Statistical Inference with applications to health data

Professor Ailish Hannigan
Department of Mathematics and Statistics
Ailish.Hannigan@ul.ie

Outline

- About me
- Sampling
- Estimation
- Hypothesis testing
- Future topics

About me

Statistics divides the study of data into three parts:

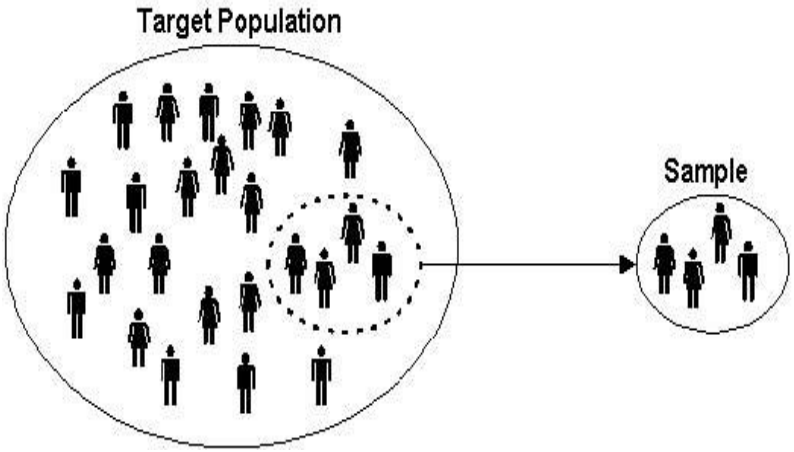
- Producing data
- Organising and describing data (Descriptive statistics)
- Drawing conclusions from data (Statistical inference)

Statistical Inference

To infer: to come to a conclusion based on evidence

Statistical inference: to come to a conclusion about a population using evidence from a sample

Sampling



- Representative sample ➡ unbiased estimates
- Large sample ➡ precise estimates

Representative sample

Who is in the sample?

How have they been selected?

Who do they represent?

Who can we generalise to?

A 1998 study falsely linked the MMR vaccine to autism and triggered a steep drop in immunisation rates. Nearly three decades later, the rumour persists.

How one controversial study shaped 20 years of menopause care

UK Biobank, big data, and the consequences of non-representativeness

biobank

Improving the health of future generations

A systematic review assessing the under-representation of elderly adults in COVID-19 trials

Unmasking the bias in facial recognition algorithms

Gender bias in artificial intelligence

Representative sample

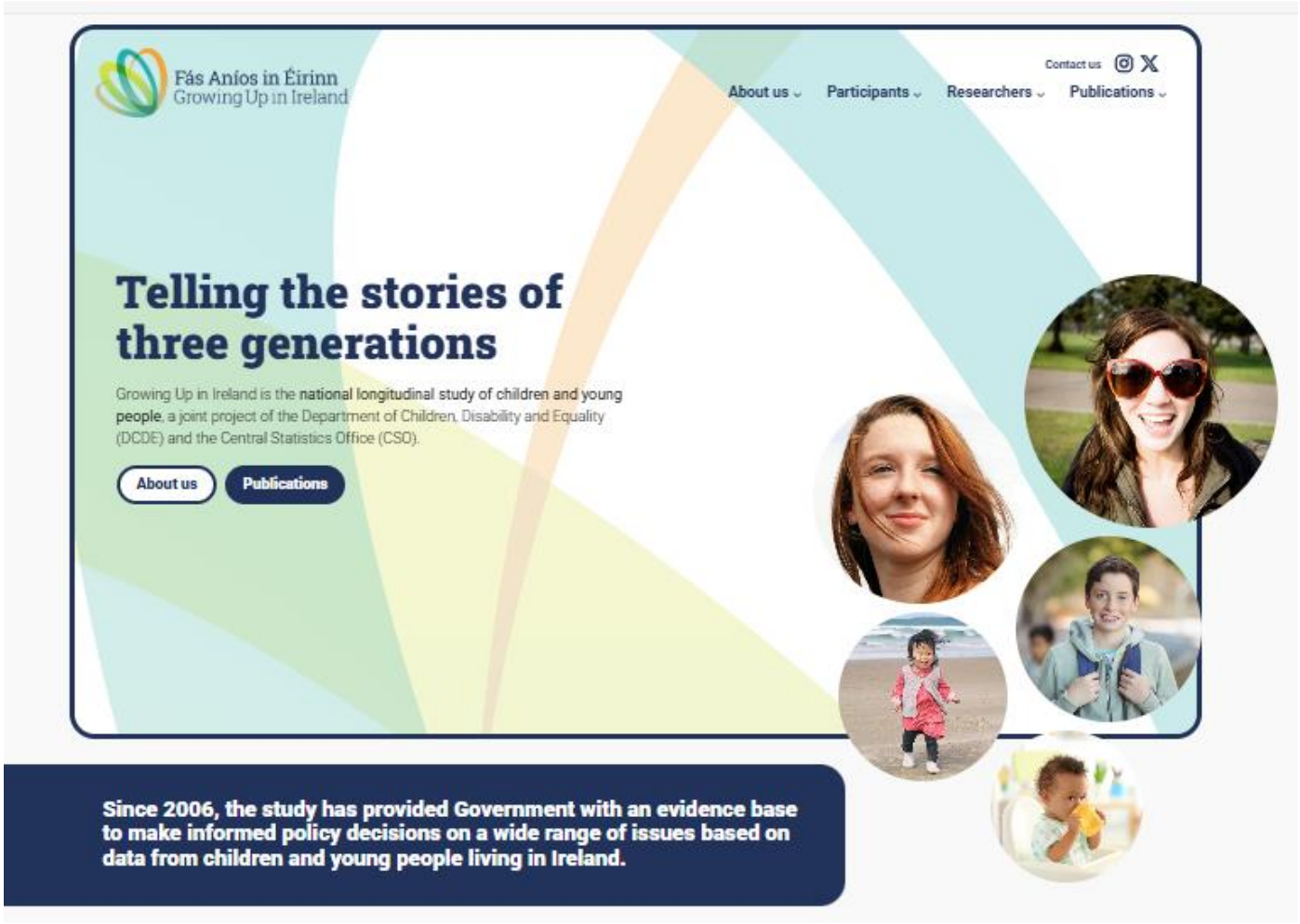
Who is in the sample?

How have they been selected?

Who do they represent?

Who can we generalise to?

Growing up in Ireland



Growing up in Ireland

- '98 Cohort started in 2008 with **8,500 children aged 9 years**. Information was collected from parents, teachers, principals and the children themselves. This cohort was revisited at age 13 years, 17/18 years, at age 20 and 25.

- '08 Cohort started in 2008 with over **11,000 9-month-olds** and their families. Follow-up waves were completed when the child was aged 3 years, 5 years, 7/8 years, 9 and at 13 years. Initial work on Cohort '08 at 17 has begun.

Sampling ‘98 cohort

- Two-stage sampling
- Sampling frame of 3,246 primary schools
- **Stratified random sample** of 1,105 primary schools

Stratified based on

- County
- Gender mix
- Disadvantaged status
- Religious denomination
- Size (total number of nine-year-old pupils)

Sampling ‘98 cohort

- Recruit all nine-year-old pupils from the selected schools up to a maximum threshold of 40 students in any one school.
- In schools which contained more than 40 nine-year-olds a random target sample of 40 pupils was selected by the school principal, with the assistance of the interviewer.

Sampling ‘98 cohort

- A total of 910 schools from the target sample of 1,105 were successfully recruited into the project, giving an overall school-level response of 82%
- Within schools, 57% of eligible children and their families returned the consent forms.

Representativeness

- Higher participation rates among better educated parents and those in higher social class categories
- Lone parent families under-represented
- Children in owner-occupied homes more likely to be represented
- **Reweighting using Census of Population**

Depression and chronic health conditions in parents of children with and without developmental disabilities: The growing up in Ireland cohort study

Stephen Gallagher [a], Ailish Hannigan [b]

RESEARCH Open Access

The impact of economic recession on the health of migrant fathers over time: results from the Growing up in Ireland longitudinal study

Nazmy Villarroel[1], Anne MacFarlane[2,3], Maria Roura[4], Alphonse Basogomba[5], Colette Bradley[6], Joseph W. LeMaster[7] and Ailish Hannigan[2,3*]

Longitudinal associations between parental incarceration and children's emotional and behavioural development: Results from a population cohort study

Daragh Bradshaw, Ailish Hannigan, Ann-Marie Creaven, Orla T. Muldoon

A Cluster Analysis of Reported Sleeping Patterns of 9-Month Old Infants and the Association with Maternal Health: Results from a Population Based Cohort Study

Aoife Hughes · Stephen Gallagher · Ailish Hannigan

Estimation

- **Point estimation**: one number which estimates the true value of the population parameter
- **Interval estimation**: a range of 'plausible' values the population parameter lies within with confidence (usually 95% confidence)

Estimating the prevalence of diabetes

In a **random sample** of 50 patients, 5 had diabetes.

Prevalence estimate: 10%
95% confidence interval: 2% to 18%

95% confident the true prevalence lies in the range 2% to 18%

Estimating the prevalence of diabetes

In a **random sample** of 500 patients, 50 had diabetes.

Prevalence estimate: 10%
95% confidence interval: 7% to 13%

Sample size

- Increasing the sample size results in a narrower confidence interval
- Narrow confidence intervals help us to draw more precise conclusions about the population and make better decisions.

Confidence intervals

‘A wide confidence interval points to lack of information, whether the difference is statistically significant or not, and is a warning against over-interpreting results from small studies.’

http://www.bmj.com/content/bmj/286/6376/1489.full.pdf

Hypothesis testing

A **statistical hypothesis** is a claim about the population that can be put to the test by using data from one or more samples

P-value

- Probability of obtaining the value of the test statistic (or even more extreme i.e. more unlikely data) if the null hypothesis were true

- **The smaller the p-value, the stronger the evidence against the null hypothesis**

When do we reject the null hypothesis?

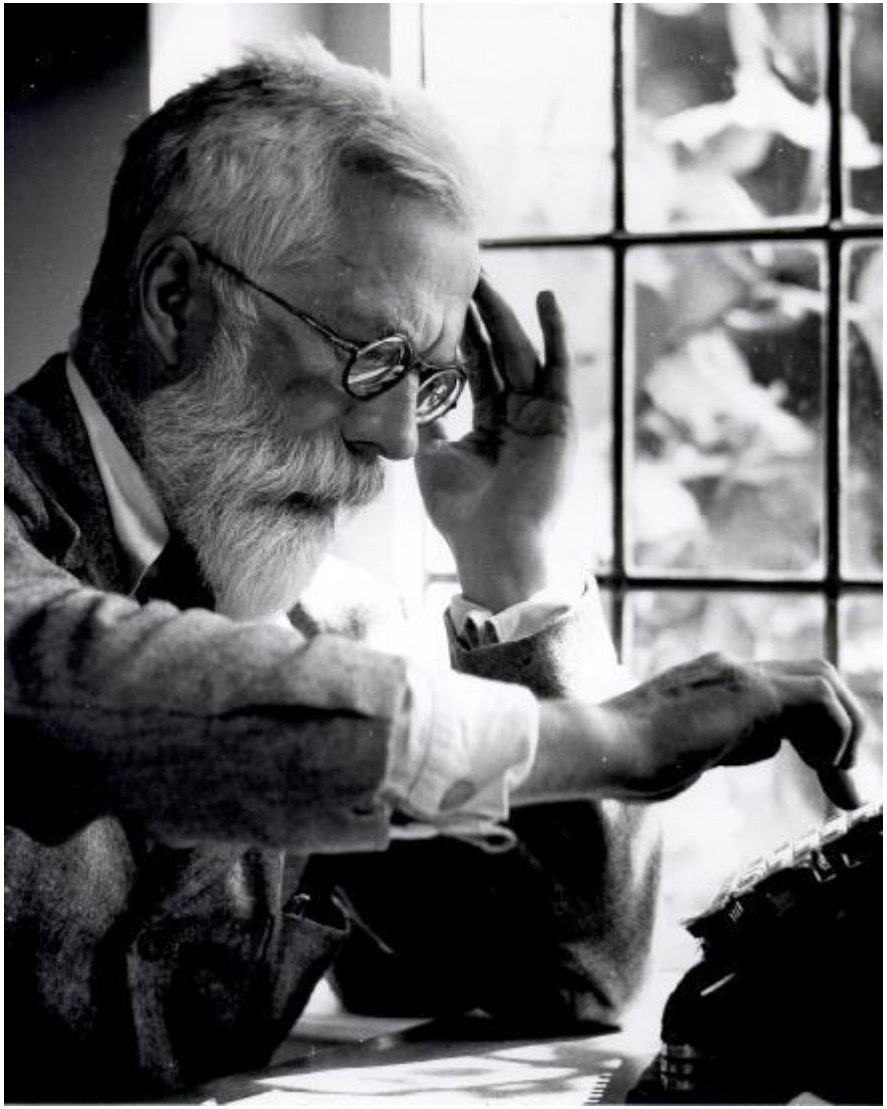
The decision level (probability) below which you reject the null hypothesis is called the significance level (usually 5% or 0.05)

- If the p-value < 0.05, we reject the null hypothesis
- This is called a **statistically significant** result

Why is the convention to use a cut-off of 0.05?

"The null hypothesis is never proved or established, but is possibly disproved"

"If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at .05 …"



R.A. Fisher (1890-1962)

Father of Statistical Inference and Statistical Genetics

From Fisher's obituary

"In character he was, let us admit it, a difficult man.

Among his wide circle of admirers he accepted homage with fair grace. With most other people he seemed to feel that his genius entitled him to more social indulgence than they were willing to concede: in short he could be, and not infrequently was, gratuitously rude."

American Statistical Association (2016)

"The p-value was never intended to be a substitute for scientific reasoning"

"A p-value, or statistical significance, does not measure the size of an effect **or the importance of a result**."

https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf

American Statistical Association (2016)

"**By itself**, a p-value does not provide a good measure of evidence regarding a model or hypothesis."

"*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."

https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf

Role of sample size in hypothesis testing

	Group 1	**Group 2**
Mean	3.5	3.4
Std. Deviation	1.2	1.3
Sample size	30	40
P-value from test comparing means=0.74		
Sample size	1000	2000
P-value from test comparing means=0.04		

Summary

- Representative sample
- Narrow confidence interval
- Careful interpretation of p-value

Future – Unstructured Data

- Text
- Images
- Video
- Audio
- Geospatial data
- Sensor data

Future – Data Science

‘a data scientist is someone who uses data to solve problems’

‘intersection of three areas: math/statistics, computation and a particular domain (e.g. biology)’

Thank you

UNIVERSITY OF LIMERICK
OLLSCOIL LUIMNIGH

University of Limerick,
Limerick, V94 T9PX,
Ireland.

Ollscoil Luimnigh,
Luimneach,
V94 T9PX, Éire.

+353 (0) 61 202020

ul.ie