Statistical Inference with R

Categorical Variables

• • •

GW Libraries Workshop

Dan Kerchner ~ February 12, 2021

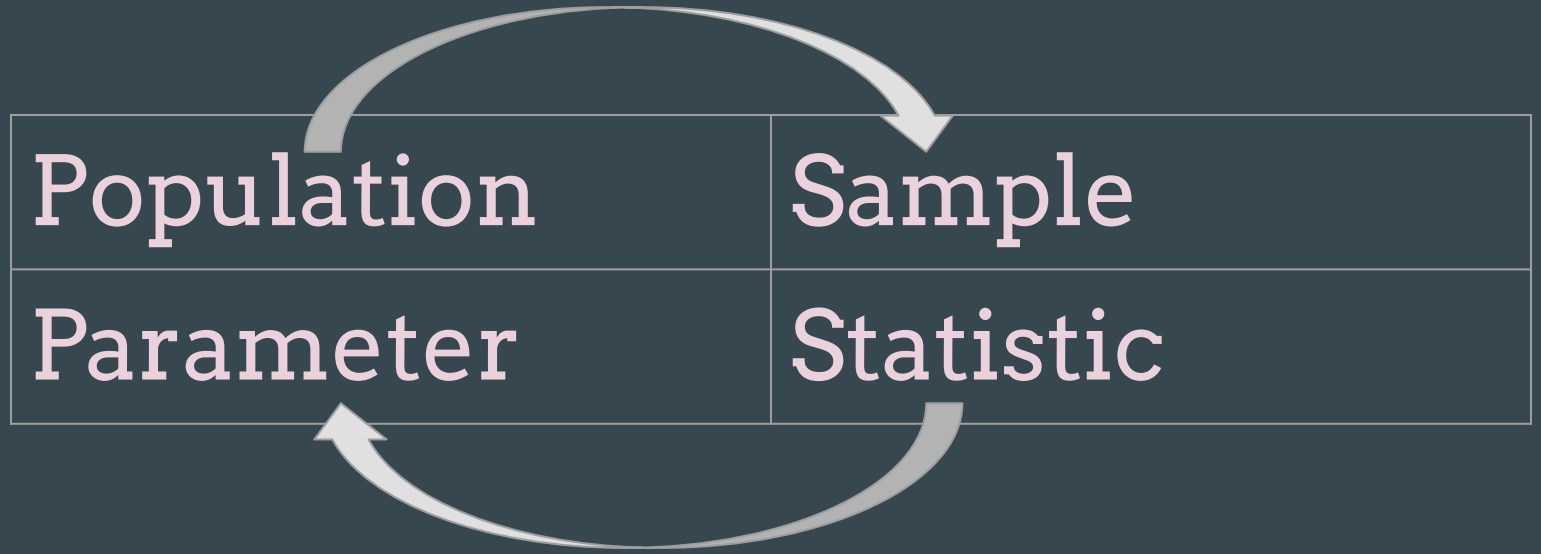
go.gwu.edu/rstats

Logistics

- Just speak up OR use the Zoom chat
- Kiri can provide individual help
- Plan for 1 brief break

Super-Brief Review of Inference for Categorical Variables

High-Level Objective



Categorical Data Analysis

- Categorical variables
  - binary/dichotomous - 2 levels
  - 3+ levels - nominal or ordinal

- Categorical data analysis
  - Response is categorical
  - Predictors may be numerical and/or categorical

Representations of Categorical Data

- # of times a category/level occurs
- Proportion/frequency of a level occurring

Proportions

Population:

A, B, AB, O, O, B, A, B,
AB, O, B, A, O, AB, O, O,
B, A, B, AB, O, O, B, A,
B, AB, O, B, A, O, AB, O,
O,B, A, B, AB, O, O, B, A,
B, AB, O, B, A, O, AB, O,
O, B, A, B, AB, O, O, B,
A, B, AB, O, B, A, O, AB,
O, O,B, A, B, AB, O, O, B,
A, B, AB, O, B, A, O, AB,
O, O, B, A, B, AB, O, O,
B, A, B, AB, O, B, A, O,
AB, O, O, O, B, O, AB, ...

Sample

A

AB

O

O

B

B

O

AB

O

O

Sample Proportions

A	20%
B	20%
AB	10%
O	50%

Categorical Predictor, (Binary) Categorical Response

Hospital	Outcome
GW	1
Georgetown	0
Sibley	1
Georgetown	1
Sibley	1
GW	0
GW	1
Sibley	0
Georgetown	1
GW	0
Sibley	1
GW	1
Georgetown	0
Sibley	1
GW	0
Georgetown	0
GW	1

	Outcome = 0	Outcome = 1
GW	3	4
Georgetown	3	2
Sibley	1	3

OR, RR, RD: Odds Ratio, Risk Ratio, Risk Difference

	Diseased	Healthy
Exposed	D_E	H_E
Not exposed	D_N	H_N

Risk Ratio (RR) = $\frac{D_E/(D_E+H_E)}{H_E/(D_E+H_E)}$

Risk Difference (RD) = $D_E/(D_E+H_E)$ - $D_N/(D_N+H_N)$

Odds Ratio (OR) = $\frac{D_E/H_E}{H_E/D_E}$

Two forms of inference

Confidence Interval

95% CI for π = (0.44, 0.49)

Hypothesis Testing

H_0: $\pi = \pi_0$ ← Null Hypothesis
H_A: $\pi \neq \pi_0$ ← Alternative Hypothesis

p-value: Chance that we are rejecting H_0 when we should not

Test of proportions

H_0: $\pi_1 = \pi_{1,0}, \pi_2 = \pi_{2,0}, \pi_3 = \pi_{3,0}$ ← Null Hypothesis

H_A: $\pi_1 \neq \pi_{1,0}, \pi_2 \neq \pi_{2,0}, \pi_3 \neq \pi_{3,0},$ ← Alternative Hypothesis

Inference with Odds Ratios, Risk Ratios, Risk Differences

H_0: OR = 1 (or RR = 1 or RD = 0) ← Null Hypothesis

H_A: OR ≠ 1 (or RR ≠ 1 or RD ≠ 0) ← Alternative Hypothesis

Prerequisites ~ Assumptions

For proportion test, χ^2 test, OR/RR/RD

- observations are independent
- $n \geq 5$ in each group* (observed vs. expected - depends)
- proportion of interest is not too close to 0 or 1

When the assumptions are not satisfied, we may use other approaches (Nonparametric tests, bootstrapping, etc.)

Goals

Today's Goal

- Learn to use R to read in data, estimate measures and conduct hypothesis tests for **categorical** measures
  - Checking assumptions
  - Visualizing
  - Computing p-values and confidence intervals

Today: 3 Scenarios

- Binomial variable: Single population proportion
- Multinomial variables / proportions
- Association between binomial predictor & binomial response

Today's Data Set #1

MacMahon, B., Cole, P., Lin, T. M., Lowe, C. R., Mirra, A. P., Ravnihar, B., Salber, E. J., Valaoras, V. G., & Yuasa, S. (1970). **Age at first birth and breast cancer risk**. Bulletin of the World Health Organization, 43, 209-221.

Bull. Org. mond. Santé
Bull. Wld Hlth Org. } 1970, 43, 209-221

Age at First Birth and Breast Cancer Risk *

B. MACMAHON,[1] P. COLE,[2] T. M. LIN,[3] C. R. LOWE,[4] A. P. MIRRA,[5] B. RAVNIHAR,[6] E. J. SALBER,[7] V. G. VALAORAS [8] & S. YUASA [9]

An international collaborative study of breast cancer and reproductive experience has been carried out in 7 areas of the world. In all areas studied, a striking relation between age at first birth and breast cancer risk was observed. It is estimated that women having their first child when aged under 18 years have only about one-third the breast cancer risk of those whose first birth is delayed until the age of 35 years or more. Births after the first, even if they occur at an early age, have no, or very little, protective effect. The reduced risk of breast cancer in women having their first child at an early age explains the previously observed inverse relationship between total parity and breast cancer risk, since women having their first birth early tend to become ultimately of high parity. The association with age at first birth requires different kinds of etiological hypotheses from those that have been invoked in the past to explain the association between breast cancer risk and reproductive experience.

One of the most consistently observed epidemiological characteristics of breast cancer is the inverse association between the number of children a woman has borne and her risk of developing the disease. This association has been observed in all geographic areas and ethnic groups in which it has been studied. The association has been interpreted as indicating that some concomitant of pregnancy protects against the later development of breast cancer, the amount of protection being related to the number of pregnancies.

Analyses of data from a recent international collaborative study have shown that breast cancer risk is strongly correlated with age at first pregnancy (Lowe & MacMahon, 1970; Salber, Trichopoulos & MacMahon, 1969; Valaoras et al., 1969; Yuasa & MacMahon, 1970; and Lin, Chen & MacMahon; Ravnihar, MacMahon & Lindtner; Mirra & Cole, unpublished data). These analyses were based on the women's ages at their first pregnancy, even if that pregnancy aborted. Differences between cases and controls with respect to frequency of abortion were observed in only a few centres and were in the direction which suggested increased risk associated with abortion—contrary to the reduction in risk associated with full-term births. Therefore, it seemed worth while to conduct analyses restricting attention to the age at which the first full-term birth occurred. The details are presented in this paper. The analysis has also been extended to take a more detailed account of possible interrelationships with other variables and to examine the effect of age at confinements, other than the first.

METHODS

* This study was supported by Grant E-385 A from the American Cancer Society, Grant 402-C-200 from the Boris Kidrič Fund of Yugoslavia, a grant from the Medical Research Council of Great Britain, a grant from the Ministry of Health and Welfare, Japan, Grant 5 PO1 CA 06373 from the US National Cancer Institute, a grant from the National Council for Science of China (Taiwan) and Grants R/00057, R/00062, R/00072 and C2/181/24 from the World Health Organization.

[1] Professor, Department of Epidemiology, Harvard School of Public Health, Boston, Mass., USA.

[2] Assistant Professor, Department of Epidemiology, Harvard School of Public Health, Boston, Mass., USA.

[3] Associate Professor, Department of Epidemiology, College of Medicine, National Taiwan University, Taipei, Taiwan.

[4] Mansel Talbot Professor, Department of Social and Occupational Medicine, Welsh National School of Medicine, University of Wales, Cardiff, Wales.

[5] Director, Central Cancer Registry, São Paulo, Brazil.

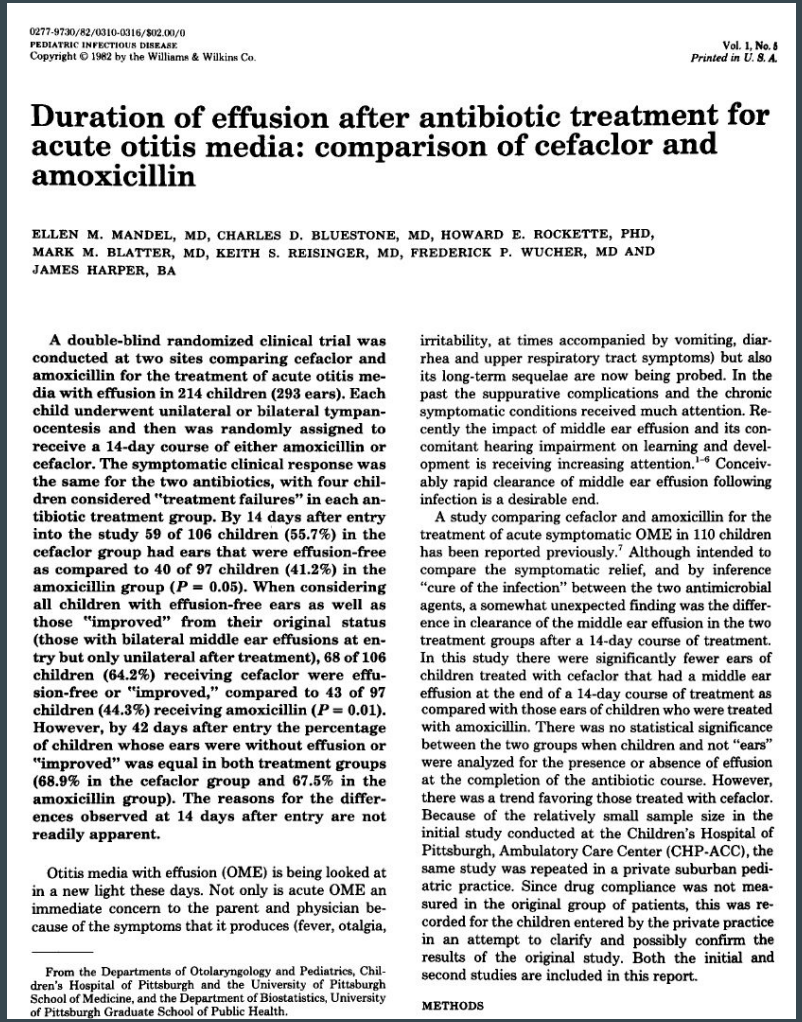
[6] Professor, Institute of Oncology, Medical Faculty, University of Ljubljana, Yugoslavia.

[7] Senior Research Associate, Department of Epidemiology, Harvard School of Public Health, Boston, Mass., USA.

[8] Professor, Department of Hygiene and Epidemiology, University of Athens, Athens, Greece.

Today's Data Set #2

Mandel, E., Bluestone, C. D., Rockette, H. E., Blatter, M. M., Reisinger, K. S., Wucher, E. P., & Harper, J. (1982). **Duration of effusion after antibiotic treatment for acute otitis media: Comparison of cefaclor and amoxicillin**. Pediatric Infectious Diseases, 1, 310–316.



0277-9730/82/0310-0316/$02.00/0
PEDIATRIC INFECTIOUS DISEASE
Copyright © 1982 by the Williams & Wilkins Co.
Vol. 1, No. 5
Printed in U. S. A.

Duration of effusion after antibiotic treatment for acute otitis media: comparison of cefaclor and amoxicillin

ELLEN M. MANDEL, MD, CHARLES D. BLUESTONE, MD, HOWARD E. ROCKETTE, PHD, MARK M. BLATTER, MD, KEITH S. REISINGER, MD, FREDERICK P. WUCHER, MD AND JAMES HARPER, BA

A double-blind randomized clinical trial was conducted at two sites comparing cefaclor and amoxicillin for the treatment of acute otitis media with effusion in 214 children (293 ears). Each child underwent unilateral or bilateral tympanocentesis and then was randomly assigned to receive a 14-day course of either amoxicillin or cefaclor. The symptomatic clinical response was the same for the two antibiotics, with four children considered "treatment failures" in each antibiotic treatment group. By 14 days after entry into the study 59 of 106 children (55.7%) in the cefaclor group had ears that were effusion-free as compared to 40 of 97 children (41.2%) in the amoxicillin group ($P = 0.05$). When considering all children with effusion-free ears as well as those "improved" from their original status (those with bilateral middle ear effusions at entry but only unilateral after treatment), 68 of 106 children (64.2%) receiving cefaclor were effusion-free or "improved," compared to 43 of 97 children (44.3%) receiving amoxicillin ($P = 0.01$). However, by 42 days after entry the percentage of children whose ears were without effusion or "improved" was equal in both treatment groups (68.9% in the cefaclor group and 67.5% in the amoxicillin group). The reasons for the differences observed at 14 days after entry are not readily apparent.

Otitis media with effusion (OME) is being looked at in a new light these days. Not only is acute OME an immediate concern to the parent and physician because of the symptoms that it produces (fever, otalgia, irritability, at times accompanied by vomiting, diarrhea and upper respiratory tract symptoms) but also its long-term sequelae are now being probed. In the past the suppurative complications and the chronic symptomatic conditions received much attention. Recently the impact of middle ear effusion and its concomitant hearing impairment on learning and development is receiving increasing attention.[1–6] Conceivably rapid clearance of middle ear effusion following infection is a desirable end.

A study comparing cefaclor and amoxicillin for the treatment of acute symptomatic OME in 110 children has been reported previously.[7] Although intended to compare the symptomatic relief, and by inference "cure of the infection" between the two antimicrobial agents, a somewhat unexpected finding was the difference in clearance of the middle ear effusion in the two treatment groups after a 14-day course of treatment. In this study there were significantly fewer ears of children treated with cefaclor that had a middle ear effusion at the end of a 14-day course of treatment as compared with those ears of children who were treated with amoxicillin. There was no statistical significance between the two groups when children and not "ears" were analyzed for the presence or absence of effusion at the completion of the antibiotic course. However, there was a trend favoring those treated with cefaclor. Because of the relatively small sample size in the initial study conducted at the Children's Hospital of Pittsburgh, Ambulatory Care Center (CHP-ACC), the same study was repeated in a private suburban pediatric practice. Since drug compliance was not measured in the original group of patients, this was recorded for the children entered by the private practice in an attempt to clarify and possibly confirm the results of the original study. Both the initial and second studies are included in this report.

From the Departments of Otolaryngology and Pediatrics, Children's Hospital of Pittsburgh and the University of Pittsburgh School of Medicine, and the Department of Biostatistics, University of Pittsburgh Graduate School of Public Health.

METHODS

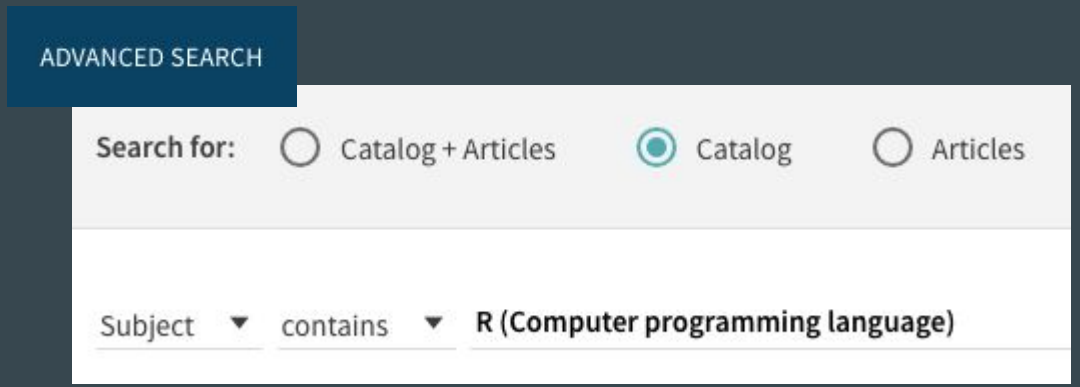
Some Handy R Links

Tutorials

- RStudio R paths: education.rstudio.com/learn/
- Data Carpentry & Software Carpentry:
  - datacarpentry.org
  - software-carpentry.org
- Linkedin Learning @ GW: go.gwu.edu/linkedinlearning
- r-tutor.com/r-introduction & r-tutor.com/elementary-statistics
- UCLA Data Analysis Examples: stats.idre.ucla.edu/other/dae/
- R Graph Gallery (w/code): r-graph-gallery.com

Books you can access for free

- Free books online - Hadley Wickham:
  - R for Data Science r4ds.had.co.nz
  - Advanced R adv-r.hadley.nz/
- Through your GW library privileges:



Reference Links

- R language (CRAN): r-project.org
- R search engine: rseek.org
- rstudio.com
  - Cheat Sheets! rstudio.com/resources/cheatsheets
- stackoverflow.com

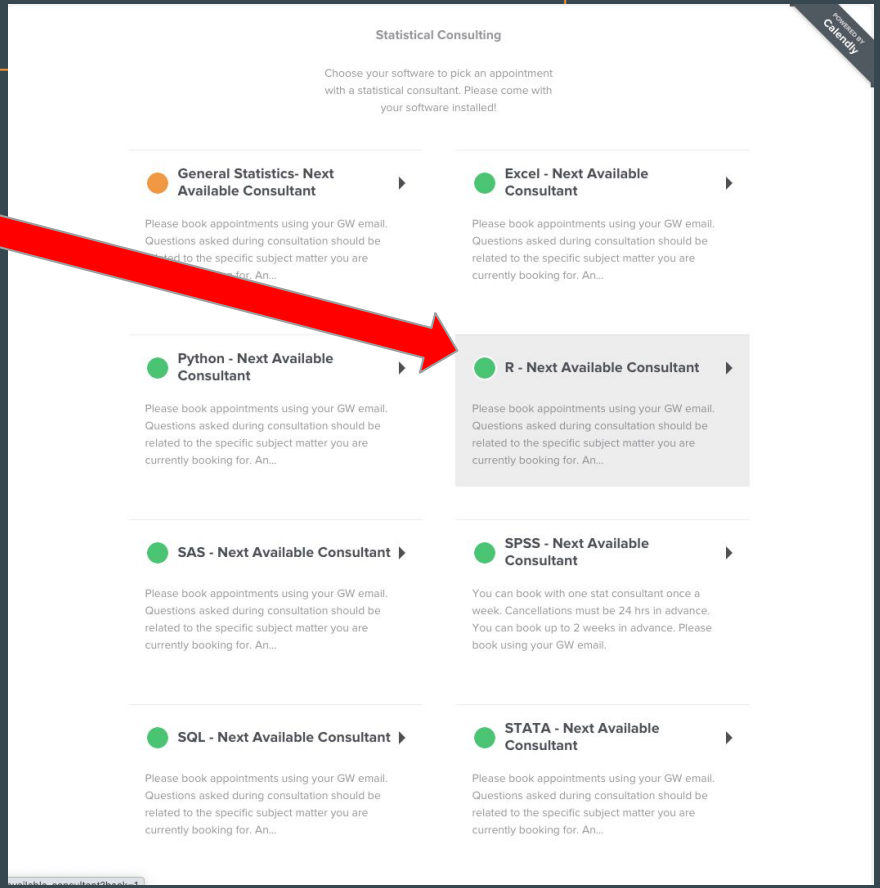
Statistics+R help @ GW

R-Statistics Appointments:
calendly.com/statistical-consulting-gw

Also...

Appointments with me:
calendly.com/kerchner

Coding consultations (Python, git, etc.):
calendly.com/gwul-coding/



Thanks!

Dan Kerchner kerchner@gwu.edu

Disovankiri Boung dboung@gwu.edu