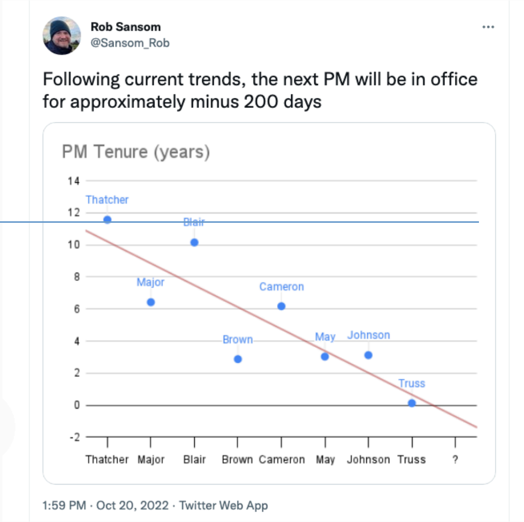
# Methods of Applied Statistics I

STA2101H F LEC9101

Week 7

October 26 2022



# Today

Start Recording

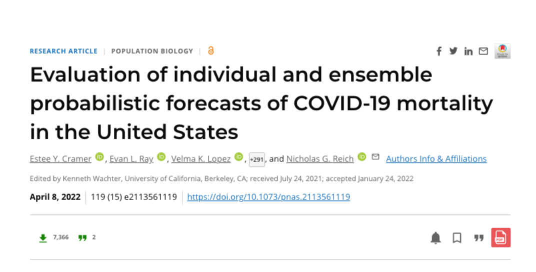
1. Upcoming events
2. Recap
3. Likelihood theory and logistic regresson
4. Observational studies and causality
5. In the News

6. Hour 3: Comments on HW 1-6 estimates of effect size, missing data

7. Office Hour Wednesday October 26 4-5 pm on Zoom

# Past

- Monday October 24 3.30-4.30 : DoSS Seminar Room 9014 (Hydro Building)
- Data Science Seminar Series
- Daniel McDonald, U Chicago
- Markov-Switching State Space Models for Uncovering Musical Interpretation



RESEARCH ARTICLE | POPULATION BIOLOGY

Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States

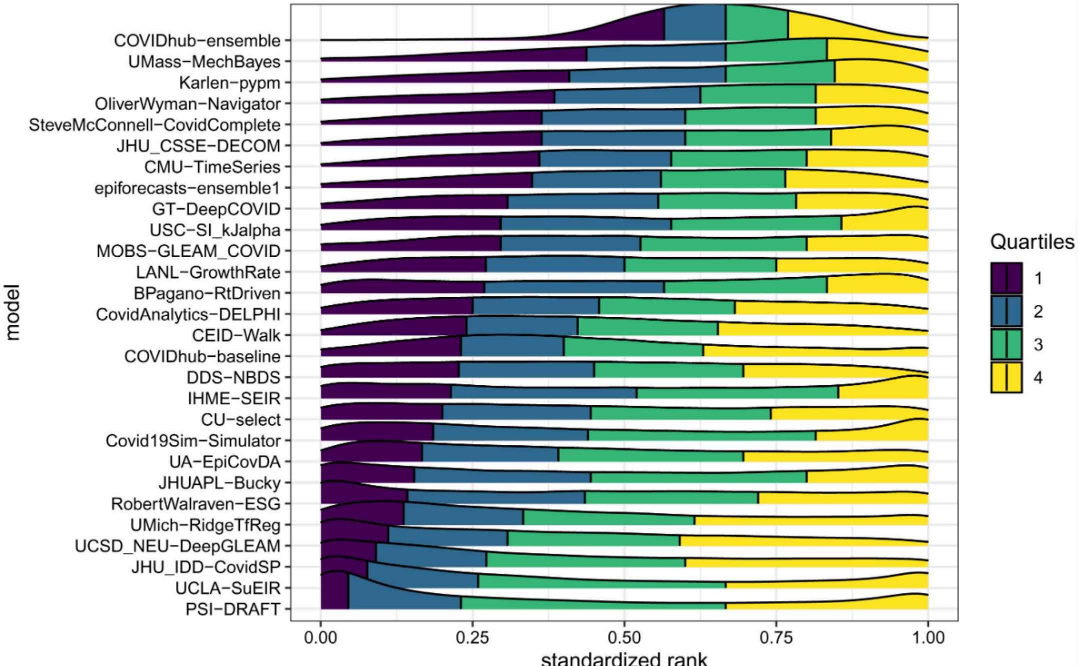
Estee Y. Cramer, Evan L. Ray, Velma K. Lopez, +291, and Nicholas G. Reich    Authors Info & Affiliations

Edited by Kenneth Wachter, University of California, Berkeley, CA; received July 24, 2021; accepted January 24, 2022

April 8, 2022 | 119 (15) e2113561119 | https://doi.org/10.1073/pnas.2113561119

7,366  2

link


COVIDhub-ensemble
UMass-MechBayes
Karlen-pypm
OliverWyman-Navigator
SteveMcConnell-CovidComplete
JHU_CSSE-DECOM
CMU-TimeSeries
epiforecasts-ensemble1
GT-DeepCOVID
USC-SI_kJalpha
MOBS-GLEAM_COVID
LANL-GrowthRate
BPagano-RtDriven
CovidAnalytics-DELPHI
CEID-Walk
COVIDhub-baseline
DDS-NBDS
IHME-SEIR
CU-select
Covid19Sim-Simulator
UA-EpiCovDA
JHUAPL-Bucky
RobertWalraven-ESG
UMich-RidgeTfReg
UCSD_NEU-DeepGLEAM
JHU_IDD-CovidSP
UCLA-SuEIR
PSI-DRAFT
model
0.00
0.25
0.50
0.75
1.00
standardized rank
Quartiles
1
2
3
4

# Upcoming

- October 27 3.30-4.30 Statistical Sciences Seminar; Room 9014, Hydro Building and online
- Mireille Schnitzer, U Montreal
  "Outcome-Adaptive LASSO for Confounder Selection With Time-Varying Treatments"

# Recap

- regression models for binomial and binary data
- examples: O-ring failure; heart disease; nodal involvment
- inference, residuals, diagnostics, analysis of deviance, nested models oct19.pdf: 20–24
- covariate classes; binary data
- model selection with

$$AIC = -2\ell(\hat\beta; y) + 2p$$

$$BIC = 2\ell(\hat\beta; y) + \log(n)p$$

# Likelihood theory

- model: $y_i \sim f(y_i;\theta), i = 1,\ldots,n; \quad \theta \in \Theta \subset \mathbb{R}^p$

independent

- joint density: $f(\underline{y};\theta) = \prod_{i=1}^n f(y_i;\theta)$
- likelihood function $L(\theta;\underline{y}) = f(\underline{y};\theta)$

# Likelihood theory

- model: $y_i \sim f(y_i; \theta), i = 1, \ldots, n; \quad \theta \in \Theta \subset \mathbb{R}^p$ independent
- joint density: $f(\underline{y}; \theta) = \prod_{i=1}^n f(y_i; \theta)$
- likelihood function $L(\theta; \underline{y}) = f(\underline{y}; \theta)$

- log-likelihood function $\ell(\theta; \underline{y}) = \log L(\theta; \underline{y}) = \sum_{i=1}^n \log f(y_i; \theta)$
- maximum likelihood estimate $\hat{\theta} = \arg\sup \ell(\theta; \underline{y})$; $\ell'(\hat{\theta}) = 0$
- Fisher information $j(\hat{\theta}) = -\ell''(\hat{\theta})$

# Likelihood theory

- model: $y_i \sim f(y_i; \theta), i = 1, \ldots, n; \quad \theta \in \Theta \subset \mathbb{R}^p$ independent
- joint density: $f(\underline{y}; \theta) = \prod_{i=1}^n f(y_i; \theta)$
- likelihood function $L(\theta; \underline{y}) = f(\underline{y}; \theta)$

- log-likelihood function $\ell(\theta; \underline{y}) = \log L(\theta; \underline{y}) = \sum_{i=1}^n \log f(y_i; \theta)$
- maximum likelihood estimate $\hat{\theta} = \arg\sup \ell(\theta; \underline{y})$; $\ell'(\hat{\theta}) = 0$
- Fisher information $j(\hat{\theta}) = -\ell''(\hat{\theta})$

- two theorems:

$$j^{1/2}(\hat{\theta})(\hat{\theta} - \theta) \xrightarrow{d} N_p(0, I)$$

asymptotically normal

- likelihood ratio statistic

$$w(\theta) = 2\{\ell(\hat{\theta}) - \ell(\theta)\} \xrightarrow{d} \chi^2_p$$

$p$ is dimension of $\theta$

$\xrightarrow{d}$ is convergence in distribution

# Likelihood Inference

- two theorems:

$$
\begin{aligned}
j^{1/2}(\hat\theta)(\hat\theta - \theta) &\xrightarrow{d} N(0, I) \\
w(\theta) = 2\{\ell(\hat\theta) - \ell(\theta)\} &\xrightarrow{d} \chi^2_p
\end{aligned}
$$

# Likelihood Inference

- two theorems:

$$j^{1/2}(\hat\theta)(\hat\theta - \theta) \quad \stackrel{d}{\to} \quad N(0, I)$$

$$w(\theta) = 2\{\ell(\hat\theta) - \ell(\theta)\} \quad \stackrel{d}{\to} \quad \chi^2_p$$

- two approximations

$$\hat\theta \quad \dot\sim \quad N_d\{\theta, j^{-1}(\hat\theta)\}$$

$$\hat\theta_k \quad \dot\sim \quad N(\{\theta_k, j^{-1}(\hat\theta)_{kk}\}$$

$$w(\theta) \quad \dot\sim \quad \chi^2_p$$

# Likelihood Inference

- two theorems:

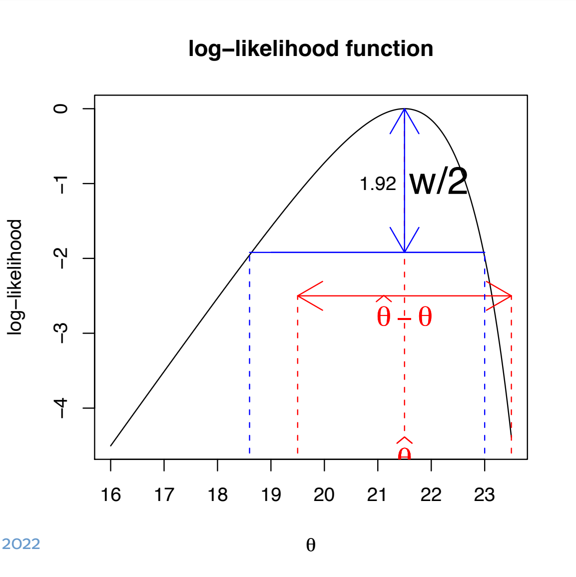
$$j^{1/2}(\hat\theta)(\hat\theta - \theta) \quad \stackrel{d}{\to} \quad N(0, I)$$
$$w(\theta) = 2\{\ell(\hat\theta) - \ell(\theta)\} \quad \stackrel{d}{\to} \quad \chi^2_p$$

- two approximations

$$\hat\theta \quad \dot\sim \quad N_d\{\theta, j^{-1}(\hat\theta)\}$$
$$\hat\theta_k \quad \dot\sim \quad N(\{\theta_k, j^{-1}(\hat\theta)_{kk}\}$$

$$w(\theta) \quad \dot\sim \quad \chi^2_p$$

- compare two models using change in likelihood ratio statistic    nested models


log–likelihood function
log–likelihood
0
–1
–2
–3
–4
1.92
w/2
$\hat{\theta} - \theta$
$\hat{\theta}$
16
17
18
19
20
21
22
23
θ

link

**STA2101: Likelihood Cheatsheet**

$Y = Y_1, \ldots, Y_n$ independently distributed with densities $f(y_i \mid x_i; \theta), \theta \in \Theta \subset \mathbb{R}^p; y_i \in \mathbb{R}$. The observations are independent, but not identically distributed, due to the dependence on the $p \times 1$ vector $x_i$. Independence is critical, but i.d. can usually be handled, so the dependence on $x_i$ below is often suppressed.

Likelihood function is the joint probability of the observations, considered as a function of the parameter

$$L(\theta; y) \propto \prod_{i=1}^{n} f(y_i \mid x_i; \theta)$$



Log-likelihood function

# Nested models

- Comparing two models:
- likelihood ratio test

$$2\{\ell_A(\hat{\beta}_A) - \ell_B(\hat{\beta}_B)\}$$

  compares the maximized log-likelihood function under model A and model B
- example
  model A: $\text{logit}(p_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \quad \beta_A = (\beta_0, \beta_1, \beta_2)$
  model B: $\text{logit}(p_i) = \beta_0 + \beta_1 x_{1i}, \quad \beta_B = (\beta_0, \beta_1)$

- when model B is nested in model A, LRT is approximately $\chi^2_\nu$ distributed under model B
- $\nu = dim(A) - dim(B)$ theory of profile likelihood

# … nested models

```
> logitmodcorrect2 <- glm(cbind(r,m-r) ~ temperature + pressure, family = binomial, data = shuttle2)

> summary(logitmodcorrect2)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.520195   3.486784   0.723   0.4698
temperature -0.098297   0.044890  -2.190   0.0285 *
pressure     0.008484   0.007677   1.105   0.2691
---
    Null deviance: 24.230  on 22  degrees of freedom
Residual deviance: 16.546  on 20  degrees of freedom
AIC: 36.106
Number of Fisher Scoring iterations: 5
```

# ... nested models

```
> logitmodcorrect2 <- glm(cbind(r,m-r) ~ temperature + pressure, family = binomial, data = shuttle2)

> anova(logitmodcorrect,logitmodcorrect2)
Analysis of Deviance Table

Model 1: cbind(r, m - r) ~ temperature
Model 2: cbind(r, m - r) ~ temperature + pressure
  Resid. Df Resid. Dev Df Deviance
1        21     18.086
2        20     16.546  1   1.5407
```

# ...nested models

- Model A: $\mathrm{logit}(p_i) = \beta_0 + \beta_1 \mathtt{temp}_i + \beta_2 \mathtt{pressure}_i$
- Model B: $\mathrm{logit}(p_i) = \beta_0 + \beta_1 \mathtt{temp}_i$
- nested: Model B is obtained by setting $\beta_2 = 0$
- Under Model B, the change in deviance is (approximately) an observation from a $\chi^2_1$
- $\Pr(\chi^2_1 \geq 1.5407) = 0.22$; this is a $p$-value for testing $H_0 : \beta_2 = 0$
- so is $1 - \Phi\{\frac{\hat\beta_2}{\widehat{s.e.}(\hat\beta_2)}\} = 1 - \Phi(1.105) = 0.27$

ELM-1 p.30

# Binomial likelihood

- model

# Binomial likelihood

- model

- likelihood

# Binomial likelihood

- model

- likelihood

- log-likelihood

# Binomial likelihood

- model

- likelihood

- log-likelihood

- score function

# Binomial likelihood

- model
- likelihood
- log-likelihood
- score function
- maximum likelihood estimate

# Binomial likelihood

- model
- likelihood
- log-likelihood
- score function
- maximum likelihood estimate
- Fisher information

# … binomial likelihood

- model $y_i \sim Bin(n_i, p_i), i = 1, \ldots, m$ no regression

# … binomial likelihood

- model $y_i \sim Bin(n_i, p_i), i = 1, \ldots, m$ — no regression

- likelihood

# ... binomial likelihood

- model $y_i \sim Bin(n_i, p_i), i = 1, \ldots, m$

no regression

- likelihood
- log-likelihood

# … binomial likelihood

- model $y_i \sim Bin(n_i, p_i), i = 1, \ldots, m$ — no regression
- likelihood
- log-likelihood
- score function

# … binomial likelihood

- model $y_i \sim Bin(n_i, p_i), i = 1, \ldots, m$

no regression

- likelihood
- log-likelihood
- score function
- maximum likelihood estimate

# … binomial likelihood

- model $y_i \sim Bin(n_i, p_i), i = 1, \ldots, m$ no regression
- likelihood
- log-likelihood
- score function
- maximum likelihood estimate
- maximized log-likelihood function

# Special to the Binomial and Poisson

- regression model is nested in saturated model
-
$$w = 2[\ell(\hat{p}) - \ell\{p(\hat{\beta})\}] \,\dot{\sim}\, \chi^2_{m-p}$$

# Pearson’s chi-square

# Special to the binomial



- logistic regression model $p_i = p_i(\beta) = \text{expit}(x_i^{\mathrm{T}}\beta), \quad \hat{p}_i = p_i(\hat{\beta})$ is nested in the saturated model $\tilde{p}_i = y_i/n_i$

# Special to the binomial



- logistic regression model $p_i = p_i(\beta) = \text{expit}(x_i^{\mathrm{T}}\beta), \quad \hat{p}_i = p_i(\hat{\beta})$
  is nested in the saturated model $\tilde{p}_i = y_i/n_i$

- residual deviance compares fitted model to saturated model

# Special to the binomial



- logistic regression model $p_i = p_i(\beta) = \text{expit}(x_i^{\mathrm{T}}\beta), \quad \hat{p}_i = p_i(\hat{\beta})$
  is nested in the saturated model $\tilde{p}_i = y_i/n_i$

- residual deviance compares fitted model to saturated model

- under the fitted model, approximately distributed as $\chi^2_{n-p}$

# Special to the binomial

- logistic regression model $p_i = p_i(\beta) = \text{expit}(x_i^{\mathrm{T}}\beta), \quad \hat{p}_i = p_i(\hat{\beta})$
  is nested in the saturated model $\tilde{p}_i = y_i/n_i$

- residual deviance compares fitted model to saturated model

- under the fitted model, approximately distributed as $\chi^2_{n-p}$
  if each $n_i$ “large”

ELM-2 §3.2

```
> summary(Ex1018.glm)

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 40.710  on 22  degrees of freedom
 Residual deviance: 18.069  on 17  degrees of freedom
 AIC: 41.69
```

# Deviance and binary data

- if $n_i \equiv 1$, then

# Deviance residuals



```
> summary(Ex1018binom.glm)

Call:
glm(formula = cbind(r, m - r) ~ ., family = binomial, data = nodal2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4989  -0.7726  -0.1265   0.7997   1.4351
```

# Deviance residuals

`glm.diag; library(SMPracticals)`

```
> summary(Ex1018binom.glm)

Call:
glm(formula = cbind(r, m - r) ~ ., family = binomial, data = nodal2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4989  -0.7726  -0.1265   0.7997   1.4351
```

Deviance: $2\sum_{i=1}^{n}[y_i \log\{y_i/n_i p_i(\hat{\beta})\} + (n_i - y_i)\log\{(n_i - y_i)/(n_i - n_i p_i(\hat{\beta}))\}]$

# Deviance residuals

`glm.diag; library(SMPracticals)`

```
> summary(Ex1018binom.glm)

Call:
glm(formula = cbind(r, m - r) ~ ., family = binomial, data = nodal2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4989  -0.7726  -0.1265   0.7997   1.4351
```

Deviance: $2\sum_{i=1}^{n}[y_i \log\{y_i/n_i p_i(\hat{\beta})\} + (n_i - y_i)\log\{(n_i - y_i)/(n_i - n_i p_i(\hat{\beta}))\}]$

approximately $\chi^2_{n-q}$

$$r_{Di} = \pm\sqrt{}(2[y_i \log\{y_i/n_i\hat{p}_i\} + (n_i - y_i)\log\{(n_i - y_i)/(n_i - n_i\hat{p}_i)\}])$$

# Overdispersion



- $Y_i \sim Bin(n_i, p_i) \Rightarrow \mathrm{E}(Y_i) = n_i p_i, \quad \mathrm{Var}(Y_i) = n_i p_i(1 - p_i)$
- variance is determined by the mean

# Overdispersion

- $Y_i \sim Bin(n_i, p_i) \Rightarrow \mathrm{E}(Y_i) = n_i p_i, \quad \mathrm{Var}(Y_i) = n_i p_i (1 - p_i)$
- variance is determined by the mean

- ```
  bmod <- glm(cbind(survive,total-survive) ~ location + period, family = binomial,
              data = troutegg)

  summary(bmod)
   Null deviance: 1021.469  on 19  degrees of freedom
  ## Residual deviance:   64.495  on 12  degrees of freedom
  ## AIC: 157.03
  ```
  ```

# Overdispersion

- $Y_i \sim Bin(n_i, p_i) \Rightarrow \mathrm{E}(Y_i) = n_i p_i, \quad \mathrm{Var}(Y_i) = n_i p_i(1 - p_i)$
- variance is determined by the mean

- 
```
bmod <- glm(cbind(survive,total-survive) ~ location + period, family = binomial,
            data = troutegg)

summary(bmod)
 Null deviance: 1021.469  on 19  degrees of freedom
## Residual deviance:   64.495  on 12  degrees of freedom
## AIC: 157.03
```
- quasi-binomial: $\mathrm{E}(Y_i) = n_i p_i, \quad \mathrm{Var}(Y_i) = \phi n_i p_i(1 - p_i)$
- estimate $\phi$? over-dispersion parameter
- usually use $X^2/(n - p)$, where

$$X^2 = \sum \frac{(y_i - n_i\hat{p}_i)^2}{n\hat{p}_i(1 - \hat{p}_i)}$$

# Quasibinomial

overdisp.Rmd; overdisp.html

```
> step(EX1018binom.glm)

Coefficients:
(Intercept)       stage1        xray1        acid1
      -3.05         1.65         1.91         1.64

Degrees of Freedom: 22 Total (i.e. Null);  19 Residual
Null Deviance:^^I    40.7
Residual Deviance: 19.6 ^^IAIC: 39.3
```

– we can drop `age` and `grade` without affecting quality of the fit

– in other words the model can be simplified by setting two regression coefficients to zero

– several mistakes in text on pp. 491,2;

– deviances in Table 10.9 are incorrect as well `http://statwww.epfl.ch/davison/SM/` has corrected version

# … variable selection

- `step` implements stepwise regression
- evaluates each fit using AIC = $-2\ell(\hat{\beta}; y) + 2p$
- penalizes models with larger number of parameters

- we can also compare fits by comparing deviances $\longrightarrow$ binaryELM2.html

- `step` implements stepwise regression
- evaluates each fit using AIC = $-2\ell(\hat{\beta}; y) + 2p$
- penalizes models with larger number of parameters

- we can also compare fits by comparing deviances $\longrightarrow$ binaryELM2.html
- 

```
> update(Ex1018binom.glm, .~. - aged - grade)

Call:  glm(formula = cbind(rtot, total - rtot) ~ stage + xray + acid,
    family = binomial, data = nodal2)

Coefficients:
(Intercept)       stage1        xray1        acid1
      -3.05         1.65         1.91         1.64

Degrees of Freedom: 22 Total (i.e. Null);  19 Residual
Null Deviance:^^I    40.7
Residual Deviance: 19.6 ^^IAIC: 39.3
> deviance(ex1018binom)
[1] 18.06869
> pchisq(19.6-18.07, df = 2, lower = F)
[1] 0.4653
```

# AIC

- as terms are added to the model, deviance always decreases
- because log-likelihood function always increases
- similar to residual sum of squares

# AIC

- as terms are added to the model, deviance always decreases
- because log-likelihood function always increases
- similar to residual sum of squares

- Akaike Information Criterion penalizes models with more parameters
- $$AIC = 2\{-\ell(\hat{\beta}; y) + p\}$$

SM (4.57)

- comparison of two model fits by difference in $AIC$

# Measures of risk [1,2]

- see posted handout on case-control studies
- consider for simplicity binomial responses with a single binary covariate:

$$\text{logit}(p_i) \sim \beta_0 + \beta_1 z_i, \quad i = 1, \ldots, n$$

# Measures of risk [1,2]

- see posted handout on case-control studies
- consider for simplicity binomial responses with a single binary covariate:

$$\mathrm{logit}(p_i) \sim \beta_0 + \beta_1 z_i, \quad i = 1, \ldots, n$$

- no difference between groups $\iff$ odds-ratio $\equiv 1$

# ... Measures of risk

- we might be interested in risk ratio $\frac{p_1}{p_0}$ instead of odds ratio $\frac{p_1(1-p_0)}{p_0(1-p_1)}$
- also called relative risk

# ... Measures of risk

- we might be interested in risk ratio $\frac{p_1}{p_0}$ instead of odds ratio $\frac{p_1(1-p_0)}{p_0(1-p_1)}$
- also called relative risk
- if $p_1$ and $p_0$ are both small, ($y = 1$ is rare), then

$$\frac{p_1}{p_0} \approx \frac{p_1(1-p_0)}{p_0(1-p_1)}$$

- sometimes $p_1/p_0$ can be large but if $p_1$ and $p_0$ are both small the difference $p_1 - p_0$ might also be very small

# … Measures of risk

- we might be interested in risk ratio $\frac{p_1}{p_0}$ instead of odds ratio $\frac{p_1(1-p_0)}{p_0(1-p_1)}$
- also called relative risk
- if $p_1$ and $p_0$ are both small, ($y = 1$ is rare), then

$$\frac{p_1}{p_0} \approx \frac{p_1(1-p_0)}{p_0(1-p_1)}$$

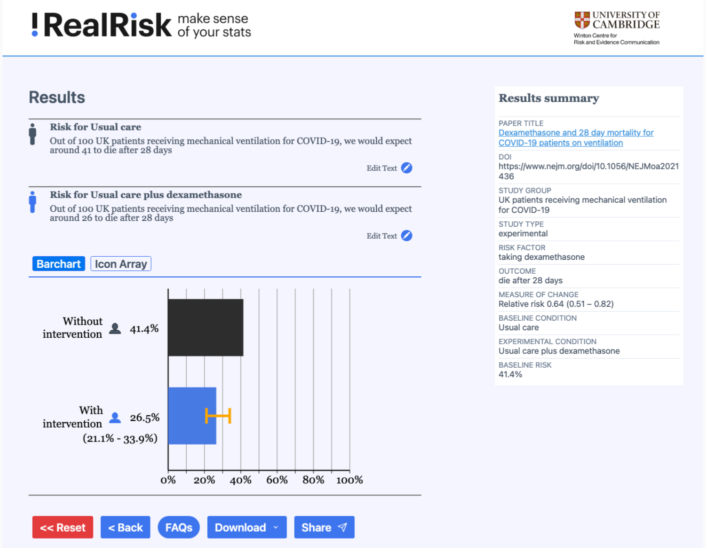
- sometimes $p_1/p_0$ can be large but if $p_1$ and $p_0$ are both small the difference $p_1 - p_0$ might also be very small
- in order to estimate the risk difference we need to know the baseline risk $p_0$

## … Measures of risk

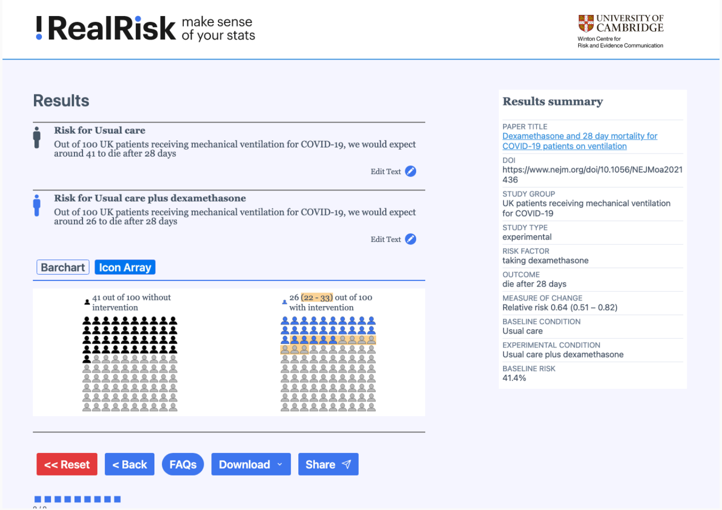
- we might be interested in risk ratio $\frac{p_1}{p_0}$ instead of odds ratio $\frac{p_1(1-p_0)}{p_0(1-p_1)}$
- also called relative risk
- if $p_1$ and $p_0$ are both small, ($y = 1$ is rare), then

$$\frac{p_1}{p_0} \approx \frac{p_1(1-p_0)}{p_0(1-p_1)}$$

- sometimes $p_1/p_0$ can be large but if $p_1$ and $p_0$ are both small the difference $p_1 - p_0$ might also be very small
- in order to estimate the risk difference we need to know the baseline risk $p_0$
- bacon sandwiches www.youtube.com/watch?v=4szyEbU94ig
- risk calculator https://realrisk.wintoncentre.uk/p1




Odds ratio 0.64; baseline risk 41.4%




Odds ratio 0.64; baseline risk 41.4%


Odds ratio 2.91; baseline risk 1/1000

Whether we sample prospectively or retrospectively, the odds ratio is the same

| | Lung cancer | |
|---|---|---|
| | 1 | 0 |
| | cases | controls |
| smoke = 1 (yes) | 688 | 650 |
| smoke = 0 (no) | 21 | 59 |
| | 709 | 709 |

$$\text{retro: } OR = \frac{(688/709)/(21/709)}{(650/709)/(59/709)} = \frac{688 \times 59}{650 \times 21} = 2.97$$

$$\text{prosp: } OR = \frac{\{688/(688+650)\}/\{650/(688+650)\}}{21/(21+59)/\{59/(21+59)\}} = \frac{688 \times 59}{650 \times 21} = 2.97$$

# Types of observational studies

- secondary analysis of data collected for another purpose
- estimation of some feature of a defined population

could in principle be found exactly

- tracking across time of such features
- study of a relationship between features, where individuals may be examined
  - at a single time point
  - at several time points for different individuals
  - at different time points for the same individual
- census
- meta-analysis: statistical assessment of a collection of studies on the same topic

# In the News

Report on Business

SATURDAY, OCTOBER 22, 2022 | GLOBEANDMAIL.COM

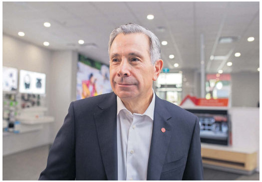
IAN McGUGAN
The question investors can no longer ignore: How do you recession-proof a portfolio? B11

ENERGY
Chemical engineer Peter Guthrie to be Alberta's next energy minister B3

PERSONAL FINANCE
Don't worry, young adults, CPP and EI will be there when you need them B11

INVESTING
Equity strategist buys health care stocks, moves away from banks B13

COVER STORY



## Back to basics

Almost a year into his tenure at Rogers Communications, CEO Tony Staffieri shares what dominates his focus: service and performance. **Andrew Willis** reports B6

PHOTO BY SHAY CONROY/THE GLOBE AND MAIL

How Europe is trying to build a future free of fossil fuels during an energy crisis

ADAM RADWANSKI BRUSSELS

Shopify has a growing problem with customer retention, analysis reveals

CHRIS HANNAY, TEMUR DURRANI, MAHIMA SINGH

SPORTS

HOCKEY Bryan Trottier's memoir shows a gentle macho positivity, Cathal Kelly says B16

BASEBALL Toronto Blue Jays agree to three-year deal with manager John Schneider B17

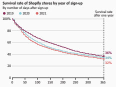
SOCCER Women's World Cup draw to set the stage for 32-team tournament next year B28

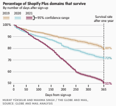
WHAT'S THE EXPORT IMPACT?

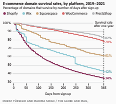
SATURDAY, OCTOBER 22, 2022 | THE GLOBE AND MAIL | REPORT ON BUSINESS | B7

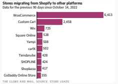
## Shopify: Company's customer survival rate is substantially lower than its rivals, data show

FROM B1

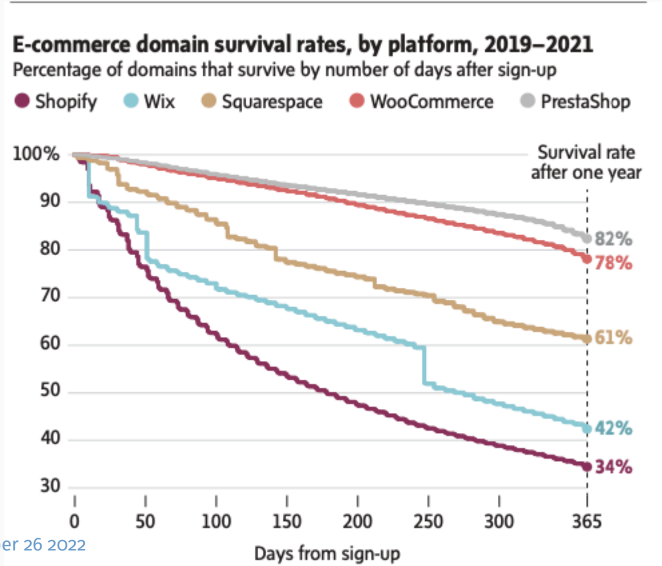



# Shopify



E-commerce domain survival rates, by platform, 2019–2021
Percentage of domains that survive by number of days after sign-up
Shopify
Wix
Squarespace
WooCommerce
PrestaShop
100%
90
80
70
60
50
40
30
Survival rate after one year
82%
78%
61%
42%
34%
0
50
100
150
200
250
300
365
Days from sign-up