# Statistics and graphs

CSW

Susan Stepney

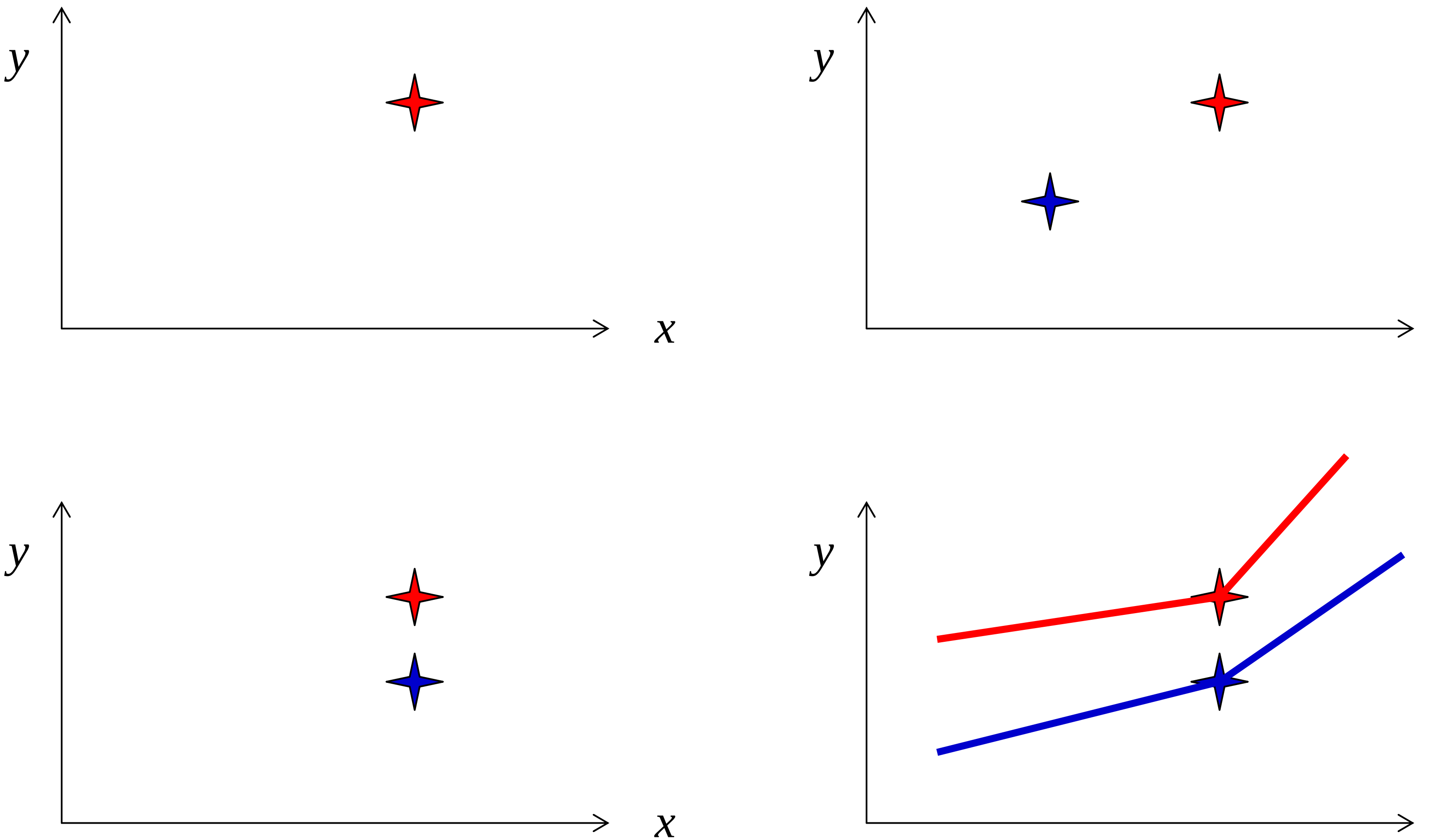
Department of Computer Science
University of York

# recap

- Null hypothesis $H_0$ is assumed true, unless ***rejected*** by the experimental data

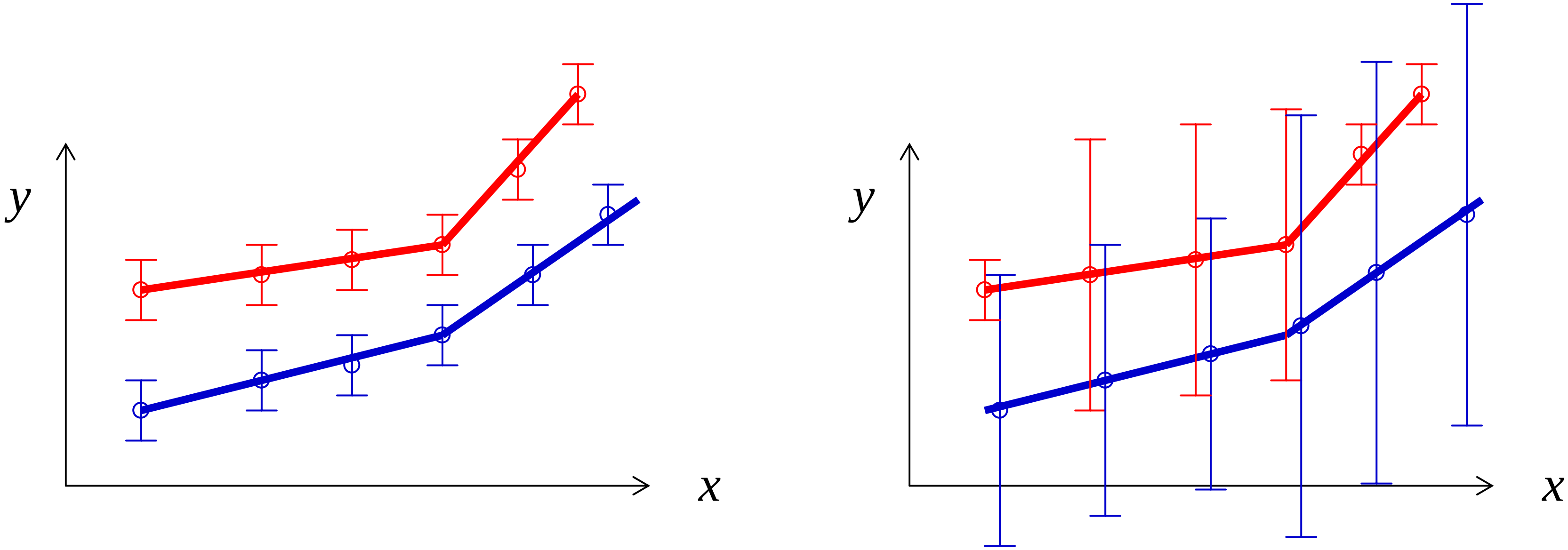
- $H_0$ = "all swans are white"
  - rejected by a counter-example
  - I see a black swan, and reject $H_0$
- $H_0$ = "no effect on readability"
  - rejected by *statistically significant* evidence against
  - I see a certain degree of improvement, and reject $H_0$
    - reject only at a given *confidence level*
  - how *much* improvement do I need?
    - how *much* confidence can I have?

# What do these tell you?



# it depends…

- everything varies with the data sample somehow
- is the difference *statistically significant*?

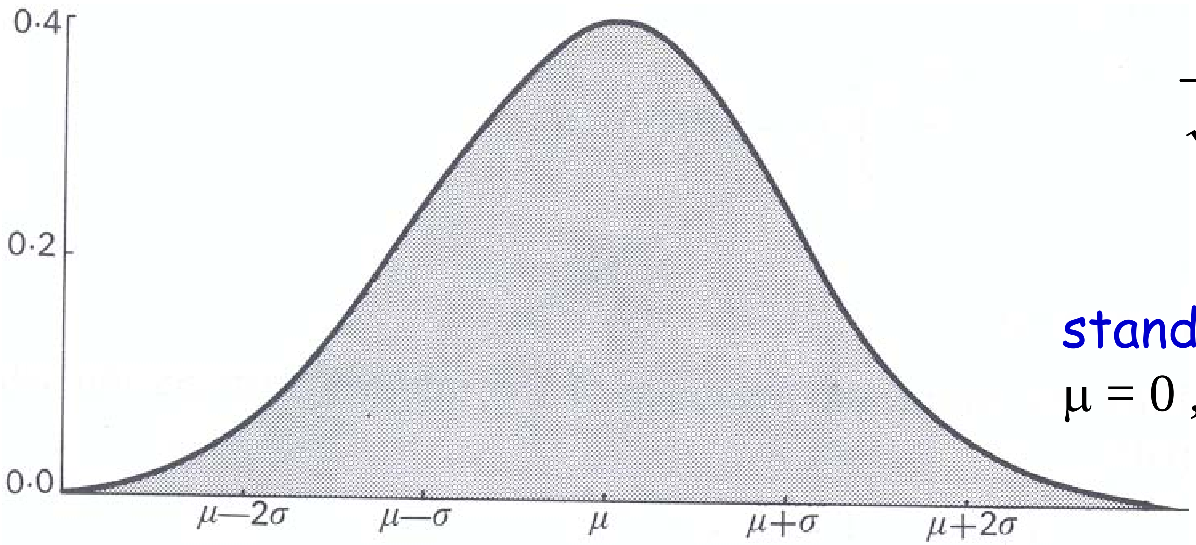


# some definitions

- population size $N$ of items $\{ x_i \}_{i = 1..N}$
- population mean $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$
- population standard deviation (RMS) $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mu - x_i\right)^2}$

- sample size $n$ usually << $N$ of items $\{ x_j \}_{j = 1..n}$
- sample mean $m = \frac{1}{n}\sum_{i=1}^{n} x_i$
- sample standard deviation $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(m - x_i\right)^2}$
  - note the $n - 1$
  - because derived from *estimated* sample mean
    - one fewer “degrees of freedom”

# Normal (Gaussian) probability distribution




$$\frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

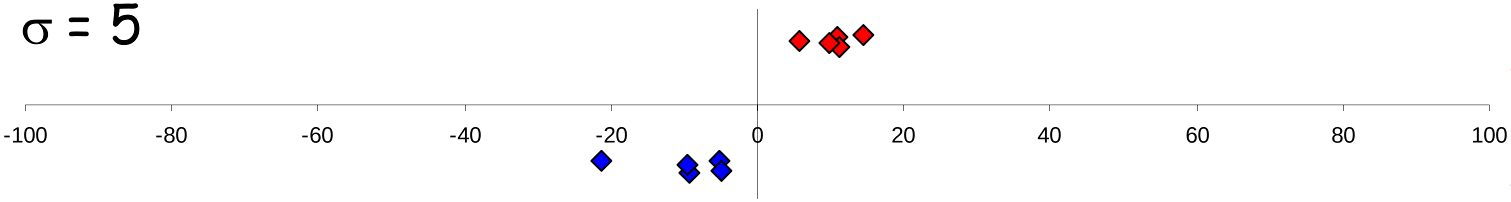
standard normal distribution,
$\mu = 0$ , $\sigma = 1$ :

$$\frac{1}{\sqrt{2\pi}}\exp\left(-x^2/2\right)$$

- 68.3% lies within $\pm$ 1 standard deviation of the mean
- 95.5% lies within $\pm$ 2 s.d.
- 99.7% (essentially "all") lies within $\pm$ 3 s.d.
  - it is (often incorrectly!) assumed that the underlying distribution is normal
  - there are certain tricks if it isn't

# normal distributions ; $\mu = \pm 10$ ; $n = 5$

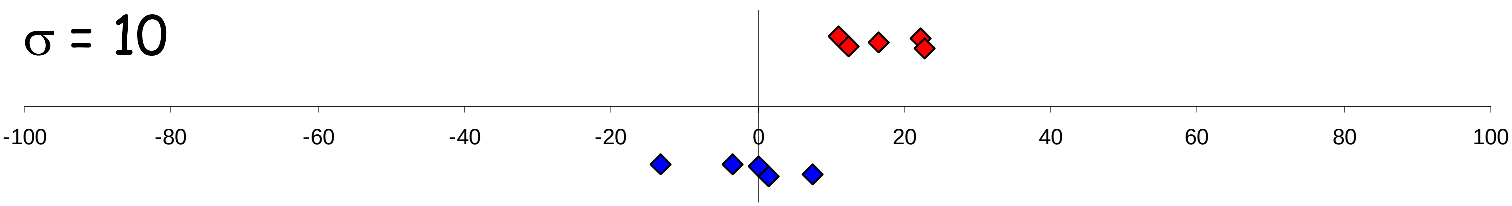
$\sigma = 5$




$m = 10.3$, $s = 3.1$

$m = -10.1$, $s = 6.6$

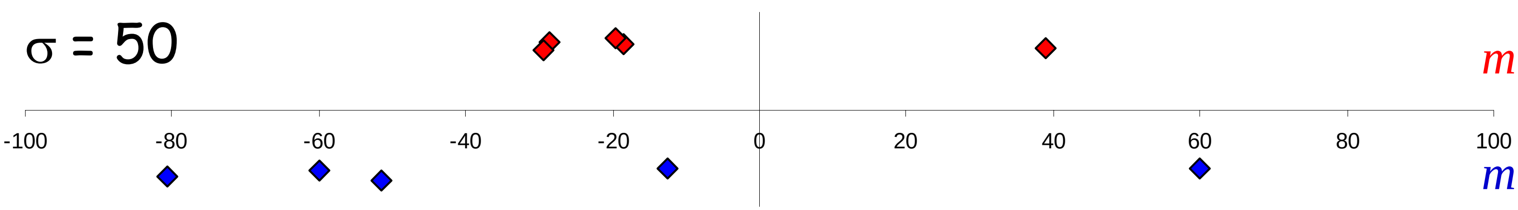
$\sigma = 10$




$m = 17.0$, $s = 5.4$

$m = -1.5$, $s = 7.6$

$\sigma = 50$

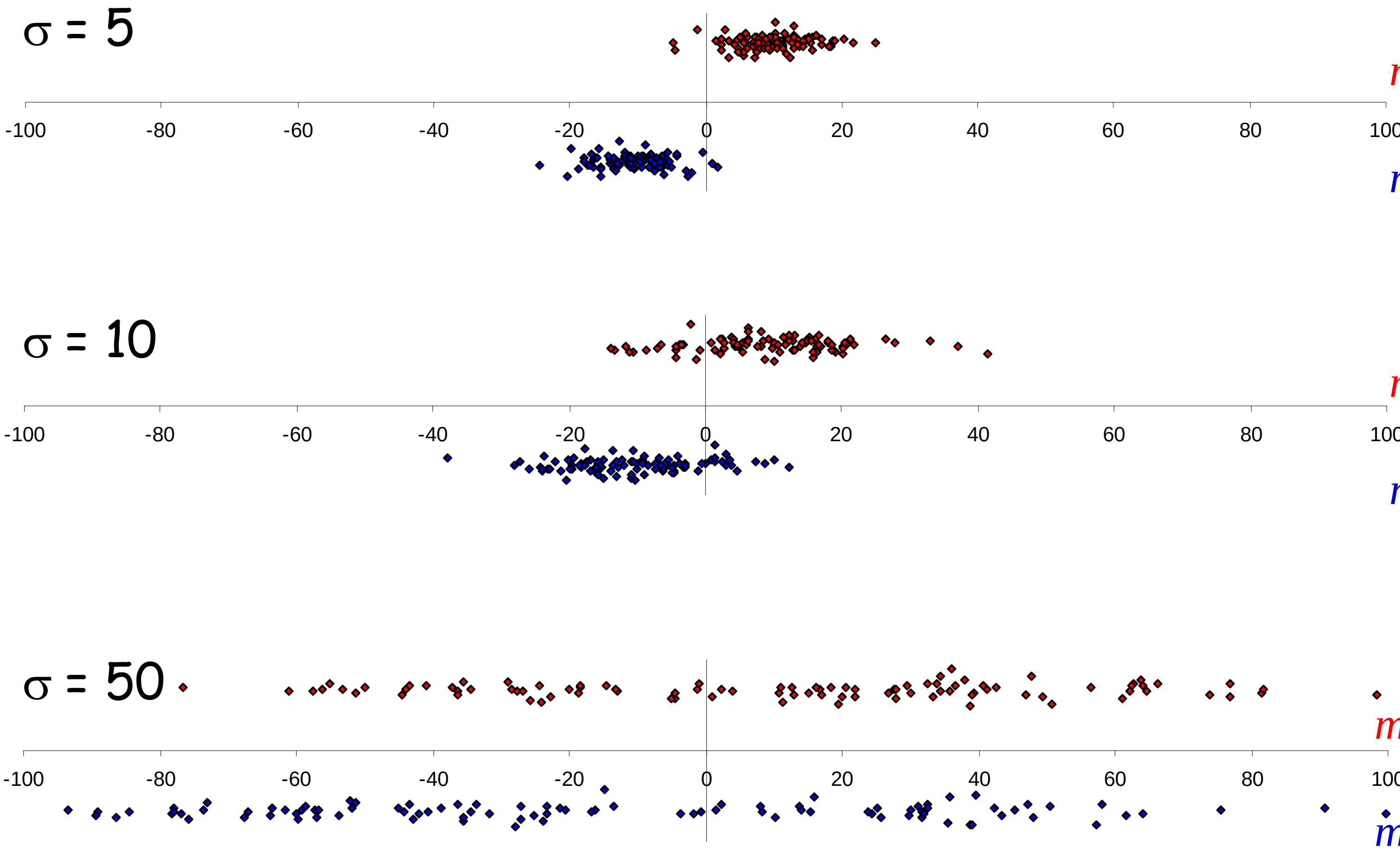



$m = -11.5$, $s = 28.7$

$m = -28.9$, $s = 55.5$

[inspired by Wineberg & Christensen, tutorial, CEC 2004]

# normal distributions ; $\mu = \pm 10$ ; $n = 100$




$m$ = 9.9, $s$ = 5.1

$m$ = -10.6, $s$ = 4.6

$m$ = 9.0, $s$ = 10.3

$m$ = -10.7, $s$ = 9.3

$m$ = 16.3, $s$ = 44.9

$m$ = -12.6, $s$ = 51.5

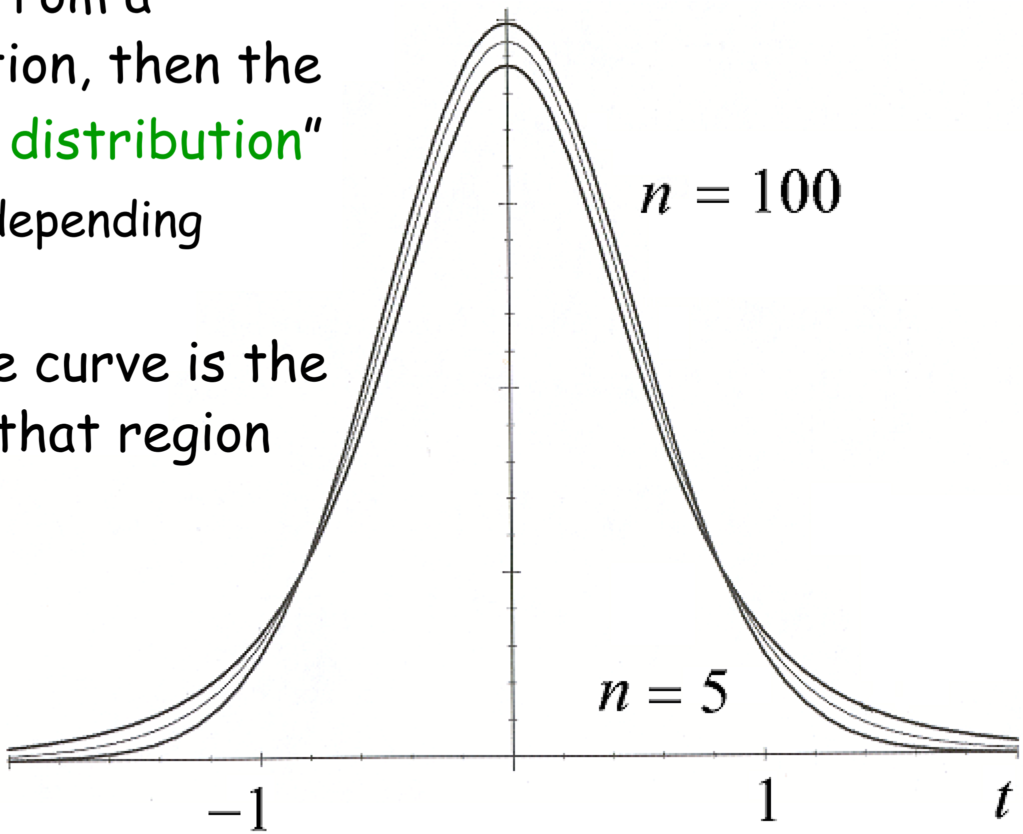
# estimating the true mean from a sample

- started with a normal population, mean $\mu$ and s.d. $\sigma$
- generated *samples* drawn from this, $n$ = 5, $n$ = 100
- calculated the sample mean $m$ and s.d. $s$
  - the sample mean is different from the true mean
  - can we use it to *estimate* the true mean?
    - because, in real cases, we *don't know* the true mean!
- intuitively:
  - the smaller the standard deviation, the closer to the true mean
  - the bigger the sample, the closer to the true mean

# $t$ distribution

- we have a sample with mean $m$, from a population with mean $\mu$
  - *statistics* : take samples from a standard normal distribution, then the sample means $m_i$ have a "$t$ distribution"
    - *family* of distributions, depending (weakly) on $n$
  - area under a region of the curve is the probability that $m$ lies in that region

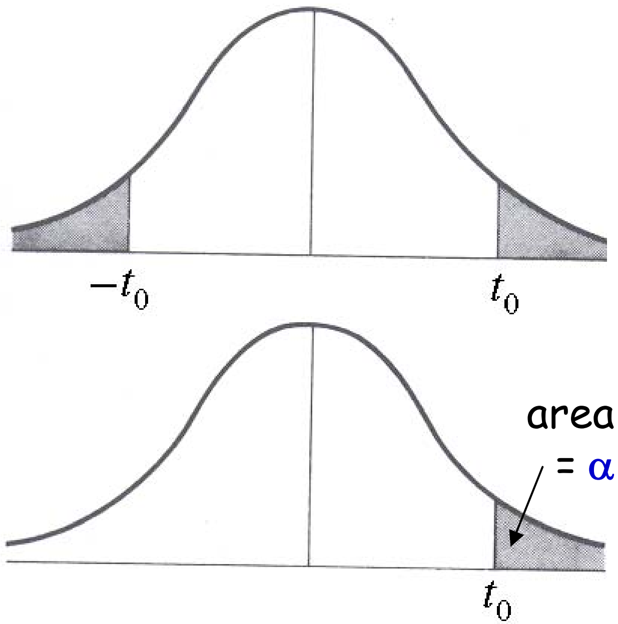
$$t = \frac{m - \mu}{s/\sqrt{n}}$$




- also called the "Student $t$ distribution", after the pen-name of statistician W. S. Gossett

# confidence levels

- "95% confidence" : only $\alpha$ = 5% of the area under the curve lies in the tail(s)
  - only 5% chance that the mean lies outside the bounds
  - "99% confidence" : $\alpha$ = 1% lies in the tail(s)
- books of tables of the value of $t_0$ for different confidence levels $\alpha$ and different sample sizes $n$
  - one tailed test: $t_\alpha(n-1)$
  - two tailed test: $t_{\alpha/2}(n-1)$
- or MS-Excel function
  - one tailed test: **TINV(2*$\alpha$, $n-1$)**
  - two tailed test: **TINV($\alpha$, $n-1$)**



# confidence levels (two tailed) $\mu$ = 10, $\sigma$ = 10

$$t_0 = \frac{m - \mu_0}{s / \sqrt{n}}$$

- n = 5

  measured mean $m$ = 17.0, standard deviation $s$ = 5.4

  95% confidence (1 in 20 chance of being wrong):

  TINV(0.05,4) = 2.78 ; $\mu = m \pm 2.78 * s$ / root $n$

  $\mu$ = 17.0 ± 6.7 ; **10.3 ≤ $\mu$ ≤ 23.7**

  99% confidence: $\mu$ = 17.0 ± 11.2 ; 5.8 ≤ $\mu$ ≤ 28.2

  99.9% confidence: $\mu$ = 17.0 ± 20.9 ; -3.9 ≤ $\mu$ ≤ 37.9

$$\mu_0 = m \pm t_0 \frac{s}{\sqrt{n}}$$

- n = 100

  measured mean $m$ = 9.0, standard deviation $s$ = 10.3

  95% confidence: $\mu$ = 9.0 ± 2.1 ; 6.9 ≤ $\mu$ ≤ 11.1

  99% confidence: $\mu$ = 9.0 ± 2.7 ; 6.3 ≤ $\mu$ ≤ 11.7

  99.9% confidence: $\mu$ = 9.0 ± 3.5 ; 5.5 ≤ $\mu$ ≤ 12.5

# are two means different?

- use the (Student) $t$ test, to calculate the probability $p$ that two sample means $m_1$ and $m_2$ are the same

- calculate the $t$ statistic $$t_0 = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

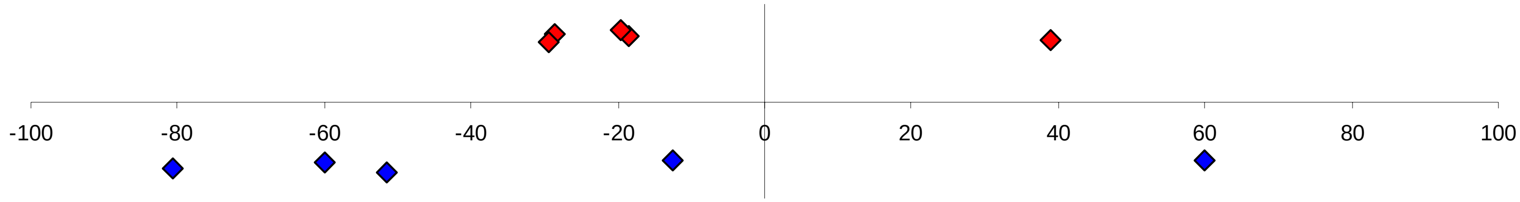
- calculate the probability that the means are same, which is the area under the $t$ distribution between $[-t_0, t_0]$
  - calculate in MS-Excel: $p = \mathbf{TDIST}(t_0 , n_1 + n_2 - 2 , 2)$

# confidence levels (two tailed) $\mu = \pm 10$, $\sigma = 50$

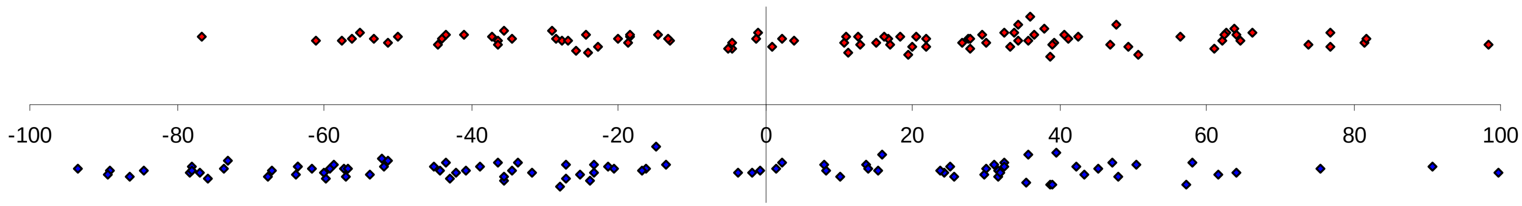
- $n = 5$
  - measured : $m_1 = -11.5$, $m_2 = -28.9$, $s_1 = 28.7$, $s_2 = 55.5$
    $t_0 = 0.624$; p = TDIST(0.624, 8, 2) = 0.55
  - probability the two means are *same* = **55%**
    - would *you* bet money on it?

- $n = 100$
  - measured : $m_1 = 16.3$, $m_2 = -12.6$, $s_1 = 44.9$, $s_2 = 51.5$
    $t_0 = 4.23$; p = TDIST(4.23, 198, 2) = $3.5 \times 10^{-5}$
  - probability the two means are *same* = **0.0035%**
    - I'd put money on them being different !

# reminder : $\mu = \pm 10$ ; $\sigma = 50$ ; $n = 5, 100$




$p$(means different) = 45%




$p$(means different) = 1

# “significant” does not mean “important”

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

# effect size : Cohen's $d$

$$d = t\sqrt{\frac{2}{n-1}}$$

- measure of ***importance***
  - here assume $n_1 = n_2 = n$

  0.2 : a small effect
    - the data's *dispersion* is bigger than the difference in the means

  0.5 : a medium effect

  0.8 : a big effect
    - may be worth getting excited about
- example: $\sigma$ = 50, $\boldsymbol{n}$ = 100, $\boldsymbol{t}_0$ = 4.23
    - probability that means are the same = **0.0035%**
    - but has $d$ = 0.6 : so don't get *too* excited by it
- example: $\sigma$ = 10, $\boldsymbol{n}$ = 100, $\boldsymbol{t}_0$ = 14.2
    - probability that means are the same = **0.0%**
    - and has $d$ = 2.0
      - so it's a big effect, too

[J. Cohen. *Statistical power analysis for the behavioral sciences*, 2nd edn. 1988. http://web.uccs.edu/lbecker/Psy590/es.htm]

# non-parametric tests

- all this so far assumes a normal probability distribution
  - some distributions are very affected by far "outliers"
    - long "tails" can dominate the mean, and make the standard deviation meaningless
  - there are tricks you can play if your distribution isn't normal
    - Central Limit Theorem, …
- can use ***non-parametric tests***
  - "no parameters" – no assumption about the *shape* of the probability distribution function
  - uses ***rank ordering*** instead of values
    - intuition: a lowest ranked outlier sample is just "the worst"; is doesn't matter how *much* worse than the rest it is

# median : the “non-parametric mean”

- mean : item with average *value*
  - mean {-20, 1, 2, 3, 4} = -2
  - mean {-20, 1, 2, 3} = -3.5
- median : item with average *rank*
  - rank the items in order, and pick the middle one
  - median {-20, 1, 2, 3, 4} = 2
    - if there is an even number of data items, average the two values
  - median {-20, 1, 2, 3} = 1.5
- median = 50$^{th}$ percentile
  - quartiles (25$^{th}$ and 75$^{th}$ percentiles) are a non-parametric measure of *spread*

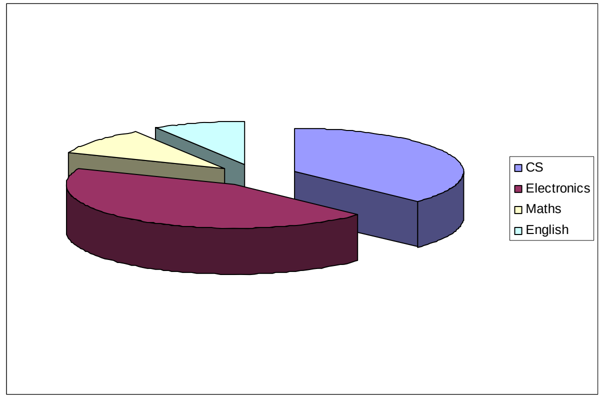
# are two medians different?

- there are non-parametric versions of the $t$ test to check if two ranked samples are significantly different
- also analogues of confidence tests, etc
- if you need to use these – check with an expert!

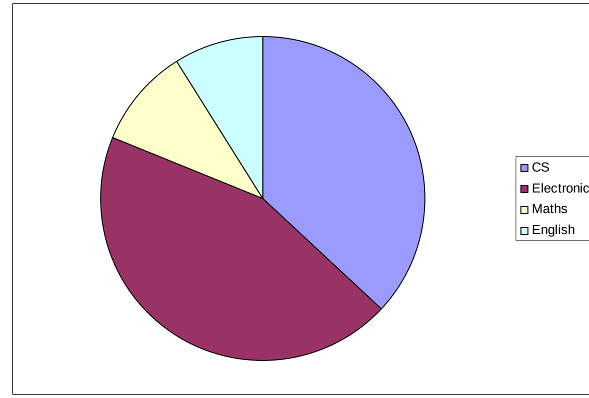
# is your picture worth a kilo-word?

| | |
|---|---|
| CS | 37 |
| Electronics | 44 |
| Maths | 10 |
| English | 9 |

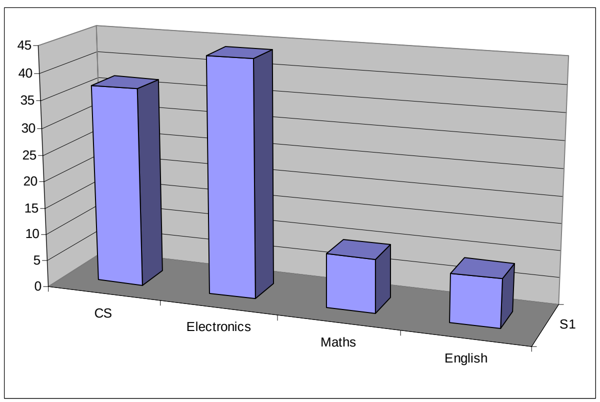
best !
for *this* data
(get more data?)



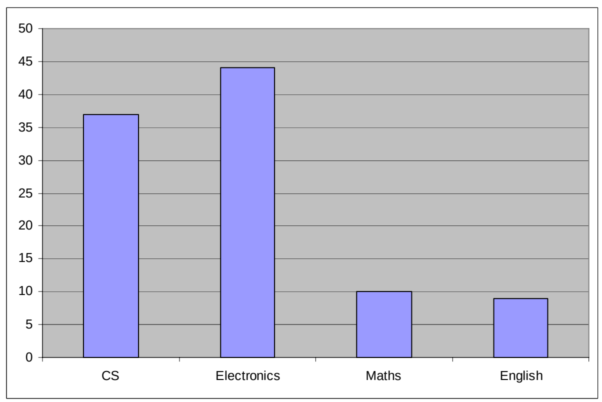

F kduwmxqn



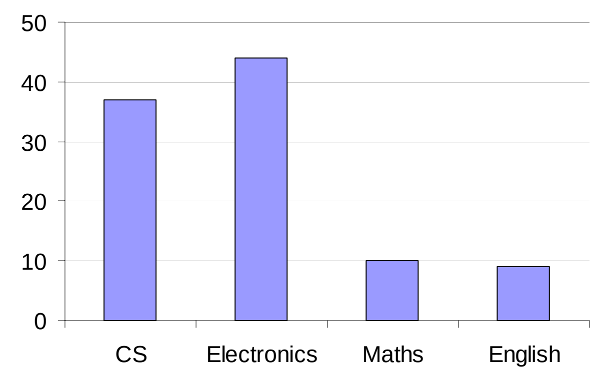

vpdoo glvwlqfwlrq revfxuhg




P ruh F kduwmxqn




Cluttered

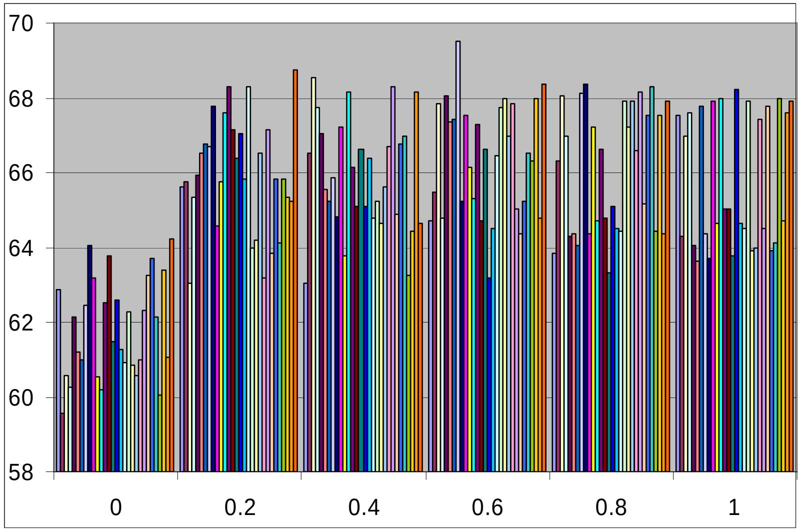


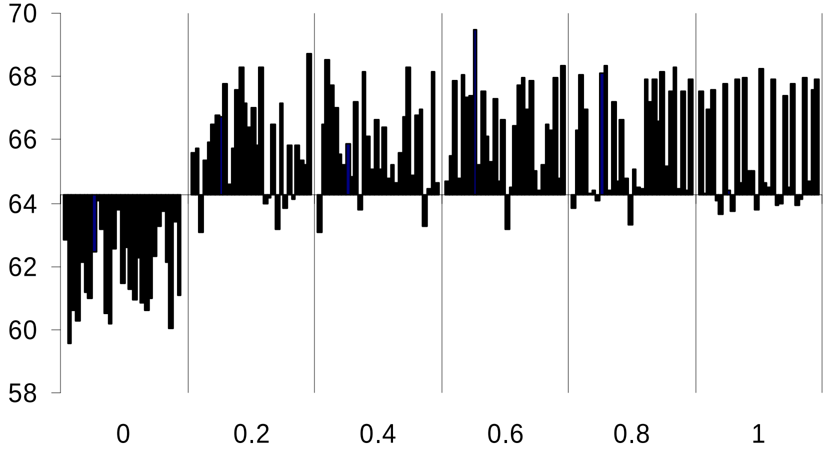

better

[Tufte. *The Visual Display of Quantitative Information.* Graphics Press, 1983]

# 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 |

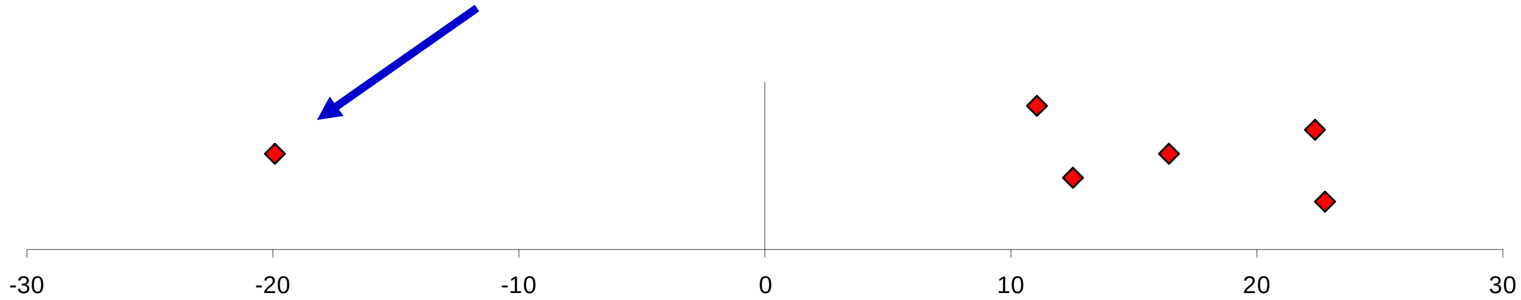



[Clark, Jacob, Stepney. Secret Agents Leave Big Footprints. *GECCO 2003*]

# plot all your data, to see outliers

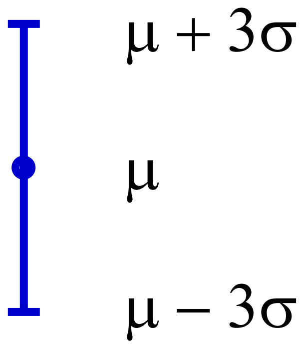
- you have some “anomalous” data

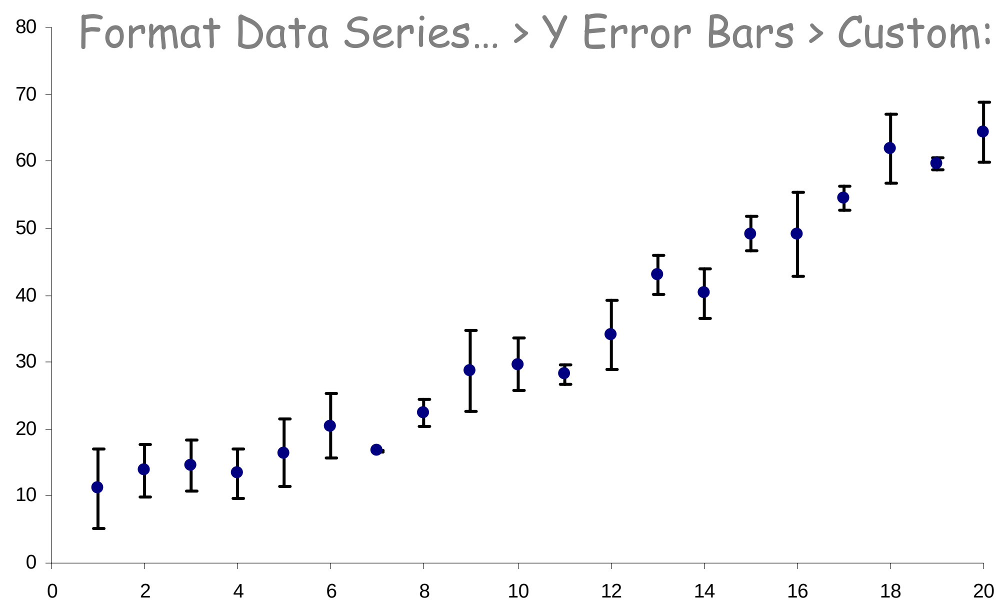



- don’t just discard it as an “outlier” – *understand* it!
  - is it just a statistical fluctuation?
    - a once-in-a-blue-moon “six sigma” outlier?
  - is it an error in the experimental design or implementation?
    - fix the problem, and rerun *all* the experiments
  - is it in interesting unexpected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

# 3-sigma error bars

- can use a ***scatterplot*** to show *all* the data
- can also *summarise* the data with a *statistic*
  - show spread as well as average, with *error bars*
  - normal "3-sigma" error bars: encompass ~ 99.7% of the data



# "box-and-whisker" error bars

- median, quartiles
  - John W. Tukey. *Exploratory data analysis.* Addison Wesley, 1977

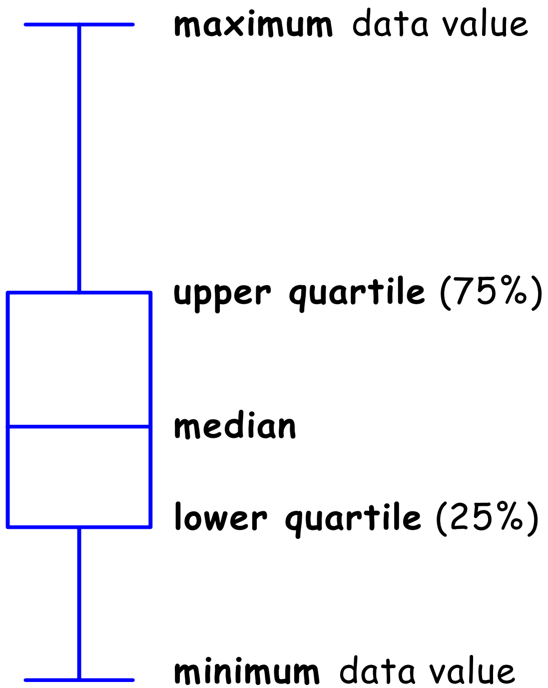
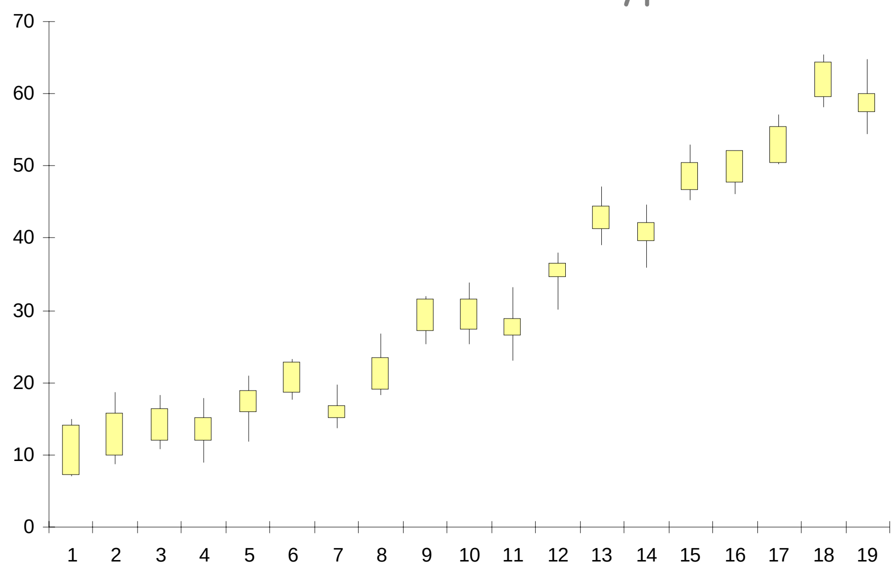
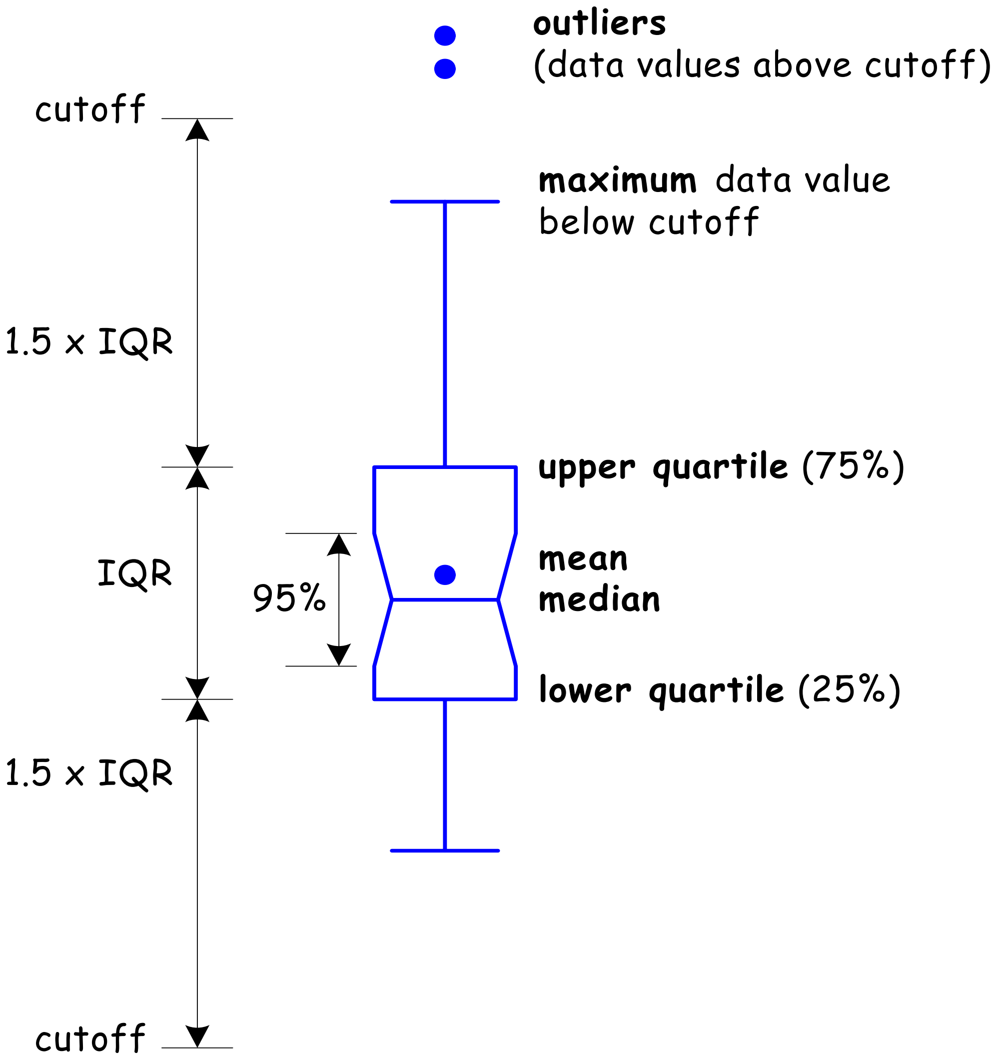



Chart Wizard > Standard Types > Stock:



# “deluxe” box-and-whisker bars

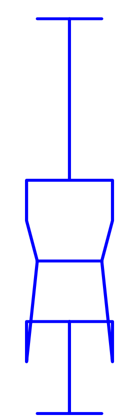



use a “cutoff” to highlight outliers

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
- small samples may have “folded back” notches outside the IQR:



# summary : statistics

- don't present single data points, or even just the mean
  - show the data *spread*
- do use $t$ distribution to calculate *confidence levels*
  - try to use *at least* the 99% level, preferably 99.9%
- do calculate *importance* as well as *statistical significance*
- do plot your data
  - *with* error bars ; *without* chartjunk

- *lots* more to experimental design, statistics and graphs
  - other measures of fit
  - "factorial" designs for controlling multiple variables
  - "small multiples", "stem-and-leaf" plots, …