UNIVERSITY OF CANBERRA

Survey Research and Design in Psychology

Lecture 9 - Power and effect sizes

Dr Ananthan Ambikairajah

University of Canberra

Focus/Overview for today

UNIVERSITY OF CANBERRA

- Significance testing
- Inferential decision making
- Statistical power
- Effect sizes
- Confidence intervals
- Publication bias
- Academic integrity
- Statistical method guidelines

Overview of significance testing

UNIVERSITY OF CANBERRA

- Logic
- History
- Criticisms
- Decisions
- Practical significance

The logic

UNIVERSITY OF CANBERRA

- How many heads in a row would I need to throw before you would protest that something “wasn’t right”?

The logic

UNIVERSITY OF CANBERRA

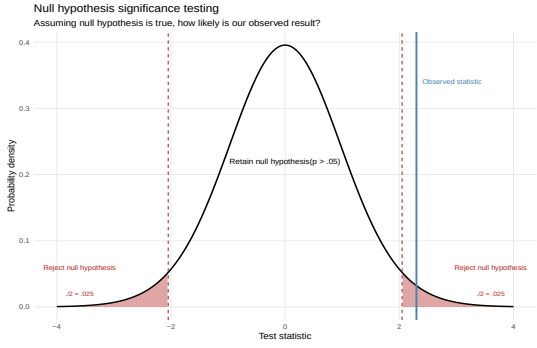
- Based on the statistical properties of sample data, we can extrapolate the probability of the observed relationships occurring in the target population.
- In so doing, we are assuming that the sample data is representative and that the data meets the assumptions associated with the inferential test.

The logic

UNIVERSITY OF CANBERRA

- Null hypothesis (H_0) = no effect
- Alternative hypothesis (H_a) = there is an effect
- Start by assuming that H_0 is true, and then evaluate how likely the observed data are under the null hypothesis
- Select a critical value (alpha level) - we usually go with .05
- Do you have a directional hypothesis? If so, you can use 1 tailed test - but this increases power
  - Calculate the effect and its p-value to determine the likelihood of H_0 in the target population. Is it bigger or smaller than the alpha value?
  - If it is smaller, then the effect is statistically significant: there is less than a 5% chance (if $\alpha = .05$) that the relationship is observed under assumption that null hypothesis is true - reject null hypothesis.
    - If it is bigger, then fail to reject H_0
- Researchers tolerate some false positives (critical α) to make a probability-based decision about H_0

The logic


Null hypothesis significance testing
Assuming null hypothesis is true, how likely is our observed result?
Probability density
0.4
0.3
0.2
0.1
0.0
Observed statistic
Retain null hypothesis(p > .05)
Reject null hypothesis
.025
Reject null hypothesis
.025
−4
−2
0
2
4
Test statistic

The history

UNIVERSITY OF CANBERRA

- Developed by Ronald Fisher (1920s - 1930s)
- To determine which agricultural methods yielded greater output
- Were variations in output between two plots attributable to chance or not?
- Agricultural research designs could not be fully experimental because natural variations such as weather and soil quality could not be fully controlled
- Therefore, it was needed to determine whether variations in the dependent variable were due to the independent variables or to chance

The history

UNIVERSITY OF CANBERRA

- Significance testing spread to other fields, including social sciences
- Spread was aided by the development of computers and training
- It became widely used during the 2nd half of the 20th century
- So widely used that in the latter 20th century, it attracted critique for over-use and misuse

Criticisms of significance testing

UNIVERSITY OF CANBERRA

- Critiqued as early as 1930
- Cohen's (1980s-1990s) critique helped a critical mass of awareness to develop
- Led to changes in publication guidelines and teaching about over-reliance on significance testing and alternative and adjunct techniques

Criticisms

1. The null hypothesis is rarely true
2. Significance testing provides:

- A binary decision (yes or no)
- But mostly we are interested in the size of the effect i.e. how much of an effect?

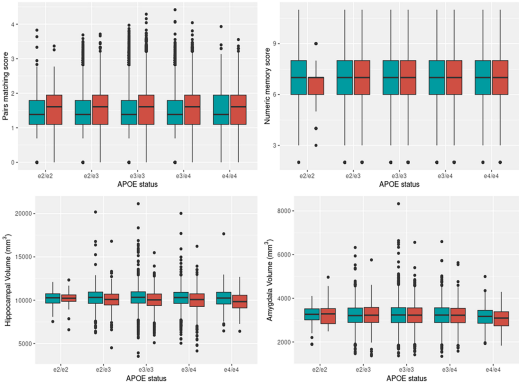
3. Statistical vs. practical significance
4. Significance is a function of effect size, sample size and α
5. Statistical significance simply means that the observed effect (relationship or differences) are unlikely to be due to chance
6. Statistical significance can be evident for very small (trivial) effects if sample size and/or critical alpha are large enough

Pratical significance

UNIVERSITY OF CANBERRA

- Practical significance is about whether the effect is large enough to be of value in a real world sense:
  - Is an effect worth being concerned about?
  - Is the effect noticeable or worthwhile?

HRT, APOE, Age and Brain



Ambikairajah et al. (2024)

Received: 20 June 2023 | Revised: 12 January 2024 | Accepted: 16 January 2024

DOI: 10.1002/hbm.26612

RESEARCH ARTICLE

WILEY

Investigating the synergistic effects of hormone replacement therapy, apolipoprotein E and age on brain health in the UK Biobank

Ananthan Ambikairajah[1,2,3] | Mizanur Khondoker[4] | Edward Morris[5] | Ann-Marie G. de Lange[6,7,8] | Rasha N. M. Saleh[4,9] | Anne Marie Minihane[4,10] | Michael Hornberger[4]

[1]Discipline of Psychology, Faculty of Health, University of Canberra, Canberra, Australian Capital Territory, Australia

[2]Centre for Ageing Research and Translation, Faculty of Health, University of Canberra, Canberra, Australian Capital Territory, Australia

[3]National Centre for Epidemiology and Population Health, Australian National University, Canberra, Australian Capital Territory, Australia

[4]Norwich Medical School, University of East Anglia, Norwich, UK

[5]Norfolk and Norwich NHS Trust, Norwich, UK

[6]Department of Clinical Neurosciences, Lausanne University Hospital (CHUV) and University of Lausanne, Lausanne, Switzerland

[7]Department of Psychology, University of Oslo, Oslo, Norway

[8]Department of Psychiatry, University of Oxford, Oxford, UK

[9]Department of Clinical and Chemical Pathology, Faculty of Medicine, Alexandria University, Alexandria, Egypt

[10]Norwich Institute of Healthy Ageing, Norwich, UK

Correspondence
Ananthan Ambikairajah, Discipline of Psychology, Faculty of Health, University of Canberra, Building 12, 11 Kirinari St, Canberra, ACT 2617, Australia.
Email: ananthan.ambikairajah@canberra.edu.au

Abstract

Global prevalence of Alzheimer's Disease has a strong sex bias, with women representing approximately two-thirds of the patients. Yet, the role of sex-specific risk factors during midlife reduction hormone replacement therapy (HRT) and their

Estrogen use, *APOE*, and cognitive decline

Evidence of gene–environment interaction

K. Yaffe, MD; M. Haan, DrPH; A. Byers, MPH; C. Tangen, DrPH; and L. Kuller, MD, DrPH

Article abstract—*Objective:* *APOE*-ε4 increases the risk of cognitive decline, while elderly women who take estrogen may have less risk of cognitive decline. The authors sought to determine whether estrogen use modifies the association between *APOE*-ε4 and cognitive decline. *Method:* As part of the Cardiovascular Health Study, 3,393 Medicare-eligible women (≥65 years) were randomly selected and recruited from Sacramento County, CA; Washington County, MD; Forsyth County, NC; and Pittsburgh, PA. Cognitive testing was administered annually; the authors studied the 2,716 women with cognitive testing on ≥2 visits. They analyzed change in score on the Modified Mini-Mental State Examination (3MS) as a function of estrogen use, *APOE* genotype, and baseline common and internal carotid artery wall thickening. *Results:* A total of 297 (11%) women were current estrogen users and 336 (12%) were past estrogen users. Over the 6-year average follow-up, baseline current users declined 1.5 points on the 3MS whereas never users declined 2.7 points ($p = 0.023$). Compared with ε4-negative women, ε4-positive women had a greater adjusted hazard ratio of cognitive impairment (3MS < 80), hazard risk [HR] = 1.47; 95% CI, 1.13 to 1.90. There was an interaction between estrogen use and ε4 presence ($p = 0.037$). Among ε4-negative women, current estrogen use reduced the risk of adjusted cognitive impairment compared with never users by almost half (HR = 0.59; 95% CI, 0.36 to 0.99), whereas, it did not reduce the risk among ε4-positive women (current use, HR = 1.33; 95% CI, 0.74 to 2.42). Compared with never use, current estrogen use was associated with less internal and common carotid wall thickening in ε4-negative women but not in ε4-positive women (p for interaction < 0.05 for both). Differences remained after adjusting for age, education, race, and stroke. *Conclusions:* Estrogen use was associated with less cognitive decline among ε4-negative women but not ε4-positive women. Potential mechanisms, including carotid atherosclerosis, by which ε4 may interact with estrogen and cognition warrant further investigation. **Key words:** Estrogen—*APOE*—Cognitive decline—Elderly women.

NEUROLOGY 2000;54:1949–1953

Yaffe et al. (2000)

HRT, APOE, Age and Brain

UNIVERSITY OF CANBERRA

Saleh et al. Alzheimer's Research & Therapy (2023) 15:10
https://doi.org/10.1186/s13195-022-01121-5

Alzheimer's Research & Therapy

RESEARCH Open Access

Hormone replacement therapy is associated with improved cognition and larger brain volumes in at-risk *APOE4* women: results from the European Prevention of Alzheimer's Disease (EPAD) cohort

Rasha N. M. Saleh[1*], Michael Hornberger[1], Craig W. Ritchie[2] and Anne Marie Minihane[1]

Abstract

Background The risk of dementia is higher in women than men. The metabolic consequences of estrogen decline during menopause accelerate neuropathology in women. The use of hormone replacement therapy (HRT) in the prevention of cognitive decline has shown conflicting results. Here we investigate the modulating role of *APOE* genotype and age at HRT initiation on the heterogeneity in cognitive response to HRT.

Methods The analysis used baseline data from participants in the European Prevention of Alzheimer's Dementia (EPAD) cohort (total *n*= 1906, women= 1178, 61.8%). Analysis of covariate (ANCOVA) models were employed to test the independent and interactive impact of *APOE* genotype and HRT on select cognitive tests, such as MMSE, RBANS, dot counting, Four Mountain Test (FMT), and the supermarket trolley test (SMT), together with volumes of the medial temporal lobe (MTL) regions by MRI. Multiple linear regression models were used to examine the impact of age of HRT initiation according to *APOE4* carrier status on these cognitive and MRI outcomes.

Results *APOE4* HRT users had the highest RBANS delayed memory index score (β-*APOE**HRT interaction = 0.009) compared to *APOE4* non-users and to non-*APOE4* carriers, with 6–10% larger entorhinal (left) and amygdala (right and left) volumes (β-interaction= 0.002, 0.003, and 0.005 respectively). Earlier introduction of HRT was associated with larger right (standardized β= −0.555, p=0.035) and left hippocampal volumes (standardized β= −0.577, p=0.028) only in *APOE4* carriers.

Conclusion HRT introduction is associated with improved delayed memory and larger entorhinal and amygdala volumes in *APOE4* carriers only. This may represent an effective targeted strategy to mitigate the higher life-time risk of AD in this large at-risk population subgroup. Confirmation of findings in a fit for purpose RCT with prospective recruitment based on *APOE* genotype is needed to establish causality.

*Correspondence:
Rasha N. M. Saleh

Introduction

More than two-thirds of Alzheimer's disease (AD) patients are women [1, 2]. The recent 2022 Global Burden

Saleh et al. (2023)

HRT, APOE, Age and Brain

UNIVERSITY OF CANBERRA

ELSEVIER

Neurobiology of Aging 33 (2012) 1129–1137

NEUROBIOLOGY OF AGING

www.elsevier.com/locate/neuaging

Postmenopausal hormone therapy, timing of initiation, APOE and cognitive decline

Jae H. Kang[a,*], Francine Grodstein[a,b]

[a] Channing Lab, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
[b] Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA

Received 29 April 2010; revised 13 September 2010; accepted 9 October 2010

Abstract

Associations between postmenopausal hormone therapy (HT) and cognitive decline may depend on apolipoprotein E (APOE) status or timing of initiation. We included 16,514 Nurses' Health Study participants aged 70–81 years who were followed since 1976 and completed up to 3 telephone cognitive assessments (2 years apart), between 1995 and 2006. The tests assessed general cognition (Telephone Interview of Cognitive Status; TICS), verbal memory, and category fluency. We used longitudinal analyses to estimate differences in cognitive decline across hormone groups. APOE genotype was available in 3697 participants. Compared with never users, past or current HT users showed modest but statistically significant worse rates of decline in the TICS: the multivariable-adjusted difference in annual rate of decline in the TICS among current estrogen only users versus never users was −0.04 (95% confidence interval, −0.07 to −0.004); for current estrogen + progestin users, the mean difference was −0.05 (95% confidence interval, −0.10 to −0.002). These differences were equivalent to those observed in women who are 1–2 years apart in age. We observed no protective associations with early timing of hormone initiation. We found suggestive interactions with APOE e4 status (e.g., on TICS, *p* interaction, 0.10), where the fastest rate of decline was observed among APOE e4 carriers who were current HT users. Regardless of timing of initiation, HT may be associated with worse rates of decline in general cognition, especially among those with an APOE e4 allele.

Keywords: Cohort studies; Cognitive aging; Risk factors in epidemiology; MCI; Memory

Kang & Grodstein (2012)

HRT, APOE, Age and Brain

UNIVERSITY OF CANBERRA

Among ϵ4-negative women, those currently taking estrogen had a 1.5 ± 1.0 (95% CI) smaller 3MS point decline over 6 years compared with the never users ($p = 0.003$) (table 2). Past estrogen users' change in scores did not differ from never users ($p = 0.83$). Among ϵ4-positive women, current or past estrogen use was not associated with the amount of cognitive decline (compared with never use: $p = 0.37$ for current use; $p = 0.79$ for past use). There was an interaction between estrogen use, *APOE*-ϵ4, and cognitive decline ($p = 0.037$). After adjusting for age, education, race, and stroke history, the interaction between *APOE* and estrogen use remained but lessened somewhat in statistical significance ($p = 0.06$).

Yaffe et al. (2000)

HRT, APOE, Age and Brain

UNIVERSITY OF CANBERRA

Table 2 Cognitive outcomes scores (mean±SEM) according to HRT use and *APOE4* genotype status

	Non-E4							E4							P_{APOE}	P_{HRT}	$P_{APOE*HRT}$
	No-HRT	*n*	HRT	*n*	Total	*n*	*p*-HRT	No-HRT	*n*	HRT	*n*	Total	*n*	*p*-HRT			
MMSE total score	28.49 ±0.07	603	28.43 ±0.36	50	28.49 ±0.07	653	0.607	28.15 ±0.11	350	28.22 ±0.30	30	28.16 ±0.10	380	0.960	0.565	0.724	0.782
Dot counting score	16.60 ±0.22	389	17.05 ±1.06	32	16.62 ±0.22	421	0.726	16.24 ±0.30	235	17.44 ±0.71	21	16.32 ±0.29	256	0.848	0.953	0.942	0.710
RBANS scores																	
RBANS total scale	103.57 ±0.62	600	105.04 ±2.78	49	103.63 ±0.61	649	0.921	100.52 ±0.85	351	106.68 ±3.44	29	100.88 ±0.83	380	0.045	0.488	0.128	0.097
RBANS attention index	97.65 ±0.70	601	102.61 ±2.73	28	97.86 ±0.68	629	0.222	97.23 ±0.93	352	102.23 ±3.34	29	97.51 ±0.90	381	0.706	0.818	0.297	0.652
RBANS delayed memory index	102.09 ±0.59	602	102.07 ±2.46	28	102.09 ±0.58	630	0.757	98.29 ±0.85	352	108.37 ±2.79	29	98.85 ±0.81	381	**0.002**	0.695	0.027[a]	0.009[a]
RBANS immediate memory index	106.55 ±0.58	602	105.18 ±30	28	106.49 ±0.57	630	0.854	101.65 ±0.87	352	105.59 ±3.83	29	101.87 ±0.85	381	0.150	0.434	0.307	0.209
RBANS language index	100.10 ±0.47	602	100.79 ±2.61	28	100.13 ±0.47	630	0.752	99.30 ±0.69	353	101.50 ±2.84	29	99.42 ±0.67	382	0.303	0.399	0.536	0.311
RBANS visuo-constructional index	105.16 ±0.65	602	106.82 ±2.95	28	105.23 ±0.63	630	0.310	104.66 ±0.92	352	108.32 ±2.91	29	104.87 ±0.88	381	0.483	0.163	0.938	0.233
FMT total score	8.31 ±0.43	32	9.33 ±0.33	3	8.40 ±0.40	35	0.803	7.48 ±0.55	33	10.50 ±1.50	3	7.77 ±0.56	36	0.195	0.449	0.271	0.439
SMT total score	6.54 ±0.63	34	6.33 ±0.88	3	6.53 ±0.58	37	0.781	5.14 ±0.54	33	10.00 ±1.53	4	5.53 ±0.55	37	0.158	0.549	0.451	0.241

Mean ± SEM of cognitive test scores stratified according to *APOE* genotype and HRT use. Significant *P* values for *APOE* genotype, HRT, and *APOE**HRT are shown, using the ANCOVA model (MANCOVA for RBANS scores). *p*-HRT within each *APOE* genotype is calculated using the pairwise comparison of the estimated marginal mean with Bonferroni adjustment for multiple comparison. Age, years of education, marital status, handedness, and CDR were used as covariates. *HRT* hormone replacement therapy, *MMSE* Mini-Mental State Examination, *RBANS* Repeatable Battery for the Assessment of Neuropsychological Status. *FMT* four mountain test, *SMT* supermarket trolley test. *P*-significant <0.05. [a] insignificant after FDR correction for multiple comparison. Bold: significant after FDR correction

Saleh et al. (2023)

HRT, APOE, Age and Brain

UNIVERSITY OF CANBERRA

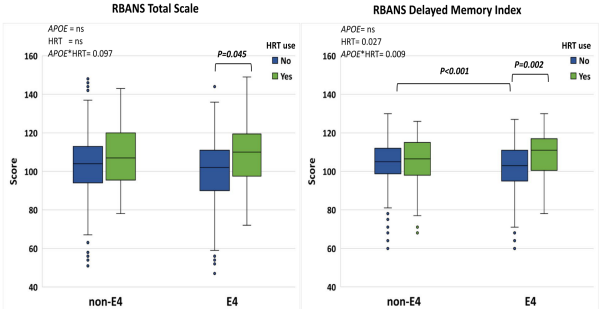


Fig. 1 Box plots showing the mean scores of RBANS total scale (left) and RBANS delayed memory index (right) in non-*APOE4* versus *APOE4* stratified according to HRT use. Pairwise comparison within each genotype group was carried out on the estimated marginal mean (within the MANCOVA model), after adjustment for age, years of education, marital status, handedness, and CDR). Statistical results in the upper left corner show *P* values of *APOE* genotype, HRT, and APOE*HRT for RBANS total scale (left) and delayed memory index (right) using the MANCOVA model. Non-*APOE4* *n*= 630 (no-HRT *n*=602, HRT *n*= 28), *APOE4* *n*= 381 (no-HRT *n*=352, HRT *n*= 29)

Saleh et al. (2023)

HRT, APOE, Age and Brain

UNIVERSITY OF CANBERRA

ELSEVIER

Neurobiology of Aging 33 (2012) 1129–1137

NEUROBIOLOGY OF AGING

www.elsevier.com/locate/neuaging

Postmenopausal hormone therapy, timing of initiation, APOE and cognitive decline

Jae H. Kang[a,*], Francine Grodstein[a,b]

[a] Channing Lab, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
[b] Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA

Received 29 April 2010; revised 13 September 2010; accepted 9 October 2010

Abstract

Associations between postmenopausal hormone therapy (HT) and cognitive decline may depend on apolipoprotein E (APOE) status or timing of initiation. We included 16,514 Nurses' Health Study participants aged 70–81 years who were followed since 1976 and completed up to 3 telephone cognitive assessments (2 years apart), between 1995 and 2006. The tests assessed general cognition (Telephone Interview of Cognitive Status; TICS), verbal memory, and category fluency. We used longitudinal analyses to estimate differences in cognitive decline across hormone groups. APOE genotype was available in 3697 participants. Compared with never users, past or current HT users showed modest but statistically significant worse rates of decline in the TICS: the multivariable-adjusted difference in annual rate of decline in the TICS among current estrogen only users versus never users was −0.04 (95% confidence interval, −0.07 to −0.004); for current estrogen + progestin users, the mean difference was −0.05 (95% confidence interval, −0.10 to −0.002). These differences were equivalent to those observed in women who are 1–2 years apart in age. We observed no protective associations with early timing of hormone initiation. We found suggestive interactions with APOE e4 status (e.g., on TICS, *p* interaction, 0.10), where the fastest rate of decline was observed among APOE e4 carriers who were current HT users. Regardless of timing of initiation, HT may be associated with worse rates of decline in general cognition, especially among those with an APOE e4 allele.

Keywords: Cohort studies; Cognitive aging; Risk factors in epidemiology; MCI; Memory

Kang & Grodstein (2012)

Science communication matters

UNIVERSITY OF CANBERRA

BBC

Home News Sport Business Innovation Culture Travel Earth Video Live

HRT could cut Alzheimer's risk in some women - early study

15 January 2023

Share

https://www.bbc.com/news/health-64276452

ScienceDaily

Your source for the latest research news

New! Sign up for our free **email newsletter**.

SD Health Tech Enviro Society Quirky

Science News *from research organizations*

HRT could ward off Alzheimer's among at-risk women

Date: January 14, 2023

Source: University of East Anglia

Summary: Hormone Replacement Therapy (HRT) could help prevent Alzheimer's Dementia among women at risk of developing the disease -- according to new research.

Share:

https://www.sciencedaily.com/search/?keyword=hrt#gsc.tab=0&gsc.q=hrt&gsc.page=1

Science communication matters

UNIVERSITY OF CANBERRA

Some dementia patients wait five years for a diagnosis, new research shows

'Confronting' diagnosis delays

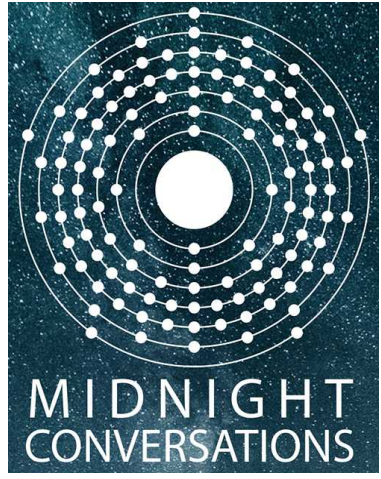
A new study from the University of Canberra's Centre for Ageing Research and Translation has found people are waiting an average of three years to be diagnosed with Alzheimer's disease — the most common form of dementia.

Dr Ananthan Ambikairajah says patients are waiting too long for a dementia diagnosis. *(ABC News: Lily Nothling)*

Lead researcher Ananthan Ambikairajah said for some other types of dementia, the delays were even longer.

ABC News Article: https://www.abc.net.au/news/2025-10-01/research-dementia-patients-waiting-five-years-for-diagnosis/105838266

ABC News Interview: https://www.youtube.com/watch?v=2q9lqAzf6rU



https://open.spotify.com/show/4SwmOkuSjIXLTzfMkjcPRh?si=982b233278454363

APA publication manual

UNIVERSITY OF CANBERRA

- APA publication manual recommendations about effect sizes, confidence intervals and power
  - 2001 - APA 5th edition (2001) recommended reporting effect sizes, power etc
  - 2009 - APA 6th edition (2009) further strengthened the requirements to use null hypothesis significance testing as a starting point and to also include effect sizes, confidence intervals and power.

Alternatives to significance testing

UNIVERSITY OF CANBERRA

"Historically, researchers in psychology have relied heavily on null hypothesis significance testing (NHST) as a starting point for many (but not all) of its analytic approaches. APA stresses that NHST is but a starting point and that additional reporting such as effect sizes, confidence intervals, and extensive description are needed to convey the most complete meaning of the results... complete reporting of all tested hypotheses and estimates of appropriate ESs and CIs are the minimum expectations for all APA journals."

— *(APA Publication Manual (6th ed., 2009, p. 33))*

Signifiance testing and p-values

- American statistical association statement on significance testing and p-values (Wasserstein & Lazar, 2016)
  1) P-values can indicate how incompatible the data are with a specified statistical model.
  2) P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
  3) Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
  4) Proper inference requires full reporting and transparency.
  5) A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
  6) By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Recommendations

UNIVERSITY OF CANBERRA

- Use traditional null hypothesis significance testing (Fisherian logic/inferential testing)
- Also use complementary techniques (effect sizes and confidence intervals)
- Emphasise practical significance
- Recognise merits and shortcomings of each approach

Summary of significance testing

UNIVERSITY OF CANBERRA

- Logic:
  - Examine sample data to determine p value that it represents a population with no effect (or some effect). It is a "bet" - at what point do you reject H_0
- History:
  - Developed by Fisher for agricultural experiments in early 20th century
  - During the 1980s and 1990s, significance testing was increasingly criticised for over-use and mis-application

Summary of significance testing

UNIVERSITY OF CANBERRA

- Criticisms:
  - Binary
  - Depends on sample size, effect size and critical alpha
  - Need practical significance
- Recommendations:
  - Whenever you report a significance test (p-level), also report an effect size and confidence intervals

Inferential decision making

UNIVERSITY OF CANBERRA

- Types of hypotheses:
  - Null hypothesis (H_0): No differences or no relationship
  - Alternative hypothesis (H_1): Differences or relationship

Inferential statistics

UNIVERSITY OF CANBERRA

- In inferential testing, a conclusion about a target population is made based on sample data. Either:
  - Do not reject H_0: p is not significant (i.e. not below the critical α)
  - Reject the H_0: p is significant (i.e. below the critical α)

Inferential statistics

UNIVERSITY OF CANBERRA

- We hope to make a correct inference based on the sample data i.e. either:
  - Do not reject H_0: correctly retain H_0 (i.e. when there is no real difference/effect in the population)
  - Reject the H_0 (Power): correctly reject H_0 (i.e. ability to detect when there is a real difference/effect in the population)

However, we risk making errors

- Type I error: Incorrectly reject H_0 (i.e. there is no difference/effect in the population)
- Type II error: Incorrectly fail to reject H_0 (i.e. there is a difference/effect in the population)

Inferential decision making table

UNIVERSITY OF CANBERRA

		Reality	
		H_0 False	**H_0 True**
Test	**Reject H_0**	Correct rejection H_0 = Power = 1 - β	Type I error = α
	Accept H_0	Type II error	Correct acceptance of H_0

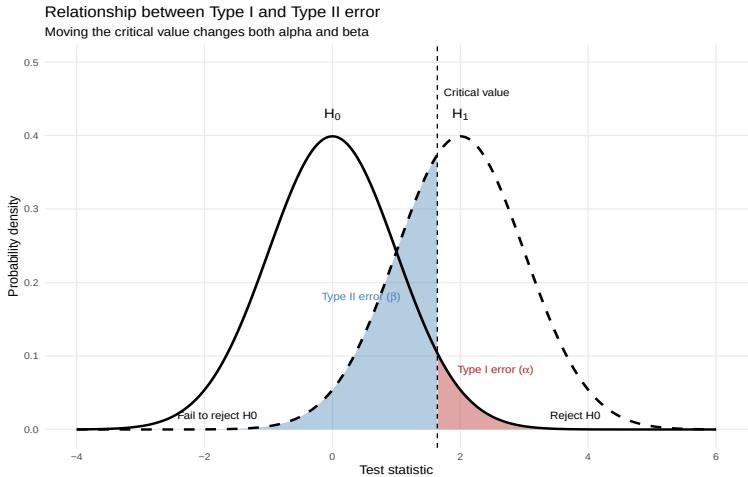
Summary of inferential decision making

UNIVERSITY OF CANBERRA

- Correct acceptance of H_0
- Correct rejection of H_0
- False rejection of H_0 (i.e., rejecting H_0 when it is true) (Type I error) $= \alpha$
- False acceptance of H_0 (i.e., not rejecting H_0 when it is false) (Type II error) $= \beta$

- Traditionally emphasis has been:
  - Too much on limiting Type I errors and
  - Not enough on limiting Type II errors
  - Balance is needed

Type I and Type II errors



Statistical power

UNIVERSITY OF CANBERRA

- The probability of correctly rejecting a false H_0 i.e getting a significant result when there is a real difference in the population

Statistical power

UNIVERSITY OF CANBERRA

		Reality	
		H_0 False	**H_0 True**
Test	**Reject H_0**	POWER	Type I error = α
	Accept H_0	Type II error	Correct acceptance of H_0

Statistical power

UNIVERSITY OF CANBERRA

- Desirable power $>$ .80
- Typical power $\sim$ .60 (in the social sciences)
- Power becomes higher when any of the folowing increase:
  - Sample size (N)
  - Critical alpha (α)
  - Effect size (Δ)

Question

- A study has a power of .80 what is the likelihood of a type II error?
  - .80
  - .20
  - .40
  - .10

		Reality	
		***H*₀ False**	***H*₀ True**
Test	**Reject *H*₀**	POWER	Type I error = α
	Accept *H*₀	Type II error	Correct acceptance of H_0

Question

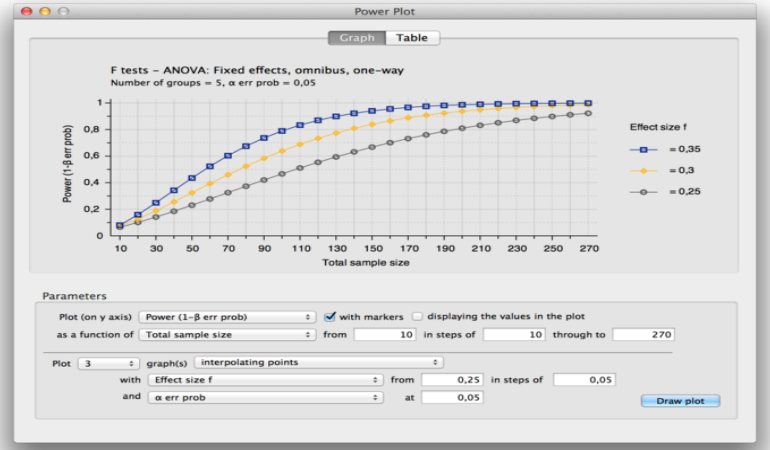
- A study has a power of .80 what is the likelihood of a type II error?
  - .80
  - **.20**
  - .40
  - .10

		Reality	
		***H*₀ False**	***H*₀ True**
Test	**Reject *H*₀**	POWER	Type I error = α
	Accept *H*₀	Type II error	Correct acceptance of H_0

Power analysis

UNIVERSITY OF CANBERRA

- Ideally, calculate expected power before conducting a study (a priori) based on:
  - Estimated sample size
  - Critical α
  - Expected or minimum effect size (e.g. from related research)
- Can also report actual power (post-hoc) in the results


Power Plot
Graph
Table
F tests – ANOVA: Fixed effects, omnibus, one-way
Number of groups = 5, α err prob = 0,05
Power (1-β err prob)
1
0,8
0,6
0,4
0,2
0
10 30 50 70 90 110 130 150 170 190 210 230 250 270
Total sample size
Effect size f
= 0,35
= 0,3
= 0,25
Parameters
Plot (on y axis) Power (1–β err prob) with markers displaying the values in the plot
as a function of Total sample size from 10 in steps of 10 through to 270
Plot 3 graph(s) interpolating points
with Effect size f from 0,25 in steps of 0,05
and α err prob at 0,05
Draw plot

Summary of statistical power

UNIVERSITY OF CANBERRA

- Power = the probability of detecting a real effect as statistically significant
- Increase power by:
  - Increasing sample size
  - Increasing critical α level
  - Increasing effect size
- Power:
  - $>$ .8 is "desirable"
  - $\sim$ .6 is more typical
- Can be calculated prospectively and retrospectively

Effect sizes

UNIVERSITY OF CANBERRA

- What is an effect size?
  - A measure of the strength (or size) of a relationship or effect
  - Where p value is reported, also present an effect size
  - "reporting and interpreting effect sizes in the context of previously reported effects is essential to good research" (Wilkinson, 1999)

Effect sizes

UNIVERSITY OF CANBERRA

- Why do we use them?
  - An inferential test may be statistically significant (i.e., the result is unlikely to have occurred by chance), but this doesn't indicate how large the effect is (the effect might be trivial).
  - On the other hand, there may be non-significant, but notable effects (especially in low powered tests).
  - Unlike significance testing, effect sizes are not influenced by N.

Commonly used effects

UNIVERSITY OF CANBERRA

- Correlation
  - r, r^2, sr^2
  - R, R^2
- Mean differences
  - Standardised mean difference e.g. Cohen's d
  - Eta squared (η^2), partial eta squared (η_p^2)

Standardised mean difference

UNIVERSITY OF CANBERRA

- The difference between two means in standard deviation units:
  - -ve = negative difference
  - 0 = no difference
  - +ve = positive difference

Standardised mean difference

UNIVERSITY OF CANBERRA

- A standardised measure of the difference between two means
  - $d = \frac{M_2 - M_1}{\sigma}$
  - $d = \frac{M_2 - M_1}{pooled\sigma}$ = e.g. Cohen's d, Hedges' g

Effect sizes

UNIVERSITY OF CANBERRA

- Cohens (1977):
  - .2 = small
  - .5 = moderate
  - .8 = large
- Wolf (1986):
  - .25 = educationally significant
  - .50 = practically significant (therapeutic)
- If your inferential test is not significant, you should still present an effect size as well as interpret this (could have come about by chance)

Effect sizes

UNIVERSITY OF CANBERRA

- Interpreting effect sizes
  - No agreed standards
  - Ultimately subjective
  - Best approach is to compare with other similar studies and use logic

Effect sizes

UNIVERSITY OF CANBERRA

- It depends on context
  - A small effect size can be impressive if, for example, a variable is:
    - Difficult to change (e.g. a personality construct) and/or
    - Very valuable (e.g. life expectancy)
  - A large effect size doesn't necessarily mean that there is any practical value, for example, if:
    - it is not related to the aims of the investigation

Power and effect sizes in psychology

UNIVERSITY OF CANBERRA

- Ward (2002) examined articles in 3 psychology journals to assess the use of statistical power and effect size measures.
  - Journal of Personality and Social Psychology
  - Journal of Consulting and Clinical Psychology
  - Journal of Abnormal Psychology
- They found that:
  - 7% of studies estimated or discuss statistical power.
  - 30% provided effect sizes.
  - Average effect size was medium
  - Current research designs typically do not have sufficient power to detect medium effect sizes.

Summary of effect sizes

UNIVERSITY OF CANBERRA

- Effect sizes = standardised difference or strength of relationship
- Inferential tests should be accompanied by effect sizes and confidence intervals
- Common bivariate effect sizes include:
  - Cohen's d
  - Correlation r

Confidence intervals

UNIVERSITY OF CANBERRA

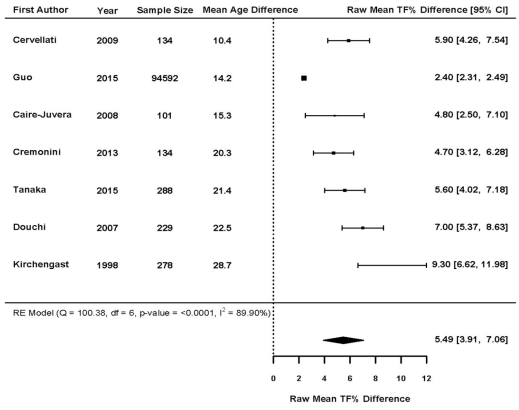
- Very useful, underutilised
- Gives "range of certainty" or "area of confidence"
  - For example, a population mean is 95% likely to be between -1.96 and +1.96 standard deviations of the sample mean
- Expressed as a:
  - Lower limit
  - Upper limit
- If we repeated the study many times and constructed an interval each time using the same method, 95% of those intervals would contain the true parameter

Confidence intervals

UNIVERSITY OF CANBERRA

- Confidence intervals can be reported for
  - Beta (unstandardised regression coefficient) in multiple linear regression
  - Means
  - Other effect sizes (e.g. r, R, d)
- Confidence intervals can be examined statistically and graphically e.g. error-bar graphs

Confidence Interval



Ambikairajah, Walsh, Tabatabaei-Jafari, et al. (2019)

Practice questions

UNIVERSITY OF CANBERRA

- If a multiple linear regression predictor has a $B = .5$, with a 95% confidence interval of .25 to .75, what should be concluded?
  a) Do not reject H_0 (i.e. that $B = 0$)
  b) Reject H_0 (i.e. that $B = 0$)

Practice questions

UNIVERSITY OF CANBERRA

- If a multiple linear regression predictor has a $B = .5$, with a 95% confidence interval of .25 to .75, what should be concluded?
  a) Do not reject H_0 (i.e. that $B = 0$)
  b) **Reject H_0 (i.e. that $\beta = 0$)**

Practice questions

UNIVERSITY OF CANBERRA

- If a multiple linear regression predictor has a $B = .2$, with a 95% confidence interval of - .2 to .6, what should be concluded?
  a) Do not reject H_0 (i.e. that $B = 0$)
  b) Reject H_0 (i.e. that $B = 0$)

Practice questions

- If a multiple linear regression predictor has a $B = .2$, with a 95% confidence interval of - .2 to .6, what should be concluded?
  a) **Do not reject H_0 (i.e. that $\beta = 0$)**
  b) Reject H_0 (i.e. that $B = 0$)

Practice questions

UNIVERSITY OF CANBERRA

- Which set of confidence intervals indicates that a mean score is significantly different from 3?
  a) 95% CI [-1, 5]
  b) 95% CI [2, 3]
  c) 95% CI [6, 9]
  d) 95% CI [-3, 4]

Practice questions

UNIVERSITY OF CANBERRA

- Which set of confidence intervals indicates that a mean score is significantly different from 3?
  a) 95% CI [-1, 5]
  b) 95% CI [2, 3]
  c) **95% CI [6, 9]**
  d) 95% CI [-3, 4]

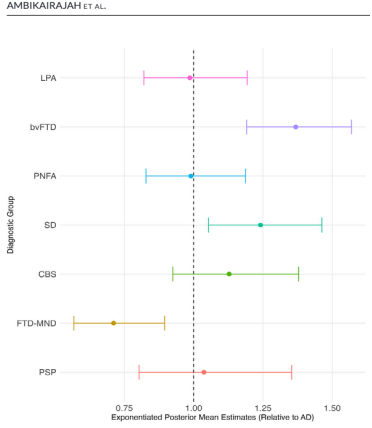
Summary of confidence intervals

UNIVERSITY OF CANBERRA

- Gives a "range of certainty" when generalising from a sample to a target population
- If we repeated the study many times and constructed an interval each time using the same method, 95% of those intervals would contain the true parameter
- Confidence intervals can be used for means, B and effect sizes
- Can be examined
  - Statistically (upper and lower limits)
  - Graphically (e.g. error-bar graphs)

Extension material - Credible Intervals

- There is a 95% probability the effect lies within this interval
  - e.g. there is a 95% probability that the true average time to dementia diagnosis for AD lies between 3.03 and 3.72 years, given the data and model.



Ambikairajah et al. (2025)

Publication bias

UNIVERSITY OF CANBERRA

- When the likelihood of publication depends on the nature and direction of results.
- Significant effects are more likely to be published
- Type I publication errors are underestimated to an extent that is: "frightening, even calling into question the scientific basis for much published literature." (Greenwald, 1975)

Publication bias

UNIVERSITY OF CANBERRA

"The extreme view of the file drawer problem is that journals are filled with the 5% of the studies that show Type I errors, while the file drawers are filled with the 95% of the studies that show non-significant results."

— *Rosenthal, 1979*

File drawer problem

- Tendency for non-significant results to be "filed away" (hidden) and not published.
- Number of null studies which would have to "filed away" in order for a body of significant published effects to be considered doubtful

Two counteracting biases

UNIVERSITY OF CANBERRA

- Low power: underestimation of real effects
- Publication bias or file-drawer effect: over-estimation of real effects

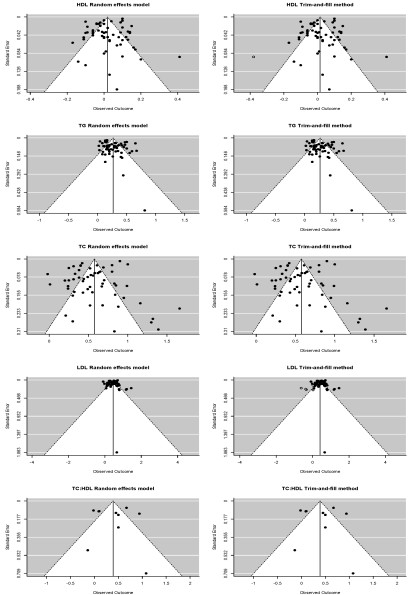
Funnel plots

UNIVERSITY OF CANBERRA

- Scatterplot of treatment effect against study size
- Precision in estimating the true treatment effect increases as sample size increases
- Studies with small sample sizes scatter more widely at the bottom of the graph (less precision)
- In the absence of publication bias the plot should resemble a symmetrical inverted funnel

Funnel plots

- Standard error decreases as sample size increases
- Example: Ambikairajah, Walsh, & Cherbuin (2019)



Publication bias

- If there is publication bias this will cause meta-analysis to overestimate effects.
- The more pronounced the funnel plot asymmetry, the more likely it is that the amount of bias will be substantial.

Counteracting the bias

UNIVERSITY OF CANBERRA

- Journal of articles in support of the null hypothesis
- Journal of negative results

Summary of publication bias

- Tendency for statistically significant studies to be published over non-significant studies
- Indicated by gap in funnel plot i.e. file-drawer effect
- Counteracting biases in scientific publishing, as there is a tendency:
  - towards low-power studies which underestimate effects
  - to publish significant effects over non-significant effects

Academic integrity

UNIVERSITY OF CANBERRA

- Richard Horton (Horton, 2015), editor of "The Lancet" (one of the world's oldest and best-known medical journals):
  1. "A lot of what is published is incorrect"
  2. "Much of the scientific literature, perhaps half, may simply be untrue. Afflicted by studies with small sample sizes, tiny effects, invalid exploratory analyses, and flagrant conflicts of interest, together with an obsession for pursuing fashionable trends of dubious importance."
  3. "Scientists too often sculpt data to fit their preferred theory of the world. Or they retrofit hypotheses to fit their data."
  4. "Our love of "significance" pollutes the literature with many a statistical fairy-tale."

Retraction watch

UNIVERSITY OF CANBERRA

- Retraction watch
  - "Authors have papers in Nature and Science retracted on the same day".
  - This was due to fabricated data.
  - The science paper was cited 240 times and the nature paper had been cited 119 times

Summary of academic integrity

UNIVERSITY OF CANBERRA

- Violations of academic integrity are most prevalent amongst those with incentives to cheat. For examples:
  - Students
  - Competitively-funded researchers
  - Commercially-sponsored researchers
- Adopt a balanced, critical approach, striving for objectivity and academic integrity

Solutions?

UNIVERSITY OF CANBERRA

- Open science at each stage of the research process

Preregistration

UNIVERSITY OF CANBERRA

- Record your:
  - Study design
  - Data collection methods (including sample size and data exclusion rules)
  - Analyses (to reduce biases associated with cherry picking/capitalizing on chance)

Preregistration

UNIVERSITY OF CANBERRA

- Preregistration ensures hypotheses are determined before running analyses
  - Preregistration separates hypothesis-generating (exploratory) from hypothesis-testing (confirmatory) research.
  - HARKing = Hypothesizing after the results are known

Replication

- Reluctance to publish replications.

Reliability

- Just how reliable are findings in psychology?
- Open science collaboration
  - 100 studies replicated
  - Replicated effect sizes were half the size of those reported in the original studies
  - 36% of findings were significant in the replications, compared to 97% reported in the papers
- Some of the findings that didn't replicate...
  - People are more likely to cheat after they read a passage informing them that their actions are determined and thus that they don't have free will.
  - People make less severe moral judgements when they've just washed their hands.
  - Partnered women are more attracted to single men when they're ovulating

Marshmallow study

UNIVERSITY OF CANBERRA

- Achievement at 15 $\sim$ marshmallow test results + **confounding variables**
- All children in initial studies were recruited from Standford University's Nursery School
- Follow up replication studies revealed socio-economic status of children explained the variance in achievement at 15 much more than the marshmallow test results (Watts et al., 2018)

Take home messages

UNIVERSITY OF CANBERRA

- Design high powered studies, or consider the use of secondary datasets
- Novel finding? Replicate it!
- Always report effect sizes, ideally with the confidence intervals that surround them

Contributions to this course

UNIVERSITY OF CANBERRA

Dr James Neill

Dr Samantha Stanley

Dr Jeroen van Boxtel

References I

Ambikairajah, A., Foxe, D., de Lange, A.-M. G., Carrick, J., Cheung, S. C., Srikanth, V. K., Hwang, Y. T., Ahmed, R. M., Burrell, J. R., & Piguet, O. (2025). A Bayesian analysis of diagnostic timelines across Alzheimer's disease, frontotemporal dementia, and other neurodegenerative conditions. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *17*(3), e70184. https://doi.org/10.1002/dad2.70184

Ambikairajah, A., Khondoker, M., Morris, E., de Lange, A.-M. G., Saleh, R. N. M., Minihane, A. M., & Hornberger, M. (2024). Investigating the synergistic effects of hormone replacement therapy, apolipoprotein E and age on brain health in the UK Biobank. *Human Brain Mapping*, *45*(2), e26612. https://doi.org/10.1002/hbm.26612

Ambikairajah, A., Walsh, E., & Cherbuin, N. (2019). Lipid profile differences during menopause: A review with meta-analysis. *Menopause*, 1. https://doi.org/10/gf8kmj

References II

Ambikairajah, A., Walsh, E., Tabatabaei-Jafari, H., & Cherbuin, N. (2019). Fat mass changes during menopause: A metaanalysis. *American Journal of Obstetrics and Gynecology*, *221*(5), 393–409.e50. https://doi.org/10/gf39q6

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*(1), 1–20. https://doi.org/10.1037/h0076157

Horton, R. (2015). Offline: What is medicine's 5 sigma? *The Lancet*, *385*(9976), 1380. https://doi.org/10.1016/S0140-6736(15)60696-1

Kang, J. H., & Grodstein, F. (2012). Postmenopausal hormone therapy, timing of initiation, APOE and cognitive decline. *Neurobiology of Aging*, *33*(7), 1129–1137. https://doi.org/10.1016/j.neurobiolaging.2010.10.007

References III

Saleh, R. N. M., Hornberger, M., Ritchie, C. W., & Minihane, A. M. (2023). Hormone replacement therapy is associated with improved cognition and larger brain volumes in at-risk APOE4 women: Results from the European Prevention of Alzheimer's Disease (EPAD) cohort. *Alzheimer's Research & Therapy*, *15*(1), 10. https://doi.org/10.1186/s13195-022-01121-5

Ward, R. M. (2002). *Highly significant findings in psychology: A power and effect size survey* (5-B; Vol. 63, p. 2630). ProQuest Information & Learning.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Watts, T. W., Duncan, G. J., & Quan, H. (2018). Revisiting the Marshmallow Test: A Conceptual Replication Investigating Links Between Early Delay of Gratification and Later Outcomes. *Psychological Science*, *29*(7), 1159–1177. https://doi.org/10/gdm9cb

References IV

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604. https://doi.org/10.1037/0003-066X.54.8.594

Yaffe, K., Haan, M., Byers, A., Tangen, C., & Kuller, L. (2000). Estrogen use, APOE, and cognitive decline: Evidence of gene–environment interaction. *Neurology*, *54*(10), 1949–1954. https://doi.org/10.1212/WNL.54.10.1949