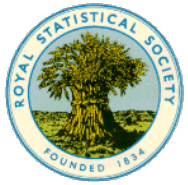


Statistics for Medical Research (Part I)

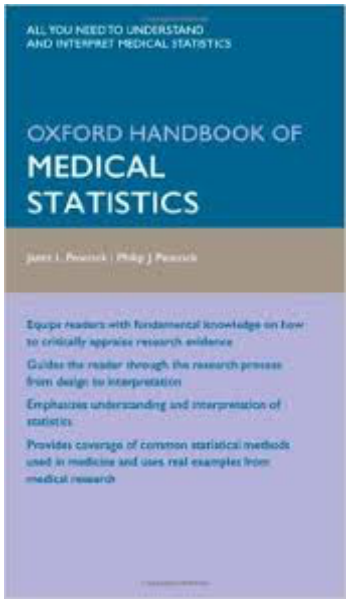
Descriptive Statistics

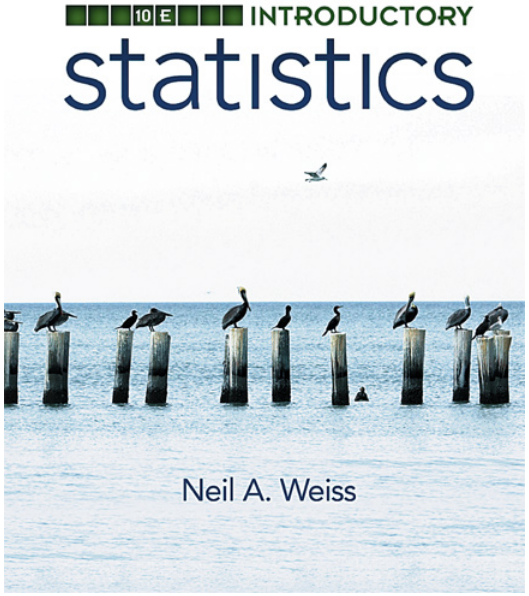
Null Hypothesis Testing

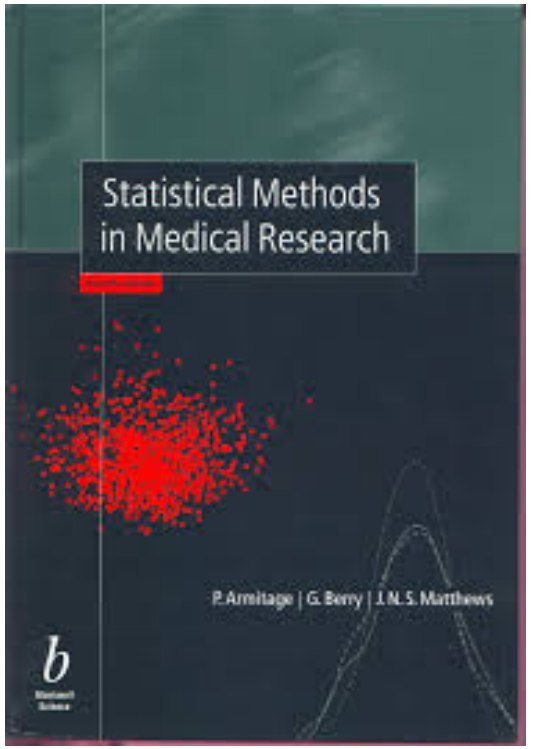
Dr. Wong Kai Choi

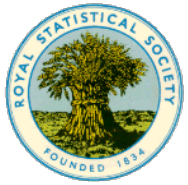
Department of Psychiatry
Pamela Youde Nethersole Eastern Hospital

19th January 2016


ALL YOU NEED TO UNDERSTAND AND INTERPRET MEDICAL STATISTICS
OXFORD HANDBOOK OF
MEDICAL
STATISTICS


10E
INTRODUCTORY
statistics
Neil A. Weiss

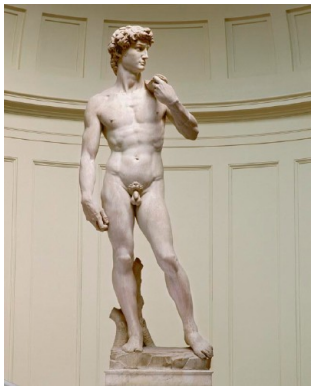

Statistical Methods
in Medical Research
P. Armitage | G. Berry | J. N. S. Matthews



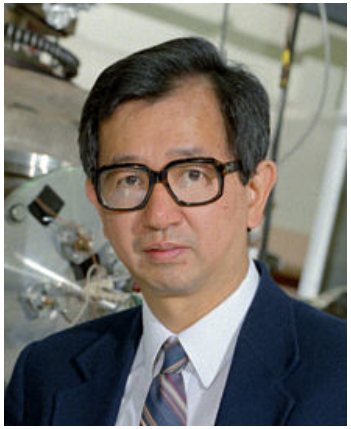
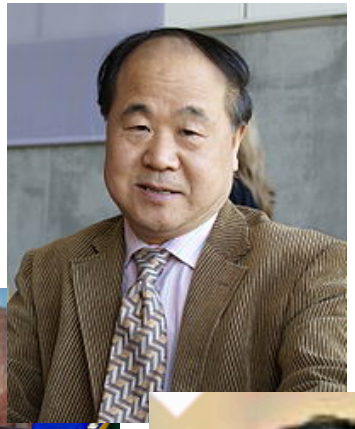
What is Statistics?

Mathematics and Statistics

Discovery of Fact

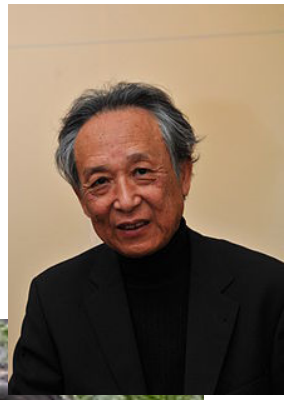
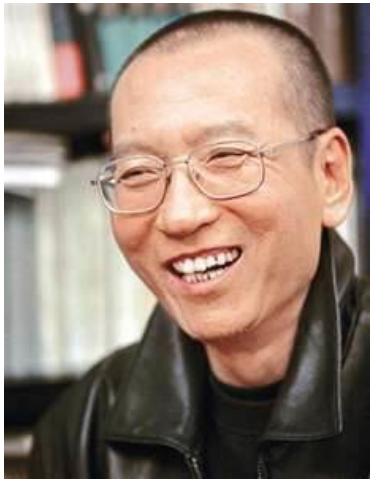


"Every block of stone has a statue inside it and it is the task of the sculptor to discover it." – Michelangelo






In the eyes of statisticians – it is a *bar chart* (6th October 2015 version)

"A mathematician is a machine for turning coffee into theorems"

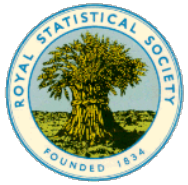
- *Paul Erdos (1913 – 1996)*

"It is not knowledge, but the act of learning, not possession but the act of getting there, which grants the greatest enjoyment."

- Carl Friedrich Gauss (1777 – 1855)

"To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of"

- *Sir R.A. Fisher*
  Indian Statistical Congress (1938)



Descriptive Statistics

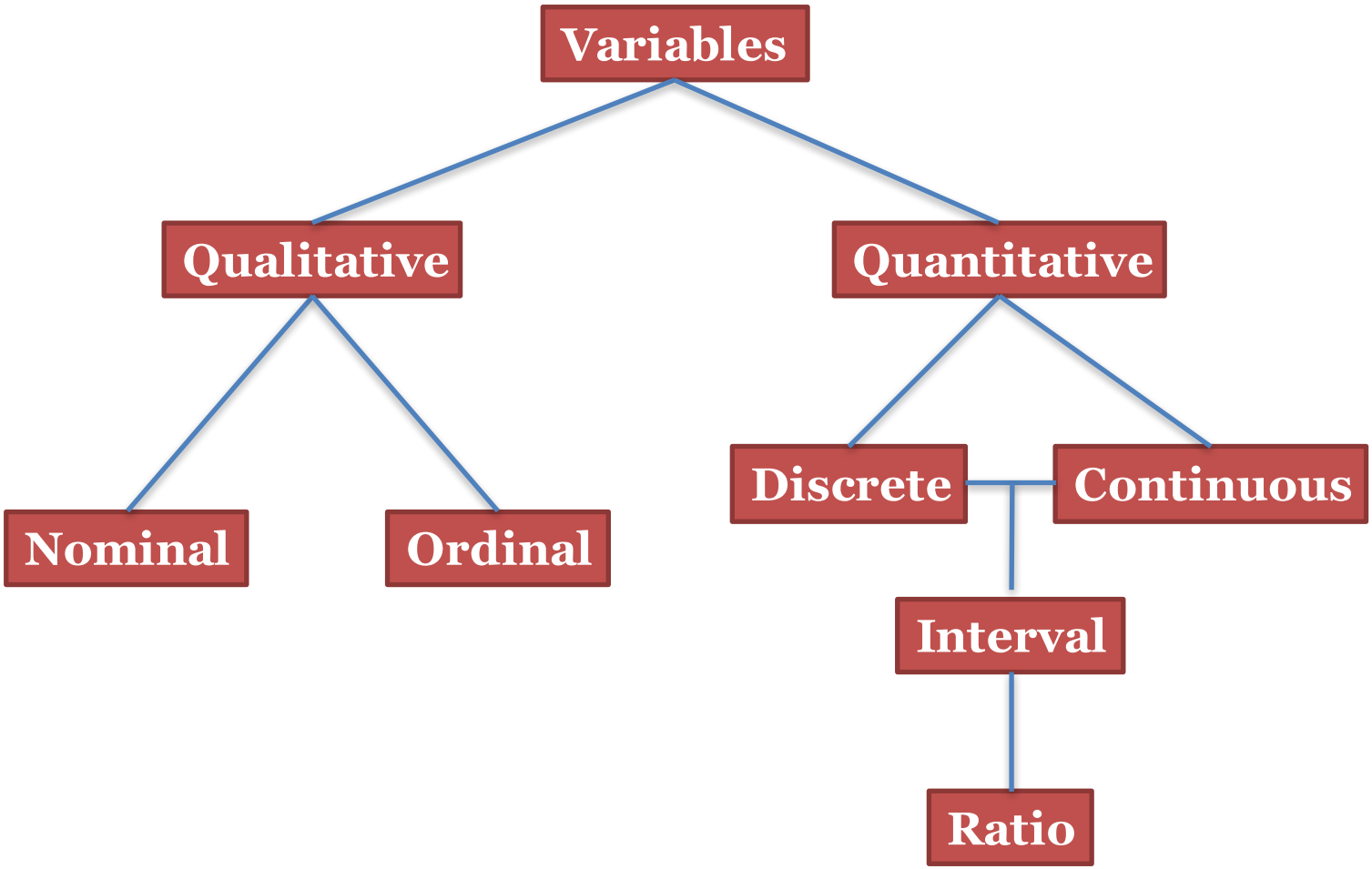
Parameter and Statistics

Parameter and Statistic

Parameter: A descriptive measure for a population

Statistic: A descriptive measure for a sample

Types of variables

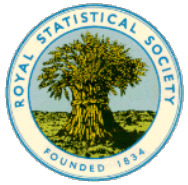


***Nominal Variables*（類別變數） –** Variables that have no numerical values

***Ordinal Variables*（有序變數） –** values whose order is significant, but no meaningful arithmetic like operations can be performed.

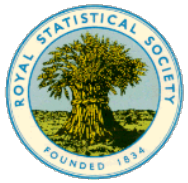
***Interval Variables*（等距變數） –** an ordinal variables that the same magnitudes of the differences between two values takes the same meaning.

***Ratio Variable*s（比率變數） –** with features of interval variables and any two values have meaningful ratio, making the operation of multiplication and division meaningful.



Descriptive Statistics

- It described the main features of a collection of data quantitatively
- It aimed at summarize the dataset
- There are 4 degrees:
  - Location
  - Spread
  - Skewness
  - Kurtosis

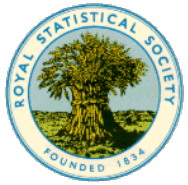


Measures of Central Location

- A measure of central location are used to pinpoint the center of a set of data values. It is often used to work out a typical value to represent the whole set of data.
- There are three commonly used measures of central location:

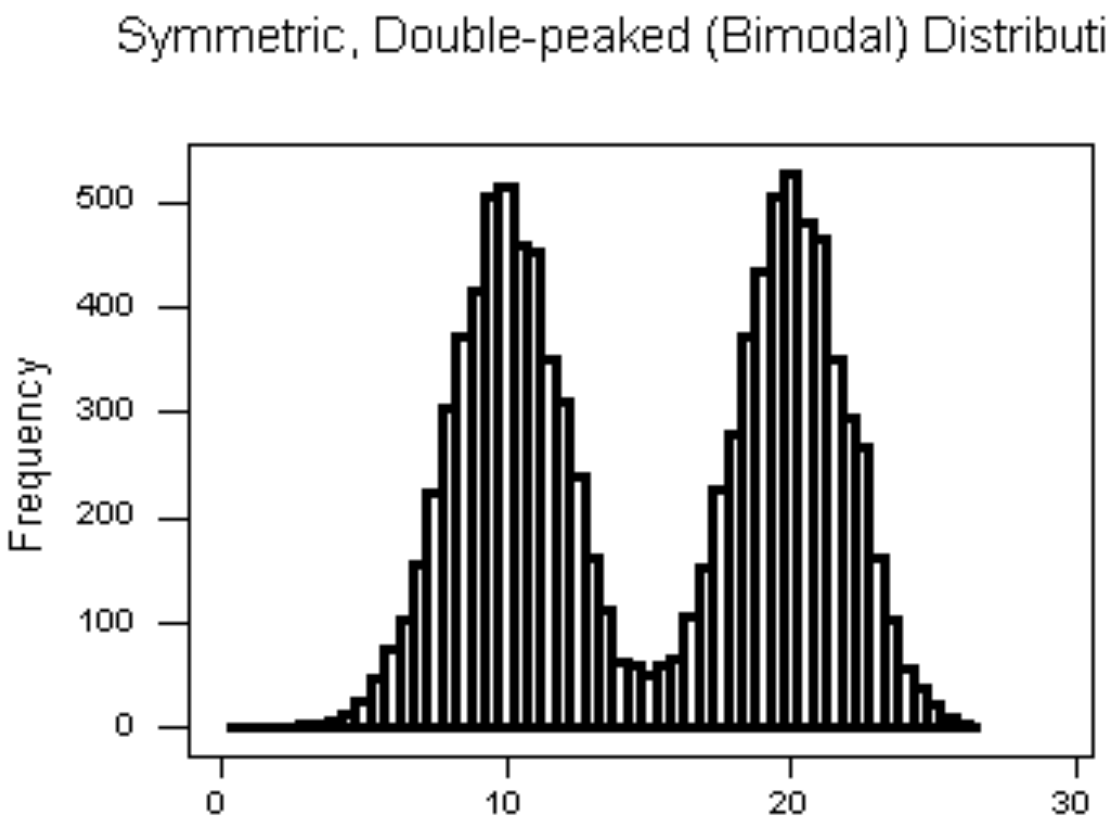
  mean, mode and median.

The mean is the most common one to be used.



Location

- It is the first degree
- Mean: the arithmetic means or expected value of random variables
- Median: the value separating the higher half of a sample from the lower half
- Mode: The most common value among the group
- In symmetrical (not only normal) distribution: mean = median = mode


Symmetric, Double-peaked (Bimodal) Distribution
Frequency
500
400
300
200
100
0
0
10
20
30

Spread

- It is a second degree
- Standard deviation: related to mean

$$\sigma = \sqrt{\mathrm{E}\left[(X - \mu)^2\right]}.$$

- Range and Percentile
- Interquartile range: related to median
  - It contain half of the sample inside the range
  - Upper quartile: separate the higher ¼ and lower ¾
  - Lower quartile: separate the lower ¼ and higher ¾

Measures of Variability

(Measures of Dispersion)

- Range
- Mean Absolute Deviation
- Variance and Standard Deviation
- Quartiles

The range

- The range of a set of measurements is the difference between the largest and the smallest measurements.
- Its major advantage is the ease with which it can be computed.
- Its major shortcoming is its failure to provide information on the dispersion of the values between the two end points.

Quartiles

Quartiles

First, arrange the data in increasing order. Next, determine the median. Then, divide the (ordered) data set into two halves, a bottom half and a top half; if the number of observations is odd, include the median in both halves.

- The **first quartile** (Q_1) is the median of the bottom half of the data set.
- The **second quartile** (Q_2) is the median of the entire data set.
- The **third quartile** (Q_3) is the median of the top half of the data set.

Mean absolute deviation

$$MAD = \frac{\sum_{i=1}^{N} |X_i - \mu|}{N}$$

Consider the data from a population set: 2, 7, 4, 12, 5.

$$MAD = \frac{|2-6| + |7-6| + |4-6| + |12-6| + |5-6|}{5}$$

$$= \frac{4+1+2+6+1}{5} = 2.8$$

The population variance

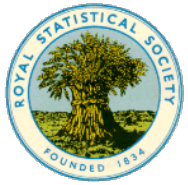
- This measure of dispersion reflects the values of ***all*** the measurements.
- The variance of a **population** of N measurements $x_1, x_2,\ldots,x_N$ having a mean μ is defined as

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

The sample variance

– The variance of a **sample** of n measurements $x_1, x_2, \ldots, x_n$ having a mean $\bar{X}$ is defined as

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

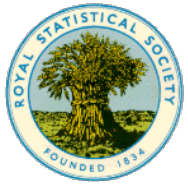


Standard Deviation

The standard deviation of a set of measurements is the square root of the variance of the measurements.

$$\text{Sample standard deviation}: s = \sqrt{s^2}$$

$$\text{Population standard deviation}: \sigma = \sqrt{\sigma^2}$$



Empirical Rules

Empirical Rule

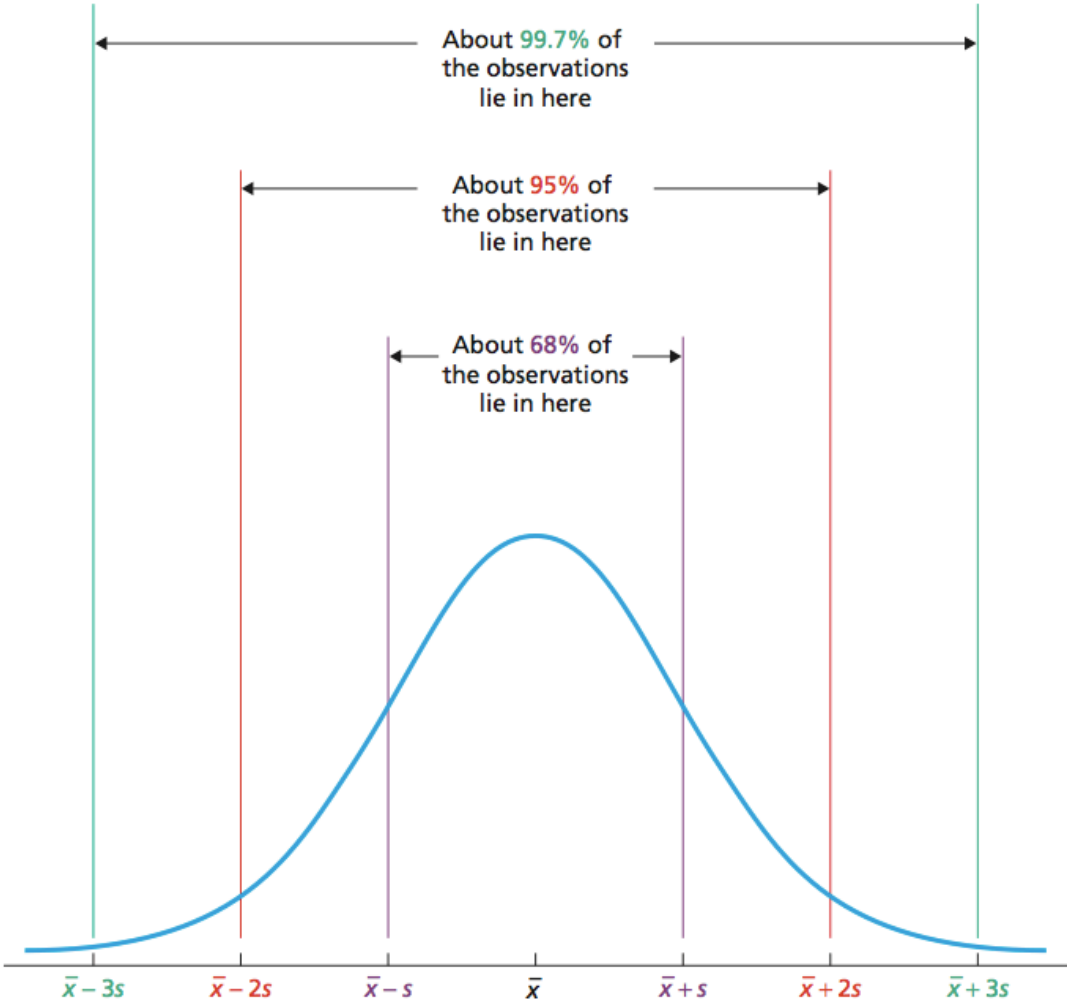
For any quantitative data set with roughly a bell-shaped distribution, the following properties hold.

Property 1: Approximately 68% of the observations lie within one standard deviation to either side of the mean, that is, between $\bar{x} - s$ and $\bar{x} + s$.

Property 2: Approximately 95% of the observations lie within two standard deviations to either side of the mean, that is, between $\bar{x} - 2s$ and $\bar{x} + 2s$.

Property 3: Approximately 99.7% of the observations lie within three standard deviations to either side of the mean, that is, between $\bar{x} - 3s$ and $\bar{x} + 3s$.

Empirical Rules



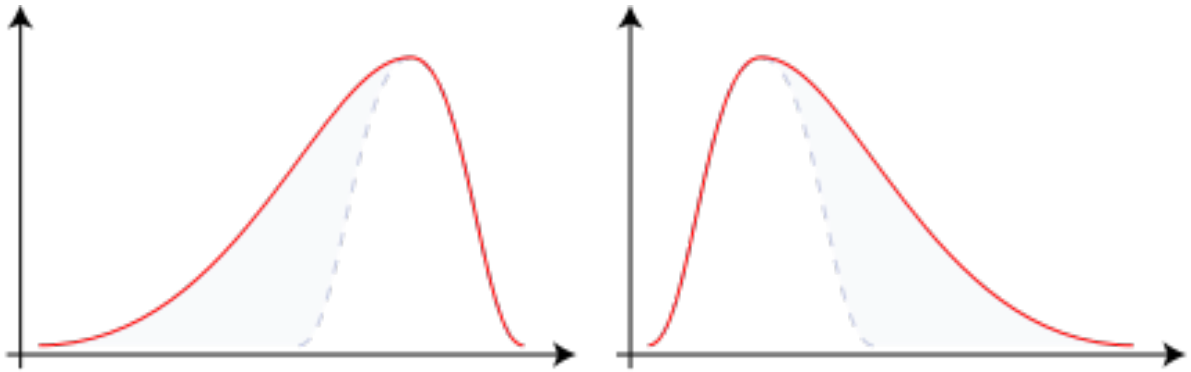
Skewness

- It is the third degree
- It measure the asymmetry of the distribution
- It is calculated by

$$\gamma_1 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{\mathrm{E}\left[(X-\mu)^3\right]}{(\mathrm{E}\left[(X-\mu)^2\right])^{3/2}} = \frac{\kappa_3}{\kappa_2^{3/2}},$$

Skewness

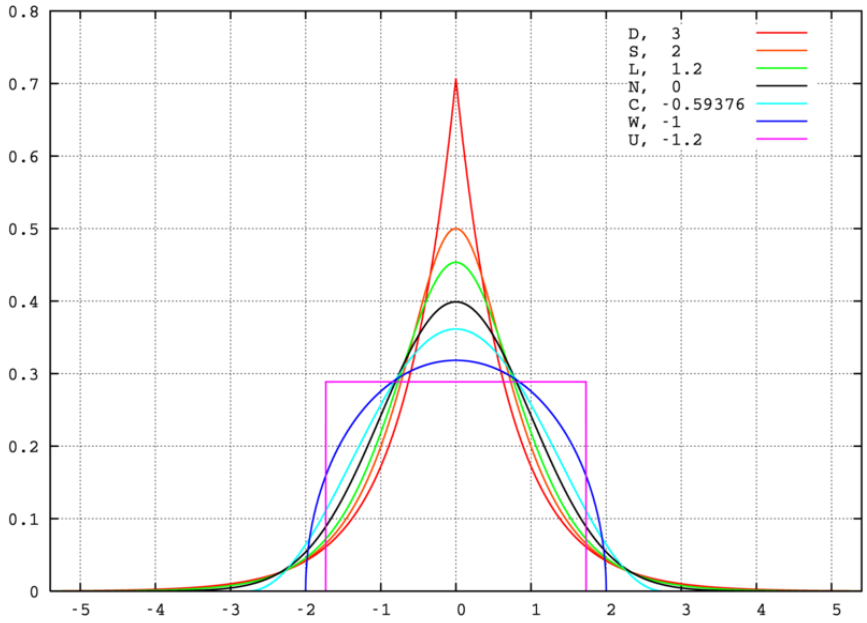
- Looked at the tail of the (unimodal) distribution
- Positive (right) skewed (mean > median > mode)
- Negative (left) skewed (mean < median < mode)

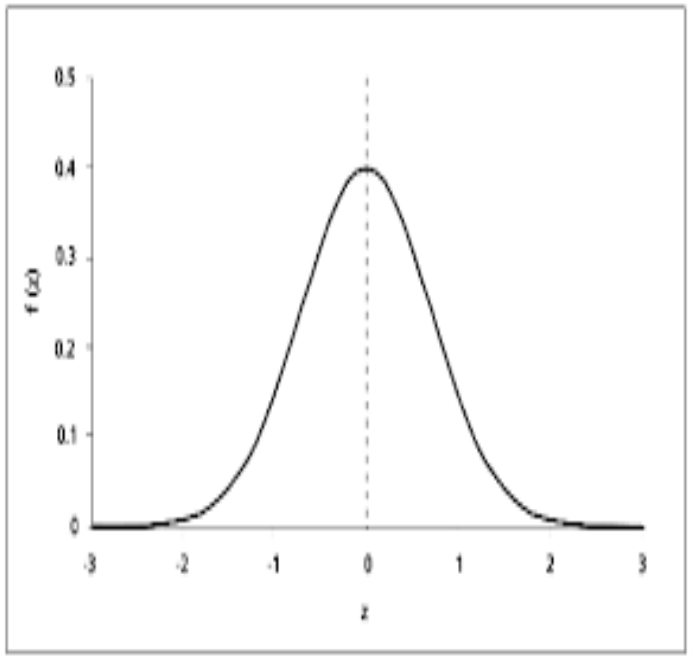


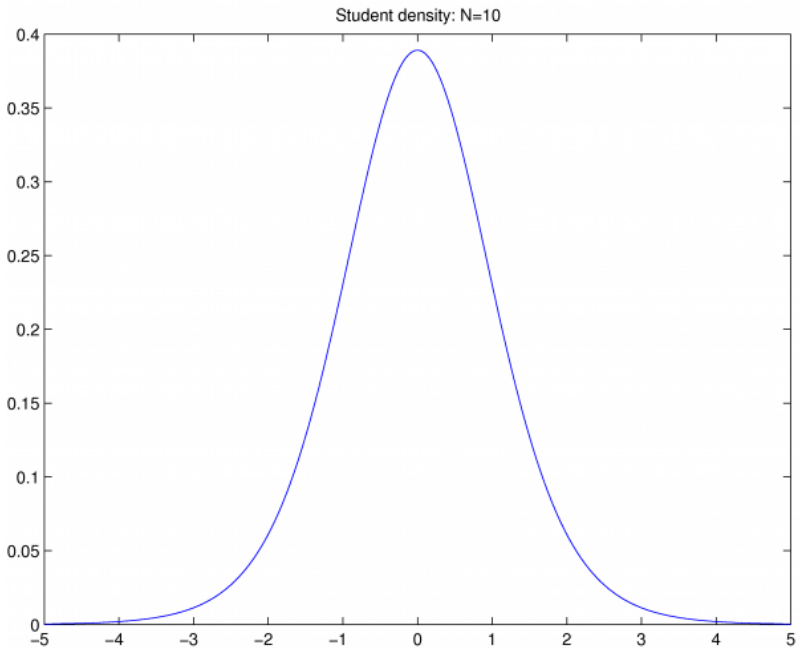
Kurtosis

- It is the fourth degree
- It is a measure of "peakedness" of the distribution

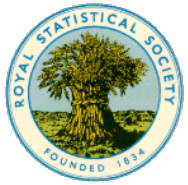
$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2} - 3$$



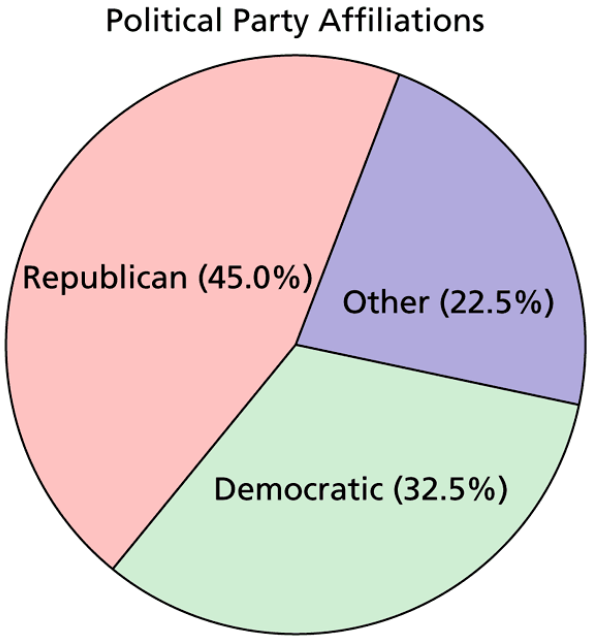


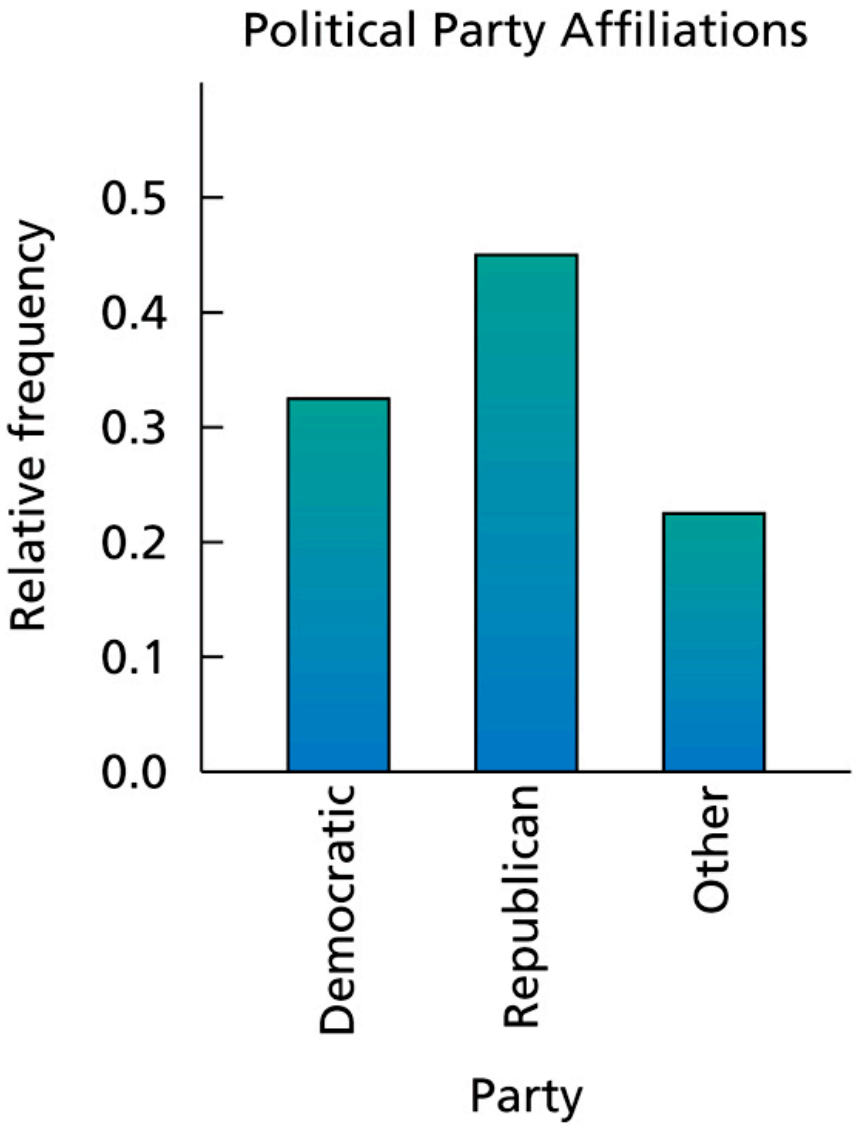


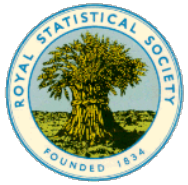
Normal Distribution			t-Distribution
Mean	=		Mean
Standard Deviation	=		Standard Deviation
Skewness	=		Skewness
Kurtosis	= 3	< 3	Kurtosis



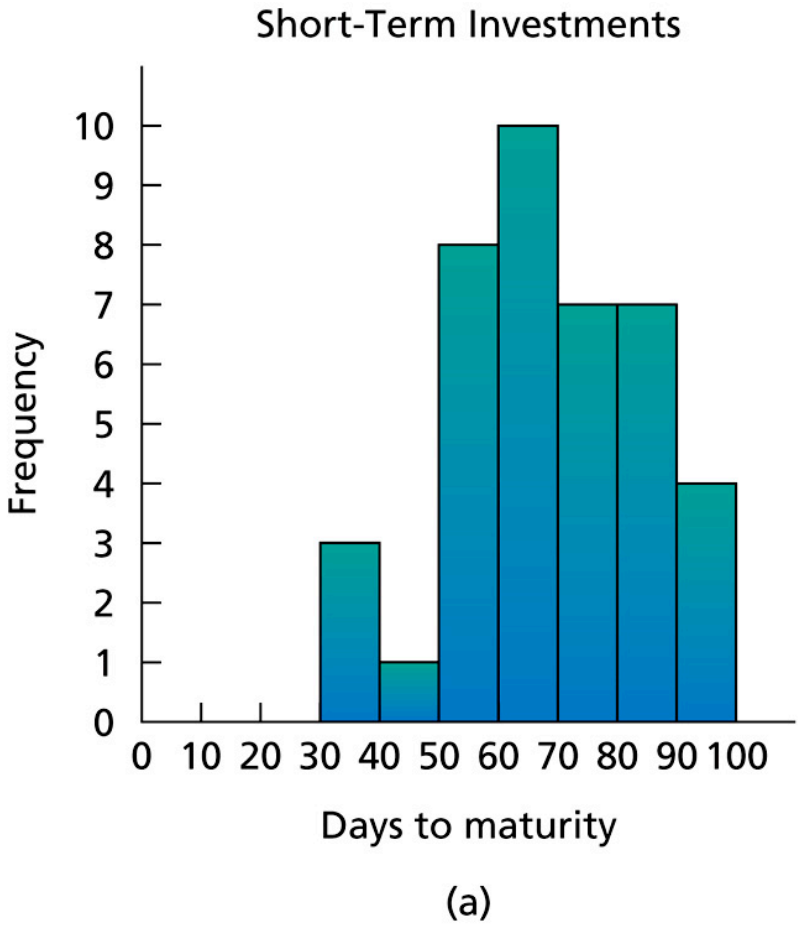
Organizing Qualitative Data

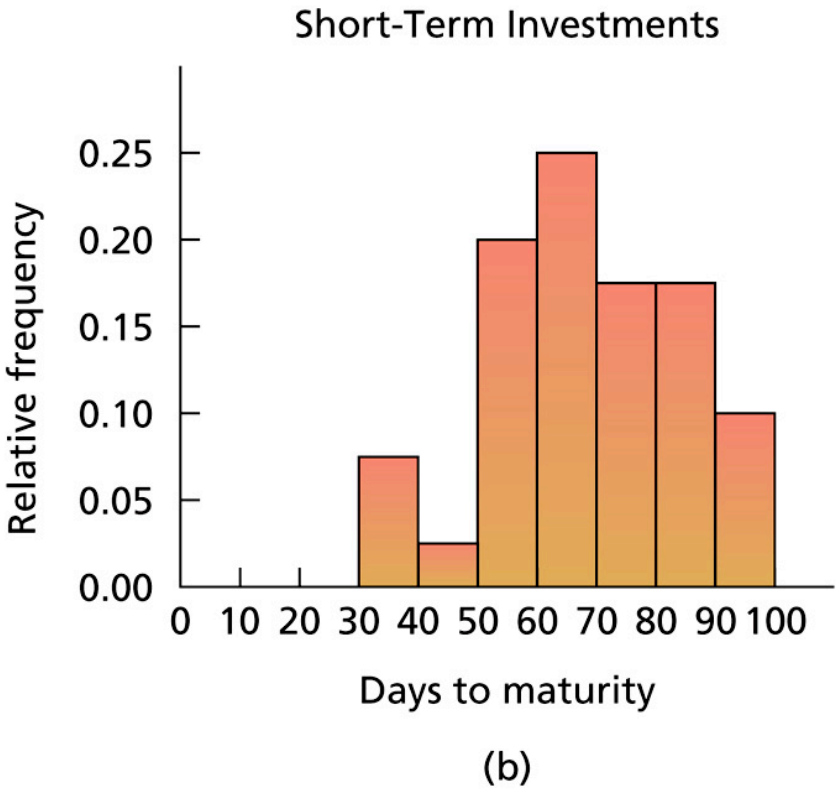

Political Party Affiliations
Republican (45.0%)
Other (22.5%)
Democratic (32.5%)


Political Party Affiliations
Relative frequency
0.0
0.1
0.2
0.3
0.4
0.5
Democratic
Republican
Other
Party

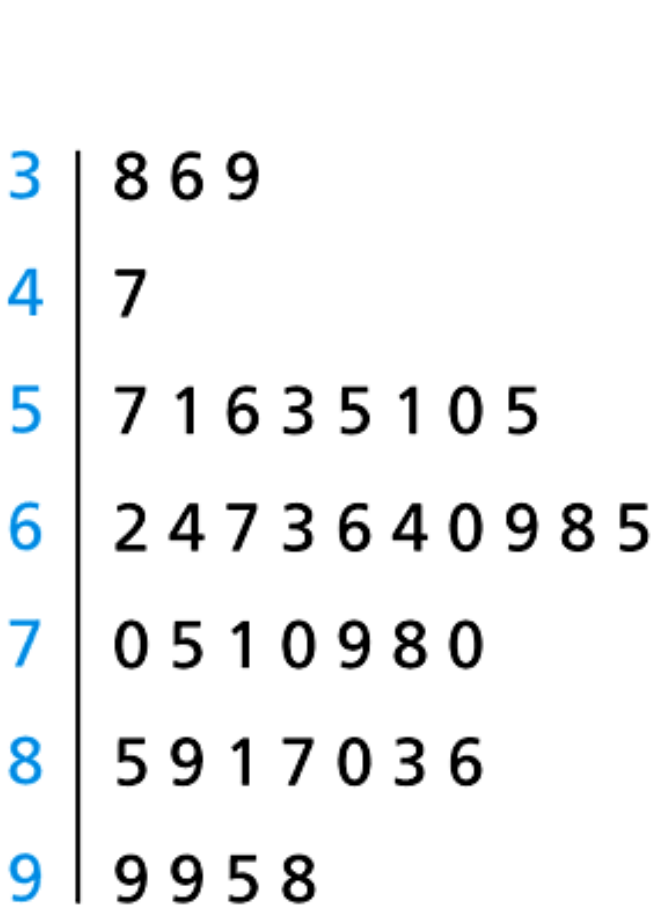


Organizing Quantitative Data


Short-Term Investments
Frequency
0
1
2
3
4
5
6
7
8
9
10
0 10 20 30 40 50 60 70 80 90 100
Days to maturity
(a)

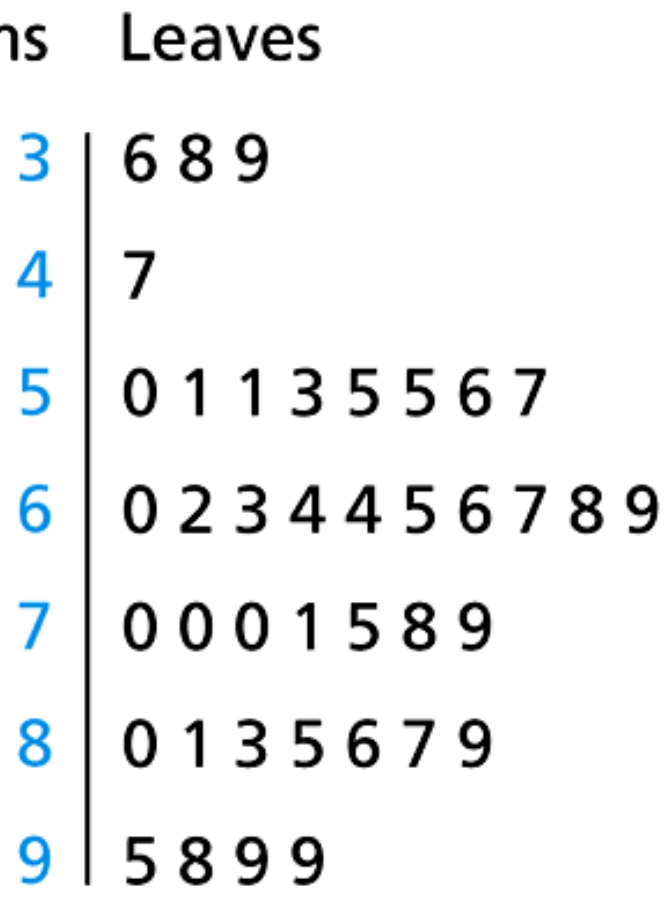

Short-Term Investments
Relative frequency
0.00
0.05
0.10
0.15
0.20
0.25
0 10 20 30 40 50 60 70 80 90 100
Days to maturity
(b)


Prices of DVD Players
190
200
210
220
230
Price ($)



3	8 6 9
4	7
5	7 1 6 3 5 1 0 5
6	2 4 7 3 6 4 0 9 8 5
7	0 5 1 0 9 8 0
8	5 9 1 7 0 3 6
9	9 9 5 8

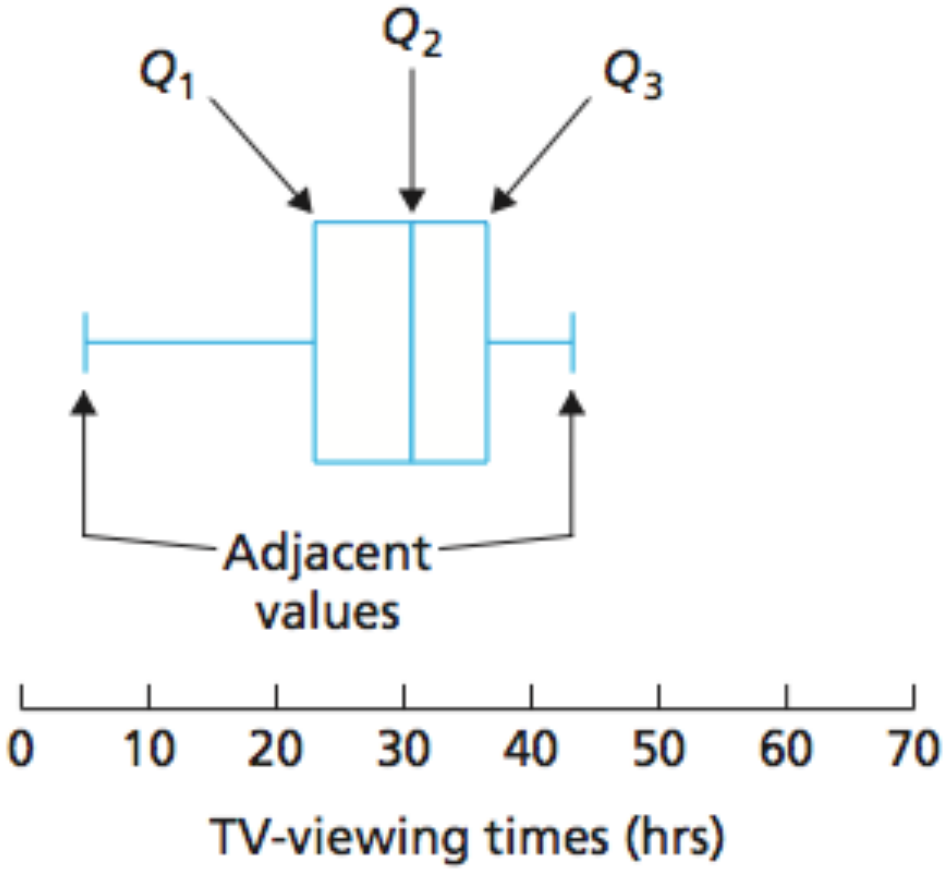
(a)

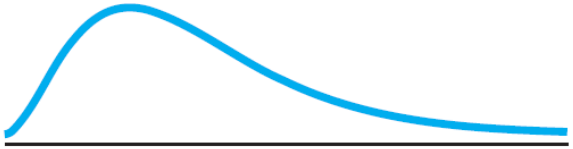


Stems	Leaves
3	6 8 9
4	7
5	0 1 1 3 5 5 6 7
6	0 2 3 4 4 5 6 7 8 9
7	0 0 0 1 5 8 9
8	0 1 3 5 6 7 9
9	5 8 9 9

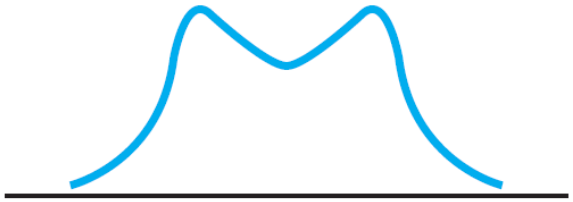
(b)

Boxplot

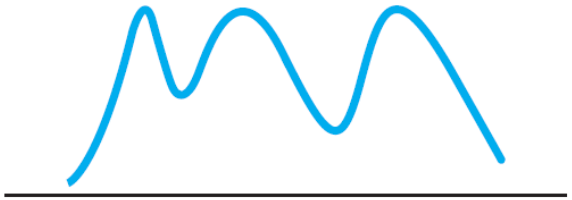

Q_1
Q_2
Q_3
Adjacent values
0
10
20
30
40
50
60
70
TV-viewing times (hrs)



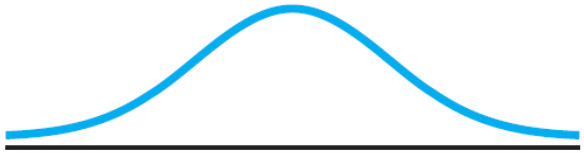
(a) Unimodal



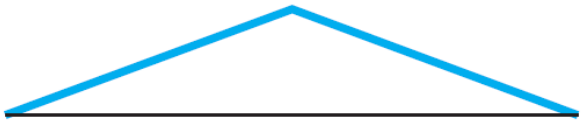
(b) Bimodal



(c) Multimodal

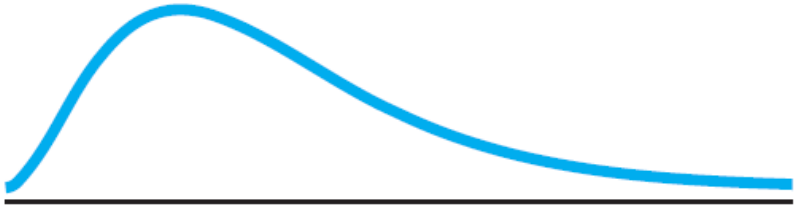


(a) Bell shaped

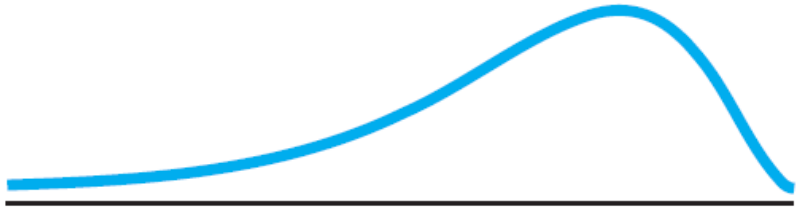


(b) Triangular

(c) Uniform (or rectangular)

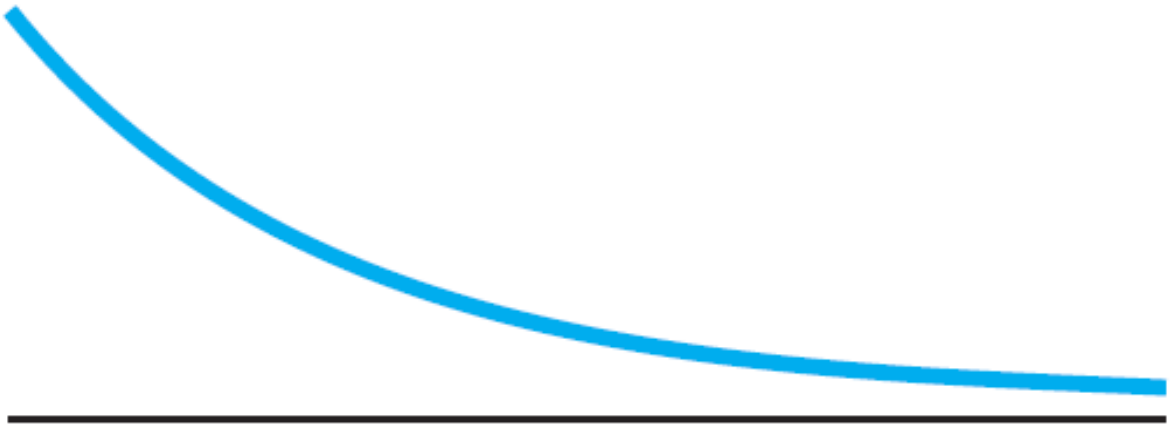


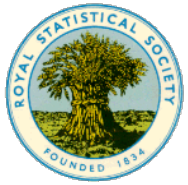
(a) Right skewed



(b) Left skewed

Reverse-J-shaped distribution





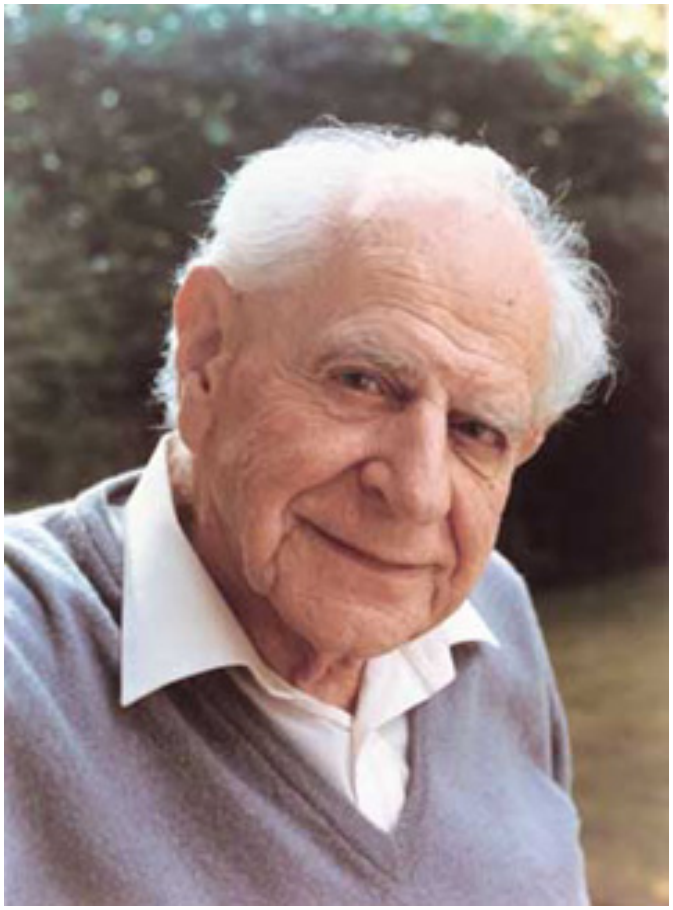
The Nature of Hypothesis Testing

Pioneers

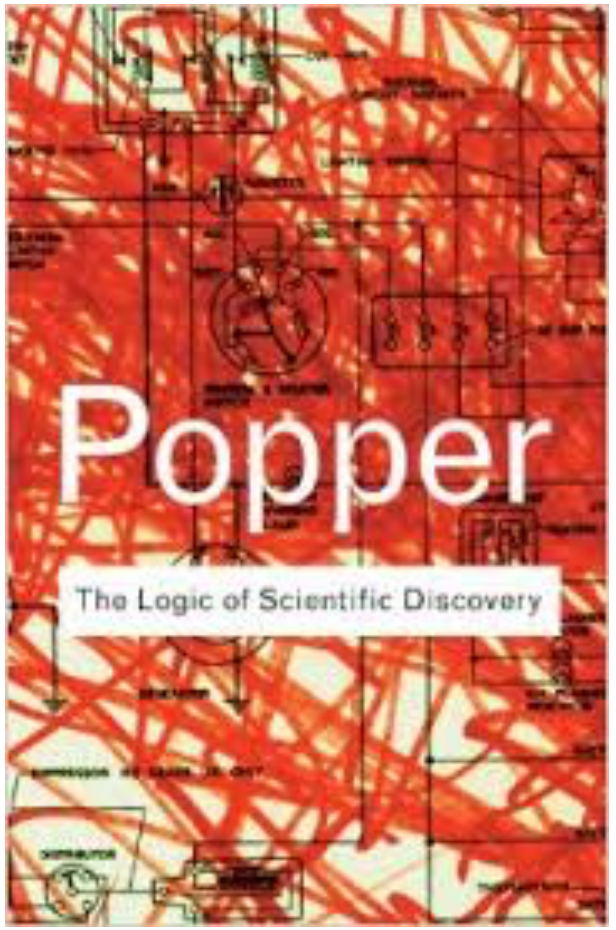
Sir Ronald Aylmer Fisher
(1890 – 1962)

William Sealy Gosset
(1876 – 1937)

Karl Pearson
(1857 – 1936)

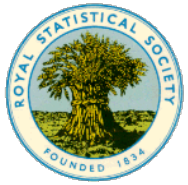


Sir Karl Raimund Popper (1902 – 1994)



The Logic of Scientific Discovery (1934)

Are all swans white?

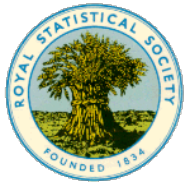


Null and Alternative Hypotheses; Hypothesis Test

Null hypothesis: A hypothesis to be tested. We use the symbol H_0 to represent the null hypothesis.

Alternative hypothesis: A hypothesis to be considered as an alternative to the null hypothesis. We use the symbol H_a to represent the alternative hypothesis.

Hypothesis test: The problem in a hypothesis test is to decide whether the null hypothesis should be rejected in favor of the alternative hypothesis.



Hypothesis test do a simple job:
REJECT or ***NOT REJECT*** the null hypothesis

Null hypothesis and Alternative hypothesis should be:
Mutually Exclusive
Collectively Exhaustive

Null Hypothesis should be ***Negation*** of
Alternative Hypothesis

Null and Alternative Hypothesis

For Unpaired Test

Null Hypothesis		Alternative Hypothesis
$\mu=\mu_0$	→	$\mu\neq\mu_0$
$\mu\geq\mu_0$	→	$\mu<\mu_0$
$\mu\leq\mu_0$	→	$\mu>\mu_0$

Null and Alternative Hypothesis

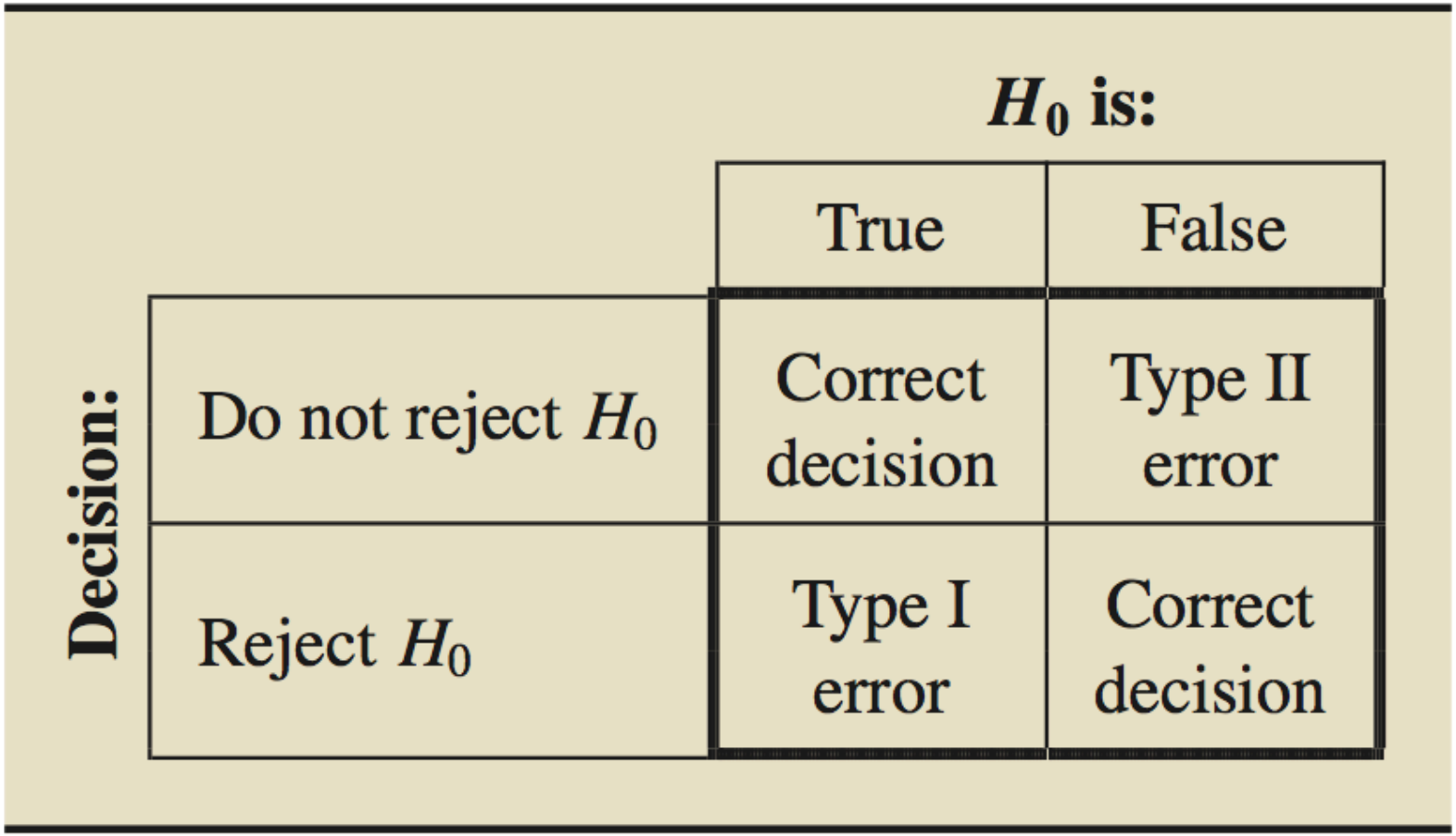
For Paired Test

Null Hypothesis		Alternative Hypothesis
$\mu_1 - \mu_2 = 0$	→	$\mu_1 - \mu_2 \neq 0$
$\mu_1 - \mu_2 \geq 0$	→	$\mu_1 - \mu_2 < 0$
$\mu_1 - \mu_2 \leq 0$	→	$\mu_1 - \mu_2 > 0$

Type I and Type II Errors

Type I error: Rejecting the null hypothesis when it is in fact true.

Type II error: Not rejecting the null hypothesis when it is in fact false.



Decision:	H_0 is: True	H_0 is: False
Do not reject H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

	Ho is True	Ho is False
Fail to Reject Ho	Innocence is correctly declared	Guilty person incorrectly set free
Reject Ho	Innocent person sent to prison	Guilt is correctly declared

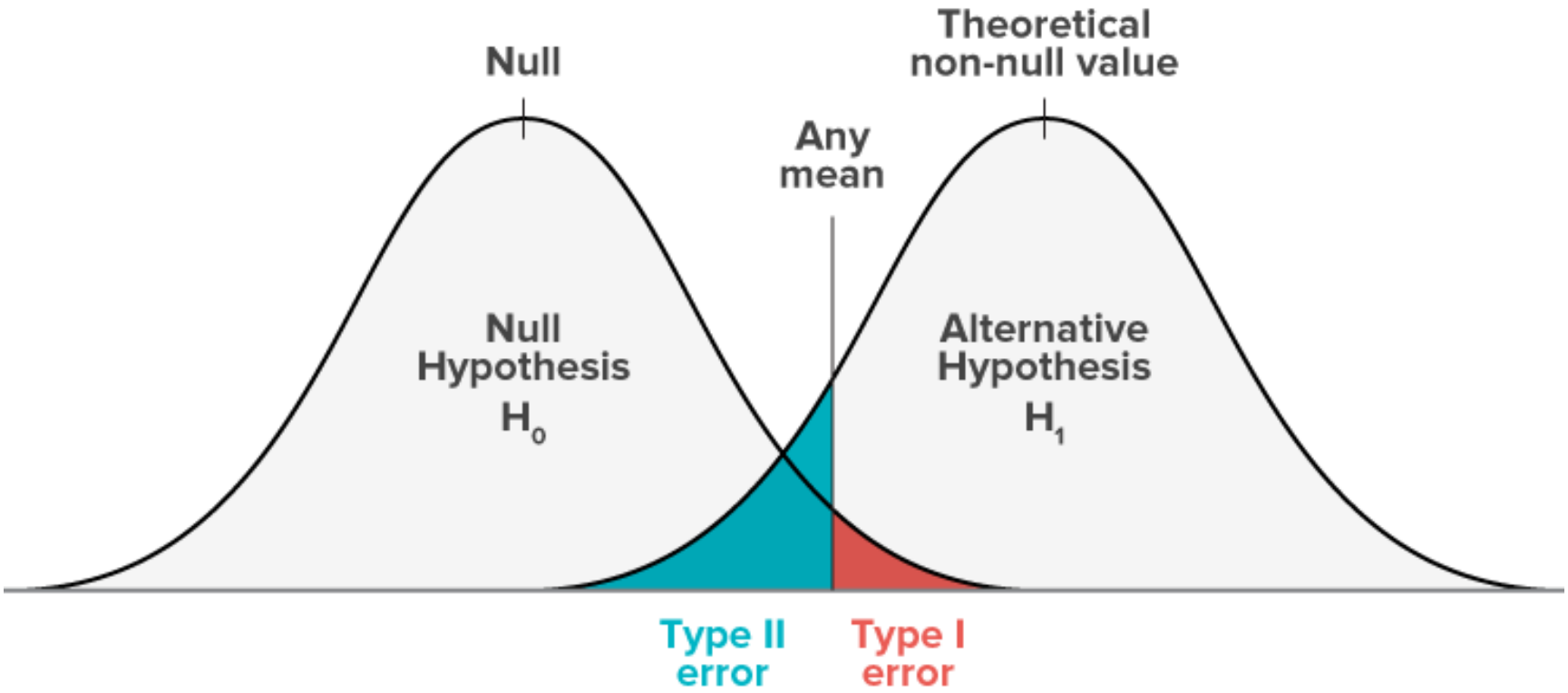
		TRUTH ABOUT THE DEFENDENT	
		Innocent	Guilty
DECISION BASED ON TRIAL EVIDENCE	Find Guilty	Type I Error	Correct decision
	Find Not Guilty	Correct decision	Type II Error

Significance Level

The probability of making a Type I error, that is, of rejecting a true null hypothesis, is called the **significance level,** α , of a hypothesis test.

Relation between Type I and Type II Error Probabilities

For a fixed sample size, the smaller we specify the significance level, α, the larger will be probability, β, of not rejecting a false null hypothesis.


Null
Theoretical non-null value
Any mean
Null Hypothesis H_0
Alternative Hypothesis H_1
Type II error
Type I error

Value of K corresponding to α and β

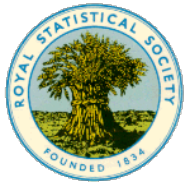
Power $(1-\beta)$	Significance level (α)		
	5%	1%	0.1%
80%	*7.8*	*11.7*	*17.1*
90%	*10.5*	*14.9*	*20.9*
95%	*13.0*	*17.8*	*24.3*
99%	*18.4*	*24.1*	*31.6*

K is a value related to sample size. When K increases, sample size will increase

Possible Conclusions for a Hypothesis Test

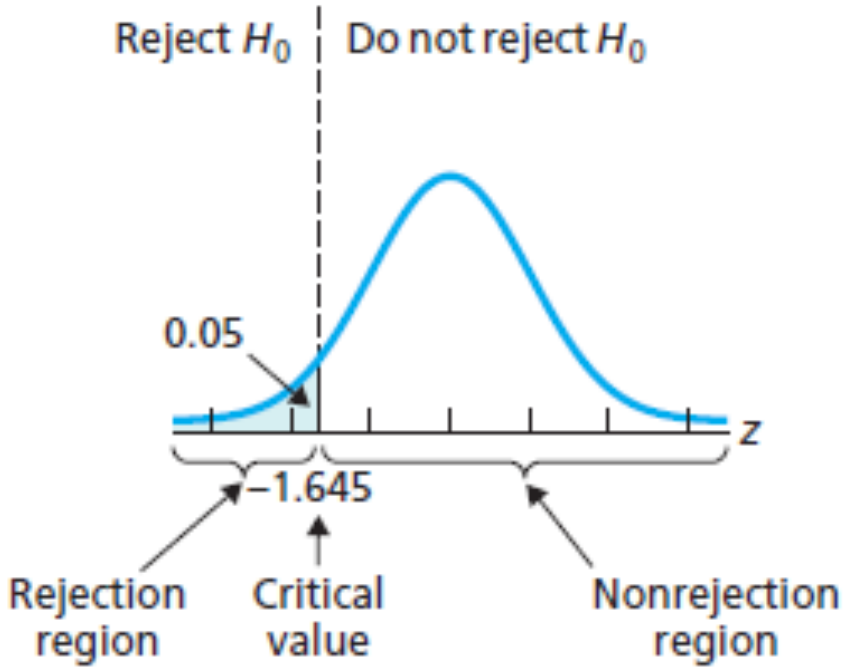
Suppose that a hypothesis test is conducted at a small significance level.

- If the null hypothesis is rejected, we conclude that the data provide sufficient evidence to support the alternative hypothesis.
- If the null hypothesis is not rejected, we conclude that the data do not provide sufficient evidence to support the alternative hypothesis.

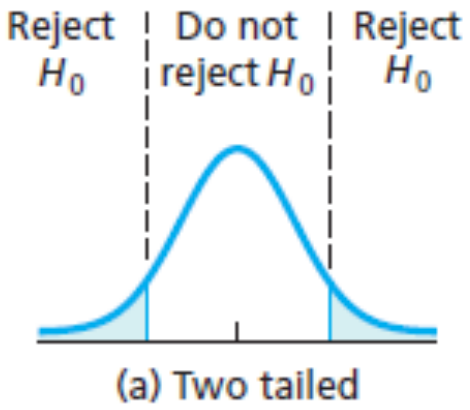


Critical-Value Approach to Hypothesis Testing

Criterion for deciding whether to reject the null hypothesis

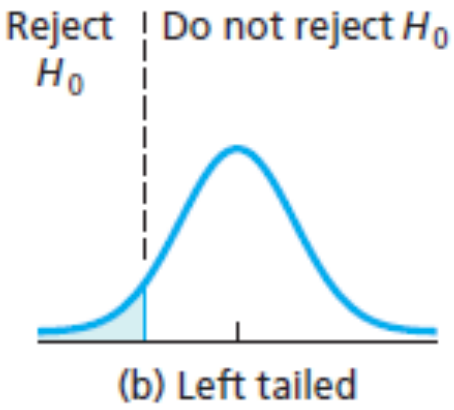


Graphical display of rejection regions for two-tailed, left-tailed, and right-tailed tests



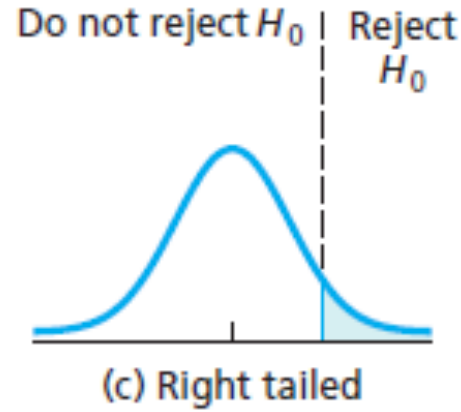
(a) Two tailed

$H_o:\mu=\mu_o$
$H_1:\mu\neq\mu_o$



(b) Left tailed

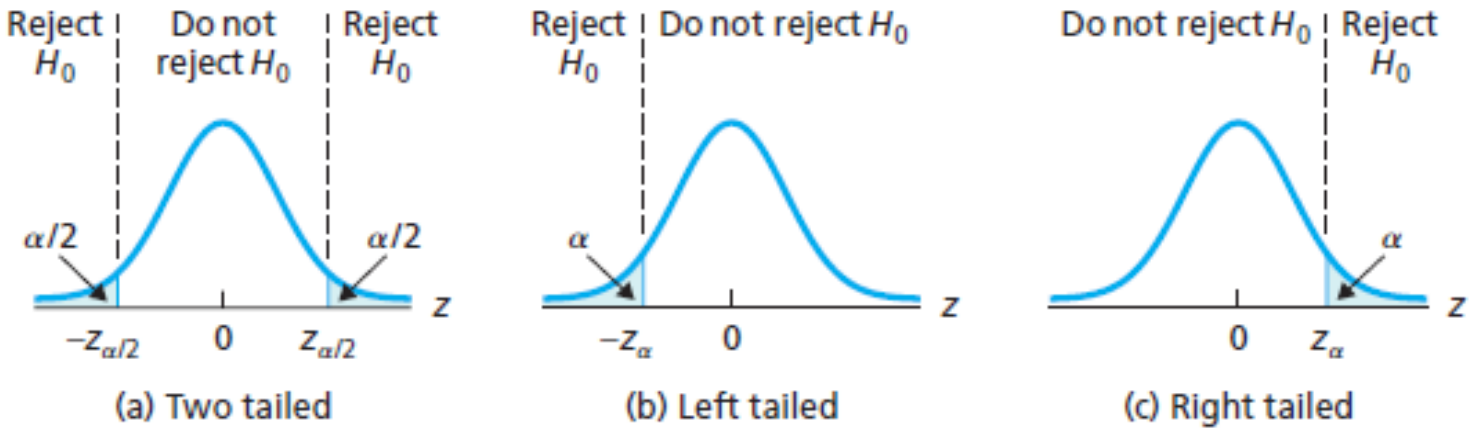
$H_o:\mu\geq\mu_o$
$H_1:\mu<\mu_o$



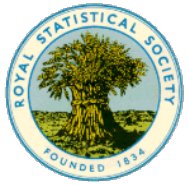
(c) Right tailed

$H_o:\mu\leq\mu_o$
$H_1:\mu>\mu_o$

Critical value(s) for a one-mean z-test at the significance level α if the test is (a) two tailed, (b) left tailed, or (c) right tailed



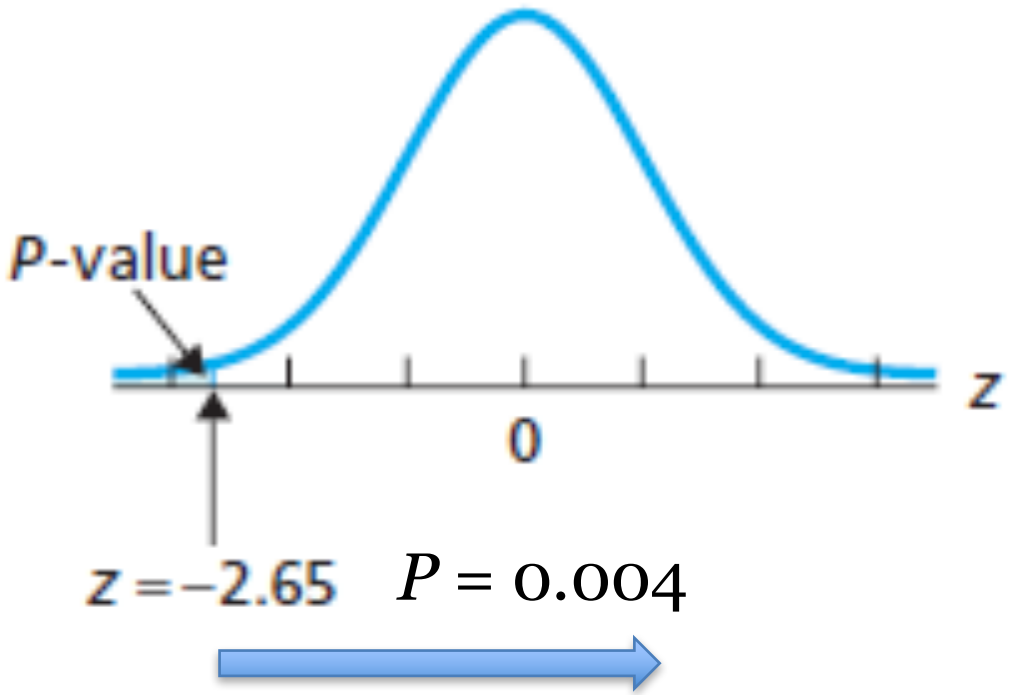
Z-value	p = 0.05	p = 0.01
Two sided test	1.959964	2.5758293
One-sided test	1.644854	2.3263479



p-Value Approach to Hypothesis Testing

P-Value

The ***P*-value** of a hypothesis test is the probability of getting sample data at least as inconsistent with the null hypothesis (and supportive of the alternative hypothesis) as the sample data actually obtained. We use the letter ***P*** to denote the *P*-value.



Null hypothesis can be rejected if significance level greater than p - value

General steps for the *P*-value approach to hypothesis testing

P-VALUE APPROACH TO HYPOTHESIS TESTING

Step 1	State the null and alternative hypotheses.
Step 2	Decide on the significance level, α.
Step 3	Compute the value of the test statistic.
Step 4	Determine the P-value, P.
Step 5	If $P \leq \alpha$, reject H_0; otherwise, do not reject H_0.
Step 6	Interpret the result of the hypothesis test.

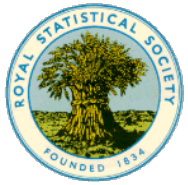
Guidelines for using the *P*-value to assess the evidence against the null hypothesis

P-value	Evidence against H_0
$P > 0.10$	Weak or none
$0.05 < P \leq 0.10$	Moderate
$0.01 < P \leq 0.05$	Strong
$P \leq 0.01$	Very strong

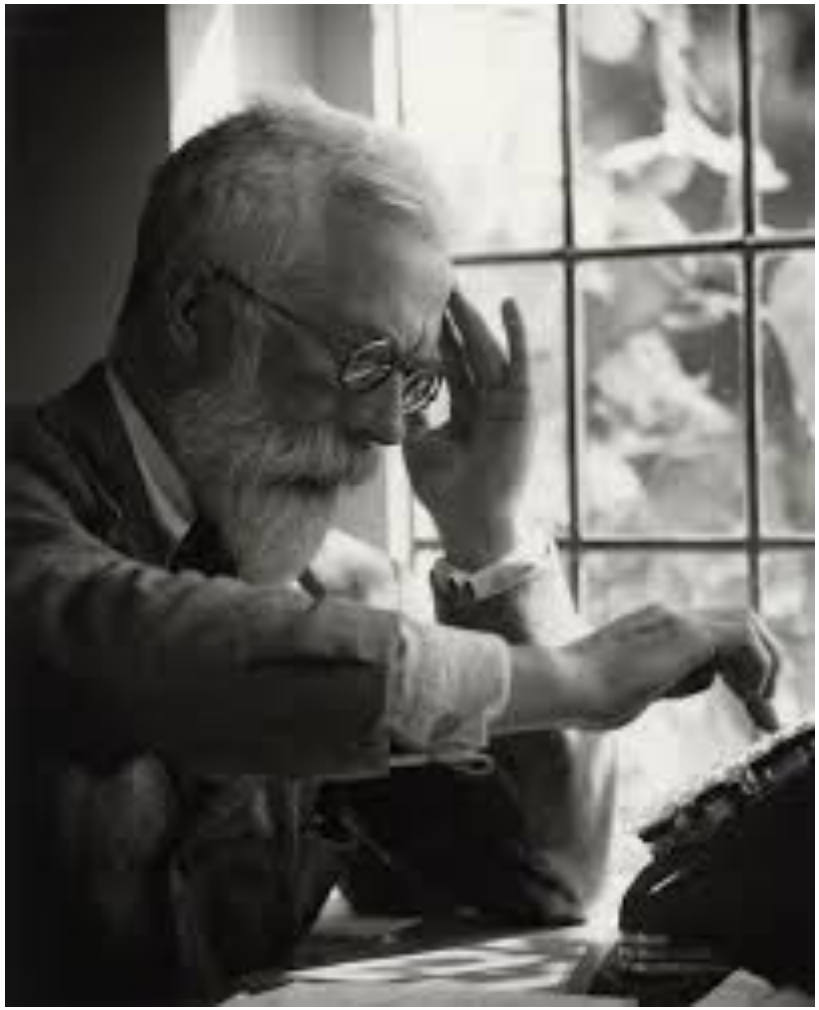
Hypothesis Tests Without Significance Levels: Many researchers do not explicitly refer to significance levels. Instead, they simply obtain the *P*-value and use it (or let the reader use it) to assess the strength of the evidence against the null hypothesis.

Null hypothesis testing only answer a true or false question

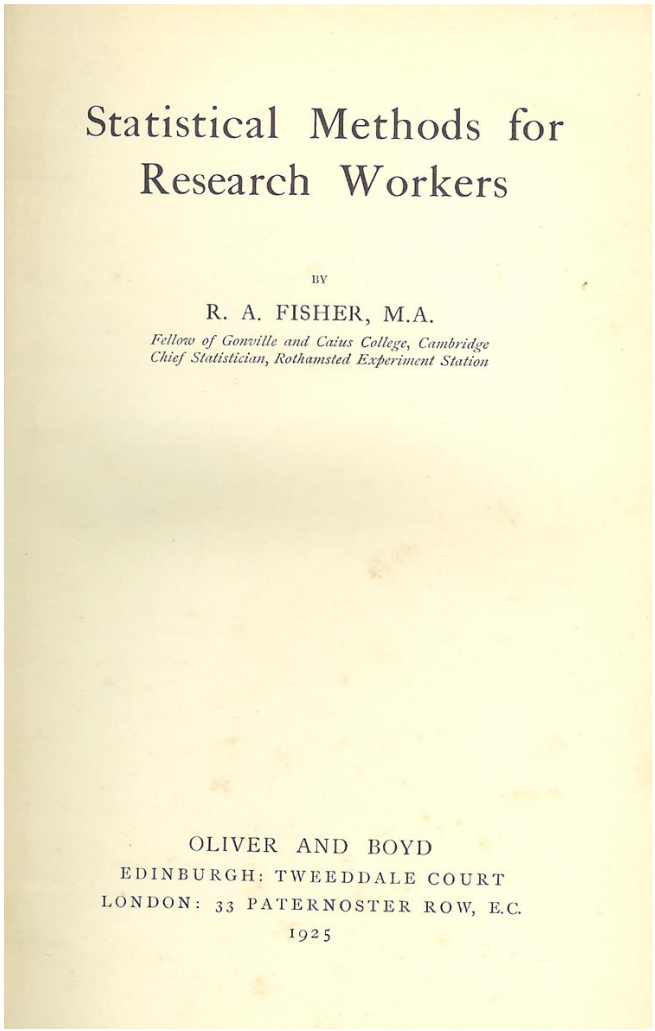
You must ask more than one “correct” questions to view a simple fact.



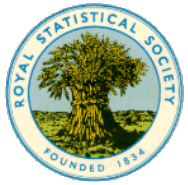
Significance level, conventionally, at 5%???



Sir Ronald Aylmer Fisher
(1890 – 1962)



Statistical Method for
Research Workers (1925)



Tables I. and II. have been constructed to show the deviations corresponding to different values of this probability. The rapidity with which the probability falls off as the deviation increases is well shown in these tables. A deviation exceeding the standard deviation occurs about once in three trials. Twice the standard deviation is exceeded only about once in 22 trials, thrice the standard deviation

Statistical Methods for Research Workers (1925), p.44

only once in 370 trials, while Table II. shows that to exceed the standard deviation sixfold would need nearly a thousand million trials. The value for which $P = \cdot 05$, or 1 in 20, is 1·96 or nearly 2 ; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant.

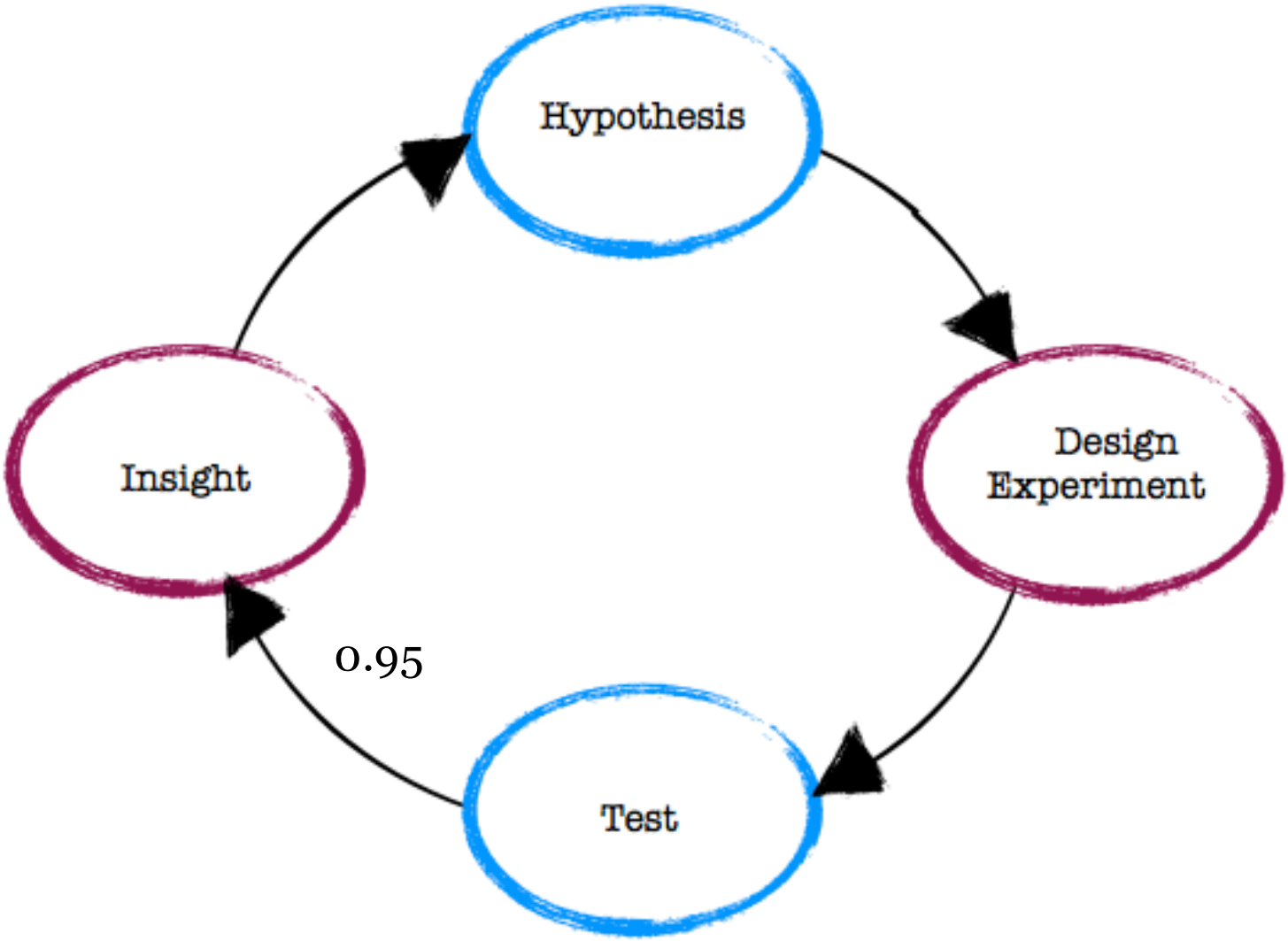
Statistical Methods for Research Workers (1925), p.45

TABLE VI

TABLE OF z. $P = \cdot 05$

Values of n_2.	Values of n_1.									
	1.	2.	3.	4.	5.	6.	8.	12.	24.	∞.
1	2·5421	2·6479	2·6870	2·7071	2·1974	2·7276	2·7380	2·7484	2·7588	2·7693
2	1·4592	1·4722	1·4765	1·4787	1·4800	1·4808	1·4819	1·4830	1·4840	1·4851
3	1·1577	1·1284	1·1137	1·1051	1·0994	1·0953	1·0899	1·0842	1·0781	1·0716
4	1·0212	·9690	·9429	·9272	·9168	·9093	·8993	·8885	·8767	·8639
5	·9441	·8777	·8441	·8236	·8097	·7997	·7862	·7714	·7550	·7368
6	·8948	·8188	·7798	·7558	·7394	·7274	·7112	·6931	·6729	·6499
8	·8355	·7475	·7014	·6725	·6525	·6378	·6175	·5945	·5682	·5371
12	·7788	·6786	·6250	·5907	·5666	·5487	·5234	·4941	·4592	·4156
24	·7246	·6123	·5508	·5106	·4817	·4598	·4283	·3904	·3425	·2749
∞	·6729	·5486	·4787	·4319	·3974	·3706	·3309	·2804	·2085	0

VII. Interclass correlations and the Analysis of Variance

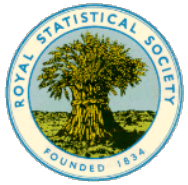

Hypothesis
Design Experiment
Test
0.95
Insight

In Series

If $\alpha = 0.05$

P (no Type I error after n cycles)
$= (1 - 0.05)^n$

When $n > 13$, $P < 0.5$



In Parallel

If we repeat the same study for 20 times, there is one type I error, that is, wrongly reject the true null hypothesis.

Usually, the result will be published because it is “significant”, “new unusual finding” and “eyeball-catching”.

in brief

A psychology journal bans p-values

Basic and Applied Social Psychology abandons the null hypothesis significance test

BASIC AND APPLIED SOCIAL PSYCHOLOGY

conclude that the observed effect is unlikely to be the result of chance, hence there is an effect. By convention, researchers say the result is "statistically significant" in that case.

But that is not enough to say the null hypothesis is false, argues Trafimow. "If the computed *p*-value gave the probability of the null hypothesis given the finding, a low value would provide solid ground for rejecting the null hypothesis. But the computed *p*-value is the probability of the finding (or a more extreme one) given that the null hypothesis is true. In other words, the computed *p*-value is the inverse probability of the one that is necessary to have solid ground for rejecting the null hypothesis."

The *BASP* ban on the NHSTP stems from this "logical invalidity", says Trafimow. He stresses that the journal is not doing away with all formal inferences: only the NHSTP and "the use of confidence intervals as an alternative way to reject null hy... is, as a way to circumv... However, the ...

This willingness to discard the null might explain some of the more surprising papers that make up the psychology literature. Columbia University's Andrew Gelman writes in an upcoming *Significance* article that: "Top journals in psychology routinely publish ridiculous, scientifically implausible claims, justified based on '$p < 0.05$:'" (See the June 2015 issue for more on this.)

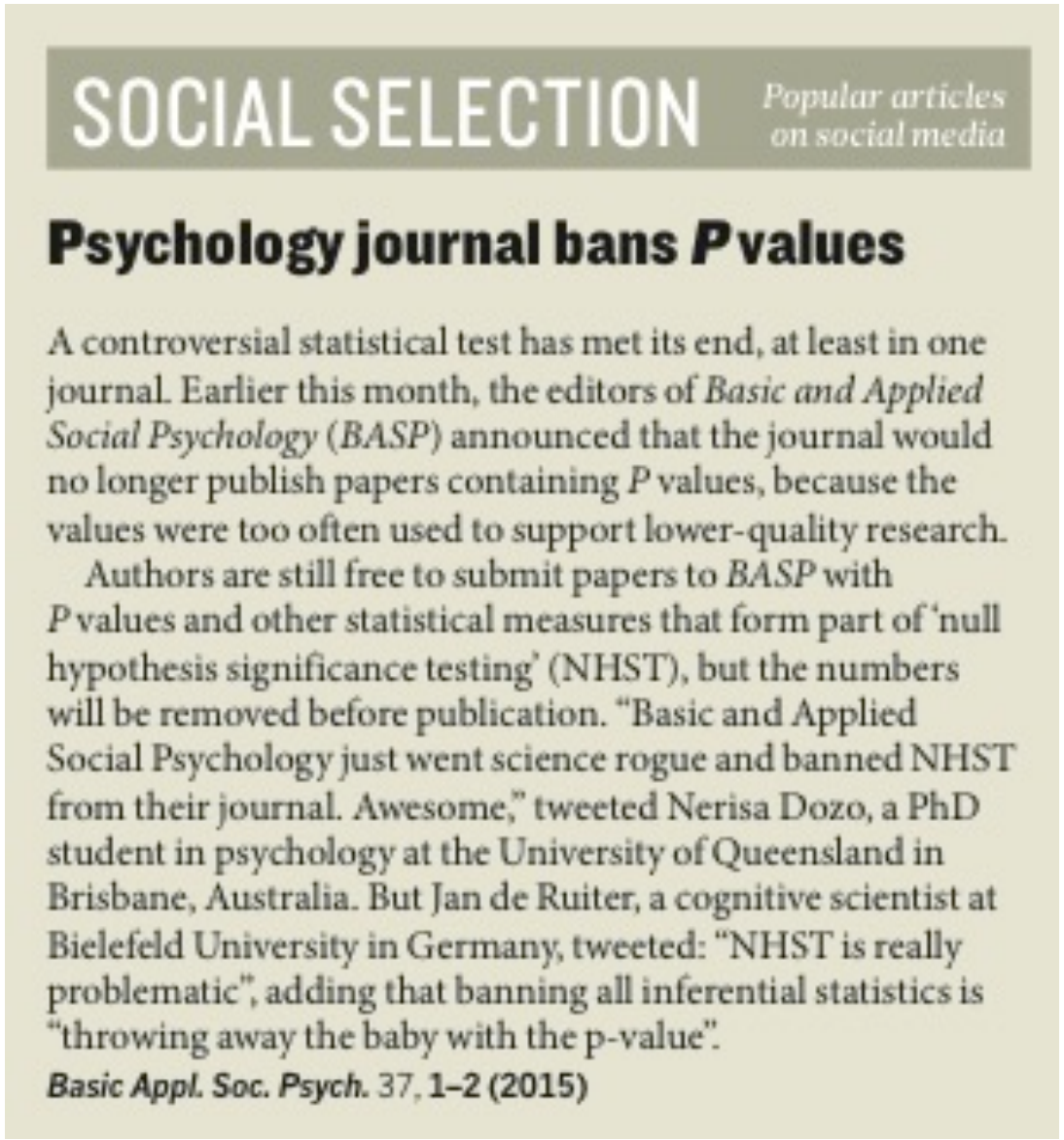
Responding to the *BASP* editorial, Gelman said: "I think standard errors, *p*-values, and confidence intervals can be very helpful in research when considered as convenient parts of a data analysis ... The problem comes when they're considered as the culmination of the analysis, as if '*p* less than 0.05' represents som... kind of proof of something. I do ... of requiring that rese... own witho...

Editorial

David Trafimow and Michael Marks
New Mexico State University

The *Basic and Applied Social Psychology* (BASP) 2014 Editorial emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid, and thus authors would be not required to perform it (Trafimow, 2014). However, to allow authors a grace period, the Editorial stopped short of actually banning the NHSTP. The purpose of the present Editorial is to announce that the grace period is over. From now on, BASP is banning the NHSTP.

Basic and Applied Social Psychology, 37: 1-2, 12 Feb 2015



SOCIAL SELECTION *Popular articles on social media*

Psychology journal bans *P* values

A controversial statistical test has met its end, at least in one journal. Earlier this month, the editors of *Basic and Applied Social Psychology* (*BASP*) announced that the journal would no longer publish papers containing *P* values, because the values were too often used to support lower-quality research.

Authors are still free to submit papers to *BASP* with *P* values and other statistical measures that form part of 'null hypothesis significance testing' (NHST), but the numbers will be removed before publication. "Basic and Applied Social Psychology just went science rogue and banned NHST from their journal. Awesome," tweeted Nerisa Dozo, a PhD student in psychology at the University of Queensland in Brisbane, Australia. But Jan de Ruiter, a cognitive scientist at Bielefeld University in Germany, tweeted: "NHST is really problematic", adding that banning all inferential statistics is "throwing away the baby with the p-value".

Basic Appl. Soc. Psych. 37, **1–2 (2015)**

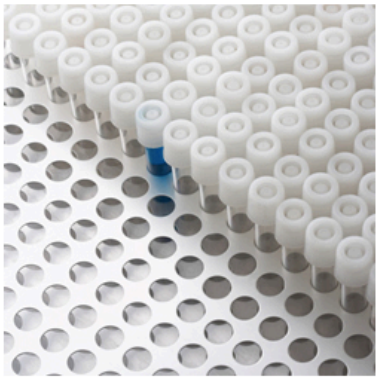
Nature (26th February 2015)

StatsLife

Home | News | Features | Opinion | Significance | Videos | Resources | Events | Jobs | Careers | Members

Academic journal bans p-value significance test

Written by Web News Editor on 05 March 2015. Posted in News



An editorial published in the academic journal *Basic and Applied Social Psychology* (BASP) has declared that the null hypothesis significance testing procedure (NHSTP) is 'invalid', and have banned it from future papers submitted to the journal.

In significance testing, a null hypothesis typically posits that changing the input to a system does not change the resulting output. To determine whether a result is statistically significant, a researcher has to calculate a p-value, which is the probability of observing an apparent effect given that the null hypothesis is true. If the p-value is less than 0.05 it is conventionally deemed a statistically significant result.

'We believe that the $p < 0.05$ bar is too easy to pass and sometimes serves as an excuse for lower quality research,' BASP authors David Trafimow and Michael Marks of Mexico State University write in their editorial. They go on to say that they hope other journals will follow suit.

Confidence intervals are also banned, since they 'do not provide a strong case for concluding that the population parameter of interest is likely to be within the stated interval.' Bayesian alternatives, however, 'are neither required nor banned' from the journal.

The use of p-values to judge whether research is 'significant' has been thought by many statisticians to be problematic, ever since the idea was introduced by Ronald Fisher in the 1920s. In 2005, Stanford School of Medicine professor John Ioannidis wrote a well-known paper, 'Why Most Published Research Findings Are False', stating that reliance on p-values alone was 'ill-founded'.

More recently, in November 2014, Royal Society Fellow David Colquhoun published 'An investigation of the false discovery rate and the misinterpretation of p-values' stating: 'If you use p=0.05 to suggest that you have made a discovery, you will be wrong at least 30% of the time'.

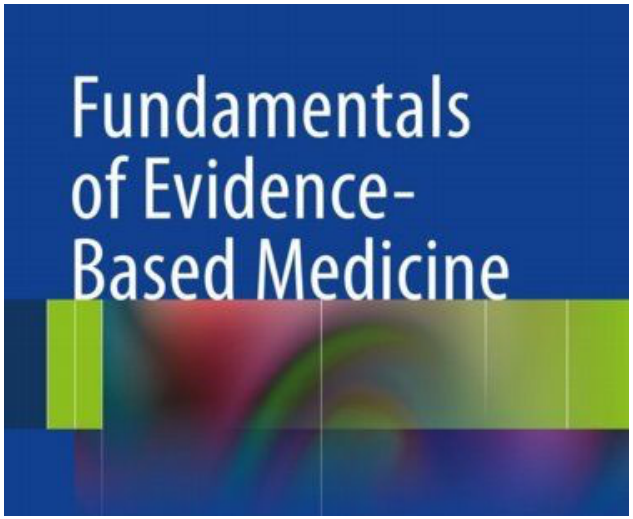
The American Statistical Association (ASA) argues that banning p-values altogether will have other 'negative consequences' in a formal statement regarding the editorial. 'The statistical community is aware of problems associated with the use and interpretation of inferential methods,' it says. 'However, the journal proposes to fall back entirely on descriptive statistics and use "larger sample sizes than is typical in much psychology research." We believe this policy may have its own negative consequences and thus the proper use of inferential methods needs to be analyzed and debated in the larger research community.'

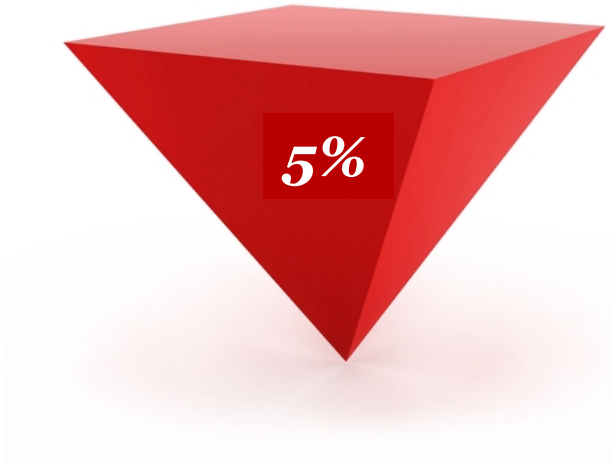
RSS president, Peter Diggle, said: 'The RSS welcomes and shares the BASP editors' concerns that inferential statistical methods are open to mis-use and mis-interpretation, but does not feel that a blanket ban on any particular inferential method is the most constructive response.'

StatsLife (5th March 2015)

In November 2014, Royal Society Fellow **David Colquhoun** published '*An investigation of the false discovery rate and the misinterpretation of p-values*' stating: '***If you use p=0.05 to suggest that you have made a discovery, you will be wrong at least 30% of the time***'.

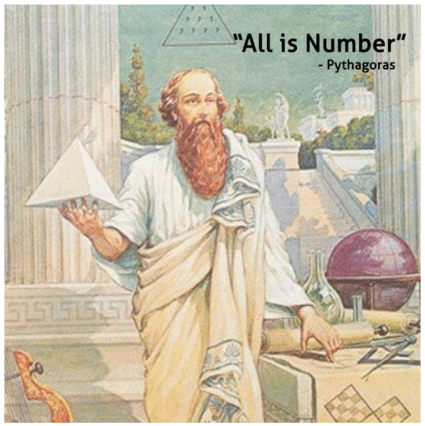
Evidence-Based medicine relies on research with 5% significance level





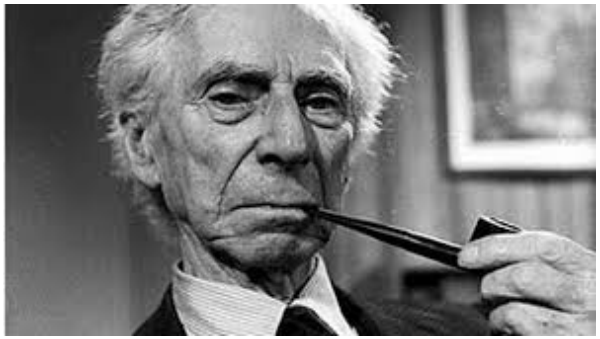
5% Significance Level itself is not Evidence Based

Three Mathematical Crises



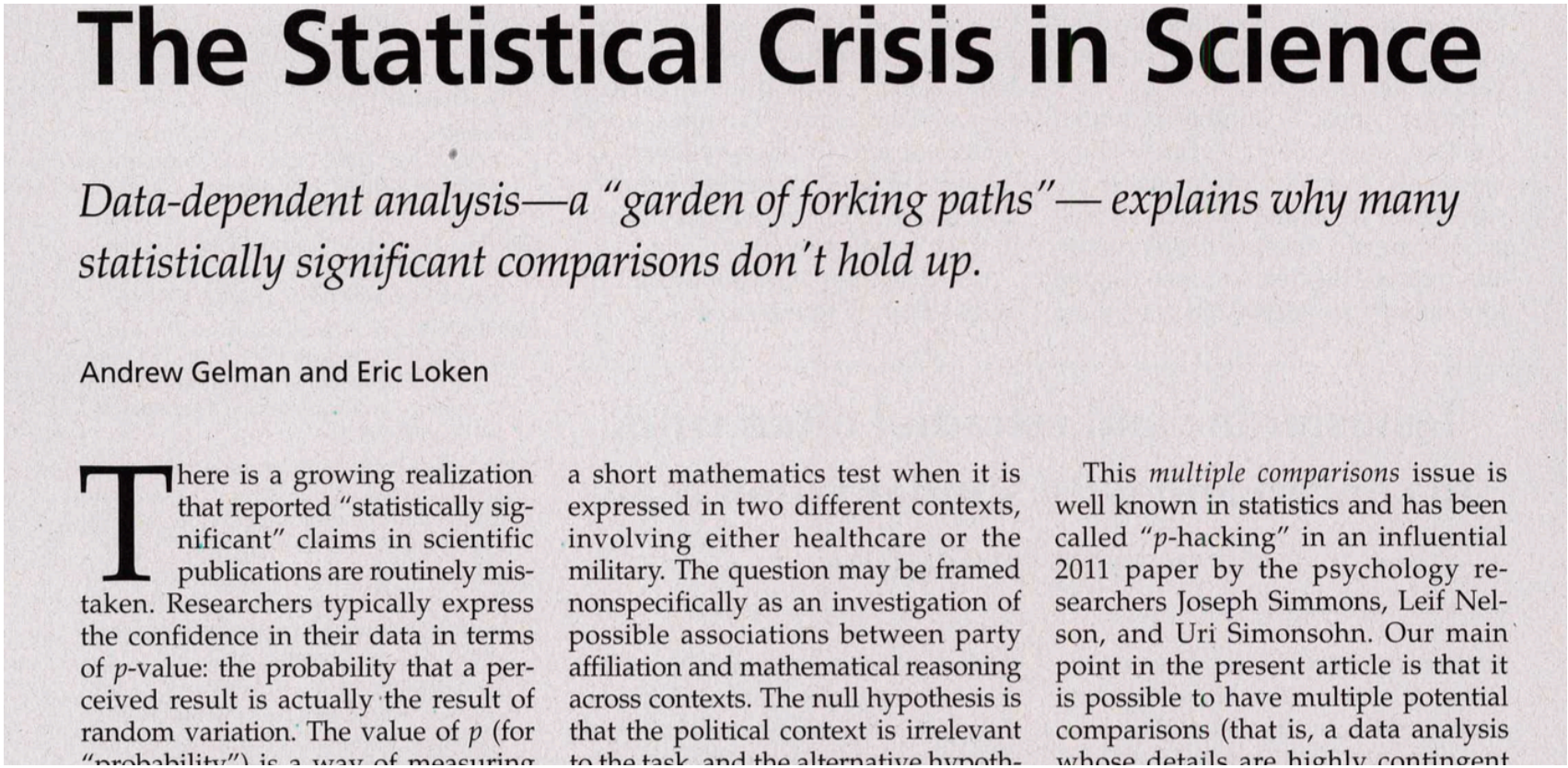
5th century BC

17th century



20th century

Frist Statistical Crisis



The Statistical Crisis in Science

Data-dependent analysis—a "garden of forking paths"— explains why many statistically significant comparisons don't hold up.

Andrew Gelman and Eric Loken

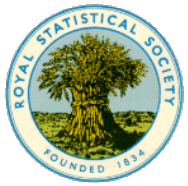
There is a growing realization that reported "statistically significant" claims in scientific publications are routinely mistaken. Researchers typically express the confidence in their data in terms of *p*-value: the probability that a perceived result is actually the result of random variation. The value of *p* (for

a short mathematics test when it is expressed in two different contexts, involving either healthcare or the military. The question may be framed nonspecifically as an investigation of possible associations between party affiliation and mathematical reasoning across contexts. The null hypothesis is that the political context is irrelevant

This *multiple comparisons* issue is well known in statistics and has been called "*p*-hacking" in an influential 2011 paper by the psychology researchers Joseph Simmons, Leif Nelson, and Uri Simonsohn. Our main point in the present article is that it is possible to have multiple potential comparisons (that is, a data analysis

American Scientist (Nov-Dec 2014)


EXIT
FROM CRISIS



If you use $p = 0.05$ to suggest that you have made a discovery, you will be wrong at least 30% of the time. If, as is often the case, experiments are underpowered, you will be wrong most of the time. This conclusion is demonstrated from several points of view. First, tree diagrams which show the close analogy with the screening test problem. Similar conclusions are drawn by repeated simulations of *t*-tests. These mimic what is done in real life, which makes the results more persuasive. The simulation method is used also to evaluate the extent to which effect sizes are over-estimated, especially in underpowered experiments. A script is supplied to allow the reader to do simulations themselves, with numbers appropriate for their own work. It is concluded that if you wish to keep your false discovery rate below 5%, you need to use a three-sigma rule, or to insist on $p \leq 0.001$. And *never* use the word 'significant'.

David Colquhoun '*An investigation of the false discovery rate and the misinterpretation of p-values*', ***Royal Society Open Science***, 19th November 2014


GLITCH
PROBLEM
DISASTER


Thank
you!