Things you were never told, did not understand, forgot, or chose to ignore in statistics

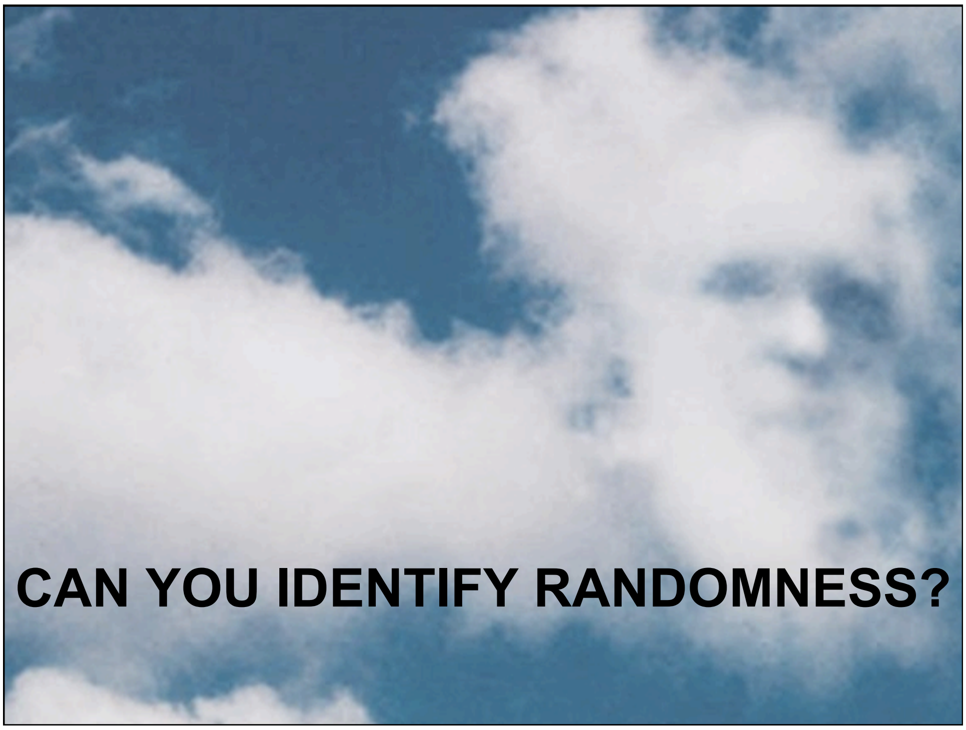
(Errors I'v made and would like you to avoid)

Keynote at 1st Conference on Advanced Probability and Statistics in Information Systems (APSIS) 2014

Magne Jørgensen
Simula Research Laboratory

The presentation is based on the following papers:

- M. Jørgensen. *The influence of selection bias on effort overruns in software development projects*, Information and Software Technology 55(9):1640-1650, 2013.
- M. Jørgensen and B. Kitchenham. *Interpretation problems related to the use of regression models to decide on economy of scale in software development*, Journal of Systems and Software, 85(11):2494-2503, 2012.
- M. Jørgensen, T. Halkjelsvik, and B. Kitchenham. *How does project size affect cost estimation error? Statistical artifacts and methodological challenges*, International Journal of Project Management, 30(7):751-862, 2012.
- M. Jørgensen, T. Dybå, D. I. K. Sjøberg, K. Liestøl. *Incorrect results in software engineering experiments. How to improve research practices.* Submitted to a journal.
- M. Jørgensen. *Fallacies and biases when adding effort estimates*, To be presented at Euromicro/SEEA, 2014.



Throw	Seq 1	Seq 2	Seq 3
1	#	O	O
2	#	#	#
3	O	O	O
4	O	#	O
5	O	#	#
6	O	O	#
7	O	O	#
8	#	O	#
9	#	O	O
10	O	#	#
11	#	O	#
18	#	#	O

Basketball or coin?

Seq. 1: 70% likely to keep previous.
*This is what most believe is the basketball player (hot hand illusion), but it is **not**.*

Seq. 3: 70% likely to change from previous.
*This is what most believe is the coin, but it is **not**. It is **not** the basketball player either.*

Seq. 2: Random sequence **and** basketball player
But, does Seq. 2 look random? Too many clusters!

My first mistake in using statistics ....

I measured an **increase** in productivity of an IT-department (function points/man-month). The management was happy, since this proved that their newly implemented processes had been successful.

Later, to my surprise, when I grouped the project into those using 4GL, those using 3GL I found a productivity **decrease** in both groups.

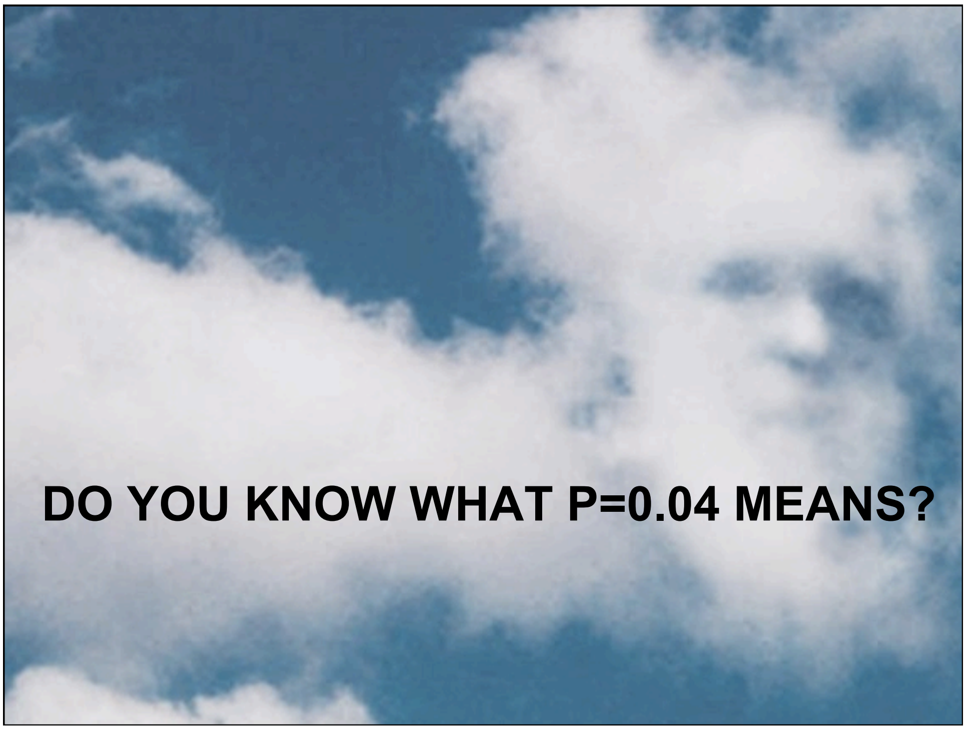
Was my analysis incorrect?

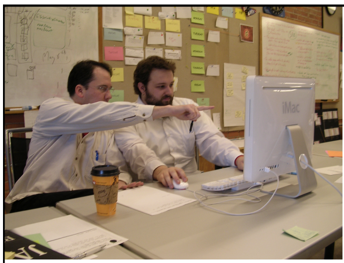
Missing variable

The increase in total productivity was caused by more and more of the work done using the higher productivity environment 4GL

All teams had decreased their productivity, but the higher productivity teams had done more of the work.

The challenge is to know whether there is a missing variable in your analysis ...

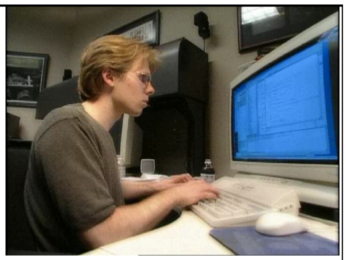




Pair vs. solo

H_0: Solo better or same

H_1: Pairs better



Result: Pairs 20% less errors than Solo, **p=0.04**

Which of the following interpretations/consequences of p=0.04 are correct (assume significance level of 0.05)?

- It is less than 5% likely that the null hypothesis (H_0) is true.
- We can accept the alternative hypothesis (H_1) with at least 95% confidence.
- An identical replication is at least 95% likely to find a significant difference. (Repeating the study 100 times, would find a statistically significant difference in the same direction about 95 times)

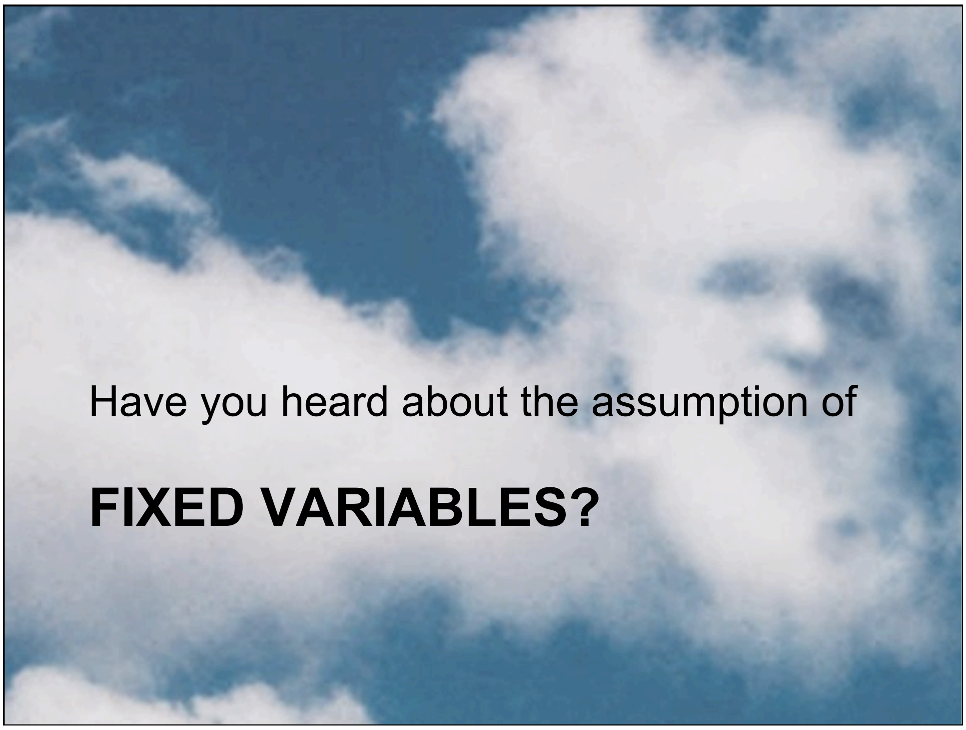
p-values are complex, unreliable values that do not answer what we should be asking about …

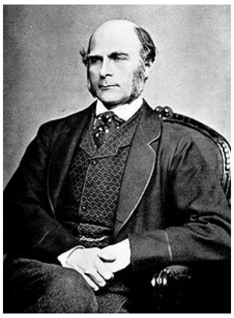
- *A p-value is **not** the probability of the hypothesis or a theory being true or false! A p-value of 0.05 may easily correspond to $p(H_0) > 20\%$.*
- *A p-value of 0.01-0.05 gives the impression of strong evidence. It is **not**!*
- *A p-value does **not** say much about how likely it is to replicate the study and find that $p<0.05$.*
- *Even with $p<0.05$, the null hypothesis may be more likely than the alternative hypothesis.*
- *The p-value examines a "yes/no" situation, while we in most cases would like to know about the effect size and its uncertainty.*

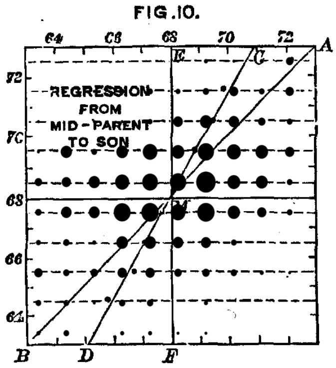
A p-value is the probability of observing the data (or more extreme data), given that H_0 is true.
We tend to mix $p(H_0 \mid D)$ with $p(D \mid H_0)$ = p-value.

Recommended reading recommending the use of confidence intervals of effect sizes

Geoff Cumming, *The new statistics: Why and How*, Psychological Science, 2014.







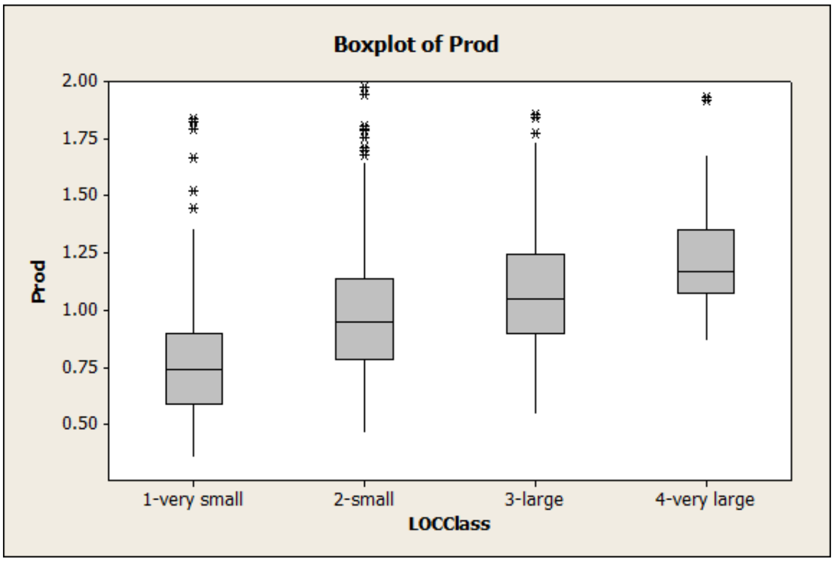
Sir Francis Galton (“*Filial regression to mediocrity*”):

- The father of regression analysis
- The first to violate the fixed variable assumption
- Identified the problem by reversing the regression

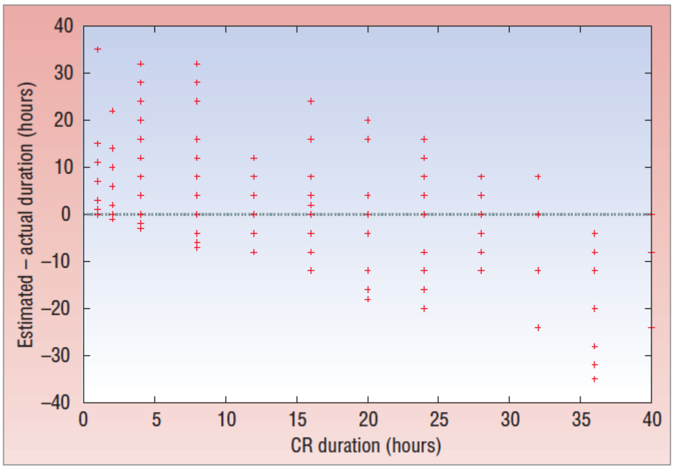
Violation of the fixed variable assumption is a problem even when we do “simple” categorical analyses

- Created a dataset with same “productivity” (lines of code per work-hour) for all “true” project sizes (“true” lines of code)
- Each measurement of lines of code was added some measurement error, e.g., due to forgetting to count lines of code, counting the same code twice, different counting practices)
  - Observed LOC = true LOC + measurement error
- Projects were divided into size groups (very small, small, large, very large) based on their lines of code.
- Do you think the mean productivity of each size category will be the same? (The “true” productivity is size indep.)

What we observe is …



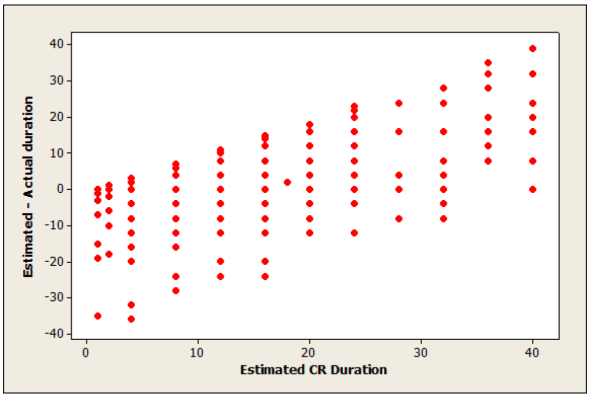
How would you interpret these data?



CR duration = Actual duration (effort) to complete a change request

Interpretation by author of paper: Larger tasks are more under-estimated.

What about these data?



They are from the exact same data set! The only difference is in the use of the **estimated** instead of **actual** duration as the task size variable.

Economy of scale? Probably not ...

(M. Jørgensen and B. Kitchenham. Interpretation problems related to the use of regression models to decide on economy of scale in software development, Journal of Systems and Software, 85(11):2494-2503, 2012.)

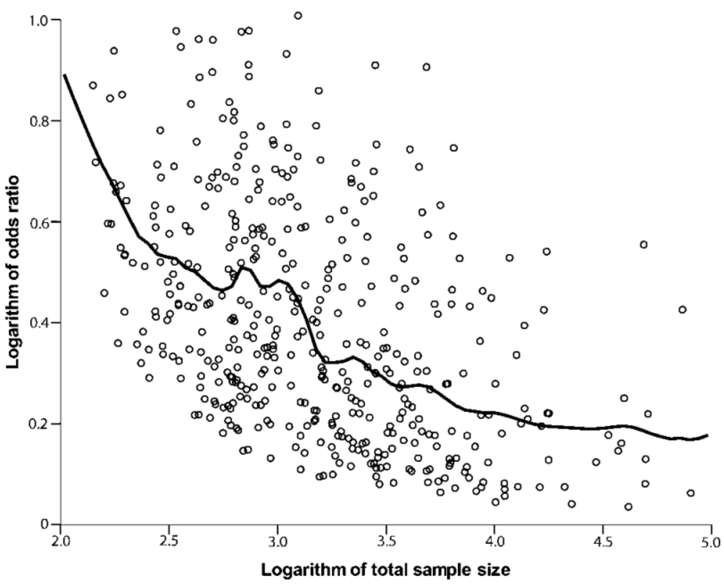
Data set	$Effort = a_1 Size^{b_1}$		$Size = a_2 Effort^{b_2}$		
	b_1	"Return on scale"	b_2	"Return on scale"	r
(Jørgensen 1997)	0.52*	EOS	0.68	Linear	0.54
(Desharnais 1988) All projects	0.94	Linear	0.53*	DOS	0.70
(Desharnais 1988) - Cobol projects	0.98	Linear	0.66*	DOS	0.81
(Desharnais 1988) - Advanced Cobol projects	0.99	Linear	0.85	Linear	0.92
(Desharnais 1988) - 4 GL projects	1.05	Linear	0.72	Linear	0.87
(Kitchenham, Pfleeger et al. 2002)	0.67*	EOS	0.78*	DOS	0.73
(Jørgensen 1995)	0.56*	EOS	0.96	Linear	0.74
"Finnish" data set[2]	0.99	Linear	0.71*	DOS	0.75
(Kemerer 1987)	0.81	Linear	0.76	Linear	0.79
(Kemerer 1987)	0.90	Linear	0.74	Linear	0.82
(Hill, Thomas et al. 2000)	1.02	Linear	0.68*	DOS	0.83
(Boehm 1981)	1.02	Linear	0.76*	DOS	0.86
(Jeffery and Stathis 1996)	0.80	Linear	0.97	Linear	0.88
(Miyazaki, Terakado et al. 1994)	O.99	Linear	0.78*	Strong DOS	0.89

EOS = Economy of scale, DOS = Diseconomy of scale

A side-effect of the quest for p<0.05

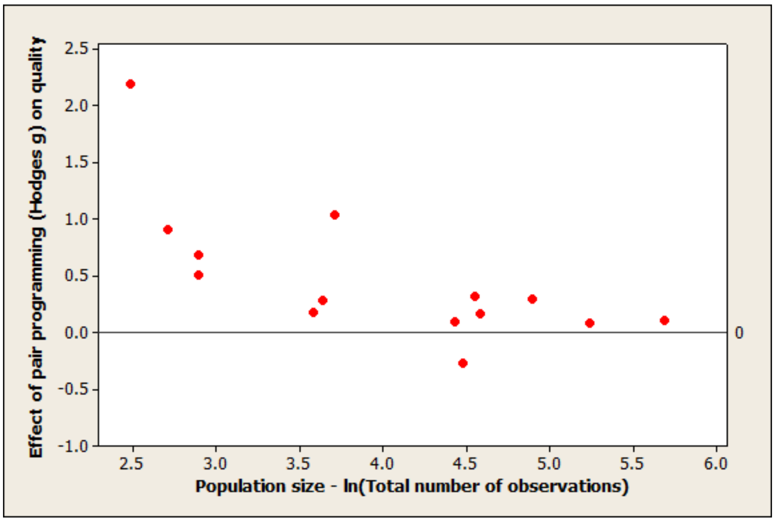
PUBLICATION BIAS GIVES INFLATED EFFECT SIZES

"Why most discovered true associations are inflated", Ioannidis, Epidemiology, Vol 19, No 5, Sept 2008



Effect sizes in studies on pair programming

Source: Hannay, Jo E., et al. "The effectiveness of pair programming: A meta-analysis." Information and Software Technology 51.7 (2009): 1110-1122.



Total publication bias (only statistically significant results are published) implies that published results has ZERO strength!

Illustration: Building a regression model

- **Data set**:
  - Effort-variable + 15 other project variables
  - Twenty software projects.
- **Regression model**:
  - Selected the best 4-variable regression model (OLS), based on ”best subset”.
  - Removed one outlier.
- **Results**:
  - R^2=76%,
  - R^2-adj=70%%,
  - R^2-pred = 56%
  - MdMRE = 28%

Not bad results...

Especially since all data were random numbers between 1 and 10!

Best subset is a rather extreme type of publication bias, but same problem with stepwise regression.

Best 4 out of 15 variable-model, means that we publish only the best out of 1365 tested models!

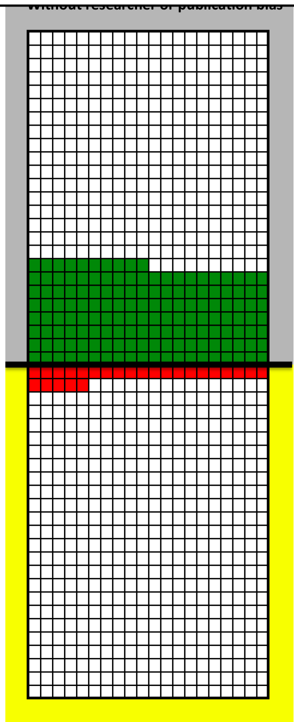
Correct test results: (150 + 475)/1000 = 62.5%

Proportion exp. stat. sign results: (150+25)/1000 = 17.5%

1000 statistical tests

Correct positive tests:

150/(150+25) = 85.7% (prob. of null hyp. being true when p<0.05 is 14.4%, not 5%)



500 true relationships

Statistical power is 30% -> 150 True positive (green)

Significance level is 5% -> 25 False positive (red)

500 false relationships

We observe about 50% p<0.05 in published SE experiments

- We should expect 17.5%
- Maximum 30%, if we only test true relationships
- Researcher and publication bias

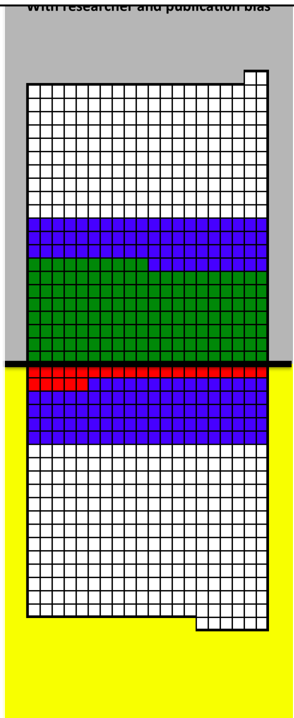
EFFECT OF ADDING 20% RESEARCHER BIAS AND 30% PUBLICATION BIAS

Correct test results: 61% (just above half of the tests)

1000 statistical tests

Correct positive tests: 65%

One third of the reported positive tests are incorrect!



Removes 78 negative tests

Researcher bias is 20% -> 70 more true positive tests (blue)

Statistical power is 30% -> 150 true positive (green)

Significance level is 5% -> 25 false positive (red)

Researcher bias is 20% -> 95 more false positive tests (blue)

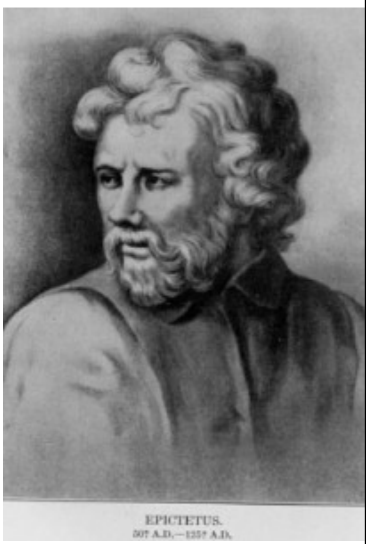
Removes 114 negative tests

LOW PROPORTION CORRECT RESULTS!

WE NEED TO IMPROVE STATISTICAL RESEARCH PRACTICES IN SOFTWARE ENGINEERING!

Last words

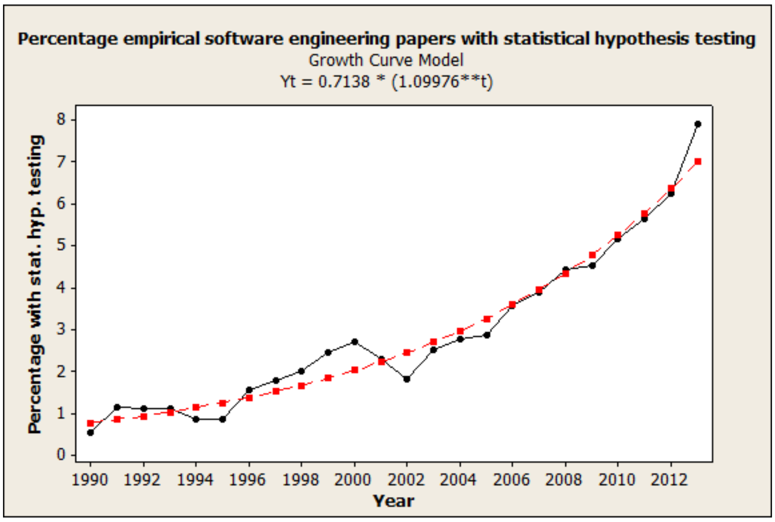
Appearances to the mind are of four kinds. Things either are what they appear to be; or they neither are, nor appear to be; or they are, and do not appear to be; or they are not, and yet appear to be. Rightly to aim in all these cases is the wise man's task.



Epictetus (AD 55-135), Discourses, Book 1, Chapter 27

BONUS MATERIAL

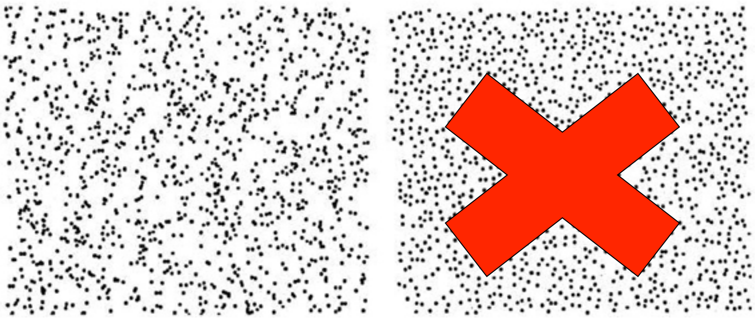
Increasing proportion of statistical hypothesis tests in software engineering papers



... the validity of your results can never be greater than that of the most questionable of your assumptions.

Vardeman & Morris (2003). Statistics and ethics: some advice for young statisticians. Am. Stat. 57, 21.

Random? None, left, right, both?



"... glow worms are gluttonous and inclined to eat anything that comes within snatching distance, so they keep their distance from each other and end up relatively evenly spaced i.e. non-randomly."

(Steven Pinker, The better angels of our nature: why violence has declined. Observations reported in Gould, 1991)

What does your probability intuition tell you?

Assume 50% hit rate, no "hot hand" (coin tossing)

Task 1: Mr X makes a sequence of five throws. Which of the following sequences is more likely to observe?

Alt. 1: Hit-Hit-Hit-Hit-Hit

Alt. 2: Hit-Miss-Hit-Hit-Miss

Answer: Same probability (the representativeness fallacy makes people believe that the first is less likely)

Task 2: Mr X makes a sequence of throws. Which of the following sequences is more likely to occur **FIRST**?

Alt. 1: Hit-Hit

Alt. 2: Miss-Hit

Example: Miss-Miss-Hit-Hit-Hit-Miss-...

→ Miss-Hit occurs first

Answer: It is three times more likely to observe Miss-Hit before Hit-Hit!

(If you don't believe me, we can make a bet where I bet 10 Euro on Alt. 2 and you 10 Euro on Alt. 1. First to win ten times, wins the 30 Euro.)

HH vs TH explained

After two throws:

A.HH (HH wins, stop)

B.HT (no-one wins, continue)

C.TH (TH wins, stop)

D.TT (no-one wins, continue)

After three throws (B-sequence)

B.1: HTH (TH wins, stop)

B.2: HTT (No-one wins, continue)

After three throws (D-sequence)

D.1: TTH (TH wins, stop)

D.2: TTT (No-one wins continue)

After four throws (B.2 sequence)

B2.1: HTTH (TH wins, stop)

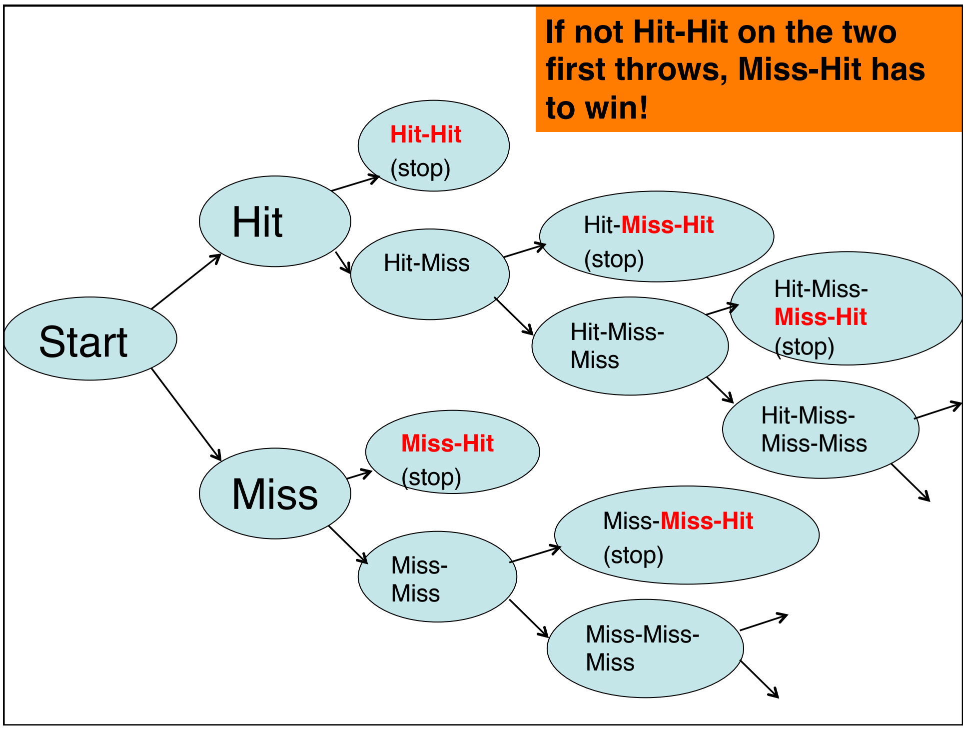
B2.2: HTTT (No-one wins, continue)

After four throws (D.2 sequence)

D2.1: TTTH (TH wins, stop)

D2.2: TTTT (No-onw wins, continue)

HH can only win when occuring on the first two throws, i.e., in only 25% of the cases!



There is no "hot hand" in basketball, but try to tell this to a basketball player ...

TABLE 1

Probability of Making a Shot Conditioned on the Outcome of Previous Shots for Nine Members of the Philadelphia 76e

Player	P(hit/3 misses)	P(hit/2 misses)	P(hit/1 miss)	P(hit)	P(hit/1 hit)	P(hit/2 hits)	P(hit/3 hits)
Clint Richardson	.50 (12)	.47 (32)	.56 (101)	.50 (248)	.49 (105)	.50 (46)	.48 (21)
Julius Erving	.52 (90)	.51 (191)	.51 (408)	.52 (884)	.53 (428)	.52 (211)	.48 (97)
Lionel Hollins	.50 (40)	.49 (92)	.46 (200)	.46 (419)	.46 (171)	.46 (65)	.32 (25)
Maurice Cheeks	.77 (13)	.60 (38)	.60 (126)	.56 (339)	.55 (166)	.54 (76)	.59 (32)
Caldwell Jones	.50 (20)	.48 (48)	.47 (117)	.47 (272)	.45 (108)	.43 (37)	.27 (11)
Andrew Toney	.52 (33)	.53 (90)	.51 (216)	.46 (451)	.43 (190)	.40 (77)	.34 (29)
Bobby Jones	.61 (23)	.58 (66)	.58 (179)	.54 (433)	.53 (207)	.47 (96)	.53 (36)
Steve Mix	.70 (20)	.56 (54)	.52 (147)	.52 (351)	.51 (163)	.48 (77)	.36 (33)
Daryl Dawkins	.88 (8)	.73 (33)	.71 (136)	.62 (403)	.57 (222)	.58 (111)	.51 (55)
Weighted means	.56	.53	.54	.52	.51	.50	.46

Gilovich, Thomas, Robert Vallone, and Amos Tversky.
"The hot hand in basketball: On the misperception of random sequences."
Cognitive psychology 17.3 (1985): 295-314.

NB: More recent studies suggest that there may be a very small "hot hand"-effect.

Instead of p-values ...

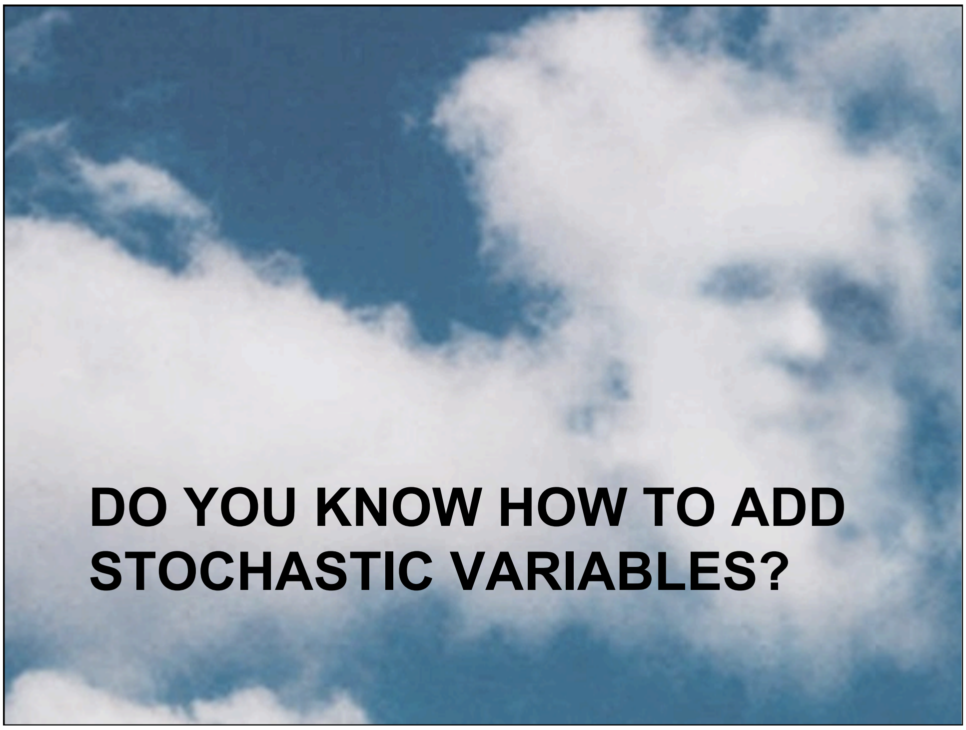
Use confidence intervals of effect size!

Example: The 95% confidence interval of the effect of pair programming on quality in our example is [2%; 38%].
(*This illustrates the true uncertainty of a finding of p=0.04 in a study with low statistical power.*)

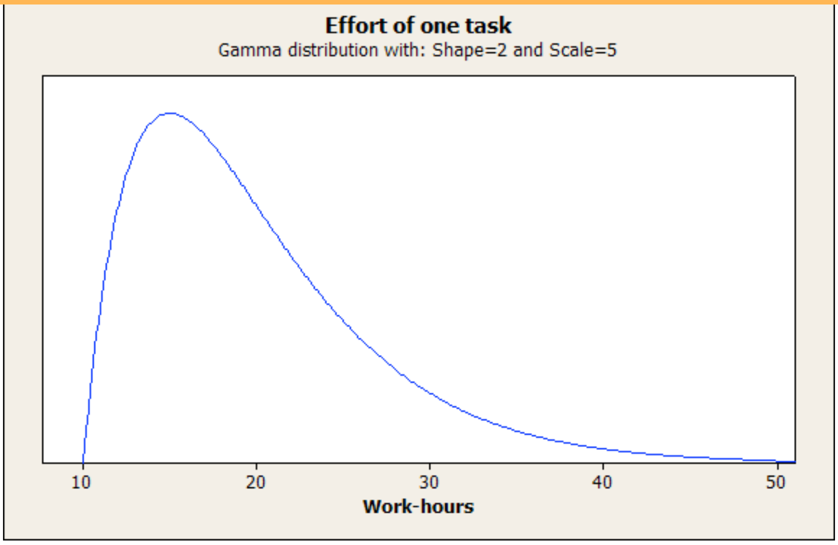
The following should replace null-hypothesis testing:

1) Formulate research questions of the type “How large is the effect?”
2) Find a good measure of effect size.
3) From the collected data, calculate the effects size and its confidence interval
4) Interpret the effect size and confidence intervals

No need for p-values!



What is the estimated most likely (total) effort of ten tasks of this type?



Most likely effort = 15 work-hours, median (p50) = 17 work-hours

With non-symmetric distributions, you can only meaningfully add the MEAN values!

Correct answer: about 200 work-hours

Typical estimate: 150 work-hours?

Result if adding ”most likely” estimates: Only 1% likely to use 150 work-hours or less. 20% likely to use less than 170 work-hours.

Simpson's paradox ("hidden variables")

The winner is "Test last"

	"Test first"	"Test last"
Total proportion of successes	78% successes (273/350)	**83% sucesses (289/350)**

The winner is "Test first"

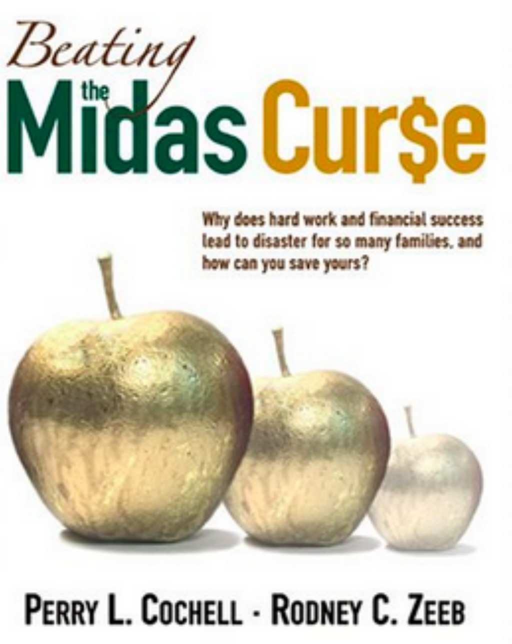
Tasks Type 1	**93% (81/87)**	87% (234/270)
Tasks Type 2	**73% (192/263)**	69% (55/80)

The organization use "test first" more frequently for tasks of Type 2 (e.g. more complex tasks), which has has a lower success rate.

- **Possible (evolutionary) reason: FALSE POSITIVES less harmful than FALSE NEGATIVES.**
- **Statistical methods can help, but can also contribute to seeing FALSE POSITIVES.**

Low statistical power
+ random variance in observed effect size
+ $p > 0.05$ makes publication less likely
= Under-representation of small effect sizes

The result: Inflated effect sizes!

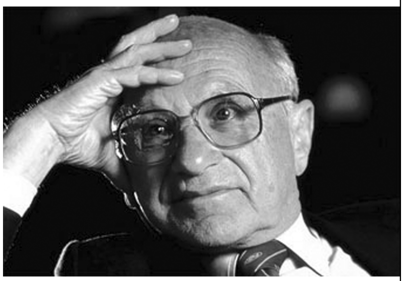


.... six out of ten affluent (rich) families will lose the family fortune by the end of the second generation.

Analyses of non-random samples (self-selected, the best 20% on a test, the projects with highest cost overrun, the developers with lowest estimates, etc.), will easily be misleading.

The more extreme the sampling, the stronger the effect of regression effects.

"*I suspect that the regression fallacy is the most common fallacy in the statistical analysis of economic data*"



Milton Friedman (Nobel prize winner in economy)

Regression towards the mean

NON-RANDOM SAMPLING

NBA Finals: Spurs hope to break Sports Illustrated cover jinx

W. Scott Bailey
Reporter/Project Coordinator-
San Antonio Business Journal
Email | Twitter | Google+ | Facebook

The national media is showing the San Antonio Spurs some love in advance of the 2013 NBA Finals, which tip off on June 6.

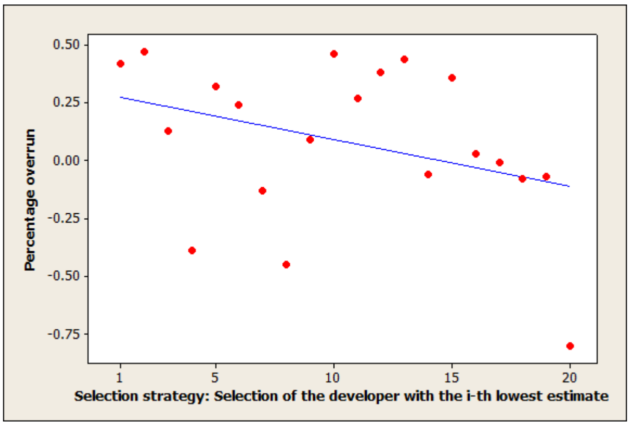
Sports Illustrated has unveiled a cover for its June 10 issue titled: "The Biggest Three."

Sports Illustrated's Chris Ballard's writes in his accompanying story that "it's hard to argue" against proclaiming the Spurs' most talented core — Tim Duncan, Tony Parker and Manu Ginobili — as the most talented trio in NBA history.

Of course, three of the five SI writers who have predicted the outcome of these

Sports Illustrated featured the Spurs' big men, Tim Duncan, Manu Ginobili and Tony Parker on the cover.

The lower the effort estimate, the higher the risk of effort overrun (the winner's curse)



Study:
20 developers estimating and completing the same five tasks

M. Jørgensen. The Influence of Selection Bias on Effort Overruns in Software Development Projects, Information and Software Technology 55(9):1640-1650, 2013.

The winner's curse

Variance shrinkage

Degree of random variation

$$estimation\ overrun = \frac{act - est}{\mu} = \frac{\mu\left(1 - \rho_{est,act}\frac{\sigma_{act}}{\sigma_{est}}\right)(1-w)}{\mu} = \left(1 - \rho_{est,act}\frac{\sigma_{act}}{\sigma_{est}}\right)(1-w)$$

Extremeness of value

M. Jørgensen. The Influence of Selection Bias on Effort Overruns in Software Development Projects, Information and Software Technology 55(9):1640-1650, 2013.

Period 1	4 GL	3 GL	Total
FP	500	2000	2500
Effort	500	4500	5000
Productivity	1.0	0.44	0.50

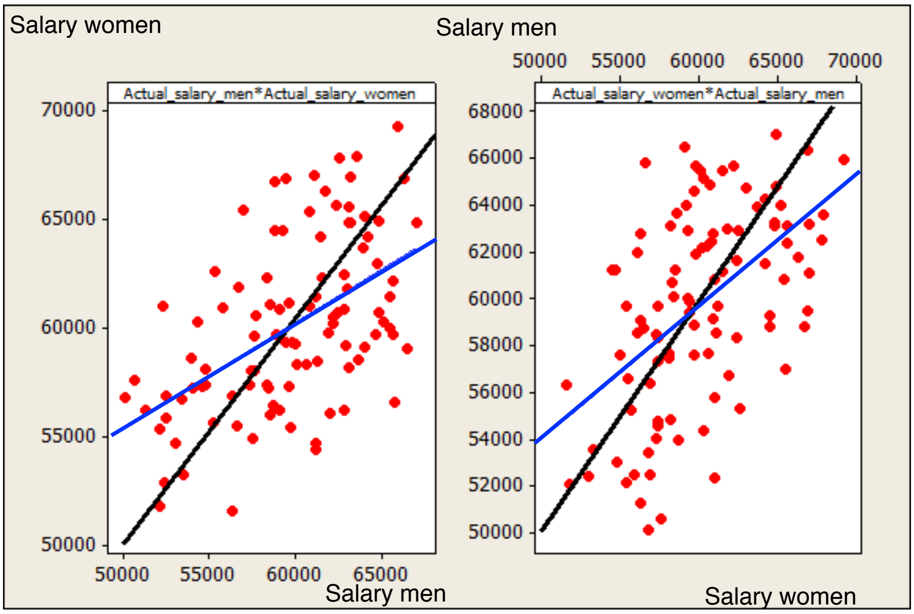
Period 2	4 GL	3 GL	Total
FP	2000	1000	3000
Effort	1800	3000	4800
Productivity	0.9	0.33	0.63
Change in productivity	-0.1	-0.11	0.13

Arithmetic "explanation": a/b + c/d ≠ (a+c)/(b+d)

How many of you know about the assumption of **fixed variables** in regression analysis, ANOVA, t-tests, ...?

Illustration: Salary discrimination?

- Assume an IT-company which:
  - Has 100 different tasks they want to complete and for each task hire one male and one female (200 workers)
  - The "base salary" of a task varies (randomly) from 50.000 to 60.000 USD and **is the same for the male and the female** completing it.
  - The actual salary is the "base salary" added a random, **gender independent, bonus**. This is done through use of a "lucky wheel" with numbers (bonuses) between 0 and 10.000.
- This should lead to (on average): Salary of female = Salary of male
- A regression analysis with female salary as the dependent variable show that the female are discriminated (less likely to get a high bonus)!
  - Salary of female = 26100 + 0.56 * Salary of male
- On the other hand, with male salary as the dependent variable, men are discriminated!?
  - Salary of male = 26900 + 0.55 * Salary of female



What to do about it

- Base regression variable inclusion on *a priori* judgment of importance
- Do not use R^2 or similar measures to assess the goodness of your prediction model
- Compare the model against reasonable alternatives.
- Test your model with
  - Same number of variables and observations
  - Reasonable distributions
  - Same process of outlier removal etc.