Test of Significance & Test of Hypothesis

Tuan V. Nguyen
Professor and NHMRC Senior Research Fellow
Garvan Institute of Medical Research
University of New South Wales
Sydney, Australia

What we are going to learn

- An example
- Test of significance
- Test of hypothesis

Unpaired t-test by R

```
g1 = c( 0.2, 0.3, 0.4, 1.1, 2.0, 2.1, 3.3, 3.8, 4.5, 4.8, 4.9, 5.0, 5.3,
   7.5, 9.8, 10.4, 10.9, 11.3, 12.4, 16.2, 17.6, 18.9, 20.7,
        24.0, 25.4, 40.0, 42.2, 50.0, 60)
g2 = c(0.2, 0.3, 0.4, 0.7, 1.2, 1.5, 1.5, 1.9, 2.0, 2.4, 2.5, 2.8, 3.6,
   4.8, 4.8, 5.4, 5.7, 5.8, 7.5, 8.7, 8.8, 9.1, 10.3, 15.6, 16.1, 16.5,
   16.7, 20.0, 20.7, 33.0)
t.test(g1, g2)
```

```
data:  g1 and g2
t = 2.0357, df = 40.804, p-value = 0.04831
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
  0.05163216 13.20239083
sample estimates:
mean of x mean of y
14.310345  7.683333
```

What does P = 0.048 mean ?

Two paradigms of test

- **Test of significance** (Ronald A. Fisher)

  Statistics is a vital part in inductive inference (drawing conclusion from sample to population, from the particular to the general)

- **Test of hypothesis** (Jerzy Neyman and Egon Pearson)

  NP dismiss the concept of inductive inference; statistics as a mechanism for making decisions and guiding behavior

Fisher' s test of significance

- Set up a null hypothesis (i.e., no difference between groups)
- Compute the probability of obtaining data if the null hypothesis is true
- Report the exact level of significance (e.g., $p = 0.051$ or $p = 0.049$). Do not use a conventional 5% level, and do not talk about accepting or rejecting hypotheses.
- Use this procedure only if you know very little about the problem at hand.

"A null hypothesis can be disproved, but never proved or established" (Fisher, 1925)

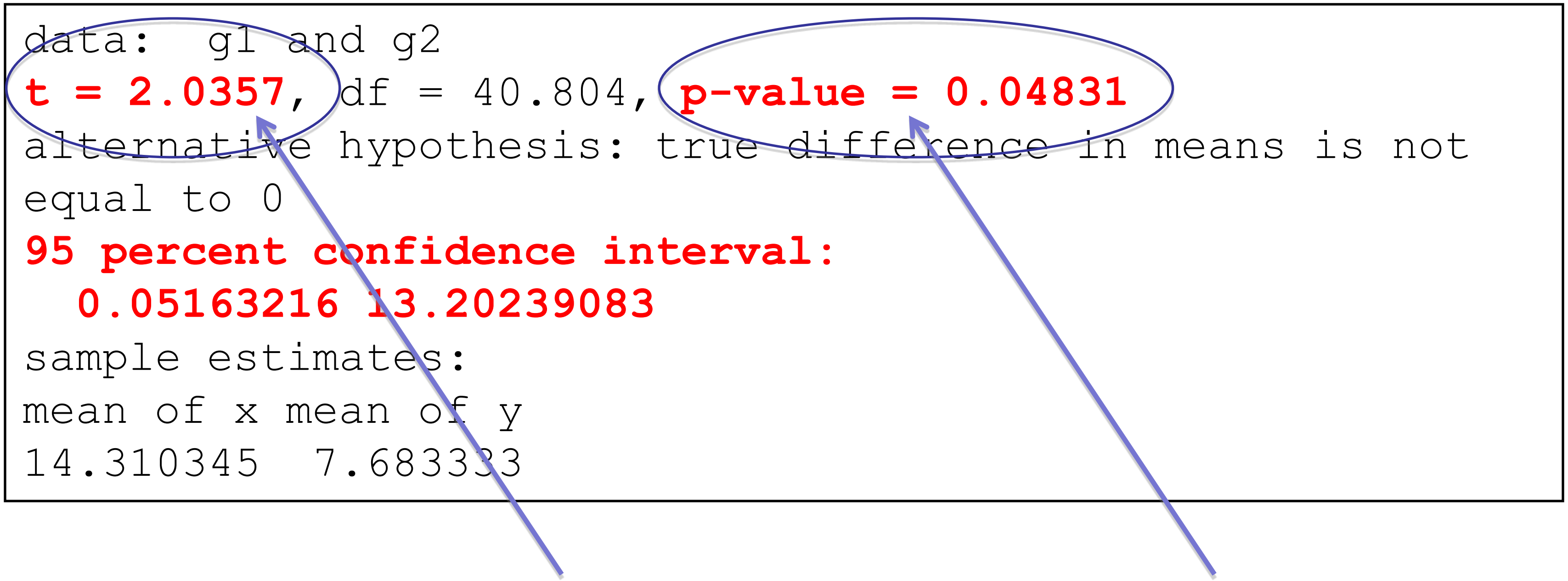
Fisher’ s test of significance

- Ho: there no NO difference between g1 and g2
- Conducted an experiment (study), obtained data
- Compute P(data | Ho is true)

P = 0.048

What does “data” mean ?

- “Data” here is the test statistic (i.e., t-statistic, z-statistic, etc)



```
data:  g1 and g2
t = 2.0357, df = 40.804, p-value = 0.04831
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
  0.05163216 13.20239083
sample estimates:
mean of x mean of y
14.310345  7.683333
```

Data

P(Data | no difference)

An example of test of significance

Study: 10 subjects were on 2 treatments (A and B); 8 patients showed B>A. Was there a real difference?

ID	A	B	B>A
1	1.00	1.02	Yes
2	0.76	0.80	Yes
3	0.89	0.85	No
4	0.70	0.73	Yes
5	0.90	0.92	Yes
6	0.88	0.93	Yes
7	0.92	0.95	Yes
8	0.80	0.82	Yes
9	0.72	0.78	Yes
10	1.10	1.08	No

An example of test of significance

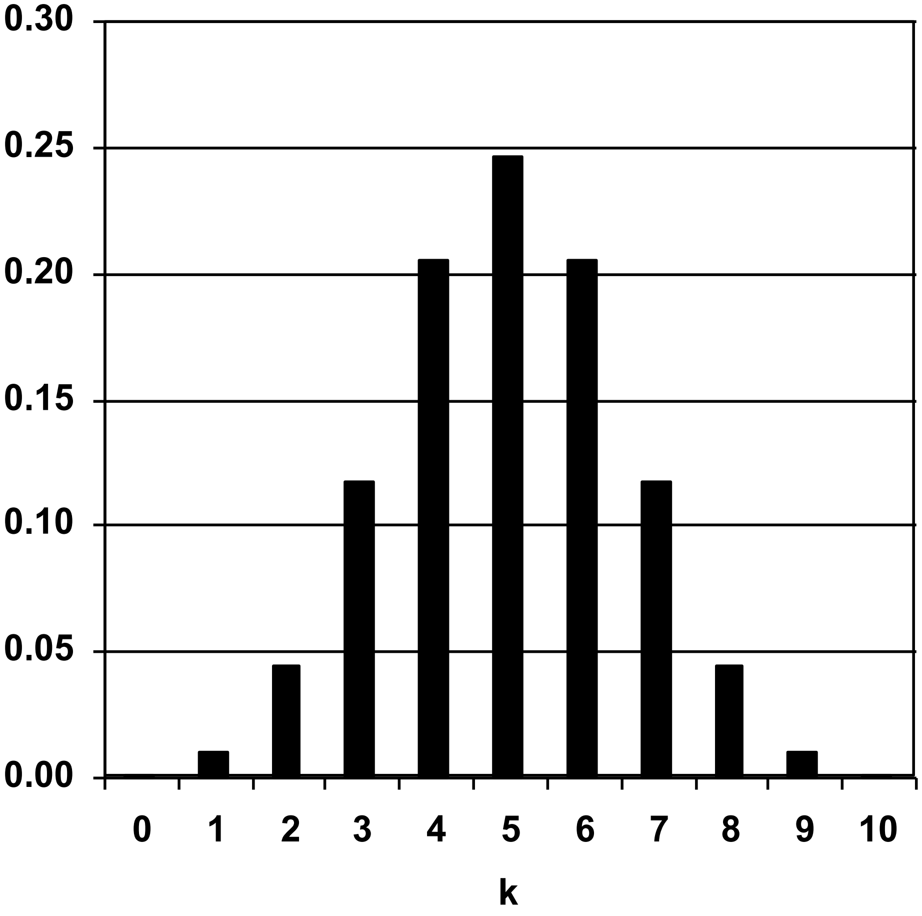
- Let p = proportion of B>A
- Null hypothesis: there was no difference (p = 0.5)
- Alternative hypothesis: there was an effect
- Under the null hypothesis, we can work out the exact probability that there was 0, 1, 2, …, 10 B>A outcomes

$$\Pr(k) = \binom{10}{k} p^k (1-p)^{10-k}$$

An example of test of significance

- Probability that there was/were 0, 1, 2, … outcomes with B>A

k	Pr(k)
0	0.0009765625
1	0.009765625
2	0.04394531
3	0.1171875
4	0.2050781
5	0.2460938
6	0.2050781
7	0.1171875
8	0.04394531
9	0.009765625
10	0.0009765625
P(k>=8)	0.054687



Neyman-Pearson model of test of hypothesis

THE TRUTH

- There is an effect
- No effect

TEST STATISTIC

- Significant
- Not significant

TRUTH	STATISTICAL TEST	Not significant
Effect	Significant	OK (1-β)
Effect	Not significant	Type II error (β)
No effect	Significant	Type I error (α)
No effect	Not significant	OK

α : significance level, 1-β : power

NP’s test of hypothesis

- Proposed by Neyman and Pearson
- Set up two hypotheses, H0 and H1,
- Decide about α (the probability of wrongly reject H1) and β (the probability of wrongly rejecting H0), and sample size before the experiment, based on subjective cost-benefit considerations.
- If the data falls into the rejection region of H0, accept H1; otherwise accept H0. Note that accepting a hypothesis does not mean that you believe in it, but only that you act as if it were true.

Rule of behaviour: we pick H0 and H1 so that "in the long run of experience, we shall not be too often wrong"

Scientific research and medical diagnosis

- Medical diagnosis
  - We don't know whether the patient has a disease (yes/no)
  - We rely on diagnostic test (+ve/-ve)
  - Two mistakes: *false positive* (patient doesnot has the disease, but the test is +ve); *false negative* (patient has the disease, but the test is –ve).
- Scientific research
  - We don't know whether there is an effect / association (yes/no)
  - We rely on statistical test (significant / non-significant)
  - Two mistakes: *type I error* (there is no effect, but the test result is significant); *type II error* (there is an effect, but the test result is non-significant).

Hybrid system

- Set up a null hypothesis of "no difference"
- Use 5% as a convention for rejecting the null
- If significant, accept your research hypothesis.

Report the result as $p < 0.05$, $p < 0.01$, or $p < 0.001$ (whichever comes next to the obtained *p*-value).

Actual procedure

Propose a hypothesis – H1

Propose a null hypothesis – H0

Collect the data – D

Compute the probability of obtaining the finding

P(D | H0)

If P(D | H0) < 0.05, reject H0, accept H1

The logics behind the hybrid system

1. If H0 is true, the data (finding) can not happen (premise 1)
2. The finding happens (premise 2)
3. Therefore, H0 is not true (Conclusion 1)
4. Either H0 or H1 must be true (by definition)
5. H0 is true (from 3)
6. Therefore H1 must be true (Conclusion 2)

Logics of hypothesis test

- The current process of hypothesis testing is a "**proof by contradiction**"

If the null hypothesis is true, then the observations are unlikely.

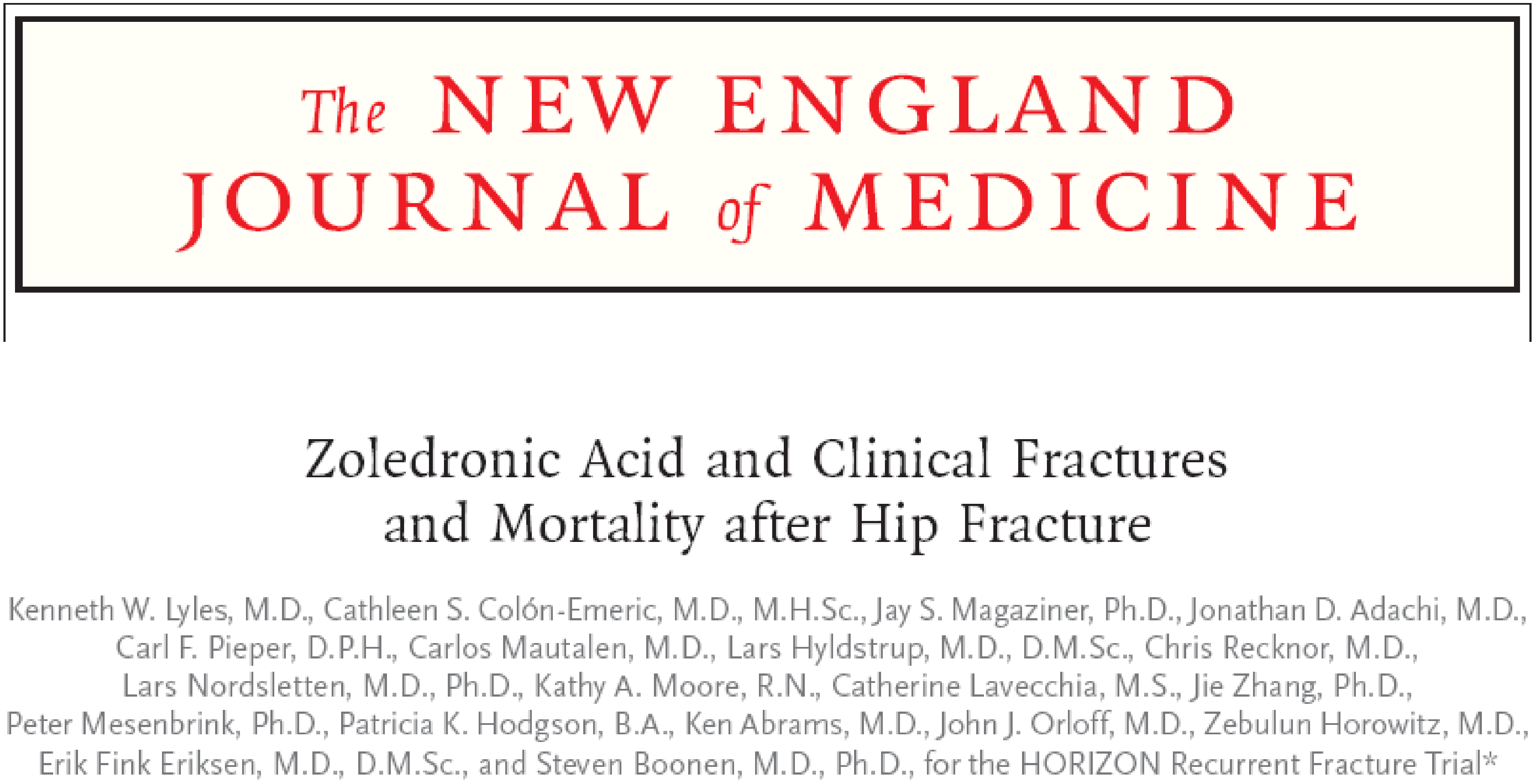
The observations occurred

Therefore, the null hypothesis is unlikely

If Tuan has hypertension, then he is unlikely to have fever.

Tuan has fever

Therefore, Tuan is unlikely to have hypertension



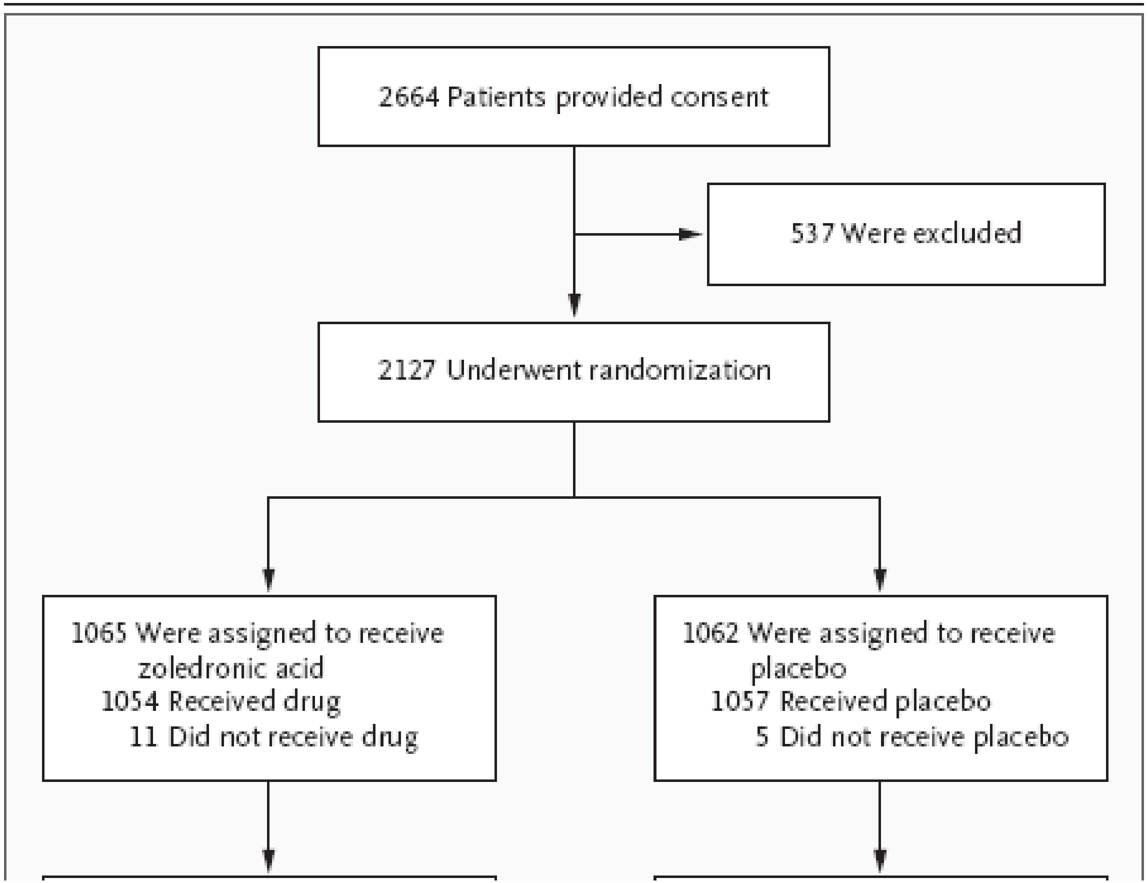
The NEW ENGLAND JOURNAL of MEDICINE

Zoledronic Acid and Clinical Fractures and Mortality after Hip Fracture

Kenneth W. Lyles, M.D., Cathleen S. Colón-Emeric, M.D., M.H.Sc., Jay S. Magaziner, Ph.D., Jonathan D. Adachi, M.D., Carl F. Pieper, D.P.H., Carlos Mautalen, M.D., Lars Hyldstrup, M.D., D.M.Sc., Chris Recknor, M.D., Lars Nordsletten, M.D., Ph.D., Kathy A. Moore, R.N., Catherine Lavecchia, M.S., Jie Zhang, Ph.D., Peter Mesenbrink, Ph.D., Patricia K. Hodgson, B.A., Ken Abrams, M.D., John J. Orloff, M.D., Zebulun Horowitz, M.D., Erik Fink Eriksen, M.D., D.M.Sc., and Steven Boonen, M.D., Ph.D., for the HORIZON Recurrent Fracture Trial*

Randomized clinical trial as a hybrid of Fisher’s test of significance and NP’s test of hypothesis

“The trial was event driven and required 211 clinical fractures to have a power of 90%. A twosided level of significance of 0.05, with two interim analyses performed with the use of an O’Brien–Fleming spending function,21 was needed to detect a 35% reduction in the rate of clinical fracture in the zoledronic acid group, as compared with the placebo group.”



2664 Patients provided consent

537 Were excluded

2127 Underwent randomization

1065 Were assigned to receive zoledronic acid
1054 Received drug
11 Did not receive drug

1062 Were assigned to receive placebo
1057 Received placebo
5 Did not receive placebo

Randomized clinical trial as a hybrid of Fisher's test of significance and NP's test of hypothesis

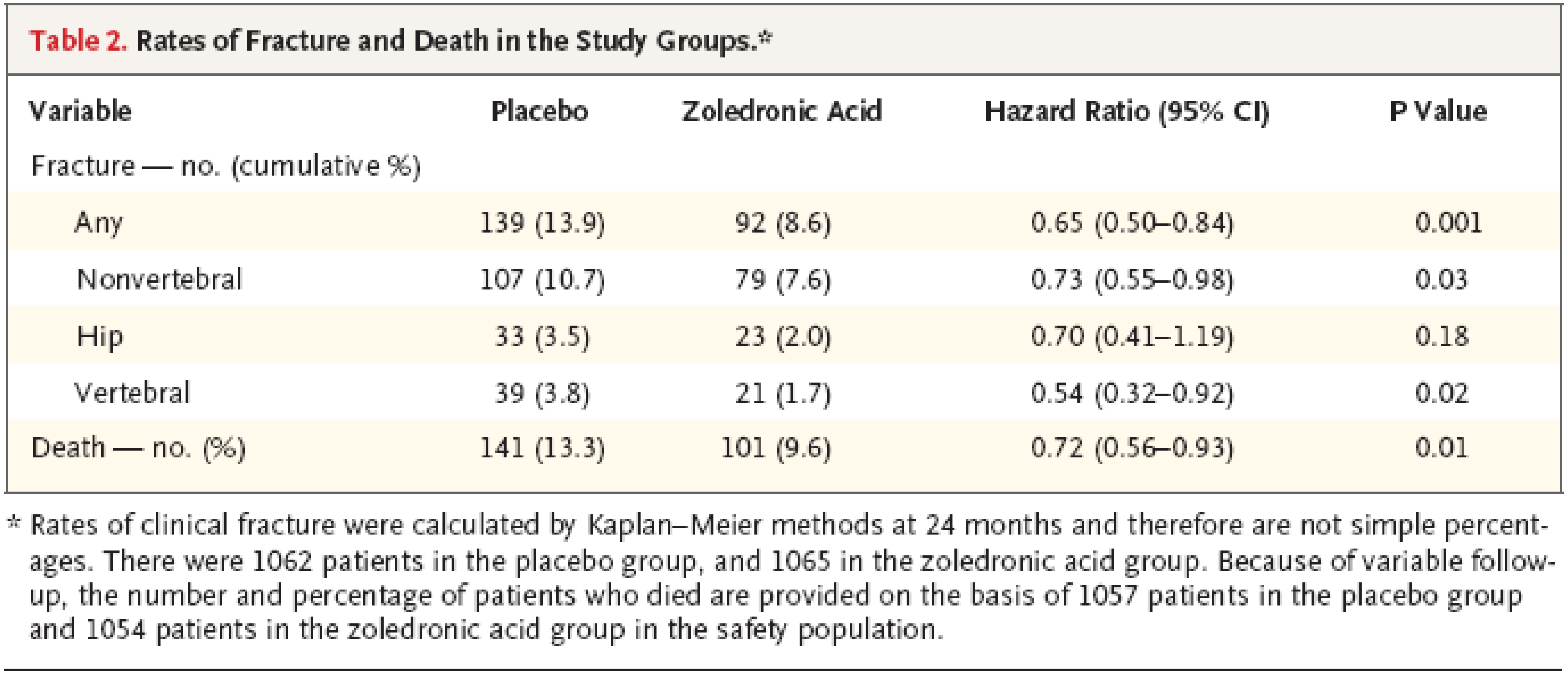


Table 2. Rates of Fracture and Death in the Study Groups.*

Variable	Placebo	Zoledronic Acid	Hazard Ratio (95% CI)	P Value
Fracture — no. (cumulative %)				
Any	139 (13.9)	92 (8.6)	0.65 (0.50–0.84)	0.001
Nonvertebral	107 (10.7)	79 (7.6)	0.73 (0.55–0.98)	0.03
Hip	33 (3.5)	23 (2.0)	0.70 (0.41–1.19)	0.18
Vertebral	39 (3.8)	21 (1.7)	0.54 (0.32–0.92)	0.02
Death — no. (%)	141 (13.3)	101 (9.6)	0.72 (0.56–0.93)	0.01

* Rates of clinical fracture were calculated by Kaplan–Meier methods at 24 months and therefore are not simple percentages. There were 1062 patients in the placebo group, and 1065 in the zoledronic acid group. Because of variable follow-up, the number and percentage of patients who died are provided on the basis of 1057 patients in the placebo group and 1054 patients in the zoledronic acid group in the safety population.

Questions

Mortality RR 0.72 (95%CI: 0.56 – 0.93); p = 0.01

- They have absolutely disproved the null hypothesis (that is, there is no effect of zoledronate).
- They have found the probability of the null hypothesis being true.
- They have absolutely proved your experimental hypothesis (that there is a difference between the population means).
- They can deduce the probability of the experimental hypothesis being true.
- They know, if they decide to reject the null hypothesis, the probability that they are making the wrong decision.
- They have a reliable finding in the sense that if, hypothetically, the study were repeated a great number of times, they would obtain a significant result on 99% of occasions

None of the answers is true!

We make mistake!

P value is NOT …

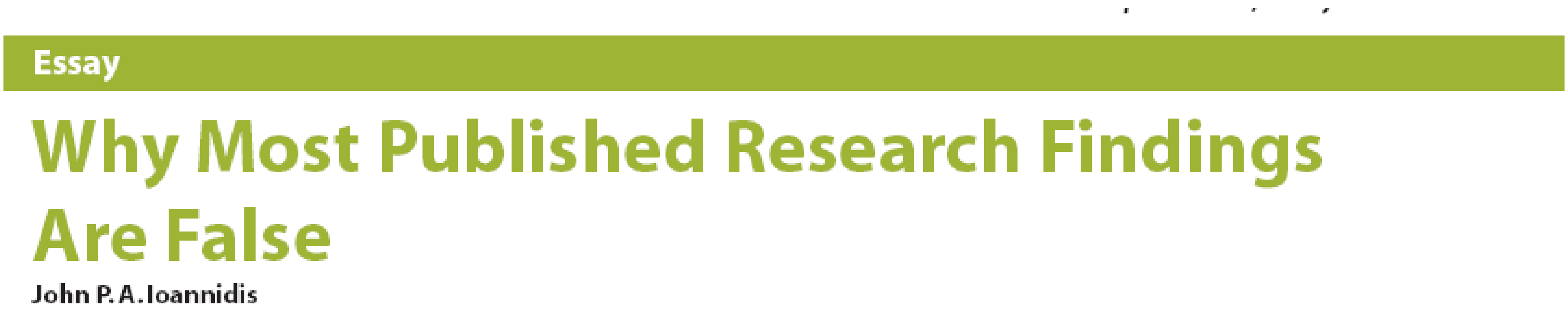
- The probability of the null hypothesis.
- The probability that you will make a Type I error if you reject the null hypothesis.
- The probability that the observed data occurred by chance.
- The probability of the observed data under the null hypothesis

Cancer risks

- Electric razors
- Broken arms (women)
- Fluorescent lights
- Allergies
- Breeding reindeer

- Being a waiter
- Owning a pet bird
- Being short
- Being tall
- Hot dogs
- Have a refrigerator!

Altman and Simon, JNCI 1992



Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Diagnosis and statistical reasoning

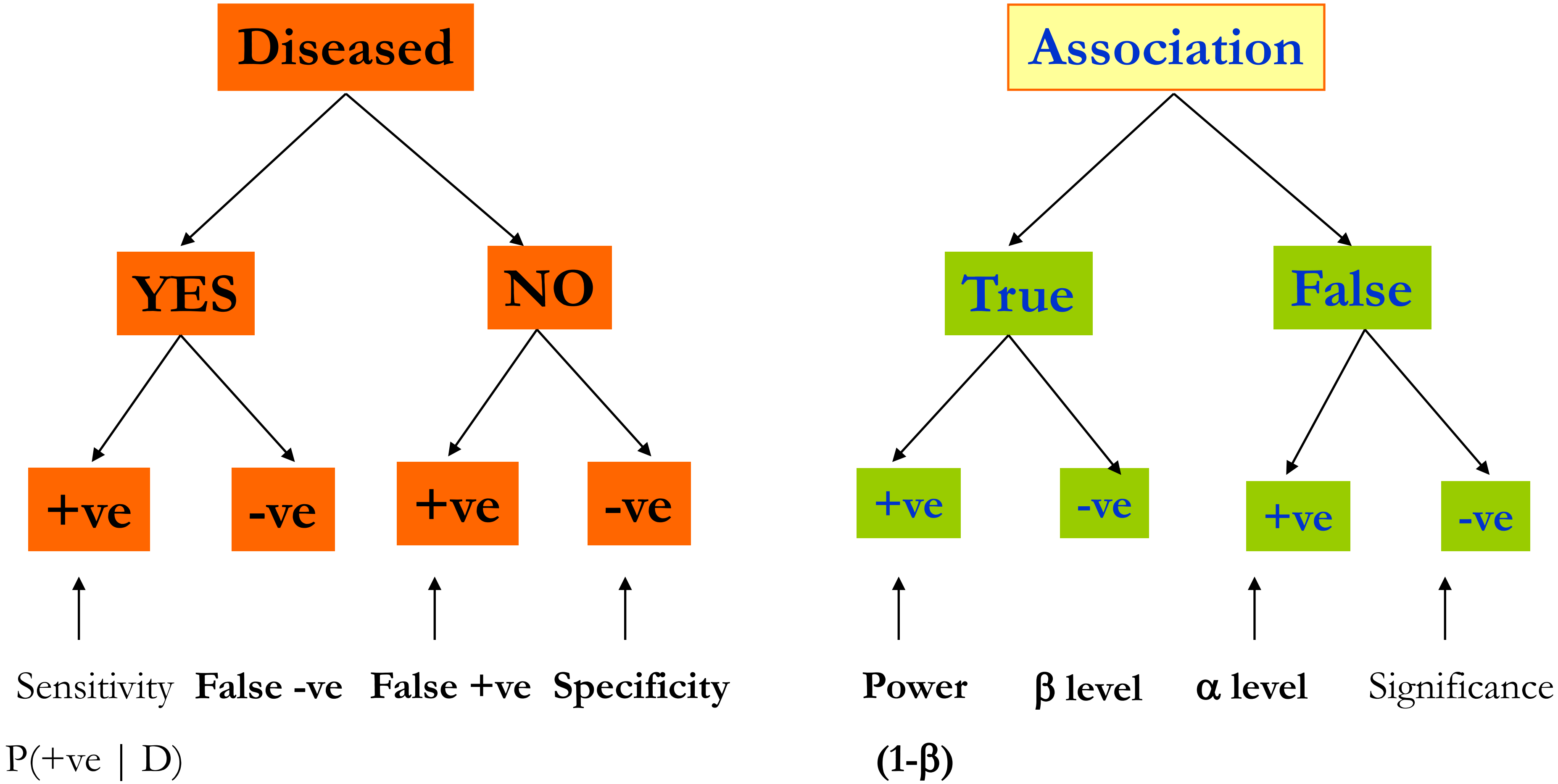
Diagnostic reasoning

	Disease status	
	Present	Absent
Test result		
+ve	True +ve (sensitivity)	False +ve
-ve	False -ve	True -ve (Specificity)

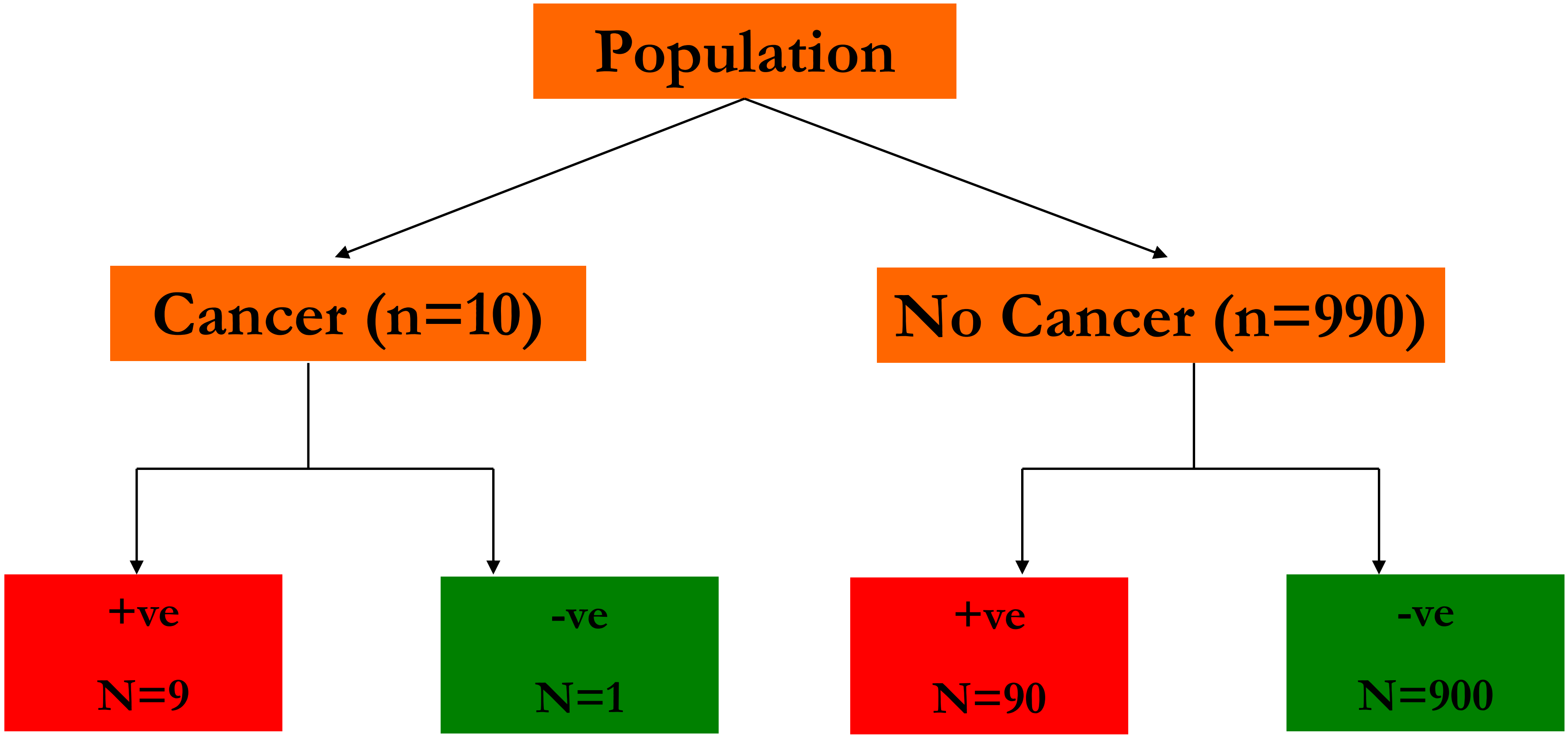
Statistical reasoning

	Effect	
	Present (Ho *not* true)	Absent (Ho is true)
Test result		
Signif.	OK $1-\beta$	Type I err. α
No Signif.	Type II err. β	OK $1-\alpha$

Diagnosis and research hypothesis



Breast cancer screening



P(+ve result | Cancer) = 9/10 = 90%

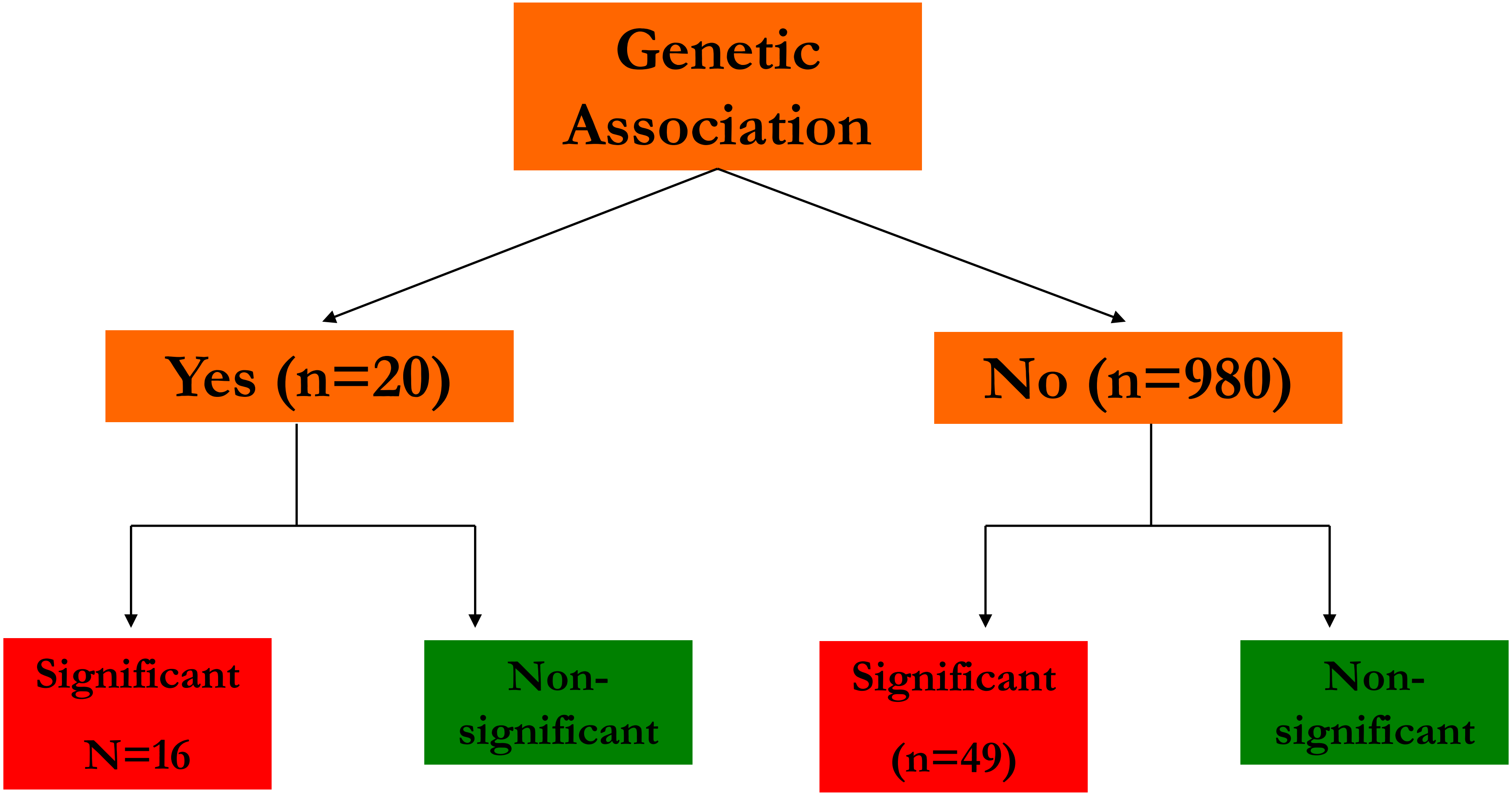
P(Cancer| +ve result) = 9/(9+90) = 9%

Consider a study …

- 20 SNPs (out of 1000 SNPs) are associated with osteoporosis
- Study power = 80%
- Type I error = 5%

- A study has been done on a SNP
- Finding: Significant association (P = 0.02)
- What is the probability that there is indeed an association

A Bayesian Interpretation



Genetic Association

Yes (n=20)

No (n=980)

Significant

N=16

Non-significant

Significant

(n=49)

Non-significant

P(True association | Significant result) = 16/(16+49) = 25%

Chance finding

About 25% of all findings with “p<0.05” should, if viewed in a scientifically agnostic light, properly be regarded as nothing more than chance findings

J. Berger (1987); R Matthews (2001)

"Half of what doctors know is wrong. Unfortunately we don't know which half."

Quoted from the Dean of Yale Medical School,

in "Medicine and Its Myths",

New York Times Magazine, 16/3/2003

Traditional statistical inference

- Traditional statistical rules are a collection of principles and conventions to avoid errors over the long run; ***they do not tell us how likely our claims are to be true, nor do they easily apply to individual results.***