Introductory Biostatistics

Understanding and reporting research results including P-values and Confidence Intervals

Julie Marsh
UWA, School of Mathematics and Statistics
Snr Research Fellow, Telethon Kids Institute

7 June 2019

Introductory Biostatistics

CONTENTS:

Government of Western Australia
Child and Adolescent Health Service

RESEARCH SKILLS SEMINAR SERIES 2019
CAHS Research Education Program

Introductory Biostatistics

Understanding and reporting research results including P-Values and Confidence Intervals

Dr Julie Marsh
Snr Research Fellow, Telethon Kids Institute

Research Skills Seminar Series | CAHS Research Education Program
Department of Child Health Research | Child and Adolescent Health Service

ResearchEducationProgram.org

Overview

- Observational vs Experimental Design (randomisation)
- Chance, Bias, & Confounding
- Statistical analyses (hypothesis testing)
- Understanding confidence intervals
- Interpreting P-vales
- Standardised reporting
- Where can I get statistical help?

2

Statisticians speak in code but they hide it in everyday language

3

Observational vs Experimental Design

4

Observational Design

An **Observational study** draws **inferences** from a sample to a population, where the independent variable is not under the control of the researcher because of the ethical concerns or logistical constraints.

Inference is where we infer (guess!) the properties of a population,

i.e. we may assume that the heights of the people in this room are normally distributed

5

Observational Design

An **Observational study** draws inferences from a **sample** to a population, where the independent variable is not under the control of the researcher because of the ethical concerns or logistical constraints.

A **sample** is a set of data collected or selected from a population using a defined procedure,

i.e. we may choose to only measure the heights of the people in the first 5 rows of the auditorium

6

Observational Design

An **Observational study** draws inferences from a sample to a **population**, where the independent variable is not under the control of the researcher because of the ethical concerns or logistical constraints.

A **statistical population** is all members of a defined group,

i.e. everyone currently sitting in the auditorium

7

Observational Design

An **Observational study** draws inferences from a sample to a population, where the **independent variable** is not under the control of the researcher because of the ethical concerns or logistical constraints.

The **independent variable** is the variable that is changed or controlled in an experimental design or simply observed in an observational design,

i.e. exposure to asbestos or an intervention, such as a drug

8

Observational Design

An **Observational study** draws inferences from a sample to a population, where the independent variable is not under the **control** of the researcher because of the ethical concerns or logistical constraints.

In **observational design**, **statistical control** often refers to the technique of separating out the effect of one independent variable on the outcome (eg. exposed or not exposed).
In an **experimental designs**, it usually refers to the use of a comparator group (i.e. placebo or standard care) and assumes all other factors are constant.

9

Experimental Design

Experimental design refers to a plan for assigning units (patients, mice, etc.) to treatments or interventions, whilst holding other factors constant. A good design addresses:

- **Association**
- **Control**
- **Variability**

10

Experimental Design

- **Association**

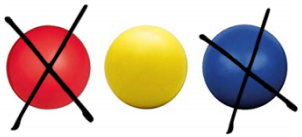
This is the relationship between the independent variable and the dependent variable (outcome, endpoint). We use the word association as it may not be possible to determine if the relationship is causal.

11

Experimental Design

- **Control**

Allows the experimenter to rule out alternative explanations due to the confounding effects of factors other than the independent variable.



12

Experimental Design

- **Variability**

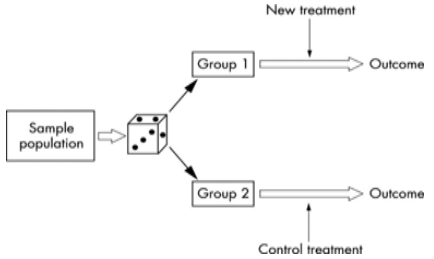
Need to minimise variability in the outcome to maximise study efficiency, which is often by reducing the number of units (e.g. patients) studied.

- Measurement variability
- Assessor variability
- Homogeneous units

13

Experimental Design

- The highest level of evidence for causality in medical research comes **Randomised Controlled Trials (RCT)**



- Patients are randomly allocated to treatment groups

Randomisation

- Often not possible to control for all factors, randomisation distributes patients with these factors equally across treatment groups
- Randomisation minimises potential **bias** from factors other than independent variable, *including some not even measured*
- Important to report baseline demographics for each treatment group in a RCT

15

Randomisation

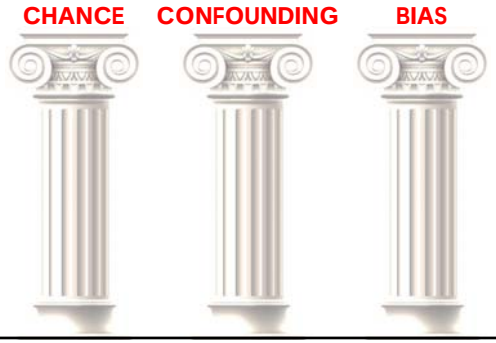
Generalisability (external validity) **of the results is determined by the inclusion/exclusion criteria**

16

Chance, Bias, and Confounding

Chance, Bias and Confounding

- The critical assessment of research results must rule out or minimise:



Chance and Bias

- **Chance**

  Statisticians focus on the likelihood that a result or relationship is caused by something other than random chance

- **Bias**

  Where a systematic error in the design, recruitment, data collection or analysis, results in an incorrect estimation of the true effect of the exposure/intervention on the outcome

19

Bias

- **Selection Bias**
  - E.g. Farmers may appear healthier than non-farmers as those in poor health may move out of the occupation
- **Recall Bias**
  - E.g. Individuals diagnosed with disease (cases) may be more likely to remember exposure that non-cases

20

Bias

- **Efficacy Bias**
  - If treatment allocation is known, then assessment of outcome by either the participant or study personnel may be distorted
- **Survival Bias**
  - Outcome is dependent on the participant surviving or tolerating treatment until the assessment date

21

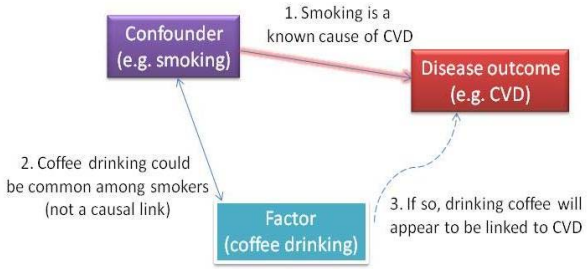
Bias

- **Treatment Allocation Bias**
  - Choice of treatment should not be related to prognostic factors
    - E.g. disease, severity, age etc.
- **Protocol/Study Deviations**
  - E.g. Cross contamination in a school-based intervention programme
  - E.g. Not taking study medication
    (hence RCT include both intent-to-treat and per-protocol analyses)

22

Confounding Factors

- Where the relationship between intervention or exposure and outcome is distorted by another variable

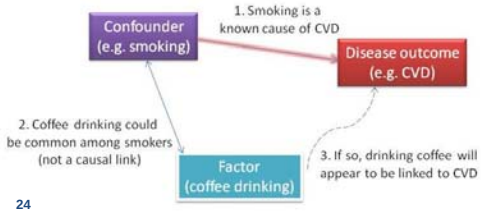


23

Confounding Factors

More formally:

- In a study of the effect of a variable X on disease Y, a **third variable, W** may be a confounder if:
  a) It is a risk factor for disease Y, and
  b) It is associated with (but does not cause) variable X



24

Confounding

- Evidence is generally reporting as differences, proportions, rates and ratios (RR, OR, HR) but **language** can **distort** the evidence
- ***Size*** of effect is **OBJECTIVE**, whereas ***Strength*** is **SUBJECTIVE**
- We often see the statement *"there is strong evidence"* when the research is reported but **you should be the judge**

25

Chance, Bias & Confounding

- **Good clinical Practice (GCP) and Experimental Design**
  - attempt to minimise bias and confounding
- **Hypothesis Testing**
  - assesses whether we could have observed the results by chance alone
- **Confidence Interval**
  - helps us understands the underlying variation in the results
- **Reporting Guidelines**
  - attempt to minimise subjective reporting

26

Statistical Analyses

27

Statistical Analysis

Hypothesis testing

"What is the probability that I would have obtained this set of data due to chance alone, under my current beliefs?"

28

Statistical Analysis

Great website that can guide you through the analysis stage if you do not have statistical support:

https://stats.idre.ucla.edu/other/mult-pkg/whatstat/

(link in your handouts)

UCLA Institute for Digital Research & Education

Search this website

HOME SOFTWARE RESOURCES SERVICES ABOUT US

CHOOSING THE CORRECT STATISTICAL TEST IN SAS, STATA, SPSS AND R

The following table shows general guidelines for choosing a statistical analysis. We emphasize that these are general guidelines and should not be construed as hard and fast rules. Usually your data could be analyzed in multiple ways, each of which could yield legitimate answers. The table below covers a number of common analyses and helps you choose among them based on the number of dependent variables sometimes referred to as outcome variables), the nature of your independent variables (sometimes referred to as predictors). You also want to

Statistical Analysis

Links take you to example code for each type of analysis: e.g. STATA 1-sample t-test

Number of Dependent Variables	Nature of Independent Variables	Nature of Dependent Variable(s)	Test(s)	How to SAS	How to Stata	How to SPSS	How to R
1	0 IVs (1 population)	interval & normal	one-sample t-test	SAS	Stata	SPSS	R
		ordinal or interval	one-sample median	SAS	Stata	SPSS	R
		categorical (2 categories)	binomial test	SAS	Stata	SPSS	R
		categorical	Chi-square goodness-of-fit	SAS	Stata	SPSS	R
	1 IV with 2 levels (independent groups)	interval & normal	2 independent sample t-test	SAS	Stata	SPSS	R

30

Example: One Sample T-Test on Stata

One sample t-test

A one sample t-test allows us to test whether a sample mean (of a normally distributed interval variable) significantly differs from a hypothesized value. For example, using the hsb2 data file, say we wish to test whether the average writing score (**write**) differs significantly from 50. We can do this as shown below.

```
ttest write=50
One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
   write |     200      52.775    .6702372    9.478586    51.45332    54.09668
------------------------------------------------------------------------------
Degrees of freedom: 199
                     Ho: mean(write) = 50
     Ha: mean < 50               Ha: mean ~= 50                 Ha: mean > 50
         t =   4.1403                t =   4.1403                   t =   4.1403
 P < t =   1.0000            P > |t| =   0.0001           P > t =   0.0000
```

The mean of the variable **write** for this particular sample of students is 52.775, which is statistically significantly different from the test value of 50. We would conclude that this group of students has a significantly higher mean on the writing test than 50.

See also

- Stata Code Fragment: Descriptives, ttests, Anova and Regression
- Stata Class Notes: Analyzing Data

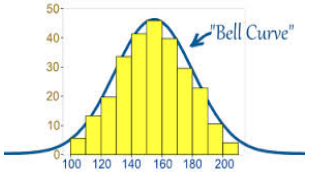
Understanding Confidence Intervals

32

Understanding Confidence Intervals

- To illustration some statistical concepts today, I will use a simple birth weight (full-term) example.
  - Birth weight is normally distributed (Gaussian distribution)
  - Sample size is 152 live births

- We will construct a confidence interval for the sample mean birth weight and perform a hypothesis test using a one-sample t-test



33

Understanding Confidence Intervals

- We collect a sample of birth weights from the offspring of mothers who exercised in pregnancy
- We wish to know if there is any difference compared to the WA population mean birth weight **population mean μ=3.4kg**

First we will look at how to describe the variability

34

Accuracy

- Accuracy of the Sample Mean depends on:
  - **Sample Size (n)**
  - **Sample Standard Deviation (s)**
- The variability associated with the sample mean is described by its

sampling distribution: $\bar{x} \sim N[\mu, \sigma^2/n]$

We estimate **σ** *(population standard deviation)* with **s**, so the standard error of the sample mean is **s/√n**

35

Understanding Confidence Intervals

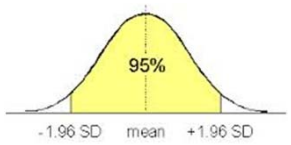
- Back to the birth weight example:
  - Sample Size **(n)** = **152 newborns**
  - Sample Mean **($\bar{x}$)** = **3.320kg**
  - Sample Standard Deviation **(s)** = **0.424**
  - Standard Error = **0.424/√152** = **0.0344**

Confidence Intervals are an easy-to-interpret way of expressing the precision with which a **Statistic** (e.g. $\bar{x}$) estimates a **Parameter** (e.g. **μ**).

36

Continued…

- Remember, for a **standard normal random variable Z**



$p(-1.96 < Z < 1.96) = 0.95$

Understanding Confidence Interval

Therefore:

$$p\left(-1.96 < \frac{(\bar{x} - \mu)}{s/\sqrt{n}} < 1.96\right) = 0.95$$

Rearranging gives us:

$$p(\bar{x} - 1.96(s/\sqrt{n}) < \mu < \bar{x} + 1.96(s/\sqrt{n})) = 0.95$$

95% confidence interval for mean birth weight:

$$\bar{x} \pm 1.96(s/\sqrt{n})$$

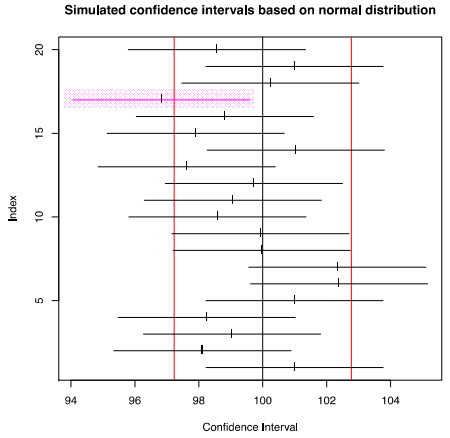
3.320 ± 1.96x0.0344 i.e. **95% CI [3.25,3.39kg]**

But how do we interpret this?

Understanding Confidence Interval

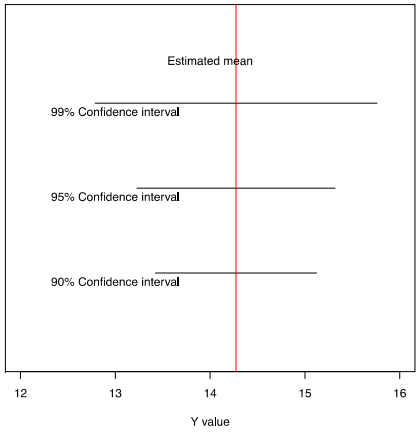
- **Intuitive:**
  - A confidence interval provides a range of plausible values for a parameter, in light of the observed data.
- **Formal:**
  - If a very large number of random samples (all same size) were collected from a population, and a 95% CI computed for each sample, then true value of the parameter would like in about 95% of these intervals

Understanding Confidence Interval



- For ***20 random samples***, based on a population (true) **mean of 100**
- We could expect around **one 95% CI** to NOT contain the true mean.
- This is our Type I Error **(α):**
  - i.e. 1 in 20, or
  - 0.05, or
  - 5%

Understanding Confidence Interval



- For constant sample size and standard error,
- The width of the confidence interval is determined by the significance level **(α)**
- **99%** CI is the **widest**
- **90%** CI is the *narrowest*

41

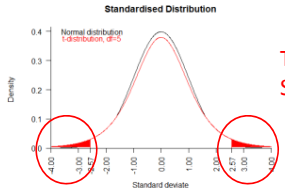
Normal compared to T-Distribution

- When sample sizes are small then it is necessary to use **Student's T-Distribution** rather than the Normal Distribution
- The *"degree of freedom"* **(df)** is simply the shape parameter for the T-Distribution

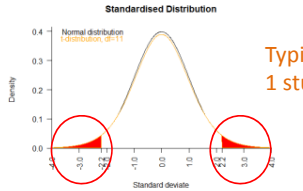
  Eg. for a **One-Sample T-Test:** ***df*=n-1** *(sample size minus 1)*

42

Normal compared to T-Distribution

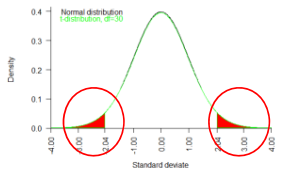


Typical mouse Study, n=6

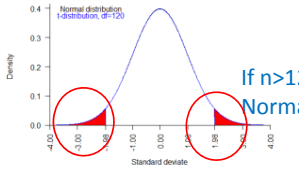


Typical phase 1 study, n=12

- Each tail region represents a **probability of 2.5%,**
- so for a **2-Tailed Test** the **significance level (α) is 5%**



Incorrect use of "CLT", n=30



If n>120 use Normal distribution

43
43

Understanding Confidence Interval

- **Standard Error** gives an idea of the **level of precision** of our estimate
- We can calculate a **Confidence Interval** based on:
  - The **Estimate**
  - Its **Standard Error**, and
  - the **Known Properties** (distribution) of the **Estimate**
- The CI gives a plausible range of values for the true population parameter

44

Interpreting P-Values

Hypothesis Testing

"What is the probability that I would have obtained this set of data due to chance alone, under my current beliefs?"

Statistical way to assist with answering questions/claims about certain population parameters by sampling from the population of interest.

Interpreting P-Values

Continuing with the birth weight example, we can write the **null (H_0) and alternative (H_1) hypotheses** as:

- **H_0** : true mean birth weigh is **3.4kg (μ=3.4)**
- **H_1** : true mean birth weight is not **3.4kg ($\mu \neq 3.4$)**
  - This is a **2-Sided Test**
  - it does not specify the direction in which we are looking for μ to depart from the value specified under the null hypothesis.

Interpreting P-Values

You may be familiar with the formula for a **one-sample t-test**:

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$\bar{x}$ = sample mean
μ = mean under the null hypothesis **(H_0)**
s = sample standard deviation
n = sample size

Interpreting P-Values

- Substituting in the values from the birth weight example:

$$z = \frac{3.320 - 3.4}{0.424/\sqrt{152}} = -2.33$$

- **Z** is likely to take values **close to zero if H_0 is correct**
- **Z** is likely to take extreme, **large positive or negative values if H_0 is incorrect**

Is **-2.33** sufficiently extreme to cause us to **reject H_0**?

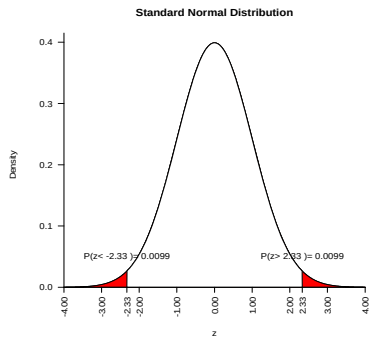
Interpreting P-Values

A **P-value** is defined to be the probability of getting a test statistic at least as extreme as the one observed, **under the assumption that H_0 is true**.

- In Hypothesis Test we **assume that H_0 is true**
  - for the birth weight example we know that the test **statistic (Z)** has a **Standard Normal Distribution, N[0,1].**
- How likely are we to see a result as least as extreme as the one observed (i.e. z < -2.33 & z > 2.33) ?

Interpreting P-Values

Hence the p-value calculation is based on the area under the curve in each tail of the standard normal distribution below:



p-value= p(Z < -2.33) + p(Z > 2.33)

= 0.0099 + 0.0099

= 0.0198

The probability in each "tail" can be obtained from statistical tables or statistical packages..

Continued…

From the hypothesis test, **p-value = 0.0198**

Since the **p-value is less than the significance level 0.05**, we reject H_0

- conclude that the mean birth weight in the exercise group is lower the the WA mean birth weight.

Is this a clinically significant result for birth weight in the exercise group?

mean: 3.32kg
95% CI (3.25,3.39kg)
p=0.02

Interpreting P-Values

- **Rejection of the null hypothesis (H_0)** is determined by the significance level (**α** is usually 5%, i.e. 0.05)
- The significance level is the probability of a **type I error** (falsely rejecting a true null hypothesis)
- Fisher (1965) did not intend the significance to be fixed at 5%, but set according to circumstances

53

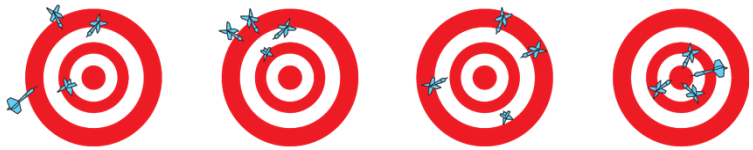
Sample Size

Why is Sample Size so important?

54

Precision

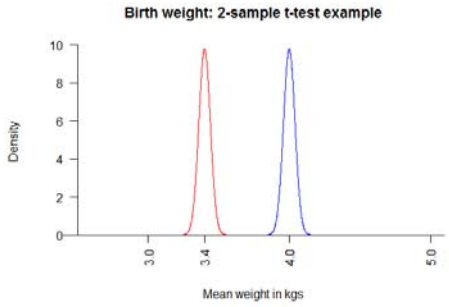
- Sample size controls the **Precision** of the conclusions by taking into account:
  - **Variability** of the measurements
  - Sensitivity of the **Statistical Method**
  - Clinically relevant **Difference** (effect size)



55

Sample Size

- Do you think there would be a significant difference between these two samples?

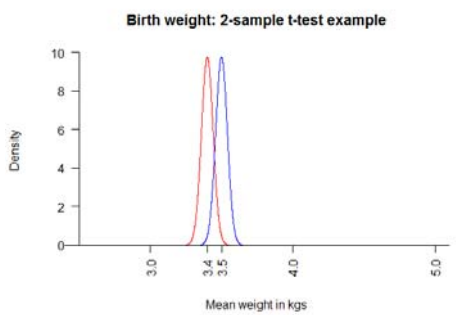


Sample Size = 150 per group
Standard Deviation = 0.5

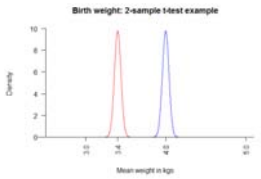
56

Sample Size

- What happens when we reduce the size of the difference between the two means?



Original plot

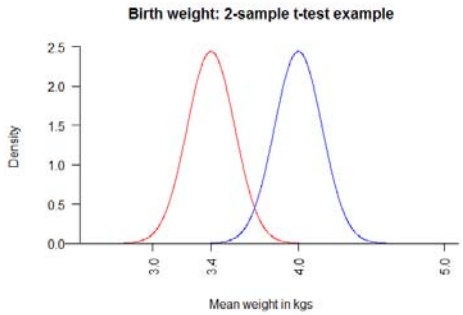


Sample Size = 150 per group
Standard Deviation = 0.5

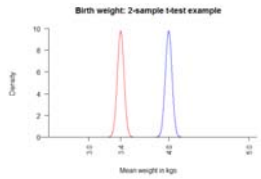
57

Sample Size

- What happens when we increase the variability?



Original plot

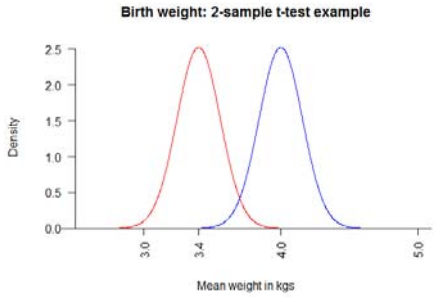


Sample Size = 150 per group
Standard Deviation = 0.5

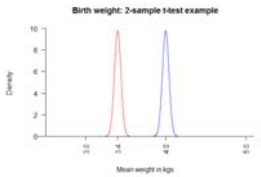
58

Sample Size

- The same happens when we decrease the sample size?
- **Standard of error = s/√n**



Original plot

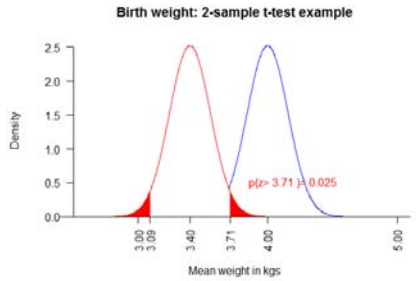


Sample Size = 10 per group
Standard Deviation = 0.5

59

Sample Size

- The **significance** level or **Type I Error** (probability of falsely rejecting the null hypothesis) is illustrated below for a **2-tailed** hypothesis test

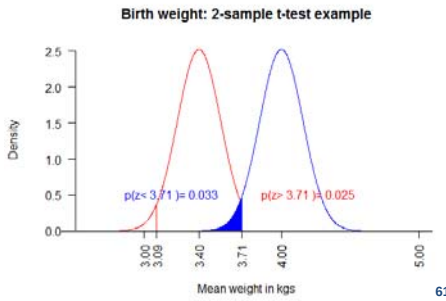


Sample Size = 10 per group
Standard Deviation = 0.5

60

Sample Size

- The statistical **power** of this test is illustrated below
  Type II Error is the probability of falsely failing to reject the null hypothesis

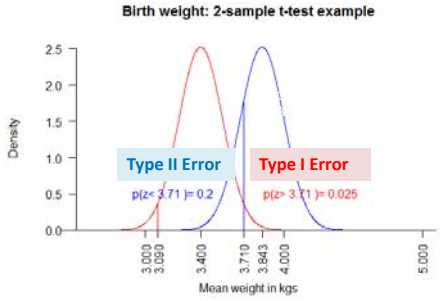


Power = 1 - Type II Error
= 1 - 0.033
= 0.967 (or 96.7%)

Sample Size = 10 per group
Standard Deviation = 0.5

Sample Size

- Usually the statistical power of a hypothesis test is defined to be **at least 80%**
- In this case we could detect a **difference of at least 0.443kg**



Power = 1 - Type II Error
= 1 - 0.2
= 0.8 (or 80%)

Sample Size = 10 per group
Standard Deviation = 0.5

Standardised Reporting

Standardised Reporting

- **P-Values** give the likelihood that an observed association could have arisen solely as a result of chance bias in sampling
- **Estimate of effect size** – gives direction and magnitude
- **Confidence intervals** are an estimate of variability in the effect size (precision)
- **Statistical significance cannot distinguish causal from non-causal associations**

Bradford-Hill Criteria for Causality

- Temporal relationship
- Strength of relationship
- Dose-response
- Consistency
- Plausibility
- Consideration of alternative explanations
- Experiment
- Specificity
- Coherence

Austin Bradford Hill: The Environment and Disease: Association or Causation? Proceedings of the Royal Society of Medicine, 58 (1965): 295-300.

Standardised Reporting

- Reporting guidelines provide
  - Checklists, depending on type of study design
  - Ensure that published research includes sufficient details for us to critically assess the quality of research

CONSORT, TREND, STROBE, REMARK, STREGA, PRISMA

These guidelines should also be consulted:

- **before you Design the Study**
- **before you Plan the Analysis**
- **before you Write the Paper**
- **before you Submit the Paper**

66

Where can I get statistical help?

67

Where can I find a Statistician?

Perth Children's Hospital:
Free advice through **Telethon Clinical Research Centre**

Telethon Kids Institute (consultancy service):
Biometrics@telethonkids.org.au

UWA (consultancy service):
consulting-cas@uwa.edu.au
The Centre for Applied Statistics, UWA, offers free advice to UWA postgraduate research students

- **More in handouts**

68

Checklist for talking to a Statistician

- Clear hypothesis
- Proposed study design
- Primary endpoint & estimate of variability
- Clinically relevant effect size
- Estimate of feasible sample size based on budget or potential annual patient recruitment
- Important confounders & source of bias
- Similar publications or systematic reviews

69

How can I learn more about statistics?

In the absence of large, randomised, well-controlled clinical trials to address every research questions we all need to increase our statistical literacy

Imagine if we updated our statistical skills as often as we updated our computer skills.

70

How can I learn more about statistics?

In person at Perth Children's Hospital:

Attend Research Skills Seminars.

In person at UWA:

The Centre for Applied Statistics provides short courses in statistics which are heavily discounted for students.

Joint Clinical-Statistical Supervision:

If one of your supervisors is a statistician then you will have unlimited access to statistical knowledge/training.

71

How can I learn more about statistics?

Online: **Data Science Specialization**

Johns Hopkins University

FAQ: You can access the course for free via https://www.coursera.org/specializations/jhu-data-science#courses.

This will allow you to explore the course, watch lectures, and participate in discussions for free. To be eligible to earn a certificate, you must either pay for enrolment or qualify for financial aid.

Links in your handouts

72

2 ADDITIONAL WEBSITES

UCLA – Institute for Digital Research and Education
Choosing the Correct Statistical Test in SAS, STAT, SPSS and R
https://stats.idre.ucla.edu/other/mult-pkg/whatstat/

John Hopkins University – Data Science Specialization
https://www.coursera.org/specializations/jhu-data-science#courses

3 STATISTICAL ANALYSIS

UCLA Institute for Digital Research & Education
https://stats.idre.ucla.edu/other/mult-pkg/whatstat/
Refer to slides 29-31

4 BRADFORD-HILL CRITERIA FOR CAUSALITY

- Temporal relationship
- Strength of relationship
- Dose-response
- Consistency
- Plausibility
- Consideration of alternative explanations
- Experiment
- Specificity
- Coherence

5 STATISTICAL SUPPORT CONTACTS

5.1 Perth Children's Hospital

Telethon Clinical Research Centre (TCRC), Department of Child Health Research
Child and Adolescent Health Service
CAHS.TCRC@health.wa.gov.au
https://pch.health.wa.gov.au/Research/For-researchers/Telethon-Clinical-Research-Centre

Biostatistics and Data Management Support through TCRC
https://cahs-healthpoint.hdwa.health.wa.gov.au/directory/research/researchers/Pages/Biostatistics.aspx
(internal Department of Health site)

5.2 Telethon Kids Institute

Consultancy Service
Biometrics@telethonkids.org.au

5.3 University of Western Australia – The Centre for Applied Statistics

**offers free advice for UWA Postgraduate Research Students*

consulting-cas@uwa.edu.au

5.4 WAHTN Clinical Trial and Data Management Centre

The Clinical Trial and Data Management Centre is a WAHTN enabling platform which aims to enhance clinical trials and related data management in Western Australia.

The platform is a WAHTN-wide entity sharing expertise in clinical trial study design (including novel designs), clinical trial conduct, data management, data-linkage, analytical techniques for clinical trial datasets, bio-repository techniques and clinical registry datasets. It facilitates the pursuit of large-scale clinical trials and translational healthcare research in WA.

https://www.wahtn.org/state-activities/clinical-trials-data-centre/

Contact:

CTDMC@curtin.edu.au
08 9266 1970

Clinical Research Support Service – EMHS Sessions
Tuesdays 9:30am – 3:00pm
EMHS Research Hub, L2 Kirkman House

30th July
6th August
8th October
3rd December

Reporting Guidelines

Estimating intervention effects:	CONSORT (randomised clinical trial) TREND (non-randomised study)
Assessing causes & prognosis:	STROBE (observational studies)
Quantifying accuracy of diagnosis & prognosis tools:	STARD (diagnostic studies) REMARK (tumour markers prognostic studies)
Testing genetic association:	STREGA (genetic observational studies)
Aggregating evidence: systematic reviews & meta-analyses	PRISMA *(previously known as QUOROM)* [MOOSE, IOM]

- **CONSORT** (CONsolidated Standards Of Reporting Trials)
  Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereux PJ, et al. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c869.

- **TREND** (Transparent Reporting of Evaluations with Non-randomised Designs)
  Des Jarlais DC, Lyles C, Crepaz N, TREND Group. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. Am J Public Health 2004;94:361-6.

- **STROBE** (STrengthening the Reporting of OBservational studies in Epidemiology)
  Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al; STROBE Initiative. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. PLoS Med 2007;4:1628-54.

- ***STARD*** *(STAndards for the Reporting of Diagnostic accuracy studies)*
  Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. BMJ2003;326:41-4.

- ***REMARK*** *(REporting recommendations for tumour MARKer prognostic studies)*
  McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM. REporting recommendations for tumour MARKer prognostic studies (REMARK). Br J Cancer2005;93:387-91.

- ***STREGA*** *(STrengthening the REporting of Genetic Associations)*
  Little J, Higgins JP, Ioannidis JP, Moher D, Gagnon F, von Elm E, et al. Strengthening the reporting of genetic association studies (STREGA): an extension of the STROBE statement. Eur J Epidemiol2009;24:37-55.

Understanding & reporting your research results

Additional Resources

Reporting Guidelines (continued)

- ***PRISMA*** *(preferred reporting items for systematic reviews and meta-analyses)*
  Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med2009;6:e1000097. [Previously known as Quality of Reporting of Meta-analyses or QUOROM]

- ***MOOSE****: Meta-analysis Of Observational Studies in Epidemiology*
  http://www.consort-statement.org/mod_product/uploads/MOOSE%20Statement%202000.pdf

- ***IOM****: Institutes of Medicine Standards for Systematic Reviews*
  http://iom.edu/Reports/2011/Finding-What-Works-in-Health-Care-Standards-for-Systematic-Reviews.aspx

Recommended basic statistics textbook

- De Veaux, R.D., Velleman, P.F., and Bock, D. (2012). Stats: Data And Models (3rd Edition), Pearson Education, Boston.

Good source for how to perform statistical analyses (SAS, SPSS, STATA & R)

- http://www.ats.ucla.edu/stat/
- If you can't find how to perform the statistical analysis on this site then you definitely need to consult a statistician!

The following pages are from the back pages of "Stats: Data And Models (2nd Edition)" and provide a great overview of common (basic) statistical formulas, methods and associated assumptions.

Quick Guide to Inference

Think			Show				Tell?
Inference about?	One group or two?	Procedure	Model	Parameter	Estimate	SE	Chapter
Proportions	One sample	1-Proportion *z*-Interval	*z*	p	$\hat{p}$	$\sqrt{\frac{\hat{p}\hat{q}}{n}}$	19
		1-Proportion *z*-Test				$\sqrt{\frac{p_0 q_0}{n}}$	20, 21
	Two independent groups	2-Proportion *z*-Interval	*z*	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$	22
		2-Proportion *z*-Test				$\sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}}$, $\hat{p} = \frac{y_1 + y_2}{n_1 + n_2}$	22
Means	One sample	*t*-Interval *t*-Test	*t* df = $n - 1$	μ	$\bar{y}$	$\frac{s}{\sqrt{n}}$	23
	Two independent groups	2-Sample *t*-Test 2-Sample *t*-Interval	*t* df from technology	$\mu_1 - \mu_2$	$\bar{y}_1 - \bar{y}_2$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	24
	n Matched pairs	Paired *t*-Test Paired *t*-Interval	*t* df = $n - 1$	μ_d	$\bar{d}$	$\frac{s_d}{\sqrt{n}}$	25
Distributions (one categorical variable)	One sample	Goodness-of-Fit	χ^2 df = *cells* − 1	$\sum \frac{(Obs - Exp)^2}{Exp}$			26
	Many independent groups	Homogeneity χ^2 Test	χ^2 df = $(r - 1)(c - 1)$				
Independence (two categorical variables)	One sample	Independence χ^2 Test					
Association (two quantitative variables)	One sample	Linear Regression *t*-Test or Confidence Interval for β	*t* df = $n - 2$	β_1	b_1	$\frac{s_e}{s_x\sqrt{n-1}}$ (compute with technology)	27
		*Confidence Interval for μ_y		μ_y	$\hat{y}_y$	$\sqrt{SE^2(b_1) \cdot (x_y - \bar{x})^2 + \frac{s_e^2}{n}}$	
		*Prediction Interval for y_y		y_y	$\hat{y}_y$	$\sqrt{SE^2(b_1) \cdot (x_y - \bar{x})^2 + \frac{s_e^2}{n} + s_e^2}$	
Inference about?	One group or two?	Procedure	Model	Parameter	Estimate	SE	Chapter

Assumptions for Inference	And the Conditions That Support or Override Them
Proportions (z)	
• **One sample**	
1. Individuals are independent.	1. SRS and $< 10\%$ of the population.
2. Sample is sufficiently large.	2. Successes and failures ≥ 10.
• **Two sample**	
1. Samples are independent.	1. (Think about how the data were collected.)
2. Data in each sample are independent.	2. Both are SRSs and $< 10\%$ of populations OR random allocation.
3. Both samples are sufficiently large.	3. Successes and failures ≥ 10 for both.
Means (t)	
• **One Sample** (df $= n - 1$)	
1. Individuals are independent.	1. SRS and $< 10\%$ of the population.
2. Population has a Normal model.	2. Histogram is unimodal and symmetric.*
• **Two independent samples** (df from technology)	
1. Samples are independent.	1. (Think about the design.)
2. Data in each sample are independent.	2. SRSs and $< 10\%$ OR random allocation.
3. Both populations are Normal.	3. Both histograms are unimodal and symmetric.*
• **Matched pairs** (df $= n - 1$)	
1. Data are matched; n pairs.	1. (Think about the design.)
2. Individuals are independent.	2. SRS and $< 10\%$ OR random allocation.
3. Population of differences is Normal.	3. Histogram of differences is unimodal and symmetric.
Distributions/Association (χ^2)	
• **Goodness of fit** (df = # of cells $- 1$; one variable, one sample compared with population model)	
1. Data are counts.	1. (Are they?)
2. Data in sample are independent.	2. SRS and $< 10\%$ of the population.
3. Sample is sufficiently large.	3. All expected counts ≥ 5.
• **Homogeneity** [df $= (r - 1)(c - 1)$; samples from many populations compared on one variable]	
1. Data are counts.	1. (Are they?)
2. Data in samples are independent.	2. SRSs and $< 10\%$ OR random allocation.
3. Samples are sufficiently large.	3. All expected counts ≥ 5.
• **Independence** [df $= (r - 1)(c - 1)$; sample from one population classified on two variables]	
1. Data are counts.	1. (Are they?)
2. Data are independent.	2. SRSs and $< 10\%$ of the population.
3. Sample is sufficiently large.	3. All expected counts ≥ 5.
Regression with *R* predictors (*t*, df $= n - k - 1$)	
• **Association** of each quantitative predictor with the response variable	
1. Form of relationship is linear.	1. Scatterplots of y against each x are straight enough. Scatterplot of residuals against predicted values shows no special structure.
2. Errors are independent.	2. No apparent pattern in plot of residuals against predicted values.
3. Variability of errors is constant.	3. Plot of residuals against predicted values has constant spread, doesn't "thicken."
4. Errors follow a Normal model.	4. Histogram of residuals is approximately unimodal and symmetric, or Normal probability plot is reasonably straight.*
Analysis of Variance (F, df depends on number of factors and number of levels in each.)	
• **Equality** of the mean response across levels of categorical predictors	
1. Additive Model (if there are 2 factors with no interaction term).	1. Interaction plot shows parallel lines (otherwise include an interaction term if possible).
2. Independent errors.	2. Randomized experiment or other suitable randomization.
3. Equal variance across treatment levels.	3. Plot of residuals against predicted values has constant spread. Boxplots (partial boxplots for 2 factors) show similar spreads.
4. Errors follow a Normal model.	4. Histogram of residuals is unimodal and symmetric, or Normal probability plot is reasonably straight.

(*Less critical as n increases)

Inference:

Confidence interval for parameter = ***statistic ± critical value × SE(statistic)***

$$\text{Test statistic} = \frac{statistic - parameter}{SD(statistic)}$$

Parameter	Statistic	*SD*(statistic)	*SE*(statistic)
p	$\hat{p}$	$\sqrt{\frac{pq}{n}}$	$\sqrt{\frac{\hat{p}\hat{q}}{n}}$
$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$	$\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$
μ	$\bar{y}$	$\frac{\sigma}{\sqrt{n}}$	$\frac{s}{\sqrt{n}}$
$\mu_1 - \mu_2$	$\bar{y}_1 - \bar{y}_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
μ_d	$\bar{d}$	$\frac{\sigma_d}{\sqrt{n}}$	$\frac{s_d}{\sqrt{n}}$
σ_ε	$s_e = \sqrt{\frac{\Sigma(y - \hat{y})^2}{n - 2}}$	(dividing $n - k - 1$ in multiple regression)	
β_1	b_1	(in simple regression)	$\frac{s_e}{s_x\sqrt{n-1}}$
μ_v	$\hat{y}_v$	(in simple regression)	$\sqrt{SE^2(b_1) \cdot (x_v - \bar{x})^2 + \frac{s_e^2}{n}}$
y_v	$\hat{y}_v$	(in simple regression)	$\sqrt{SE^2(b_1) \cdot (x_v - \bar{x})^2 + \frac{s_e^2}{n} + s_e^2}$

Pooling: For testing difference between proportions: $\hat{p}_{pooled} = \frac{y_1 + y_2}{n_1 + n_2}$

For testing difference between means: $s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$

Substitute these pooled estimates in the respective SE formulas for both groups when assumptions and conditions are met.

Chi-square: $\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp}$

ANOVA: $SS_T = \sum\sum (\bar{y}_j - \bar{\bar{y}})^2$; $MS_T = SS_T/(k - 1)$

$SS_E = \sum\sum (y_{ij} - \bar{y}_j)^2$; $MS_E = SS_E/(N - k)$

$F_{k-1,N-k} = MS_T/MS_E$

Selected Formulas

$Range = Max - Min$
$IQR = Q3 - Q1$
Outlier Rule-of-Thumb: $y < Q1 - 1.5 \times IQR$ or $y > Q3 + 1.5 \times IQR$

$$\bar{y} = \frac{\sum y}{n}$$

$$s = \sqrt{\frac{\sum (y - \bar{y})^2}{n - 1}}$$

$$z = \frac{y - \mu}{\sigma} \text{ (model based)}$$

$$z = \frac{y - \bar{y}}{s} \text{ (data based)}$$

$$r = \frac{\sum z_x z_y}{n - 1}$$

$$\hat{y} = b_0 + b_1 x \qquad \text{where } b_1 = r\frac{s_y}{s_x} \text{ and } b_0 = \bar{y} - b_1 \bar{x}$$

$$P(\mathbf{A}) = 1 - P(\mathbf{A}^C)$$

$$P(\mathbf{A} \textit{ or } \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B}) - P(\mathbf{A} \textit{ and } \mathbf{B})$$

$$P(\mathbf{A} \textit{ and } \mathbf{B}) = P(\mathbf{A}) \times P(\mathbf{B}|\mathbf{A})$$

$$P(\mathbf{B}|\mathbf{A}) = \frac{P(\mathbf{A} \textit{ and } \mathbf{B})}{P(\mathbf{A})}$$

If **A** and **B** are independent, $P(\mathbf{B}|\mathbf{A}) = P(\mathbf{B})$

$E(X) = \mu = \sum x \cdot P(x)$	$Var(X) = \sigma^2 = \sum (x - \mu)^2 P(x)$
$E(X \pm c) = E(X) \pm c$	$Var(X \pm c) = Var(X)$
$E(aX) = aE(X)$	$Var(aX) = a^2 Var(X)$
$E(X \pm Y) = E(X) \pm E(Y)$	$Var(X \pm Y) = Var(X) + Var(Y)$ if X and Y are independent

Geometric: $P(x) = q^{x-1}p$ $\quad \mu = \frac{1}{p}$ $\quad \sigma = \sqrt{\frac{q}{p^2}}$

Binomial: $P(x) = {}_nC_x p^x q^{n-x}$ $\quad \mu = np$ $\quad \sigma = \sqrt{npq}$

$$\hat{p} = \frac{x}{n} \qquad \mu(\hat{p}) = p \qquad SD(\hat{p}) = \sqrt{\frac{pq}{n}}$$

Sampling distribution of $\bar{y}$:
(CLT) As n grows, the sampling distribution approaches the Normal model with

$$\mu(\bar{y}) = \mu_y \qquad SD(\bar{y}) = \frac{\sigma}{\sqrt{n}}$$

STATISTICAL ERRORS

P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.

BY REGINA NUZZO

For a brief moment in 2010, Matt Motyl was on the brink of scientific glory: he had discovered that extremists quite literally see the world in black and white.

The results were "plain as day", recalls Motyl, a psychology PhD student at the University of Virginia in Charlottesville. Data from a study of nearly 2,000 people seemed to show that political moderates saw shades of grey more accurately than did either left-wing or right-wing extremists. "The hypothesis was sexy," he says, "and the data provided clear support." The *P* value, a common index for the strength of evidence, was 0.01 — usually interpreted as 'very significant'. Publication in a high-impact journal seemed within Motyl's grasp.

But then reality intervened. Sensitive to controversies over reproducibility, Motyl and his adviser, Brian Nosek, decided to replicate the study. With extra data, the *P* value came out as 0.59 — not even close to the conventional level of significance, 0.05. The effect had disappeared, and with it, Motyl's dreams of youthful fame[1].

It turned out that the problem was not in the data or in Motyl's analyses. It lay in the surprisingly slippery nature of the *P* value, which is neither as reliable nor as objective as most scientists assume. "*P* values are not doing their job, because they can't," says Stephen Ziliak, an economist at Roosevelt University in Chicago, Illinois, and a frequent critic of the way statistics are used.

For many scientists, this is especially worrying in light of the reproducibility concerns. In 2005, epidemiologist John Ioannidis of Stanford University in California suggested that most published findings are false[2]; since then, a string of high-profile replication problems has forced scientists to rethink how they evaluate results.

At the same time, statisticians are looking for better ways of thinking about data, to help scientists to avoid missing important information or acting on false alarms. "Change your statistical philosophy and all of a sudden different things become important," says Steven Goodman, a physician and statistician at Stanford. "Then 'laws' handed down from God are no longer handed down from God. They're actually handed down to us by ourselves, through the methodology we adopt."

DALE EDWIN MURRAY

OUT OF CONTEXT

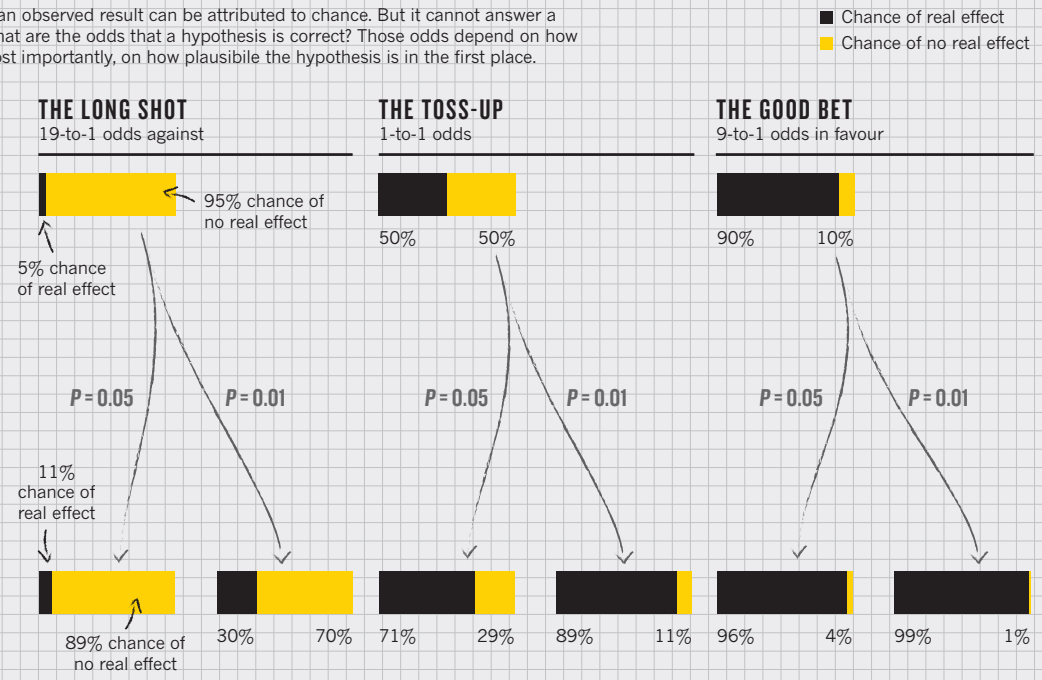
P values have always had critics. In their almost nine decades of existence, they have been likened to mosquitoes (annoying and impossible to swat away), the emperor's new clothes (fraught with obvious problems that everyone ignores) and the tool of a "sterile intellectual rake" who ravishes science but leaves it with no progeny[3]. One researcher suggested rechristening the methodology "statistical hypothesis inference testing"[3], presumably for the acronym it would yield.

The irony is that when UK statistician Ronald Fisher introduced the *P* value in the 1920s, he did not mean it to be a definitive test. He intended it simply as an informal way to judge whether evidence was significant in the

R. NUZZO; SOURCE: T. SELLKE *ET AL. AM. STAT.* **55**, 62–71 (2001)

PROBABLE CAUSE

A *P* value measures whether an observed result can be attributed to chance. But it cannot answer a researcher's real question: what are the odds that a hypothesis is correct? Those odds depend on how strong the result was and, most importantly, on how plausibile the hypothesis is in the first place.



old-fashioned sense: worthy of a second look. The idea was to run an experiment, then see if the results were consistent with what random chance might produce. Researchers would first set up a 'null hypothesis' that they wanted to disprove, such as there being no correlation or no difference between two groups. Next, they would play the devil's advocate and, assuming that this null hypothesis was in fact true, calculate the chances of getting results at least as extreme as what was actually observed. This probability was the *P* value. The smaller it was, suggested Fisher, the greater the likelihood that the straw-man null hypothesis was false.

For all the *P* value's apparent precision, Fisher intended it to be just one part of a fluid, non-numerical process that blended data and background knowledge to lead to scientific conclusions. But it soon got swept into a movement to make evidence-based decision-making as rigorous and objective as possible. This movement was spearheaded in the late 1920s by Fisher's bitter rivals, Polish mathematician Jerzy Neyman and UK statistician Egon Pearson, who introduced an alternative framework for data analysis that included statistical power, false positives, false negatives and many other concepts now familiar from introductory statistics classes. They pointedly left out the *P* value.

But while the rivals feuded — Neyman called some of Fisher's work mathematically "worse than useless"; Fisher called Neyman's approach "childish" and "horrifying [for] intellectual freedom in the west" — other researchers lost patience and began to write statistics manuals for working scientists. And because many of the authors were non-statisticians without a thorough understanding of either approach, they created a hybrid system that crammed Fisher's easy-to-calculate *P* value into Neyman and Pearson's reassuringly rigorous rule-based system. This is when a *P* value of 0.05 became enshrined as 'statistically significant', for example. "The *P* value was never meant to be used the way it's used today," says Goodman.

WHAT DOES IT ALL MEAN?

One result is an abundance of confusion about what the *P* value means[4]. Consider Motyl's study about political extremists. Most scientists would look at his original *P* value of 0.01 and say that there was just a 1% chance of his result being a false alarm. But they would be wrong. The *P* value cannot say this: all it can do is summarize the data assuming a specific null hypothesis. It cannot work backwards and make statements about the underlying reality. That requires another piece of information: the odds that a real effect was there in the first place. To ignore this would be like waking up with a headache and concluding that you have a rare brain tumour — possible, but so unlikely that it requires a lot more evidence to supersede an everyday explanation such as an allergic reaction. The more implausible the hypothesis — telepathy, aliens, homeopathy — the greater the chance that an exciting finding is a false alarm, no matter what the *P* value is.

NATURE.COM
For more on statistics, see:
go.nature.com/xlj9lr

These are sticky concepts, but some statisticians have tried to provide general rule-of-thumb conversions (see 'Probable cause'). According to one widely used calculation[5], a *P* value of 0.01 corresponds to a false-alarm probability of at least 11%, depending on the underlying probability that there is a true effect; a *P* value of 0.05 raises that chance to at least 29%. So Motyl's finding had a greater than one in ten chance of being a false alarm. Likewise, the probability of replicating his original result was not 99%, as most would assume, but something closer to 73% — or only 50%, if he wanted another 'very significant' result[6,7]. In other words, his inability to replicate the result was about as surprising as if he had called heads on a coin toss and it had come up tails.

Critics also bemoan the way that *P* values can encourage muddled thinking. A prime example is their tendency to deflect attention from the actual size of an effect. Last year, for example, a study of more than 19,000 people showed[8] that those who meet their spouses online are less likely to divorce ($p < 0.002$) and more likely to have high marital satisfaction ($p < 0.001$) than those who meet offline (see *Nature* http://doi.org/rcg; 2013). That might have sounded impressive, but the effects were actually tiny: meeting online nudged the divorce rate from 7.67% down to 5.96%, and barely budged happiness from 5.48 to 5.64 on a 7-point scale. To pounce on tiny *P* values and ignore the larger question is to fall prey to the "seductive certainty of significance", says Geoff Cumming, an emeritus psychologist at La Trobe University in Melbourne, Australia. But significance is no indicator of practical relevance, he says: "We should be asking,

'How much of an effect is there?', not 'Is there an effect?'"

Perhaps the worst fallacy is the kind of self-deception for which psychologist Uri Simonsohn of the University of Pennsylvania and his colleagues have popularized the term *P*-hacking; it is also known as data-dredging, snooping, fishing, significance-chasing and double-dipping. "*P*-hacking," says Simonsohn, "is trying multiple things until you get the desired result" — even unconsciously. It may be the first statistical term to rate a definition in the online Urban Dictionary, where the usage examples are telling: "That finding seems to have been obtained through *p*-hacking, the authors dropped one of the conditions so that the overall *p*-value would be less than .05", and "She is a *p*-hacker, she always monitors data while it is being collected."

Such practices have the effect of turning discoveries from exploratory studies — which should be treated with scepticism — into what look like sound confirmations but vanish on replication. Simonsohn's simulations have shown[9] that changes in a few data-analysis decisions can increase the false-positive rate in a single study to 60%. *P*-hacking is especially likely, he says, in today's environment of studies that chase small effects hidden in noisy data. It is tough to pin down how widespread the problem is, but Simonsohn has the sense that it is serious. In an analysis[10], he found evidence that many published psychology papers report *P* values that cluster suspiciously around 0.05, just as would be expected if researchers fished for significant *P* values until they found one.

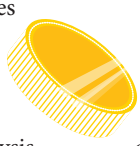
NUMBERS GAME

Despite the criticisms, reform has been slow. "The basic framework of statistics has been virtually unchanged since Fisher, Neyman and Pearson introduced it," says Goodman. John Campbell, a psychologist now at the University of Minnesota in Minneapolis, bemoaned the issue in 1982, when he was editor of the *Journal of Applied Psychology*: "It is almost impossible to drag authors away from their *p*-values, and the more zeroes after the decimal point, the harder people cling to them"[11]. In 1989, when Kenneth Rothman of Boston University in Massachusetts started the journal *Epidemiology*, he did his best to discourage *P* values in its pages. But he left the journal in 2001, and *P* values have since made a resurgence.

Ioannidis is currently mining the PubMed database for insights into how authors across many fields are using *P* values and other statistical evidence. "A cursory look at a sample of recently published papers," he says, "is convincing that *P* values are still very, very popular."

Any reform would need to sweep through an entrenched culture. It would have to change how statistics is taught, how data analysis is done and how results are reported and interpreted. But at least researchers are admitting that they have a problem, says Goodman. "The wake-up call is that so many of our published findings are not true." Work by researchers such as Ioannidis shows the link between theoretical statistical complaints and actual difficulties, says Goodman. "The problems that statisticians have predicted are exactly what we're now seeing. We just don't yet have all the fixes."

> "THE *P* VALUE WAS NEVER MEANT TO BE USED THE WAY IT'S USED TODAY."



Statisticians have pointed to a number of measures that might help. To avoid the trap of thinking about results as significant or not significant, for example, Cumming thinks that researchers should always report effect sizes and confidence intervals. These convey what a *P* value does not: the magnitude and relative importance of an effect.

Many statisticians also advocate replacing the *P* value with methods that take advantage of Bayes' rule: an eighteenth-century theorem that describes how to think about probability as the plausibility of an outcome, rather than as the potential frequency of that outcome. This entails a certain subjectivity — something that the statistical pioneers were trying to avoid. But the Bayesian framework makes it comparatively easy for observers to incorporate what they know about the world into their conclusions, and to calculate how probabilities change as new evidence arises.

Others argue for a more ecumenical approach, encouraging researchers to try multiple methods on the same data set. Stephen Senn, a statistician at the Centre for Public Health Research in Luxembourg City, likens this to using a floor-cleaning robot that cannot find its own way out of a corner: any data-analysis method will eventually hit a wall, and some common sense will be needed to get the process moving again. If the various methods come up with different answers, he says, "that's a suggestion to be more creative and try to find out why", which should lead to a better understanding of the underlying reality.

Simonsohn argues that one of the strongest protections for scientists is to admit everything. He encourages authors to brand their papers '*P*-certified, not *P*-hacked' by including the words: "We report how we determined our sample size, all data exclusions (if any), all manipulations and all measures in the study." This disclosure will, he hopes, discourage *P*-hacking, or at least alert readers to any shenanigans and allow them to judge accordingly.

A related idea that is garnering attention is two-stage analysis, or 'preregistered replication', says political scientist and statistician Andrew Gelman of Columbia University in New York City. In this approach, exploratory and confirmatory analyses are approached differently and clearly labelled. Instead of doing four separate small studies and reporting the results in one paper, for instance, researchers would first do two small exploratory studies and gather potentially interesting findings without worrying too much about false alarms. Then, on the basis of these results, the authors would decide exactly how they planned to confirm the findings, and would publicly preregister their intentions in a database such as the Open Science Framework (https://osf.io). They would then conduct the replication studies and publish the results alongside those of the exploratory studies. This approach allows for freedom and flexibility in analyses, says Gelman, while providing enough rigour to reduce the number of false alarms being published.

More broadly, researchers need to realize the limits of conventional statistics, Goodman says. They should instead bring into their analysis elements of scientific judgement about the plausibility of a hypothesis and study limitations that are normally banished to the discussion section: results of identical or similar experiments, proposed mechanisms, clinical knowledge and so on. Statistician Richard Royall of Johns Hopkins Bloomberg School of Public Health in Baltimore, Maryland, said that there are three questions a scientist might want to ask after a study: 'What is the evidence?' 'What should I believe?' and 'What should I do?' One method cannot answer all these questions, Goodman says: "The numbers are where the scientific discussion should start, not end." ■ SEE EDITORIAL P. 131

Regina Nuzzo *is a freelance writer and an associate professor of statistics at Gallaudet University in Washington DC.*

1. Nosek, B. A., Spies, J. R. & Motyl, M. *Perspect. Psychol. Sci.* **7,** 615–631 (2012).
2. Ioannidis, J. P. A. *PLoS Med.* **2,** e124 (2005).
3. Lambdin, C. *Theory Psychol.* **22,** 67–90 (2012).
4. Goodman, S. N. *Ann. Internal Med.* **130,** 995–1004 (1999).
5. Goodman, S. N. *Epidemiology* **12,** 295–297 (2001).
6. Goodman, S. N. *Stat. Med.* **11,** 875–879 (1992).
7. Gorroochurn, P., Hodge, S. E., Heiman, G. A., Durner, M. & Greenberg, D. A. *Genet. Med.* **9,** 325–321 (2007).
8. Cacioppo, J. T., Cacioppo, S., Gonzagab, G. C., Ogburn, E. L. & VanderWeele, T. J. *Proc. Natl Acad. Sci. USA* **110,** 10135–10140 (2013).
9. Simmons, J. P., Nelson, L. D. & Simonsohn, U. *Psychol. Sci.* **22,** 1359–1366 (2011).
10. Simonsohn, U., Nelson, L. D. & Simmons, J. P. *J. Exp. Psychol.* http://dx.doi.org/10.1037/a0033242 (2013).
11. Campbell, J. P. *J. Appl. Psych.* **67,** 691–700 (1982).

ResearchEducationProgram.org

ReserachEducationProgram@health.wa.gov.au