Providence SWEDISH

# Tools and Tips for Statistical Data Analysis

Jack Huber, Ph.D.
Senior Clinical Data Analyst
Providence Swedish Pharmacy

# Observations about statistics in healthcare

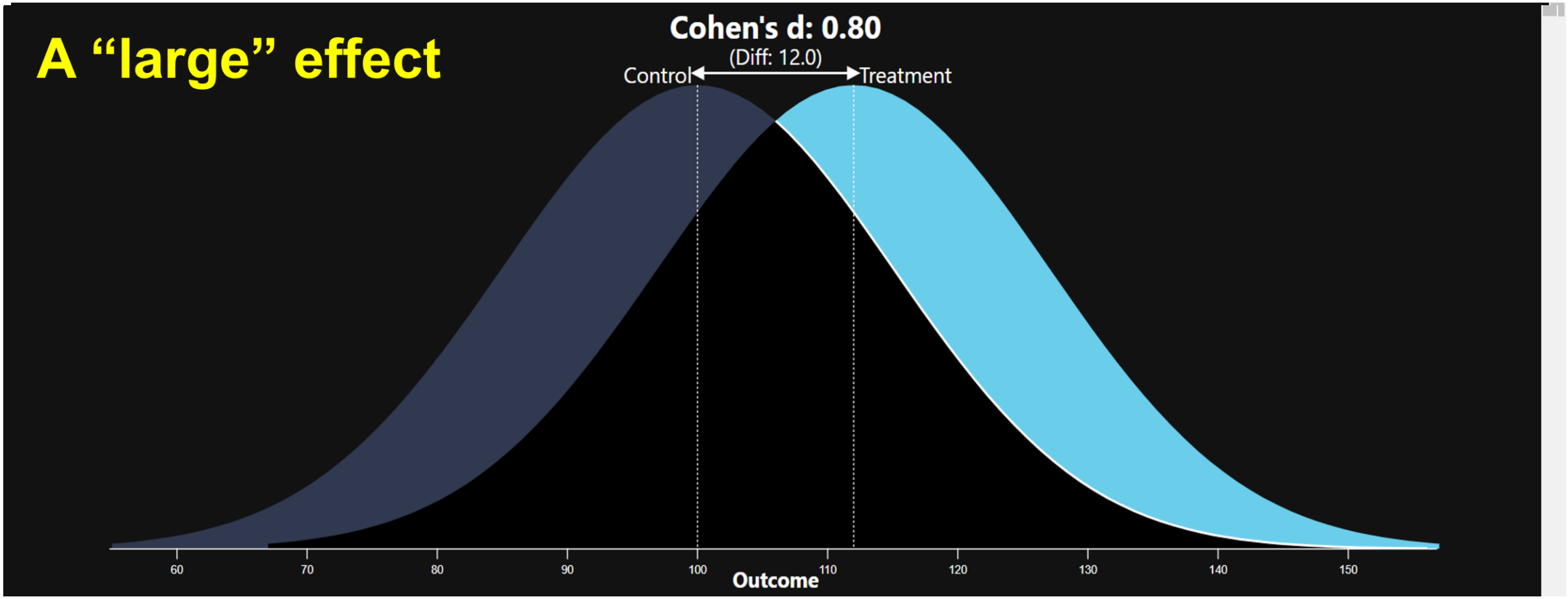
- **We value data.**
- **Statistical literacy varies.**
- **Residents need help with statistics** (Newsome et al., 2018; Windish et al., 2007)
- **Increasingly sophisticated statistics in medical journals** (Arnold et al., 2013; Horton & Switzer, 2005; Windish et al., 2007; Yi et al., 2020)
- **Growing presence of AI / machine learning methods**
- **Service**

# Where help seems needed

1. How many patients do I need for statistics?
2. What statistics should I use with my data?
3. How do I do a chi-square (or other statistic) in Excel (or other program)?

# How many patients do I need for statistics?

## Power analysis: A quick primer on effect sizes



# How many patients do I need for statistics?

## Power analysis

**Power: "the ability to find a treatment effect when it really does exist"** (Cohen, 1988)

**$\chi^2$ test of independence**
**(comparing two groups on a dichotomous outcome)**

| | Effect size | | |
|---|---|---|---|
| | "Large" (0.8) | "Moderate" (0.5) | "Small" (0.2) |
| Power | .8 | .8 | .8 |
| p-value | .05 | .05 | .05 |
| Sample size | **31** | **87** | **784** |

SOURCE: pwr package in R

# What statistics should I use?

## Understand the measurement level of the outcome variable

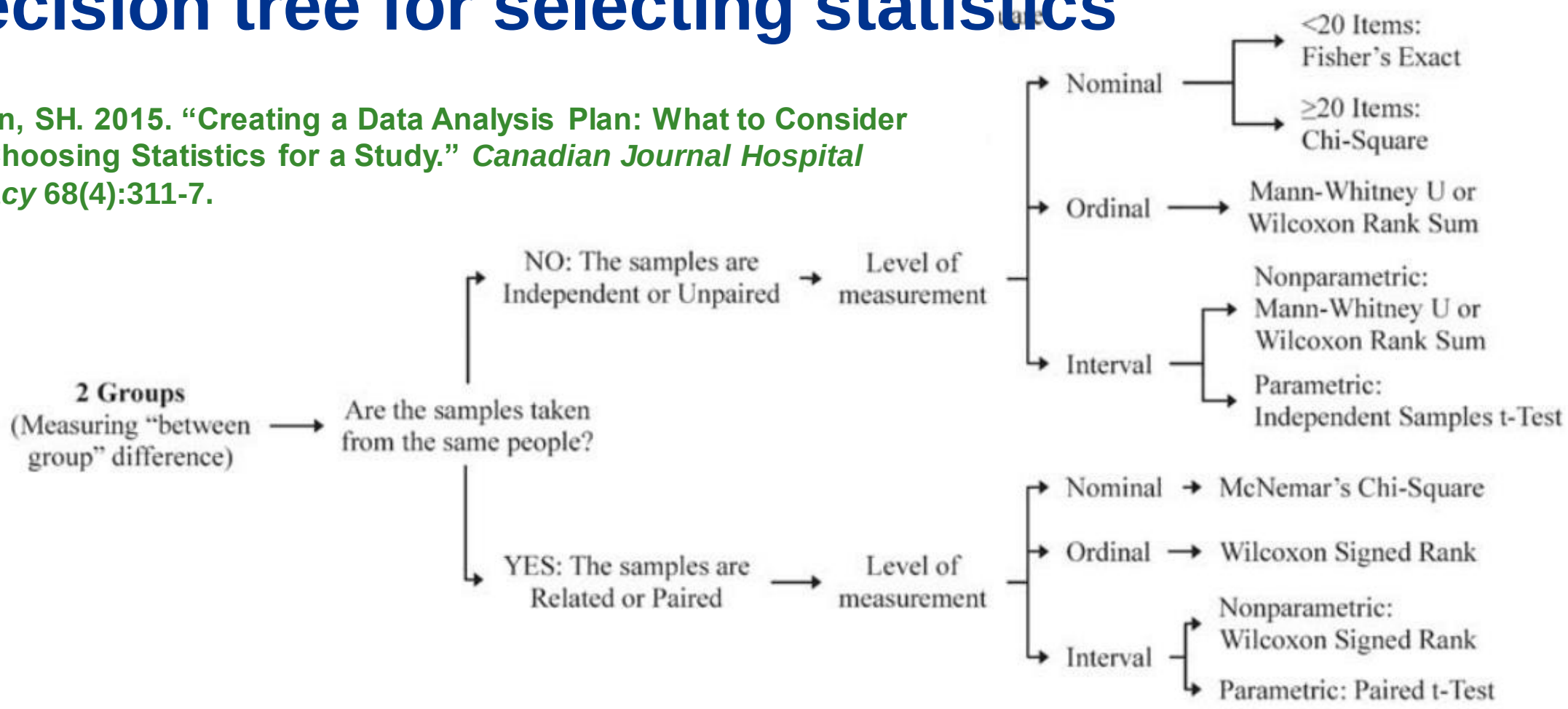
### A quick review of measurement levels (Simpson, 2015)

| Measurement level | | Simple definition | Simple examples |
|---|---|---|---|
| Categorical | Nominal / dichotomous | Patients in *unordered* categories | Mortality (Y/N) |
| | Ordinal | Patients in categories that progress in meaningful order | Counts of hospital stays<br>Levels of oxygen support |
| Continuous | Interval | Patients on a continuous scale but no true zero. Arithmetic is permitted. | Temperature |
| | Ratio | Patients on a continuous scale that includes a true zero. Arithmetic is permitted. | Blood pressure<br>Length of stay<br>Days to readmission |

# What statistics should I use?

# A decision tree for selecting statistics

Simpson, SH. 2015. "Creating a Data Analysis Plan: What to Consider When Choosing Statistics for a Study." *Canadian Journal Hospital Pharmacy* 68(4):311-7.



# What statistics should I use?

# What statistics get published?

## Most frequently used statistics in pharmacy literature

(Lee, Soin & Einarson, 2004, “Statistics in the Pharmacy Literature”, Annals of Pharmacotherapy)

**Table 2.** Summary of Inferential Statistics Found in Pharmacy Research Articles in 2001

| Category | Term or Procedure | Articles, n (%) (N = 144) | Total Statistics, % (N = 637) | Inferential Statistics, % (N = 205) |
|---|---|---|---|---|
| Parametric statistics used to test differences | Student’s *t*-test | 31 (21.5) | 4.9 | 15.1 |
| | ANOVA | 26 (18.1) | 4.1 | 12.7 |
| | Matched pairs *t*-test | 7 (4.9) | 1.1 | 3.4 |
| Post hoc and multiple comparison tests | Tukey’s HSD | 4 (2.8) | 0.6 | 2.0 |
| | Scheffe’s contrast | 2 (1.4) | 0.3 | 1.0 |
| | unspecified post hoc analysis | 3 (2.1) | 0.5 | 1.5 |
| | Bonferroni adjustment | 1 (0.7) | 0.2 | 0.5 |
| Nonparametric tests for differences | $\chi^2$ | 48 (33.3) | 7.5 | 23.4 |
| | Fisher’s exact | 12 (8.3) | 1.9 | 5.9 |
| | Wilcoxon signed-ranks | 4 (2.8) | 0.6 | 2.0 |
| | Kruskall–Wallis | 8 (5.6) | 1.3 | 3.9 |
| | Mann–Whitney U | 7 (4.9) | 1.1 | 3.4 |
| | McNemar’s | 4 (2.8) | 0.6 | 2.0 |
| | Cochrane Q | 1 (0.7) | 0.2 | 0.5 |
| Tests of association | Pearson’s r | 26 (18.1) | 4.1 | 12.7 |
| | logistic regression | 12 (8.3) | 1.9 | 5.6 |
| | multiple regression | 4 (2.8) | 0.6 | 2.0 |
| | Spearman’s rho | 3 (2.1) | 0.5 | 1.5 |
| | kappa | 2 (1.4) | 0.3 | 1.0 |

# What statistics should I use?

# What statistics get published?

## Most frequently published statistics in *NEJM, Lancet, JAMA, Nature*

(Yi et al., 2015, "Statistical Use in Clinical Studies: Is there Evidence of a Methodological Shift?" PLOS One)

| | Overall (N = 838) | 1990 (N = 301) | 2000 (N = 314) | 2010 (N = 223) | $X^2$ | p* |
|---|---|---|---|---|---|---|
| Descriptive# | 100 | 100 | 100 | 1000 | --- | --- |
| ANOVA | 47.0 | 49.3 | 47.1 | 45.2 | 5.636 | 0.060 |
| t test | 36.3 | 35.0 | 36.3 | 37.2 | 1.345 | 0.520 |
| Chi-square | 32.8 | 37.1 | 30.9 | 31.6 | 5.236 | 0.062 |
| Survival analysis*** | 28.5 | 15.3 | 23.6 | 43.4 | 56.279 | 0.001 |
| Non-parametric test* | 28.0 | 23.1 | 33.2 | 26.2 | 6.961 | 0.031 |
| Correlation analysis*** | 27.0 | 17.9 | 23.9 | 36.9 | 25.755 | 0.001 |
| Simple linear regression*** | 23.4 | 15.7 | 20.7 | 31.9 | 20.784 | 0.001 |
| Cox models*** | 16.0 | 7.7 | 13.6 | 24.6 | 29.404 | 0.001 |
| Logistic regression* | 15.3 | 12.3 | 15.6 | 17.3 | 7.686 | 0.021 |
| Fisher exact | 11.0 | 10.0 | 12.0 | 10.6 | 1.707 | 0.426 |

*Chi-square test for differences among years

#Includes means, standard deviations, median, percentages, etc.

# Statistics in medical literature: two conclusions

(Simpson, 2015)

1. **Medical journals publishing increasingly sophisticated statistics, AND**
2. **You can still publish with these comparatively simpler statistics:**
   - *t* tests
   - Contingency table tests (chi square $\chi^2$ and Fisher exact test)
   - Simple correlation analyses
   - Simple regression analyses

# How do I do a chi-square (in Excel)?

## A quick review of the statistic

We observe patients fill categories

How would they fill the variables were independent?

Calculate difference between observed and expected

Sum difference between observed and expected and evaluate probability

| OBSERVED COUNTS | | | | EXPECTED COUNTS | | | OBSERVED - EXPECTED | | (OBSERVED-EXPECTED)^2 / EXPECTED | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Treatment Group | | | | | | | | | | |
| Mortality | PRE | POST | Total | PRE | POST | Total | PRE | POST | PRE | POST | Total |
| Alive | 6 | 6 | 12 | 4.8 | 7.2 | 12 | 1.2 | -1.2 | 1.44 | 1.44 | 2.88 |
| Expired | 6 | 12 | 18 | 7.2 | 10.8 | 18 | -1.2 | 1.2 | 1.44 | 1.44 | 2.88 |
| Total | 12 | 18 | 30 | 12 | 18 | 30 | | | | $\chi^2$ | 5.76 |
| | | | | | | | | | | p | 0.36 |

## How to do a $\chi^2$ test in Excel template

# Step 1. Format your raw data for PivotTable

- All patients in the same data array
- Column with codes identifying patient groups

- Outcome variable

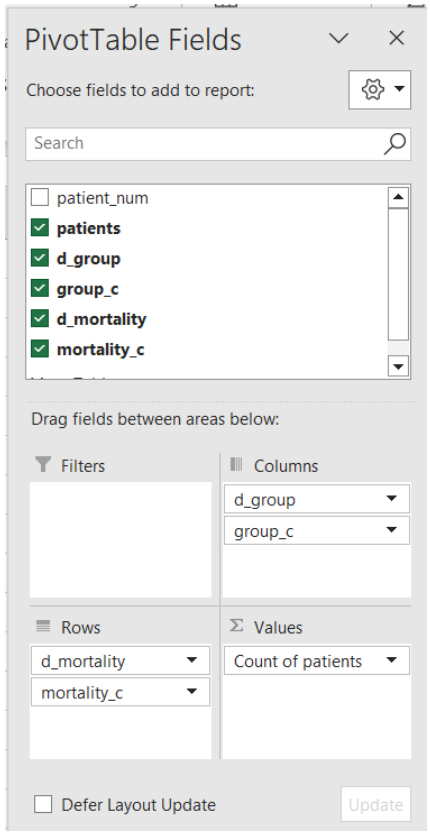
| patient_num | patients | d_group | group_c | d_mortality | mortality_c |
|---|---|---|---|---|---|
| 1 | 1 | 0 | PRE | 0 | Alive |
| 2 | 1 | 0 | PRE | 1 | Expired |
| 3 | 1 | 0 | PRE | 0 | Alive |
| 4 | 1 | 0 | PRE | 1 | Expired |
| 5 | 1 | 0 | PRE | 0 | Alive |
| 6 | 1 | 1 | POST | 1 | Expired |
| 7 | 1 | 1 | POST | 1 | Expired |
| 8 | 1 | 1 | POST | 1 | Expired |
| 9 | 1 | 1 | POST | 1 | Expired |
| 10 | 1 | 1 | POST | 0 | Alive |

Sample Chi-square data and formulae.xlsx

## How to do a $\chi^2$ test in Excel template

# Step 2. Use PivotTable to get observed counts

| Count of patients | | d_group | group_c | |
|---|---|---|---|---|
| | | 0 | 1 | Grand Total |
| d_mortality | mortality_c | PRE | POST | |
| 0 | Alive | 6 | 6 | 12 |
| 1 | Expired | 6 | 12 | 18 |
| Grand Total | | 12 | 18 | 30 |




**PivotTable settings**

Values → patients → Field settings → Count

Design → Report Layout → Show in Tabular Form

Sample Chi-square data and formulae.xlsx

## How to do a $\chi^2$ test in Excel template

# Step 3. Expected counts, residuals, $\chi^2$, p, calculated for you

**Observed counts**

**Formulas calculate for you**

**Expected counts**

Patients expected if treatment and outcome are independent

**Residual (Observed – Expected)**

Difference between observed and expected patient counts

**Square residuals**

Difference between observed and expected patient counts

**$\chi^2$ and p-value**

| | PRE | POST | |
|---|---|---|---|
| Count of Patient | Column Labels | | |
| Row Labels | 1 | 2 | Grand Total |
| 1 | 6 | 5 | 11 |
| 2 | 5 | 7 | 12 |
| 3 | 4 | 6 | 10 |
| 4 | 5 | 2 | 7 |
| Grand Total | 20 | 20 | 40 |
| | | | |
| 1 | 5.5 | 5.5 | 11 |
| 2 | 6 | 6 | 12 |
| 3 | 5 | 5 | 10 |
| 4 | 3.5 | 3.5 | 7 |
| | 20 | 20 | 40 |
| | | | |
| 1 | 0.5 | -0.5 | |
| 2 | -1 | 1 | |
| 3 | -1 | 1 | |
| 4 | 1.5 | -1.5 | |
| | | | |
| | | | Sum |
| 1 | 0.045 | 0.045 | 0.091 |
| 2 | 0.167 | 0.167 | 0.333 |
| 3 | 0.200 | 0.200 | 0.400 |
| 4 | 0.643 | 0.643 | 1.286 |
| | | $X^2$ | 2.110 |
| | | p-value | 0.550 |

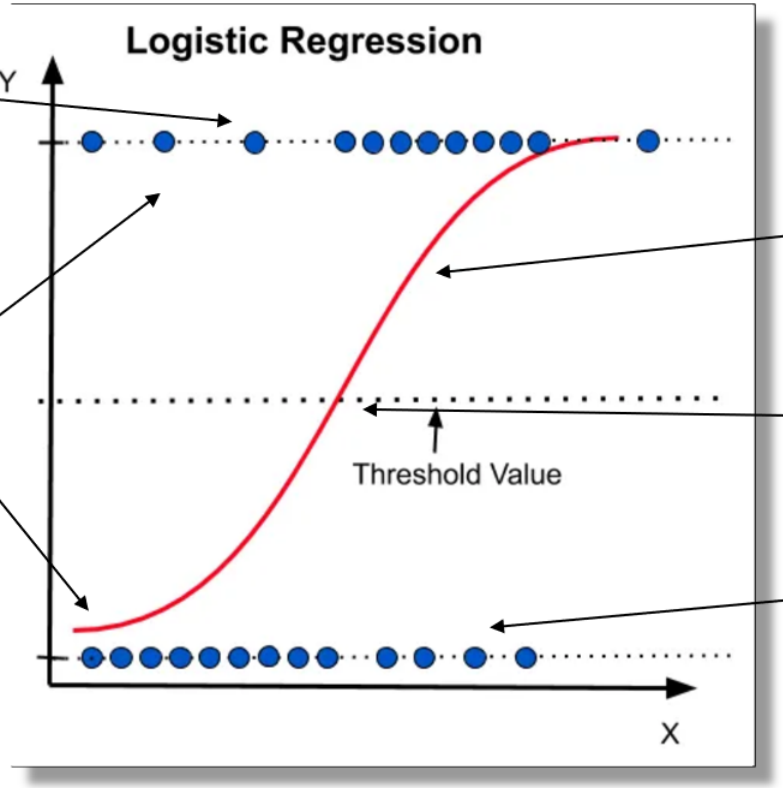
Sample Chi-square data and formulae.xlsx

# How do I do a logistic regression?

## A gentle review of the statistic

Regression is using model to (imperfectly) fit observations

Outcome is dichotomous




Functional form is logistic curve

Point of interest is where odds of outcome change

Predictors can be dichotomous or continuous

# Why logistic regression?

- Can isolate relationship between predictor and outcome controlling for additional explanatory variables
- Can examine interactions between multiple predictors on outcome variable (Does the effect of X on Y depend on the level of W?)

# How do I do a logistic regression?

## An R implementation

Quarto / R Markdown document

results_delete_me.qmd

Render on Save | Render

Source | Visual

text

```
table2

```

Finally, @tbl-table3 reports adjusted odds ratios for mAb ... ved fr
hospitalization and mortality on mAb treatment. Odds of 28 ... lizati
group, but treated patients had less than half the odds of dying after the fir
0.43 (0.26-0.68), *p* < .001).

code chunk

```{r}
#| label: tbl-table3
#| tbl-cap: "Adjusted Odds Ratios"
#| message: false
#| echo: false

library(gtsummary)
library(kableExtra)
library(Hmisc)

# ---------- REGRESSIONS ---------- #

# ---------- Rehospitalization model ---------- #
readmit.model <- glm(
  d_readmit28 ~
    d_mab_c +
    pandemic_phase,
  data = m.data,
  family = "binomial"
) %>% tbl_regression(exponentiate = TRUE) %>% bold_p()

# ---------- Mortality model ---------- #
mort.model <- glm(
  d_mortality ~
    d_mab_c +
    pandemic_phase,
  data = m.data,
  family = "binomial"
) %>% tbl_regression(exponentiate = TRUE) %>% bold_p()

# ---------- Table 3 ---------- #
table3 <-
  tbl_merge(
    tbls = list(readmit.model, mort.model),
    tab_spanner = c("**28-Day Rehospitalization**", "**Mortality**")
  )
table3

```

HTML output

Providence SWEDISH



Table 2: Primary and Secondary Outcomes for Primary Matched Cohort

| Characteristic | Untreated, N = 1,220[1] | mAb treated, N = 1,220[1] | p-value[2] |
|---|---|---|---|
| Primary outcomes | | | |
| *28-Day Readmission to ED or Hospital Inpatient* | 95 (7.8%) | 112 (9.2%) | 0.2 |
| *Mortality within Study Period* | 55 (4.5%) | 24 (2.0%) | <0.001 |
| Secondary outcomes | | | |
| *Total inpatient hospital days* | 1.0025 (8.6627) | 1.1590 (12.7330) | 0.2 |
| *Total ICU days* | 0.1671 (2.4999) | 0.1454 (2.4283) | 0.3 |
| *Total days on ventilator* | 0.1269 (2.3108) | 0.1180 (2.1186) | 0.3 |

[1] n (%)

[2] Pearson's Chi-squared test

Finally, Table 3 reports adjusted odds ratios for mAb treatment derived from logistic regressions of 28-day hospitalization and mortality on mAb treatment. Odds of 28-day rehospitalization did not vary significantly by treatment group, but treated patients had less than half the odds of dying after the first ED visit than untreated patients (aOR = 0.43 (0.26-0.68), $p < .001$).

Table 3: Adjusted Odds Ratios

| | 28-Day Rehospitalization | | | Mortality | | |
|---|---|---|---|---|---|---|
| Characteristic | OR[1] | 95% CI[1] | p-value | OR[1] | 95% CI[1] | p-value |
| mAb treated | 1.20 | 0.90, 1.61 | 0.2 | 0.43 | 0.26, 0.68 | **<0.001** |
| Pandemic phase at 1st ED visit | 0.71 | 0.61, 0.82 | **<0.001** | 0.72 | 0.58, 0.91 | **0.004** |

[1] OR = Odds Ratio, CI = Confidence Interval

# Additional resources

Visualizations of statistical concepts at https://rspsychologist.com/

Statistical decision trees in Simpson (2015)

statology.org

quarto.org

SharePoint page for this presentation: Tools and Tips for Statistical Data Analysis

# References


Arnold, L.D., Braganza, M., Salih, R., & Colditz, G.A. (2013). Statistical trends in the *Journal of the American Medical Association* and implications for training across the continuum of medical education. PLoS ONE 8(10): e77301.

Cohen, J. 1988. Statistical power analysis for the behavioral sciences. Lawrence Erlbaum.

Horton, N.J. & Switzer, S.S. (2005). Statistical methods in the journal. New England Journal of Medicine 353:1977-9.

Lee CM, Soin HK, Einarson TR. Statistics in the pharmacy literature. Annals of Pharmacotherapy. 2004 Sep;38(9):1412-8.

Newsome, C., Ryan, K. Bakhireva, L., & Sarangarm, P. (2019). Breadth of statistical training among pharmacy residency programs across the United States. Hospital Pharmacy 53(2): 101-106.

Ou, F.-S., Le-Rademacher, J.G., Ballman, K.V., Adjei, A.A., & Mandrekar, S.J. (2020). Guidelines for statistical reporting in medical journals. Journal of Thoracic Oncology 15(1): 1722-6.

Panos, G.D. & Boeckler, F.M. (2023). Statistical analysis in clinical and experimental medical research: Simplified guidance for authors and reviewers. Drug Design, Development and Therapy 2023:17, 1959-1961.

Simpson SH. Creating a Data Analysis Plan: What to Consider When Choosing Statistics for a Study. Can J Hosp Pharm. 2015 Jul-Aug;68(4):311-7. doi: 10.4212/cjhp.v68i4.1471. PMID: 26327705; PMCID: PMC4552232.

Windish, D.M., Huot, S.J., & Green,M.L. (2007). Medical residents' understanding of the biostatistics and results in the medical literature. JAMA 298(9): 1010-1022.

Yi, D., Ma, D., Li, G., Zhou, L., Xiao, Q., Zhang, Y., Liu, X., Chen, H., Pettigrew, J.C., Yi, D., Liu, L., & Wu, Y. (2015). Statistical use in clinical studies: Is there evidence of a methodological shift? PLoS ONE 10(10): e0140159.

# Thank you!