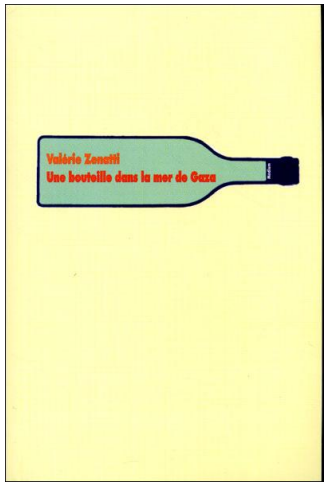
About statistical testing...

Sébastien Déjean
math.univ-toulouse.fr/~sdejean

I2MC
8 mars 2018

Some quotes to begin

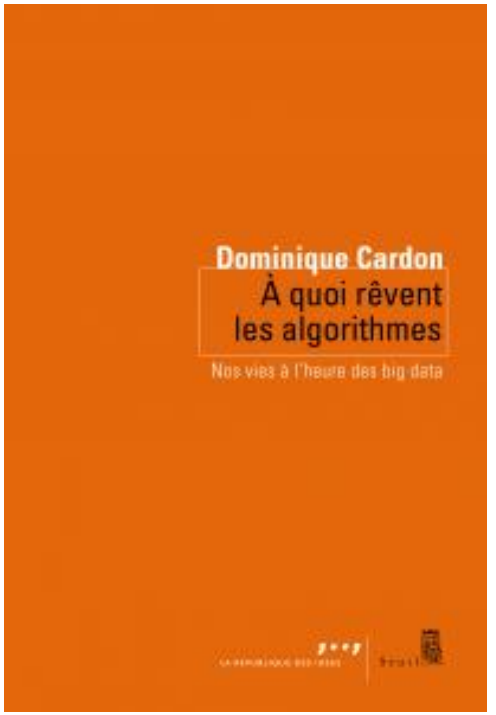


Mais les probabilités, les statistiques, c'est bon pour les maths, la biologie, ce sont des chiffres sur du papier.

Delphine Zenatti, *Une bouteille dans la mer de Gaza*

… face au déploiement de la société des calculs, il est nécessaire d'encourager la diffusion d'une culture statistique vers un public plus large que celui des seuls spécialistes.

Dominique Cardon, *A quoi rêvent les algorithmes. Nos vies à l'heure des big data.*



T'échappes à la police, pas aux statistiques.
Jean-Jacques Goldman, paroles de la chanson *Des vies*

Some quotes to begin

Après nous avoir convaincu de leur objectivité fondamentale, il ne reste plus aux chiffres qu'à nous amener doucement à penser qu'ils en déterminent le monopole. Dès lors, une forme de hiérarchie gagne l'argumentation et le raisonnement : contenir quelques chiffres qualifie automatiquement votre discours, même si personne ne prend la peine de comprendre vraiment ce qu'ils signifient, voire même s'ils sont sans rapport avec le sujet traité ! A contrario, de ce fait, toute argumentation purement textuelle semble dépréciée […] comme si le raisonnement et la rigueur ne pouvait exister hors des chiffres.

Lorsqu'on invoque les mathématiques pour garantir des résultats qui ne dépendent que des choix faits au départ, on trompe le lecteur et d'une certaine façon, on contraint cette discipline scientifique à blanchir des hypothèses douteuses. Les mathématiques sont alors prises en otage, ni plus ni moins. […] L'outil mathématique fait son travail, que l'hypothèse soit plausible ou non, qu'elle soit légitime ou non. En aucun cas, il n'assume la garantie des hypothèses sur lesquelles on le fait travailler. Un outil reste un outil.

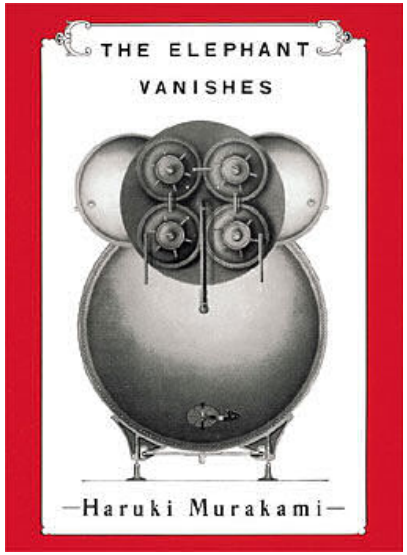
Outline

- Basics of statistical testing (inferential statistics)
- Practical aspects
- Visualisation
- Focus on two-factor ANOVA

Basics of statistical testing

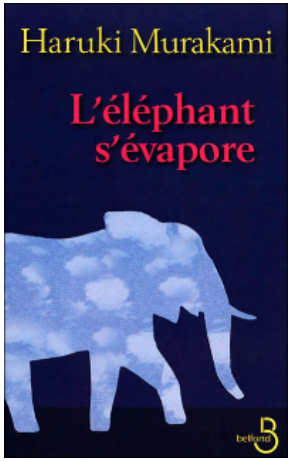
(inferential statistics)

Inferential statistics



I'm 26 years old, I work in the department of goods control […]. It would be impossible to control them carefully one by one. Consequently, we limit to pulling a few shoes buckles and to snack a few cakes ***as samples****.* (Translated from the french version)

The kangaroo Communiqué
short story from the collection *The elephant vanishes*
Haruki Murakami



J'ai 26 ans, je travaille dans le département du contrôle des marchandises […]. Il serait impossible de les contrôler soigneusement une à une […]. Par conséquent, on se borne à tirer sur quelques boucles de chaussures, à grignoter quelques gâteaux à titre d'échantillon.

Le communiqué du kangourou
nouvelle tirée du recueil *L'éléphant s'évapore*
Haruki Murakami

Inferential statistics: a story

In a factory producing gingerbread, the process performed to check the softness of the final product consists in cutting a bread into slices, then folding a slice till it breaks, and measuring the angle when the slice breaks (it's a destructive test).

The rule is that a slice of good gingerbread must break at 50° (fictive value): if the angle is lower than 50°, the bread is to dried, if it is greater than 50°, the bread is too soft. Every batches must be validated before being commercialised.

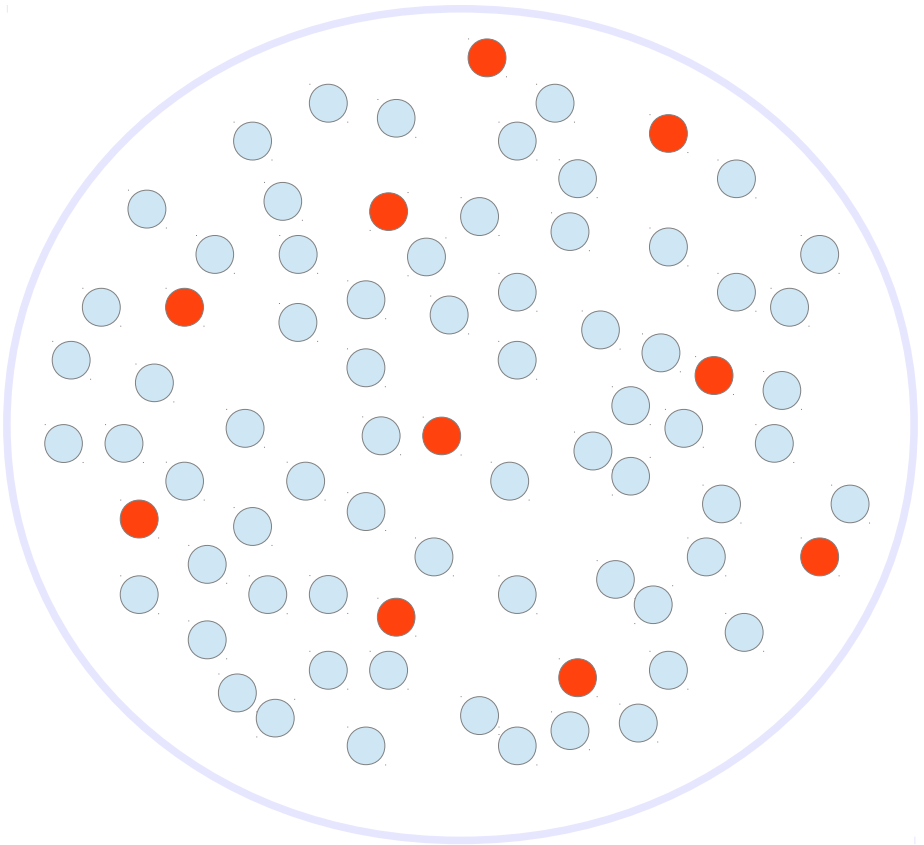
A broken slice cannot be sold as well as an improper bread (breaking angle $\neq$ 50°).

In such conditions, it is impossible to check every products (destructive test). It is mandatory to perform required measures on a sample representative of the population of gingerbreads (from the same batch). Regarding the experimental design, some basic rules must be followed: avoid taking the first or the last products build one day; or on a same production line if severals operate in parallel. The average breaking angle calculated from the sample is an estimator of this angle for the products manufactured in the same conditions (same batch).

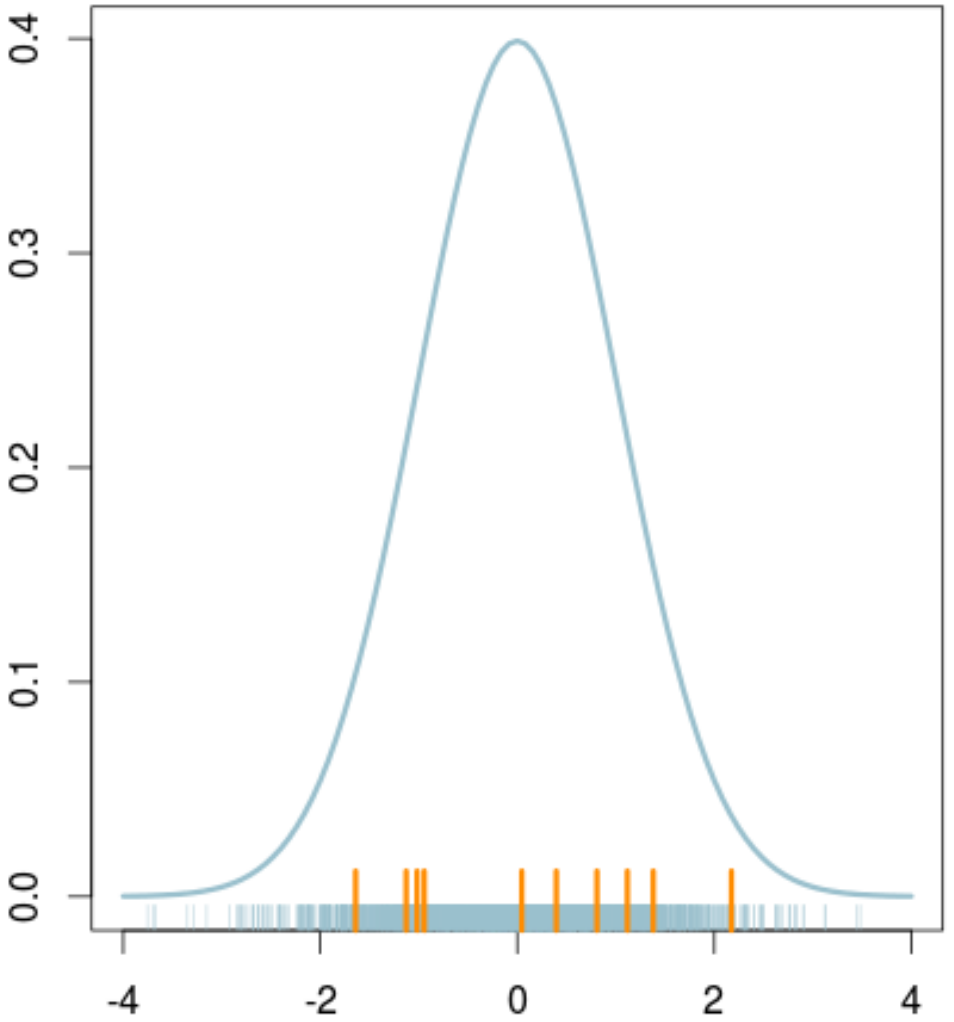
Inferential statistics: vocabulary

- Draw conclusions at a **population** level from informations collected on a **sample**.
- Survey (sample), census (population), representative sample...
- Quantitative informations at the population level are **estimated** from observations on samples.
- Measures performed on a sample are **observations** of the random variable representing the phenomenon at the population level.

Inferential statistics



Population
Sample

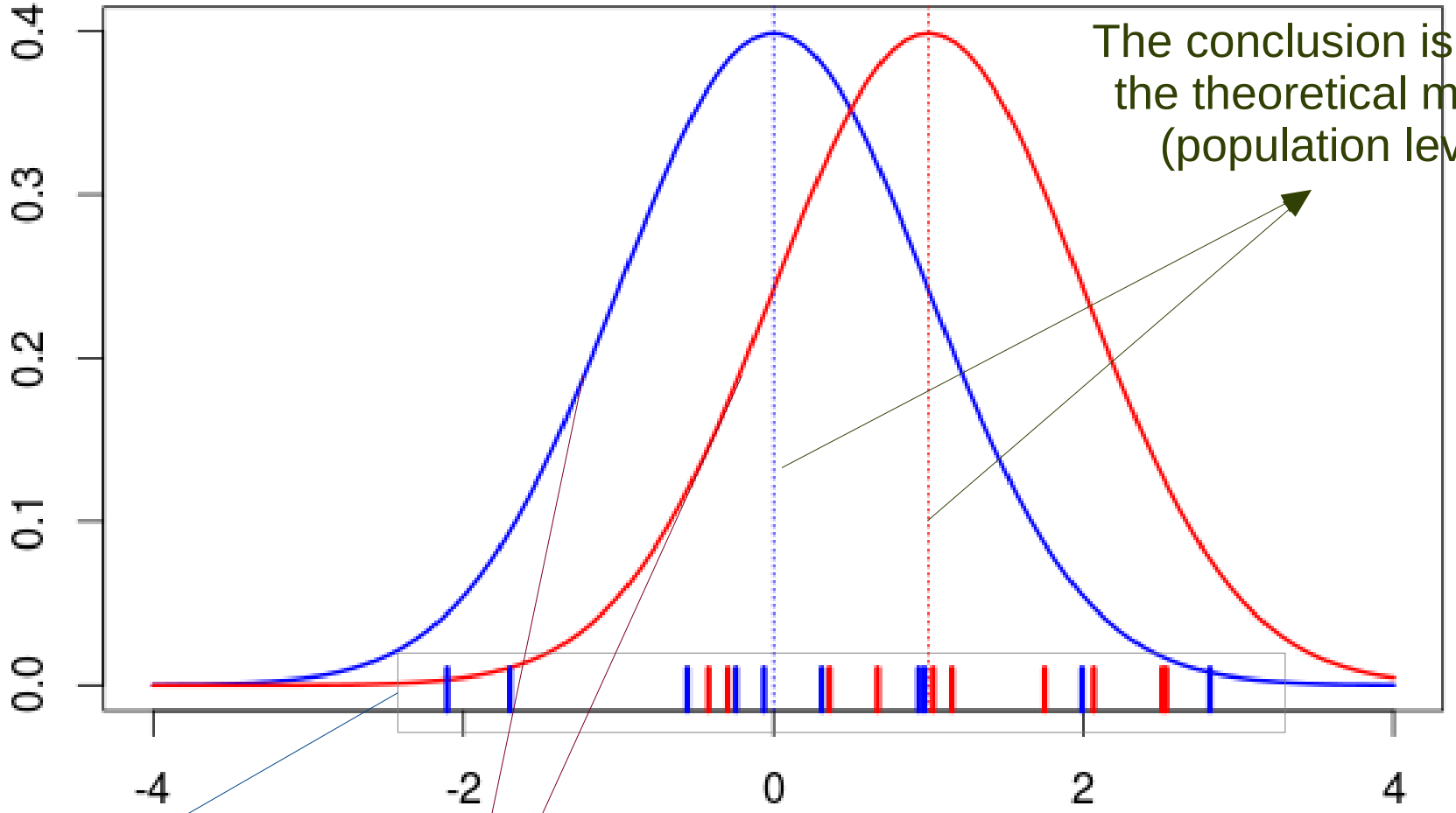


- **Confidence interval**: calculate, from a sample, an interval in which the mean (for instance), of the population should be (with a confidence level).

- **Statistical test**: does what is observed on a sample make it possible to invalidate an hypothesis made on the population?

Statistical hypothesis testing

Example: Student's test to compare 2 means



The conclusion is about the theoretical means (population level)

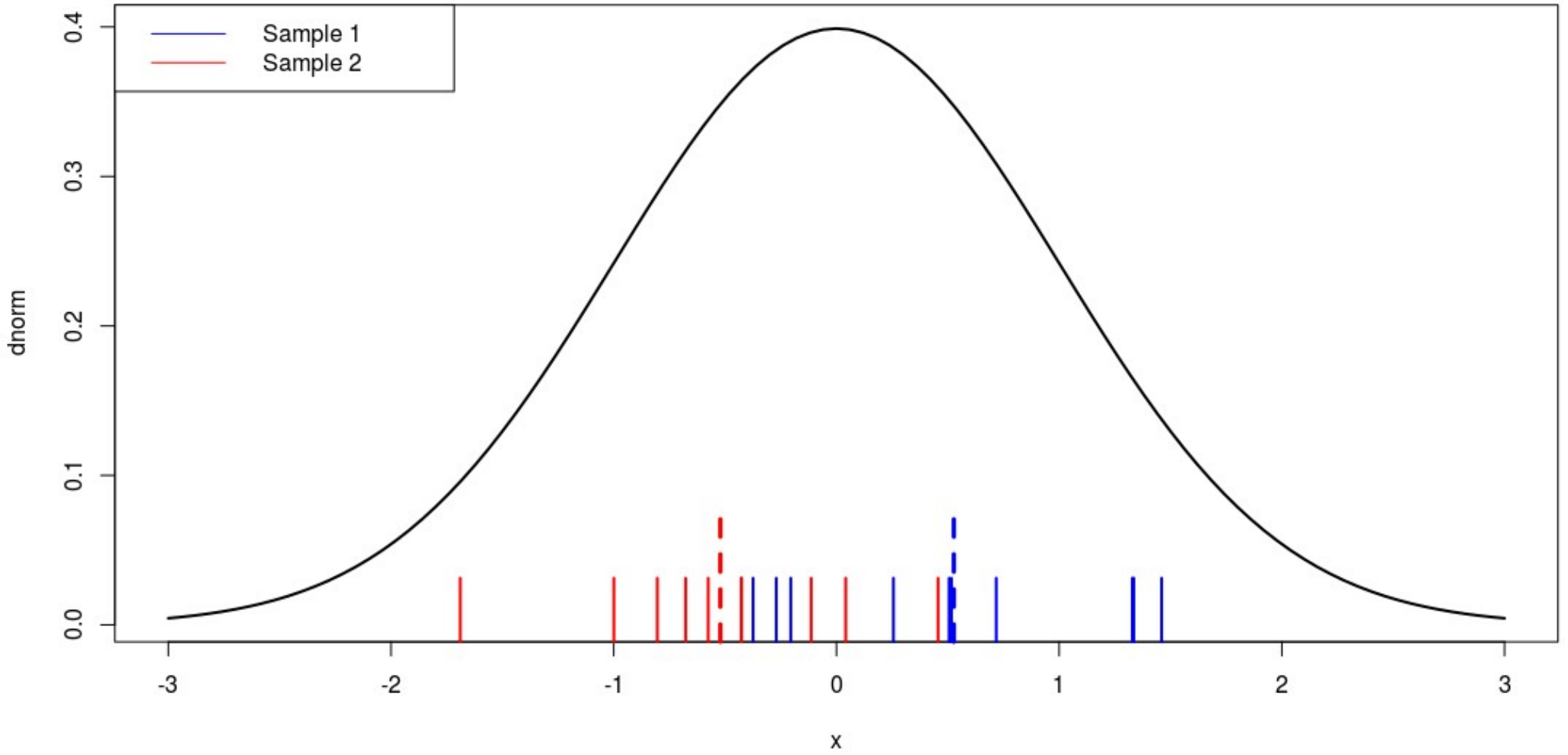
Observations

Assumptions necessary to perform the test

False positive?

2 independent samples from **the same population** can lead to a wrong conclusion.

```
        Welch Two Sample t-test
data:  matrice[indice, 1:10] and matrice[indice, 11:20]
t = 3.6523, df = 17.61, p-value = 0.001878
alternative hypothesis: true difference in means is not equal to 0
```



P-value

*

p-value

“We teach it because it’s what we do;
we do it because it’s what we teach.”

Q: Why do so many colleges and grad schools teach p = 0.05?

A: Because that’s still what the scientific community and journal editors use.

Q: Why do so many people still use p = 0.05?

A: Because that’s what they were taught in college or grad school.

George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College

p-value

An unhealthy obsession with p-values is ruining science

http://www.vox.com/2016/3/15/11225162/p-value-simple-definition-hacking

"The proportion of papers that use p-values is going up over time, and the most significant results have become even more significant over time."

John Ioannidis

Though statisticians have long been pointing out problems with "significance doping" and "P-dolatory" (the "worship of false significance") ***journals have increasingly relied on p-values to determine whether a study should be published.***

"It's this number that looks like you could use it to make a decision that might otherwise be difficult to make or require a whole lot more effort to make,"

“The p-value was never intended to be a substitute for scientific reasoning,”

Ron Wasserstein, Executive director of the American Statistical Association

p-value

Good luck trying to find a really clear definition of a p-value.

Not Even Scientists Can Easily Explain P-values
`http://fivethirtyeight.com/features/not-even-scientists-can-easily-explain-p-values/`

The ASA's Statement on p-Values: Context, Process, and Purpose
`http://amstat.tandfonline.com/doi/pdf/10.1080/00031305.2016.1154108`

"Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data (for example, the sample mean difference between two compared groups) would be equal to or more extreme than its observed value."

I called Rebecca Goldin, the director for Stats.org and a professor at George Mason University, for help parsing that still ***perplexing definition****.*

Digression

p-value ⟷ sondage

recherche scientifique ⟷ Journalisme politique

Le Parisien abandonne les sondages politiques pendant la campagne
http://www.leparisien.fr/flash-actualite-politique/le-parisien-abandonne-les-sondages-politiques-pendant-la-campagne-03-01-2017-6520437.php

Article qui pourrait être ré-intitulé d'après le premier article cité en "*Une obsession malsaine pour les sondages ruine le journalisme (politique)*" ou d'après le second "*Même les sondeurs (ceux que l'on voit à la télé, pas les "vrais") ne savent pas expliquer facilement les résultats d'un sondage*"

Digression

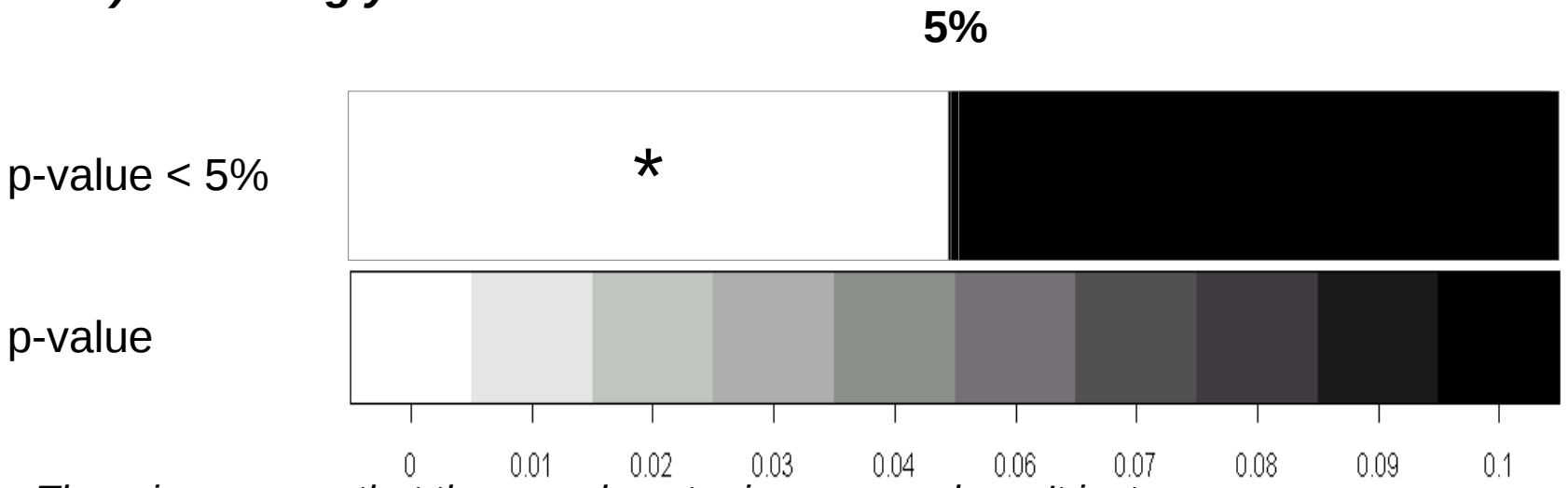
Le Parisien abandonne les sondages politiques pendant la campagne

*Le directeur des rédactions du Parisien/Aujourd'hui en France Stéphane Albouy a annoncé mardi sur France Inter que le quotidien ne commanderait plus de sondages politiques, une "pause" pendant la campagne pour "**se concentrer sur le journalisme de terrain**".*
"C'est une réflexion qu'on a mené depuis quelques temps déjà, notamment après le Brexit et l'élection de Donald Trump", explique-t-il à l'AFP, ajoutant que le journal ne commandait plus de sondages depuis plusieurs semaines déjà.
"Ce n'est pas une question de défiance envers les sondeurs mais une façon de travailler différemment que nous voulons tester pour la suite de la campagne", poursuit-il.
*Il souhaite notamment éviter "ce côté course de petits chevaux où on se focalise sur qui prend la première position" afin de "**se concentrer sur le fond, sur les programmes**".*
Il ne s'interdit pas toutefois de commenter les sondages commandés par d'autres médias.
Consommateur de sondages, le titre y consacre "quelques dizaines de milliers d'euros par an", selon Stéphane Albouy, qui insiste sur le fait qu'il ne s'agit pas avec cette "pause" de réaliser des économies.
*"On peut entendre les critiques qui nous sont faites, à nous, médias, d'être coupés d'une forme de réalité. **Nous allons privilégier le terrain**", explique-t-il, rappelant que le journal s'appuie sur un réseau de 140 journalistes déployés en Ile-de-France.*
*"**Déployer ces journalistes sur le terrain, cela coûte plus cher que les sondages, et nous oblige aussi à être plus exigeants**", estime-t-il.*

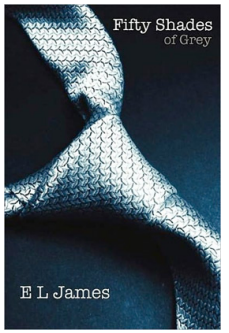
P-value and *

A true story, extract from a referee's comment on a submitted article:

Second, the statistical methods performed are confusing and interpretation of significance is improper. Details about the stat's need to be moved to the methods section. Commenting on the level of statistical significance based on the p-value is incorrect. ***A p-value is either less than alpha value (rejecting null hypothesis) or it is not (retaining null hypothesis); a smaller p-value does not indicate that something has greater or stronger significance. Please delete adjectives (i.e. slightly, strongly, etc.) accordingly.***



There is a reason that the speedometer in your car doesn't just read "slow" and "fast" -- Frank Harrell (warning about the use of cutoffs after logistic regression) R-help (February 2011)



Roman de E.L. JAMES sorti en 2011

Album de Jean-Jacques Goldman sorti en 1987

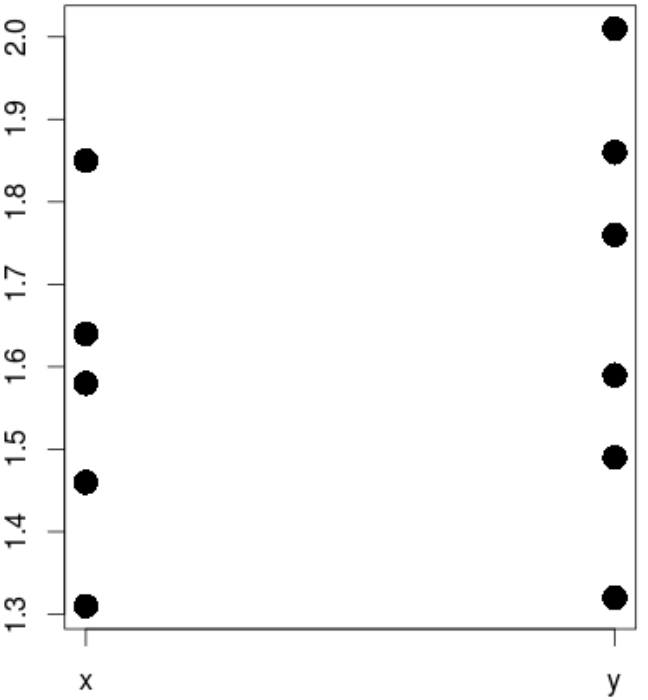
Why to present here the Aphrodite of Milos? Because she is widely known for the mystery of her missing arms and a french expression said « Les bras m'en tombent » to express stupefaction. English would say: *One could have knocked me down with a feather*! or *my mouth was agape* or *I'm gobsmacked...*

Practical aspects

Quizz

- Data

```
X  1.31 1.46 1.85 1.58 1.64
Y  1.49 1.32 2.01 1.59 1.76 1.86
```



- Question : is there a difference between x and y?

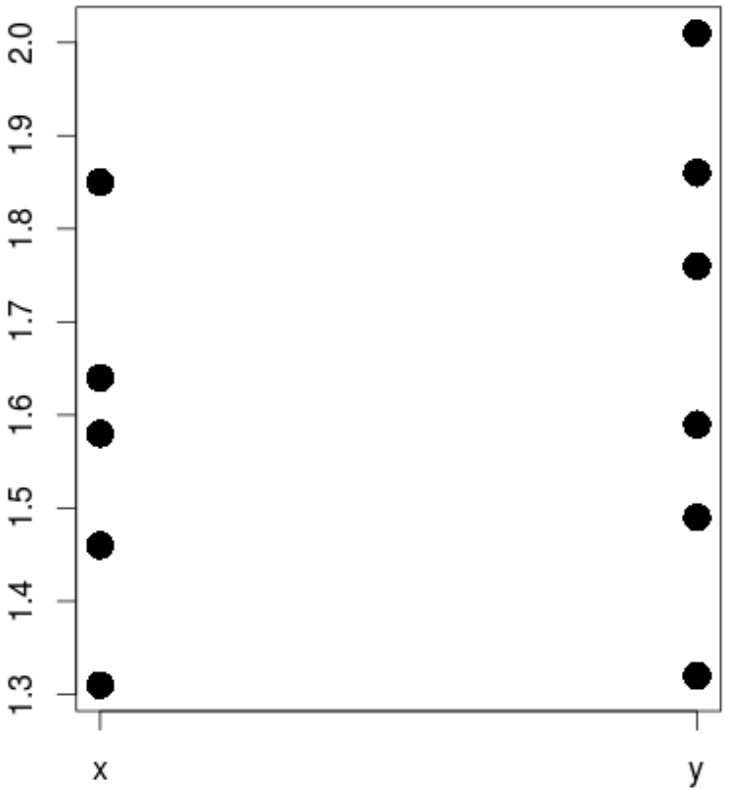
☞ Should I use a paired test?

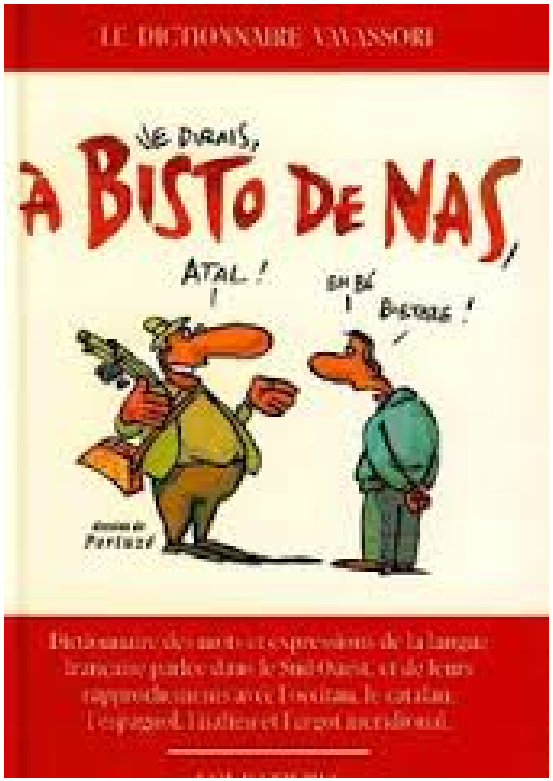
NO, the data are not paired because we don't have the same number of samples

R agrees!

```
x <- c(1.31,1.46,1.85,1.58,1.64)
y <- c(1.49,1.32,2.01,1.59,1.76,1.86)

wilcox.test(x,y, paired=TRUE)
Erreur dans wilcox.test.default(x, y,
paired = TRUE) :
  'x' et 'y' doivent avoir la même longueur
t.test(x,y, paired=TRUE)

t.test(x,y, paired=TRUE)
Erreur dans complete.cases(x, y) :
  les arguments n'ont pas tous la même
taille
```

Any feeling with the result?





☞ could you give an estimation « au pif (*) » of the p-value?

(*) au pif = a bisto de nas : adv. approximately ; roughly ; off the top of one's head ; roughly estimating ; at a ballpark estimate (US)

Let's test now!

☞ Yes, but, can I use the t-test?

YES NO

Do not use any test!

Or use Wilcoxon test

And the result is...

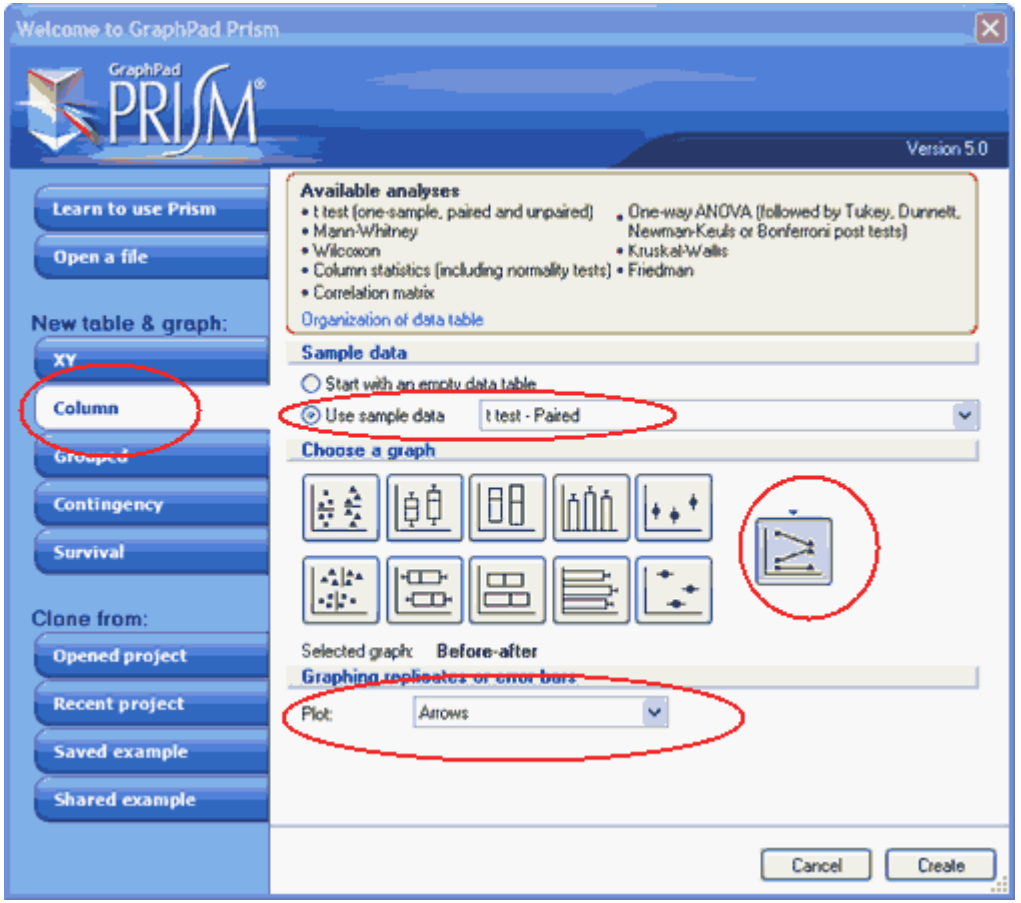
```
> wilcox.test(x,y)
Wilcoxon rank sum test
data:  x and y
W = 10, p-value = 0.4286
alternative hypothesis: true location shift is not equal to 0
```

- The 2 tests agree... cool! And what if they don't?
- Find why...

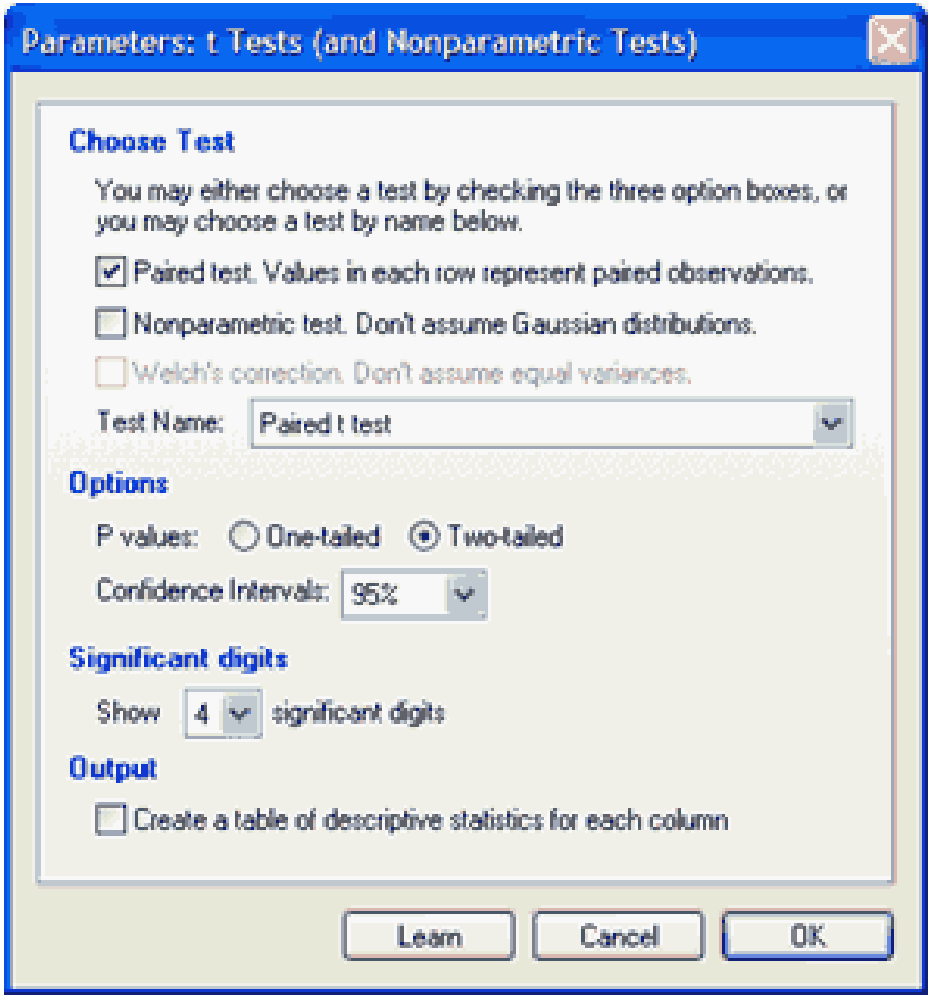
```
> t.test(x,y, var.equal=TRUE)
Two Sample t-test
data:  x and y
t = -0.7381, df = 9, p-value = 0.4792
alternative hypothesis: true difference in means is not equal to 0 95 percent
confidence interval:
 -0.4213783  0.2140450
sample estimates:
mean of x mean of y
 1.568000  1.671667
```

http://www.graphpad.com/guides/prism/5/user-guide/

- Comparing two groups (t tests ...)
  - Key concepts: t tests and related nonparametric tests
    - Q&A: Entering t test data
    - Choosing a t test
    - Q&A: Choosing a test to compare two groups
  - Unpaired t test
    - How to: Unpaired t test from raw data
    - How to: Unpaired t test from averaged data
    - Interpreting results: Unpaired t
    - Graphing tips: Unpaired t
    - Advice: Don't pay much attention to whether error bars overlap
    - Analysis checklist: Unpaired t test
  - Paired t test
    - How to: Paired t test
    - Testing if pairs follow a Gaussian distribution
    - Interpreting results: Paired t
    - Graphing tips: Paired t
    - An alternative to paired t test: Ratio t test
    - Analysis checklist: Paired t test
  - Mann-Whitney test
    - How to: Mann-Whitney test
    - Interpreting results: Mann-Whitney test
    - The Mann-Whitney test doesn't really compare medians
    - Analysis checklist: Mann-Whitney test
  - Wilcoxon matched pairs test
    - How to: Wilcoxon matched pairs test
    - Results: Wilcoxon matched pairs test
    - Analysis checklist: Wilcoxon matched pairs test



GraphPad Prism



Un-paired data?

X	18	21	16	22	19	24	17	20	23	12
Y	22	25	17	24	18	29	20	23	21	16

☞ Should I use a paired test?

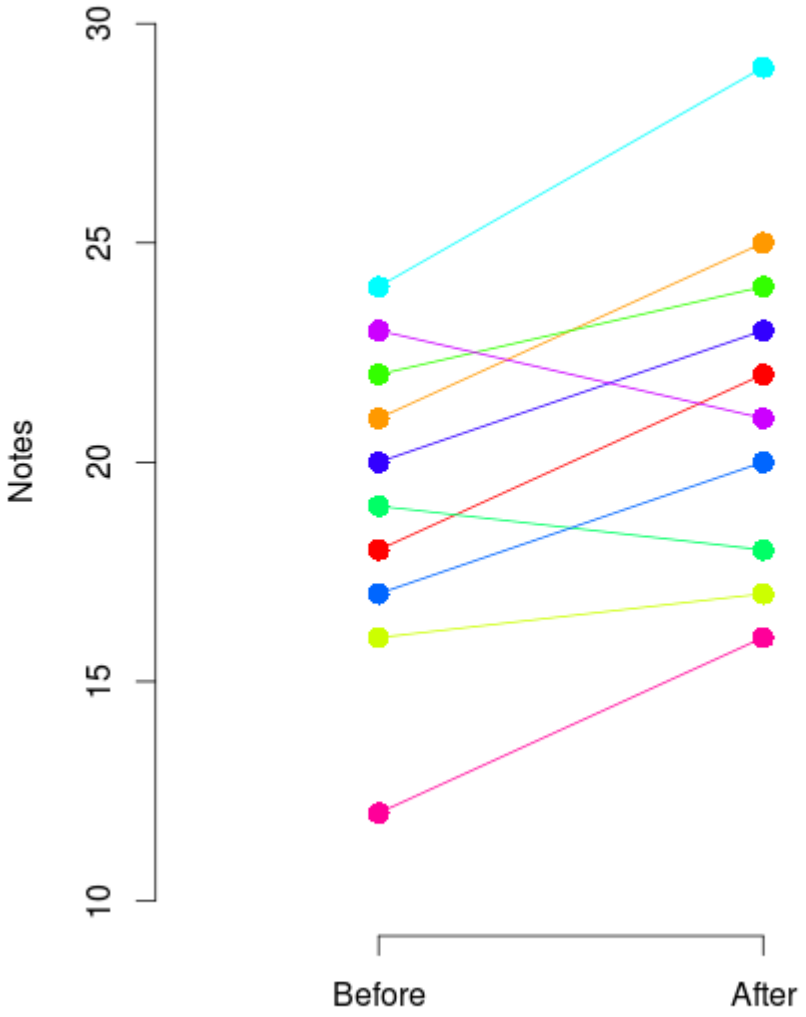
?

Equal size samples is a necessary but not […] condition!

Replace […] with: sufficient

Paired-data!

	Before	After
Louise	18	22
Léo	21	25
Emma	16	17
Gabriel	22	24
Chloé	19	18
Adam	24	29
Lola	17	20
Timéo	20	23
Inès	23	21
Raphaël	12	16



Results

```
> wilcox.test(x,y, paired=TRUE)
Wilcoxon signed rank test with continuity correction

data:  x and y
V = 5, p-value = 0.02428
alternative hypothesis: true location shift is not equal to 0
```

```
> t.test(x,y, paired=TRUE)
Paired t-test

data:  x and y
t = -3.1461, df = 9, p-value = 0.01181
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.953766 -0.646234
sample estimates:
mean of the differences
                  -2.3
```

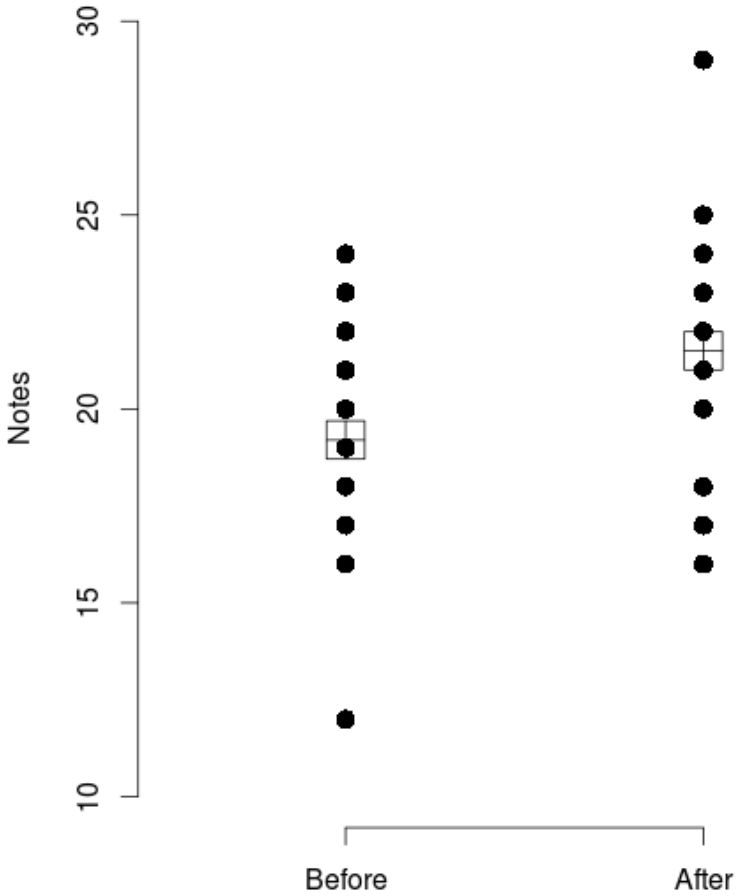
And if I am wrong?

> wilcox.test(x,y, **paired=FALSE**)
Wilcoxon rank sum test with continuity correction

data: x and y
W = 35, p-value = **0.2716**
alternative hypothesis: true location shift is not equal to 0

> t.test(x,y, **paired=FALSE**)
Two Sample t-test

data: x and y
t = -1.3529, df = 18, p-value = **0.1928**
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.871567 1.271567
sample estimates:
mean of x mean of y
     19.2   21.5



To put it in a nutshell (*)

- Paired / independant data: not to be mistaken !

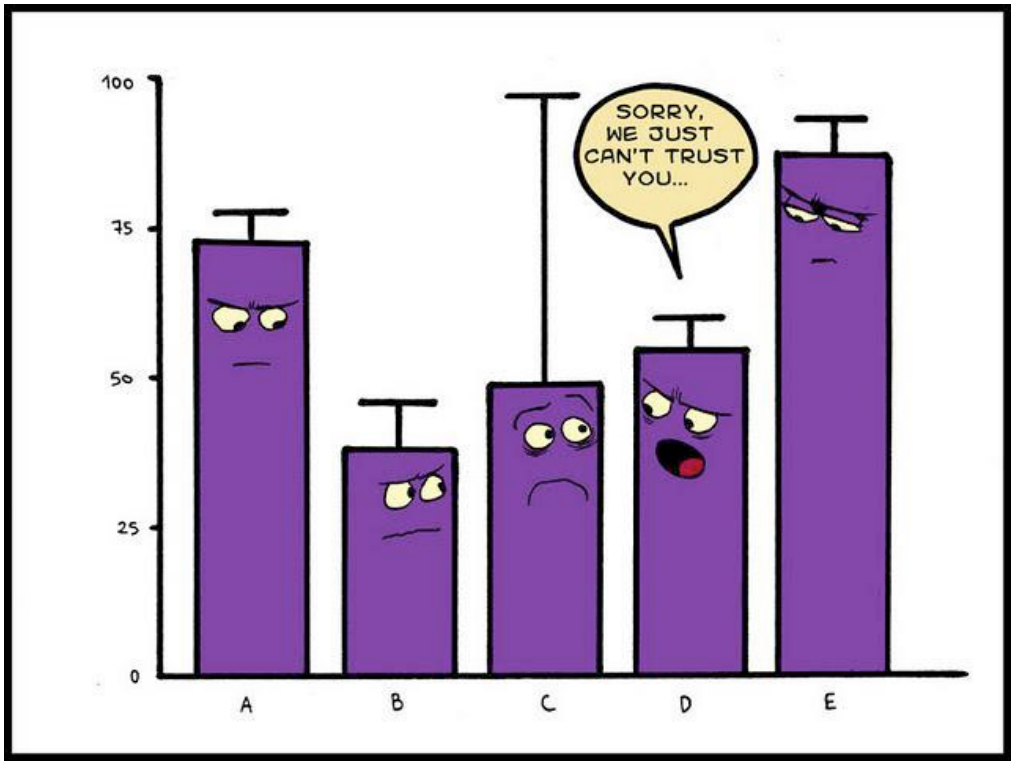
- Student, Wilcoxon: do what the referee wants!

(*) *To put it in a nutshell* : bref

However...

However, if n is very small (for example n = 3), rather than showing error bars and statistics, it is better to simply plot the individual data points.

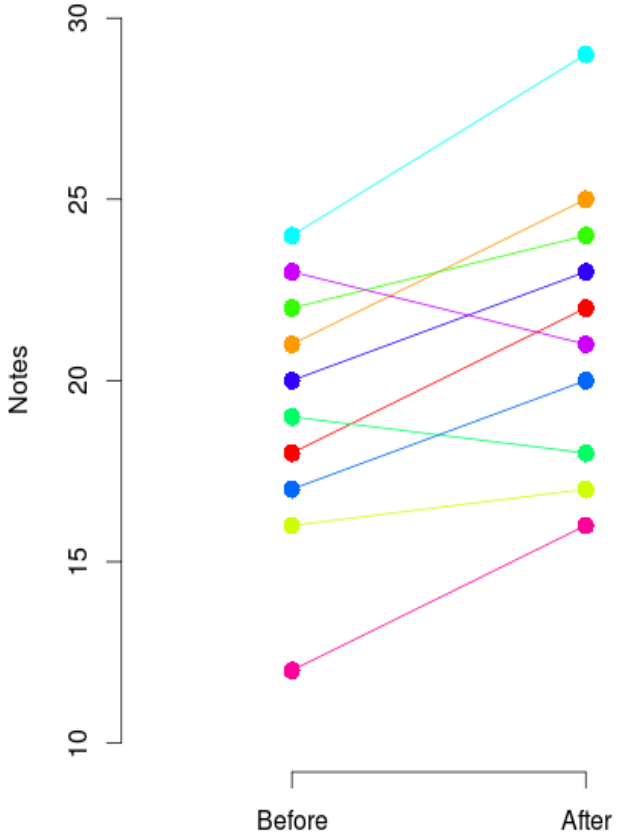
Cumming, G., Fidler, F., & Vaux, D. L. (2007).
Error bars in experimental biology.
The Journal of Cell Biology, 177(1), 7–11.

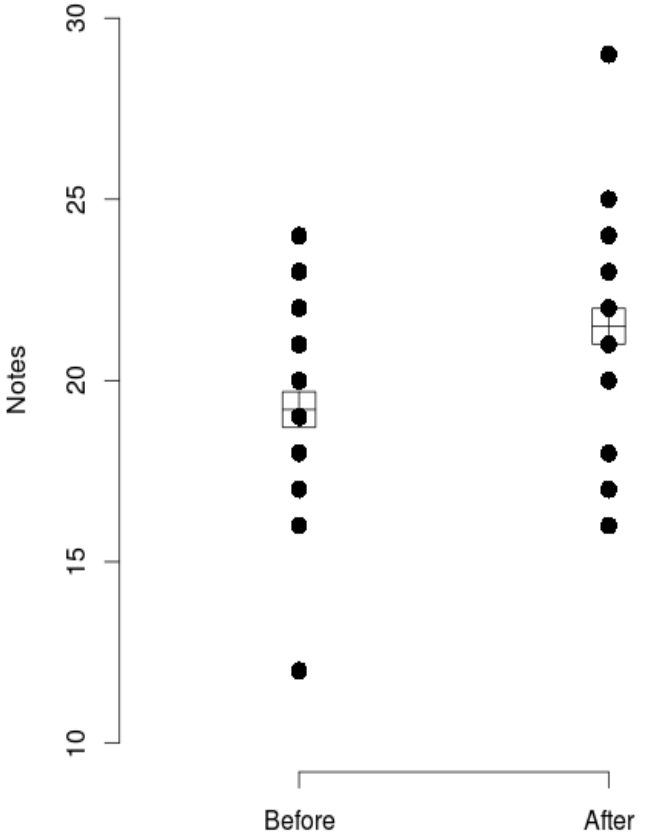


LES STATISTIQUES
C'EST PAS AUTOMATIQUE

Parlez-en à un-e statisticien-ne

Visualisation





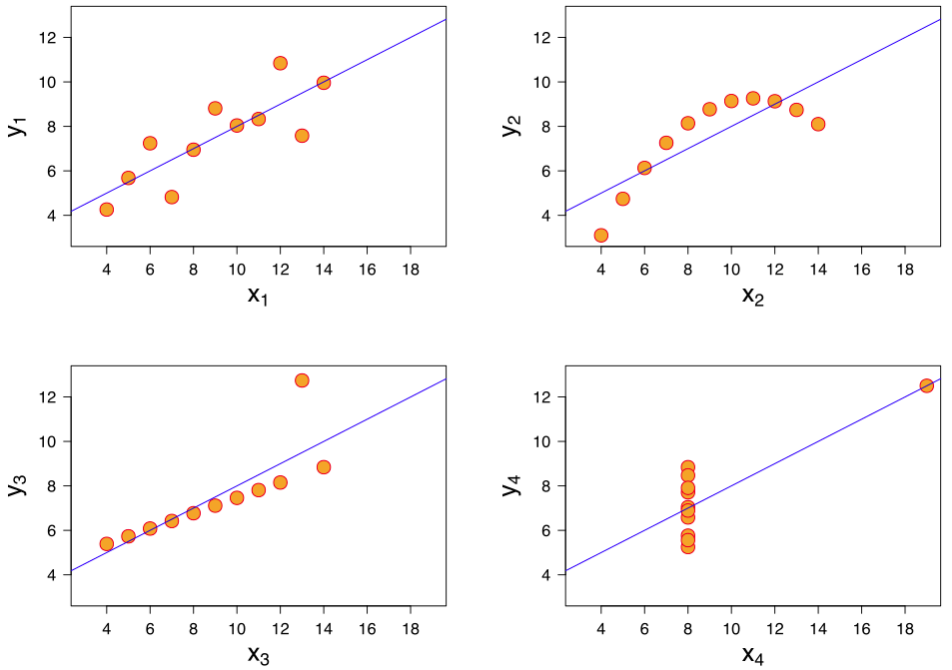
Why to visualise?

...make both calculations and graphs. Both sorts of output should be studied; each will contribute to understanding.

F. J. Anscombe, 1973

Anscombe's quartet

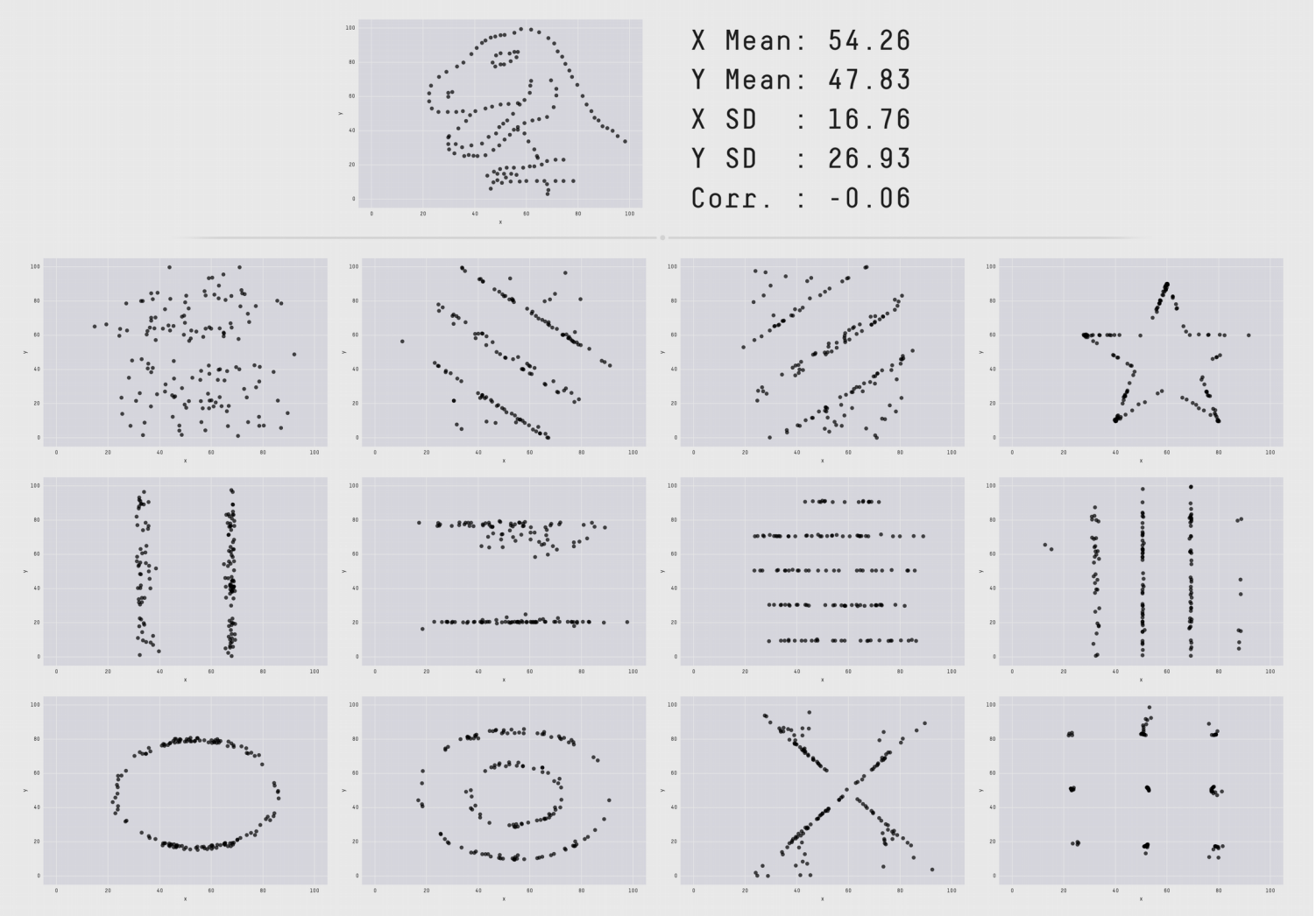
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



In the 4 cases :

	X	Y
Moyenne	9	7.5
Variance	11	4.125
Corrélation	0.816	

Same stats, different graphs...



Matejka, J., & Fitzmaurice, G. (2017, May). **Same stats, different graphs**: Generating datasets with varied appearance and identical statistics through simulated annealing. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (pp. 1290-1294). ACM.

www.autodeskresearch.com/publications/samestats

Visualisation and test!

Is there a difference between WT and Mut regarding the variable Vx?

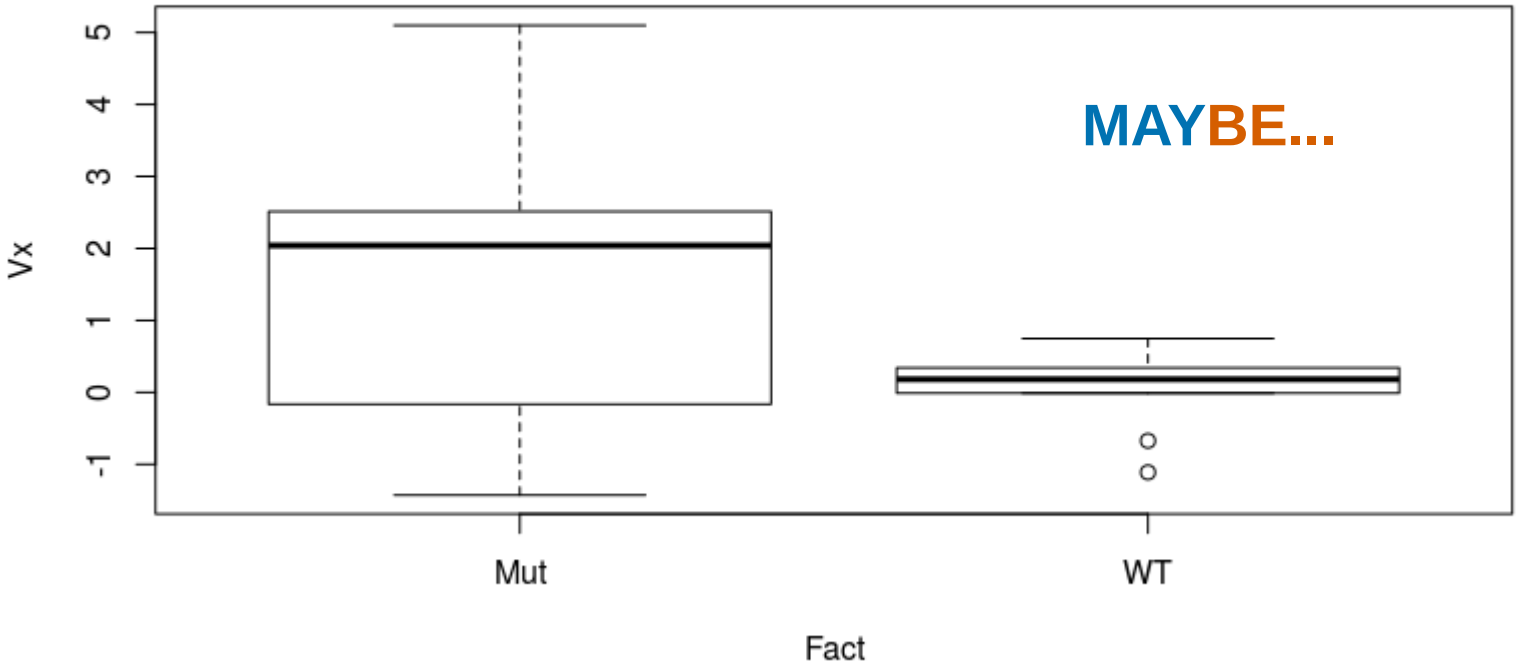
	Vx	Fact
1	-1.11	WT
2	-0.01	WT
3	0.20	WT
4	-0.67	WT
5	0.05	WT
6	0.17	WT
7	0.34	WT
8	0.24	WT
9	0.54	WT
10	0.75	WT
11	2.51	Mut
12	-0.43	Mut
13	2.09	Mut
14	2.21	Mut
15	4.36	Mut
16	-0.17	Mut
17	-1.43	Mut
18	1.99	Mut
19	0.50	Mut
20	5.10	Mut

```
> t.test(Vx~Fact)
        Welch Two Sample t-test
data:  Vx by Fact
t = 2.3854, df = 10.269, p-value = 0.03765
```

YES!

```
> wilcox.test(Vx~Fact)
    Wilcoxon rank sum test
data:  Vx by Fact
W = 72, p-value = 0.1051
```

NO!?



Visualisation and test!

Data: 1 factor, 2 quantitative variables

	factor	Vx	Vy
1	WT	2.0	2.00
2	Mut	3.0	2.50
3	WT	4.5	3.50
4	Mut	5.0	3.25
5	Mut	5.5	3.30
6	WT	6.0	4.30
7	Mut	7.0	4.20
8	WT	8.0	5.10
9	Mut	8.5	4.80
10	Mut	9.0	5.00
11	WT	10.0	6.00
12	WT	11.0	6.50

Does the factor influence Vx and Vy?

Vx

```
> t.test(Vx~fact)
        Welch Two Sample t-test
data:  Vx by fact
t = -0.34852, df = 8.7078, p-value = 0.7357
```

```
> wilcox.test(Vx~fact)
```

NO

```
    Wilcoxon rank sum test

data:  Vx by fact
W = 16, p-value = 0.8182
```

NO

Vy

```
> t.test(Vy~fact)

    Welch Two Sample t-test
```

NO

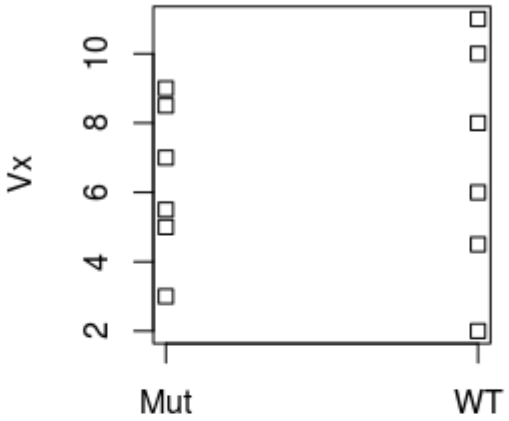
```
data:  Vy by fact
t = -0.91815, df = 8.1062, p-value = 0.385
```

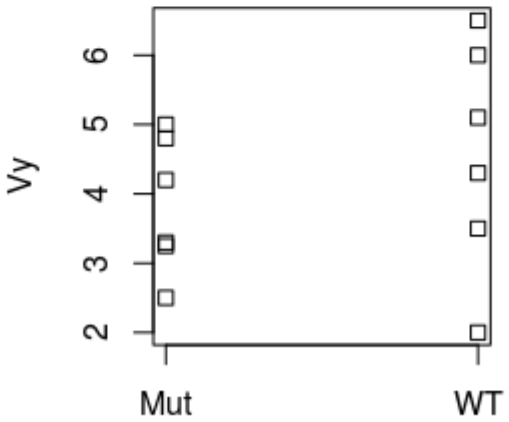
```
> wilcox.test(Vy~fact)

    Wilcoxon rank sum test

data:  Vy by fact
W = 11, p-value = 0.3095
```

NO

Visualisation and test!

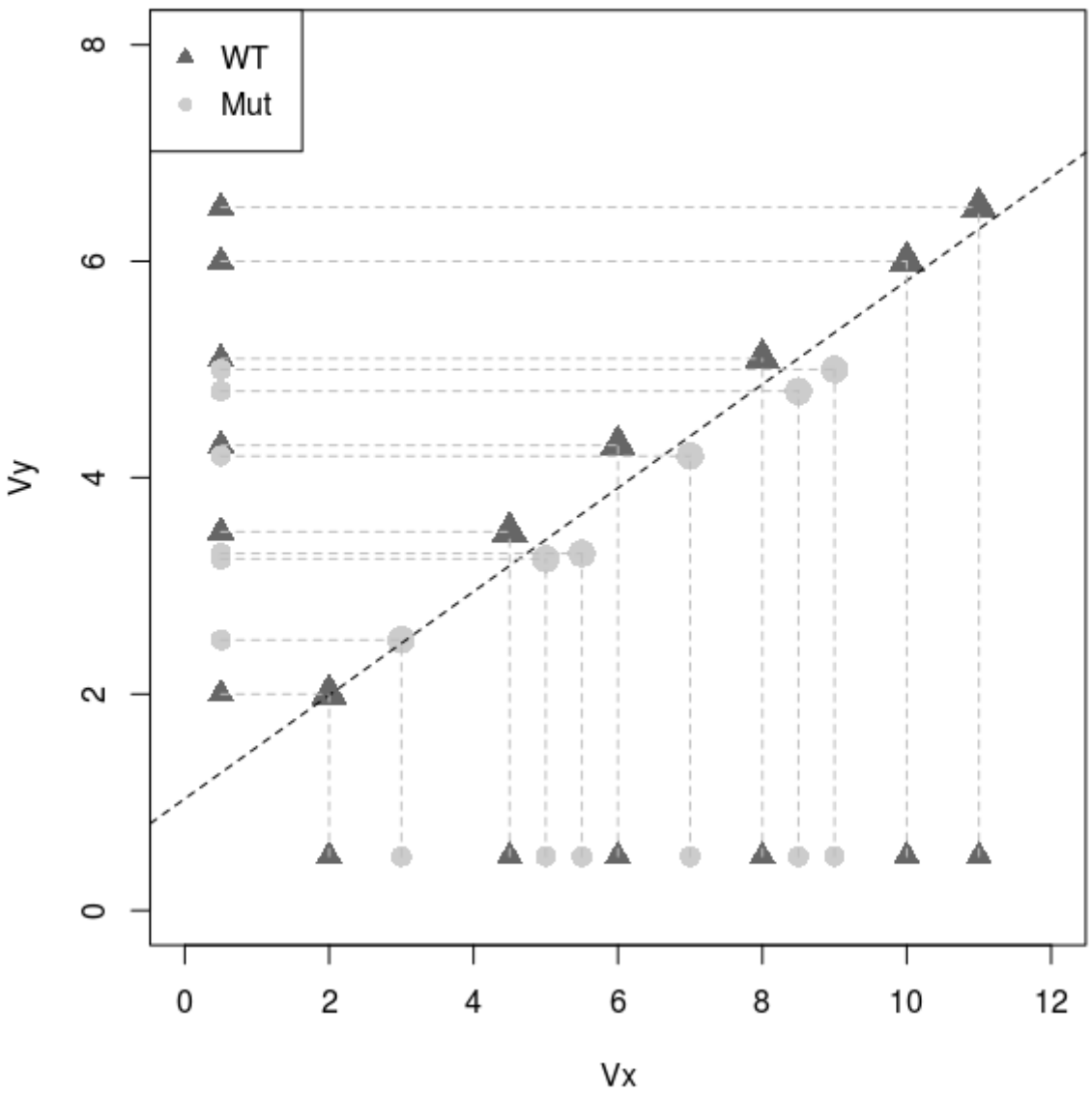
Stripcharts agree





	factor	Vx	Vy
1	WT	2.0	2.00
2	Mut	3.0	2.50
3	WT	4.5	3.50
4	Mut	5.0	3.25
5	Mut	5.5	3.30
6	WT	6.0	4.30
7	Mut	7.0	4.20
8	WT	8.0	5.10
9	Mut	8.5	4.80
10	Mut	9.0	5.00
11	WT	10.0	6.00
12	WT	11.0	6.50

But, what about a 2D scatterplot?



Two factor ANOVA

Two factor ANOVA

Id	genotype	treatment	X1	X2	X3	X4
1	WT	CTRL	10.4	10.4	10.1	10.1
2	WT	CTRL	10.5	10.5	10.2	10.2
3	WT	CTRL	9.6	9.6	9.8	9.8
4	WT	CTRL	9.5	9.5	9.9	9.9
5	WT	CTRL	10.0	10.0	10.0	10.0
6	WT	Treat	6.4	6.4	5.1	8.1
7	WT	Treat	6.5	6.5	5.2	8.2
8	WT	Treat	5.6	5.6	4.8	7.8
9	WT	Treat	5.8	5.8	4.9	8.9
10	WT	Treat	6.0	6.0	5.0	8.0
11	Mut	CTRL	12.1	10.3	5.1	5.1
12	Mut	CTRL	12.2	10.6	5.2	5.2
13	Mut	CTRL	11.8	9.7	4.8	4.8
14	Mut	CTRL	11.9	9.4	4.9	4.9
15	Mut	CTRL	12.0	10.0	5.0	5.0
16	Mut	Treat	8.1	6.3	10.1	10.1
17	Mut	Treat	8.2	6.6	10.2	10.2
18	Mut	Treat	7.8	5.5	9.8	9.8
19	Mut	Treat	7.9	5.9	9.9	9.9
20	Mut	Treat	8.0	6.0	10.0	10.0

ANOVA table

X1

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
genotype	1	19.40	19.4	192.600	2.44e-10	***
treatment	1	78.80	78.8	782.179	5.17e-15	***
genotype:treatment	1	0.00	0.0	0.045	0.835	

X2

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
genotype	1	0.00	0.00	0.0	1	
treatment	1	77.62	77.62	413.4	7.42e-13	***
genotype:treatment	1	0.00	0.00	0.0	1	

X3

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
genotype	1	0.0	0.00	0	1	
treatment	1	0.0	0.00	0	1	
genotype:treatment	1	125.0	125.00	5000	<2e-16	***

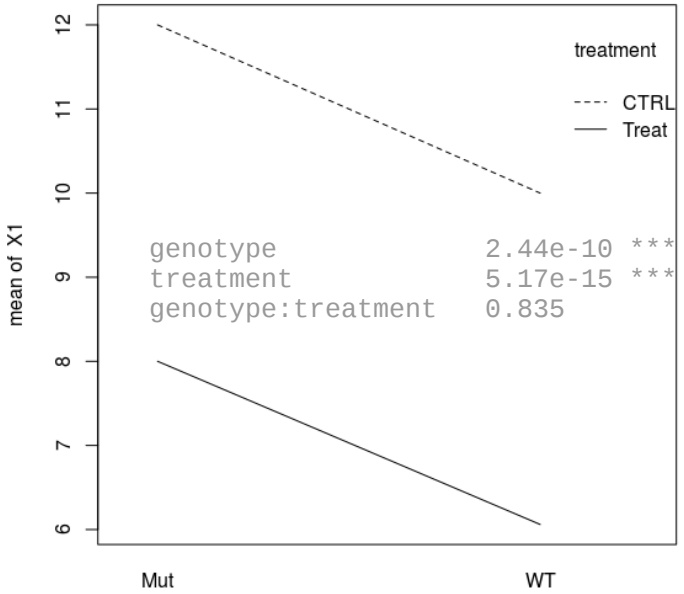
X4

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
genotype	1	12.8	12.80	204.8	1.54e-10	***
treatment	1	12.8	12.80	204.8	1.54e-10	***
genotype:treatment	1	57.8	57.80	924.8	1.38e-15	***

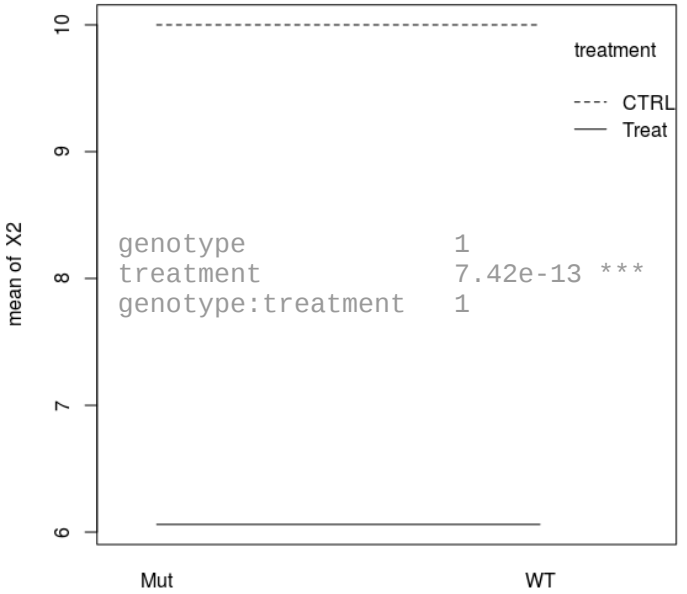
Interaction plot

psychstat3.missouristate.edu/Documents/MultiBook3/Mlt08.htm

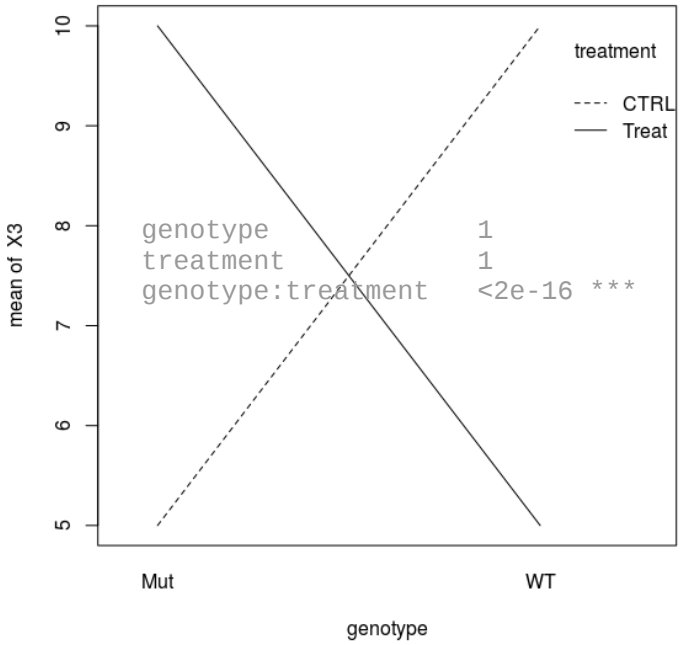
X1



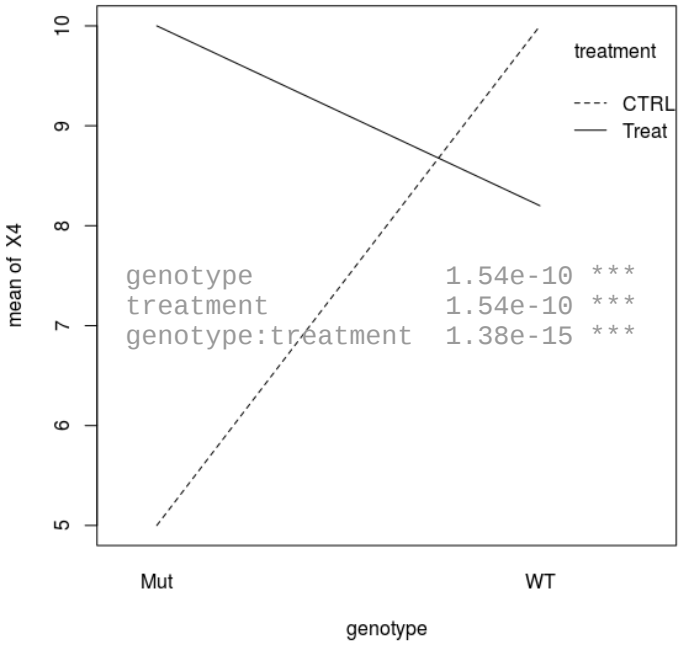
X2



X3



X4



Conclusion

en.wikipedia.org/wiki/Blind_men_and_an_elephant