Reminders

Upcoming due dates
Mon Feb 02 Quiz 4
Wed Feb 04 Project Proposal

Repo invites: Click accept in email!

Statistical Inference 1

Data Science in Practice

Jason G. Fleischer, PhD Dept. of Cognitive Science UC San Diego

We want to learn something about this

....but we can only *actually* collect data from this

CC licensed clip art from pixabay.com

Random samples are random!
They differ from each other and the population!

CC licensed clip art from pixabay.com

NIH Public Access
Author Manuscript
Epidemiology. Author manuscript; available in PMC 2014 January 01.

Published in final edited form as:
Epidemiology. 2013 January ; 24(1): 23–31. doi:10.1097/EDE.0b013e3182770237.

The Effect of Air Pollution Control on Life Expectancy in the United States: An Analysis of 545 US counties for the period 2000 to 2007

Andrew W. Correia,
Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, HSPH Building 2, 4th Floor, Boston, MA 02115

C. Arden Pope III,
Department of Economics, Brigham Young University, 142 Faculty Office Building, Provo, UT 84602

Douglas W. Dockery,
Departments of Environmental Health and Epidemiology, Harvard School of Public Health, 655 Huntington Avenue, HSPH Building 1, 1301B, Boston, MA 02115

Yun Wang,
Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, HSPH Building 2, 4th Floor, Boston, MA 02115

Majid Ezzati, and
MRC-HPA Centre for Environment and Health and Department of Epidemiology and Biostatistics, Imperial College London, Norfolk Place, St Mary's Campus, London W2 1PG

Francesca Dominici
Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, HSPH Building 2, 4th Floor, Boston, MA 02115, fdominic@hsph.harvard.edu, P: (617) 432-1056; F: (617)-739-1781

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3521092/

Is there a relationship between air pollution control and lifespan?

A decrease of 10 μg/m3 in the concentration of PM2.5 was associated with an increase in mean life expectancy of 0.35 years SD= 0.16 years, p = 0.033). This association was stronger in more urban and densely populated counties.

What if we want to know the effect of air pollution on everyone in the United States?

The sample:
545 US counties

#	State	Total Number of Counties	Total Area
1	Texas	254	268,596 mi^2
2	Georgia	159	59,425 mi^2
3	Virginia	133	42,775 mi^2
4	Kentucky	120	40,408 mi^2
5	Missouri	115	69,707 mi^2
6	Kansas	105	82,278 mi^2
7	Illinois	102	57,914 mi^2
8	North Carolina	100	53,819 mi^2
9	Iowa	99	56,273 mi^2
10	Tennessee	95	42,144 mi^2

3143 counties in the US

CA is #27

58 counties on 163k mi^2

Random Sampler

How would you want to select those 545 counties?

What criteria are ideal?

What do you think they did?

All counties with with available matching PM2.5 data for 2000 and 2007 from the EPA's Air Quality System. Includes both metropolitan and non-metro counties

States typically decide where monitors are placed based on areas of relatively high population and/or areas believed to have relatively higher pollutant concentrations. Each state is responsible for developing its own monitoring plan, which is then reviewed and revised every five years. Aug 28, 2023

United States Environmental Protection Agency (.gov)
https://www.epa.gov › outdoor-air-quality-data › who-de...

Who decides where monitors get placed? | US EPA

Distributions of land area and population size among US counties

Random Sampler

How might conclusions differ if…

the sample locations were instead completely spatially at random across the USA?

The sample came from the 1980s or the 2020s instead of the 2000s?

Approaches to Inference

CORRELATION

COMPARISON OF MEANS

REGRESSION

NON-PARAMETRIC TESTS

CORRELATION

ASSOCIATION BETWEEN VARIABLES

i.e. Pearson Correlation, Spearman Correlation, chi-square test

COMPARISON OF MEANS

DIFFERENCE IN MEANS BETWEEN VARIABLES

i.e. t-test, ANOVA

REGRESSION

DOES CHANGE IN ONE VARIABLE MEAN CHANGE IN ANOTHER?

I.e. simple regression, multiple regression

NON-PARAMETRIC TESTS

FOR WHEN ASSUMPTIONS IN THESE OTHER 3 CATEGORIES ARE NOT MET

i.e. Wilcoxon rank-sum test, Wilcoxon sign-rank test, sign test

Dependent Variable
Independent Variable

weaker relationship

stronger relationship

stronger relationship = higher correlation

The *larger* the independent variable value, the *larger* the dependent variable tends to be

The *smaller* the independent variable value, the *smaller* the dependent variable tends to be

The *taller* the father, the taller his son tends to be

The *shorter* the father, the shorter his son tends to be

The *smaller* the independent variable value, the *larger* the dependent variable tends to be

The *larger* the independent variable value, the *smaller* the dependent variable tends to be

The *lower* the number of absences, the *higher* the students' grades tend to be

The more absences, the lower the students' grades

Pearson's *r* :

linear correlation between two variables

takes values [-1,1]

Correlation is how close the data are to being in a line…
BUT IT HAS NOTHING TO DO WITH THE SLOPE

https://www.spss-tutorials.com/pearson-correlation-coefficient/

r = 0.7
r = 0.3
r = 0
r = -0.7
r = 0.3
r = 0

https://www.guessthecorrelation.com/

Which of the following is the Pearson correlation coefficient (*r*) for this relationship?

https://forms.gle/M4Q3hFkeoLKCixUX8

Correlation != Causation

http://www.tylervigen.com/spurious-correlations

What about correlation between categorical variables?

- Make a contingency table …

Sex \ Handedness	Right-handed	Left-handed	Total
Male	43	9	52
Female	44	4	48
Total	87	13	100

- Then calculate phi (based on chi-squared statistic) …

Phi coefficient [edit]

Main article: Phi coefficient

A simple measure, applicable only to the case of 2 × 2 contingency tables, is the phi coefficient (φ) defined by

$$\phi = \pm\sqrt{\frac{\chi^2}{N}},$$

where χ^2 is computed as in Pearson's chi-squared test, and N is the grand total of observations. φ varies from 0 (corresponding to no association between the variables) to 1 or −1 (complete association or complete inverse association), provided it is based on frequency data represented in 2 × 2 tables. Then its sign equals the sign of the product of the main diagonal elements of the table minus the product of the off–diagonal elements. φ takes on the minimum value −1.0 or the maximum value of +1.0 if and only if every marginal proportion is equal to 0.5 (and two diagonal cells are empty).[2]

- Yes there are ways to do this in pandas/numpy and other libraries!

CORRELATION

ASSOCIATION BETWEEN VARIABLES

i.e. Pearson Correlation, Spearman Correlation, chi-square test

COMPARISON OF MEANS

DIFFERENCE IN MEANS BETWEEN CONDITIONS

i.e. t-test, ANOVA

REGRESSION

DOES CHANGE IN ONE VARIABLE MEAN CHANGE IN ANOTHER?

I.e. simple regression, multiple regression

NON-PARAMETRIC TESTS

FOR WHEN ASSUMPTIONS IN THESE OTHER 3 CATEGORIES ARE NOT MET

i.e. Wilcoxon rank-sum test, Wilcoxon sign-rank test, sign test

t-test:

tests for difference in means between groups

William Sealy Gosset (13 June 1876 – 16 October 1937) was an English statistician, chemist and brewer who served as Head Brewer of Guinness and Head Experimental Brewer of Guinness and was a pioneer of modern statistics. He pioneered small sample experimental design and analysis with an economic approach to the logic of uncertainty. Gosset published under the pen name **Student** and developed most famously Student's t-distribution – originally called Student's "z" – and "Student's test of statistical significance".[1]

Contents [hide]

1 Life and career
2 See also
3 Bibliography
4 References
5 Further reading
6 External links

Life and career [edit]

Born in Canterbury, England the eldest son of Agnes Sealy Vidal and Colonel Frederic Gosset, R.E. Royal Engineers, Gosset attended Winchester College before matriculating as Winchester Scholar in natural sciences and mathematics at New College, Oxford. Upon graduating in 1899, he joined the brewery of Arthur Guinness & Son in Dublin, Ireland; he spent the rest of his 38-year career at Guinness.[1][2]

Gosset had three children with Marjory Gosset (née Phillpotts). Harry Gosset (1907–1965) was a consultant paediatrician; Bertha Marian Gosset (1909–2004) was a geographer and nurse; the youngest, Ruth Gosset (1911–1953) married the Oxford mathematician Douglas Roaf and had five children.

In his job as Head Experimental Brewer at Guinness, the self-trained Gosset developed new statistical methods – both in the brewery and on the farm – now central to the design of experiments, to proper use of significance testing on repeated trials, and to analysis of economic significance (an early instance of decision theory interpretation of statistics) and more, such as his small-sample, stratified, and repeated balanced experiments on barley for proving the best yielding varieties.[3] Gosset acquired that knowledge by study, by trial and error, by cooperating with others, and by spending two terms in 1906–1907 in the Biometrics laboratory of Karl Pearson.[4] Gosset and Pearson had a good relationship.[4] Pearson helped Gosset with the mathematics of his papers, including the 1908 papers, but had little appreciation of their importance. The papers addressed the brewer's concern with small samples; biometricians like Pearson, on the other hand, typically had hundreds of observations and saw no urgency in developing small-sample methods.[2]

Gosset's first publication came in 1907, "On the Error of Counting with a Haemacytometer," in which – unbeknownst to Gosset aka "Student" – he rediscovered the Poisson distribution.[3] Another researcher at Guinness had previously published a paper containing trade secrets of the Guinness

William Sealy Gosset

William Sealy Gosset (aka *Student*) in 1908 (age 32)

Born	13 June 1876 Canterbury, Kent, England
Died	16 October 1937 (aged 61) Beaconsfield, Buckinghamshire, England
Other names	Student
Alma mater	New College, Oxford, Winchester College
Known for	Student's t-distribution, statistical significance, design of experiments, Monte Carlo method, quality control, Modern synthesis, agricultural economics, econometrics
Children	5, including Isaac Henry Gosset
Scientific career	

Do the heights between males and females differ?

https://ourworldindata.org/human-height

t-test Assumptions

1. Data are continuous
2. Normally distributed
3. Equal variance b/w groups (but can use Welch's test!)
4. Not paired (will talk more about this later)

Do the heights between males and females differ?

variance is equal(ish) between groups (57 vs 50cm)

data are normally distributed

data are continuous

sample size affects statistic

N=10,000

Do the heights between males and females differ?

t-statistic: -95.6

p-value << 0.001

95% CI for true difference in means
[-5.43, -5.21]

Yes.

p-value : the probability under the **null hypothesis** of getting measurements as extreme as the observed results by chance alone

THIS IS NOT TYPE 1 ERROR RATE

https://forms.gle/6MCyp7qFsaHgGKi5A

If we flip a coin 16 times and record the number of heads....
….and then do that 1M times

...this is what happens by chance alone.

~ 40% of the time you will see 8 heads & 8 tails
proportion
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
number of heads

p-value : the probability of getting the observed results (or results more extreme) by chance alone
proportion
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
number of heads

0.4
0.3
0.2
0.1
0.0
proportion
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
number of heads
In your trial, you get 10 heads.
p-value : the probability of getting the observed results (or results more extreme) by chance alone

0.4
0.3
0.2
0.1
0.0
proportion
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
number of heads
The results you saw (10 heads) *and* all results more extreme than what you saw are highlighted here.
p-value : the probability of getting the observed results (or results more extreme) by chance alone

The probability of getting 10 heads *or something more extreme* is calculated by adding the number of times these all occurred, divided by the total number of head flips
p-value : the probability of getting the observed results (or results more extreme) by chance alone
count
4e+05
3e+05
2e+05
1e+05
0e+00
2
218
5,87
7
60,73
1
60,766
5,973
5,973
208
2
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
number of heads

The probability of getting 10 heads *or something more extreme* is
# of 10 or more extreme flips / total flips
p-value : the probability of getting the observed results (or results more extreme) by chance alone
count
4e+05
3e+05
2e+05
1e+05
0e+00
2
218
5,877
60,731
60,766
5,973
208
2
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
number of heads

The probability of getting 10 heads *or something more extreme* is
# of 10 or more extreme flips / total flips
(2 + 218 + 5,877 + 60,731 + 60,766 + 5,973 + 208 + 2) / 1x10^6
= 133,777 / 1x10^6
p-value : the probability of getting the observed results (or results more extreme) by chance alone
count
4e+05
3e+05
2e+05
1e+05
0e+00
2
218
5,877
60,731
60,766
5,973
208
2
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
number of heads

The probability of getting 10 heads *or something more extreme* is
# of 10 or more extreme flips / total flips
(2 + 218 + 5,877 + 60,731 + 60,766 + 5,973 + 208 + 2) / 1x10^6
= 133,777 / 1x10^6
= 0.133 (13.3%)
p-value : the probability of getting the observed results (or results more extreme) by chance alone
count
4e+05
3e+05
2e+05
1e+05
0e+00
2
218
5,877
60,731
60,766
5,973
208
2
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
number of heads

The probability of getting 10 heads *or something more extreme* is
# of 10 or more extreme flips / total flips
(2 + 218 + 5,877 + 60,731 + 60,766 + 5,973 + 208 + 2) / 1x10^6
= 133,777 / 1x10^6
= 0.133 (13.3%)
p-value : the probability of getting the observed results (or results more extreme) by chance alone
count
4e+05
3e+05
2e+05
1e+05
0e+00
2
218
5,877
60,731
60,766
5,973
208
2
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
number of heads

p-value : 0.133

0.4
0.3
0.2
0.1
0.0
proportion
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
number of heads
What if you observed 16 heads??
p-value : the probability of getting the observed results (or results more extreme) by chance alone

What would be the p-value of you flipping 16 heads?

A
< 0.13

B
> 0.13

The interpretation of p-values in NHST

- Null hypothesis statistical testing (NHST) means we assume there is no difference
- P-value represents
  - What's the likelihood of getting values this weird **given an assumption that A and B come from the same distribution**? We call this the null hypothesis assumption or H_0
  - A p-value is $P(D \mid H_0)$
- Historically because some old dead guy said so, we take a value of p=0.05 as a threshold of "significance"
- This represents a 1/20 chance of this data under the null assumption
- What happens if you ask 20 questions using NHST?
  - Take a look at this: https://xkcd.com/882/

The interpretation of p-values in NHST

- When your data are very inconsistent with the null hypothesis, P values can’t determine which of the following two possibilities is more probable:
  - The null hypothesis is true, but your sample is unusual due to random sampling error.
  - The null hypothesis is false
- When your data show the null has a high probability, then we can’t determine which of the following is more probable:
  - The null hypothesis is false, but your sample is unusual due to random sampling error.
  - The null hypothesis is true

The interpretation of p-values in NHST

- P-value is NOT the chance of a Type I error! This is a common mistake!
  - Type 1 error: rejecting the null, when it is really true
    - AKA a false positive result
    - Notated as $P(H_0 \mid D)$
  - Bayesian viewpoint: What's the prior probability of the null?
    - If this is the first study ever on a topic then we know less about the truth than if this is the 36th study on a topic
    - $P(H_0 \mid D) = P(D \mid H_0)P(H_0)$
  - Empirically a p=0.05 is usually a **MUCH BIGGER Type 1** error rate
    - Can't know in reality, but simulation / math studies tell us that we can expect something like for p=0.05 a Type 1 error rate of somewhere like 23 to 50%

Difference in Means

Why would a t-test *not* be appropriate for these data?

A
Not normally distributed

B
Unequal variances

C
Small sample size

D
Data are not continuous

Would a t-test find a significant difference in means?

A
t-test not appropriate

B
Yes

C
No

D
Need more information

https://twitter.com/brankomilan/status/406105955080757249

Paired data

Independent samples t-test

Is there a **difference** between **two groups**

Paired samples t-test

Is there a **difference** in a **group** between **two points in time**

https://datatab.net/tutorial/paired-t-test

Do the heights between males and females differ?

variance is equal(ish) between groups (57 vs 50cm)

data are normally distributed

data are continuous

sample size affects statistic

N=10,000

Cohen's *d*

Cohen's *d* is defined as the difference between two means divided by a standard deviation for the data

Effect sizes!