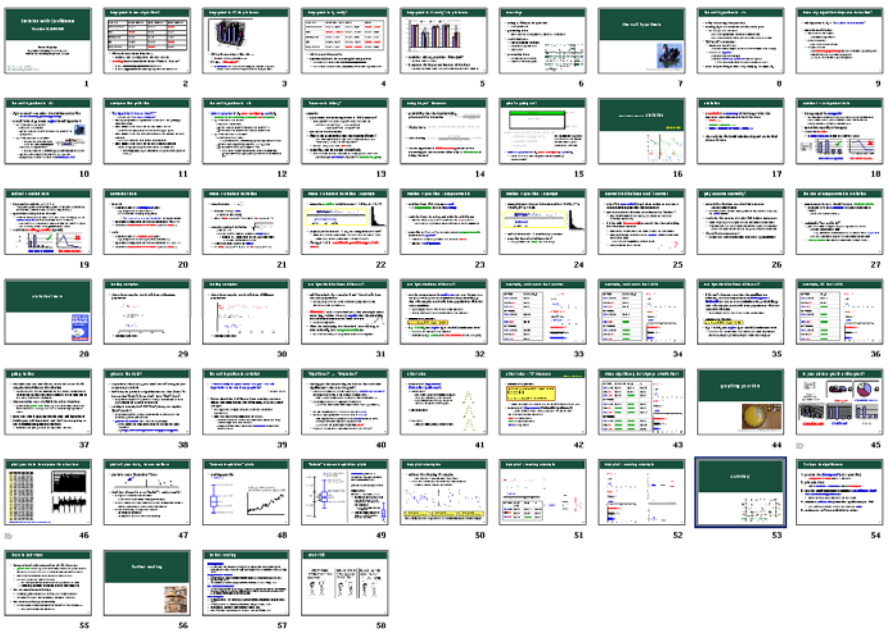


box plot : running example





summary



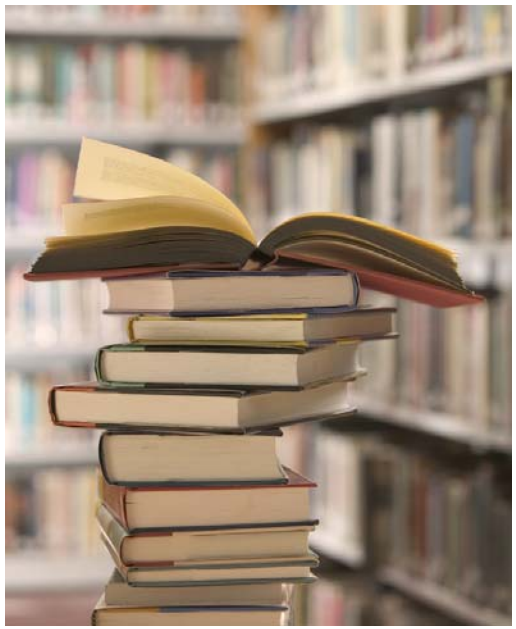
5 steps to significance

1. present the ***data spread*** (s.d. / quartiles)
   - ***not*** just an average (mean / median)
2. plot your data
   - *with* error bars / whiskers ; *without* chartjunk
3. use the ***rank-sum test*** to calculate ***confidence levels for statistical significance***
   - or the KS test, if the medians are the same
4. calculate ***effect size*** (scientific significance) as well
   - and don't get excited unless the effect is ***large***
5. make your raw data available for others

there is *lots* more

- here, covered only comparison of two data sets
  - paired data tests, eg, sets of (before, after) data pairs, to test whether a change has had a statistically significant effect
  - other non-parametric tests for these other cases
  - and the subject is still advancing
    - the non-parametric *A* effect size test published in 2000
    - bootstrap, jackknife resampling feasible with computers
- lots on experimental design
  - choosing your confidence levels, and number of runs
  - "factorial" designs for controlling multiple variables
- lots more on data presentation
  - scatter plots ; "small multiples" ; "stem-and-leaf" plots, …
    - rich presentation of a lot of data

further reading



further reading

null hypothesis

- S. Axelsson. The Base-Rate Fallacy and its Implications for the Difficulty of Intrusion Detection. *Proc. 6th ACM Conf Computers and Communication Security*. 1999

nonparametric statistics

- Sidney Siegel. *Nonparametric Statistics for the Behavioral Sciences*, 2nd edn. McGraw-Hill, 1988
- W. J. Conover. *Practical Nonparametric Statistics*, 3rd edn. Wiley, 1999

the *A* effect size measure

- Andra Vargha, Harold D. Delaney. A critique and improvement of the "CL" common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, **25**(2):101-132, 2000

anti-chartjunk

- Edward R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, 1983
- Edward R. Tufte. *Envisioning Information*. Graphics Press, 1990
- Darrell Huff. *How to Lie with Statistics*. Pelican, 1954
- John W. Tukey. *Exploratory data analysis*. Addison Wesley, 1977