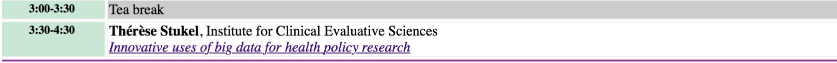
Today

- HW 1: due February 4, 11.59 pm.
- Matched case-control studies
- In the News: “High water mark: the rise in sea levels may be accelerating” Economist, Jan 17
- Big Data for Health Policy 3:30 - 4:30
  222 College St Room 230



3:00-3:30	Tea break
3:30-4:30	**Thérèse Stukel**, Institute for Clinical Evaluative Sciences *Innovative uses of big data for health policy research*

Thank you!

Alexander Stringer

```
aggregate(nodal[,c(1,2)],
            by=as.list(nodal[,-c(1,2)]),
            FUN=sum
            )
```

Shahriar Shams

```
library(SMPracticals); data(nodal)

#for some weird reason the \ddply" command doesn’t work
#on the nodal dataset, had to create nodal2
#(I see no difference between them though)

nodal2=data.frame(m=nodal$m,
                r=nodal$r,
                aged=nodal$aged,
                stage=nodal$stage,
                grade=nodal$grade,
                xray=nodal$xray,
                acid=nodal$acid)


require(plyr)
nodal3=ddply(nodal2, .(aged, stage,grade,xray,acid), summarize, m=sum(m), r=sum(r))
nodal3
```

R to the future

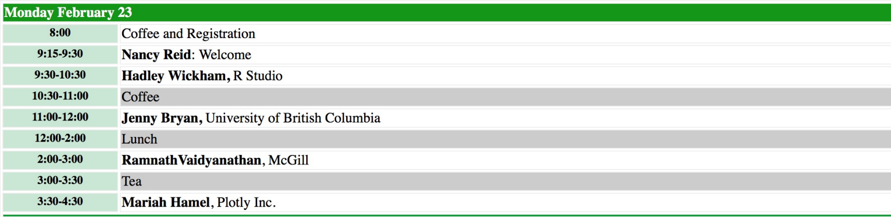
`plyr / dplyr`

Dianne Cook: *Data Visualization and Statistical Graphics in Big Data Analysis* recommends:

- Plots: `ggplot2, ggvis, animint, shiny`
- Reproducibility: `knitr`
- Data scraping: `dplyr, Rcpp, rvest`

Hadley Wickham:

`dplyr`



Monday February 23	
8:00	Coffee and Registration
9:15-9:30	**Nancy Reid**: Welcome
9:30-10:30	**Hadley Wickham,** R Studio
10:30-11:00	Coffee
11:00-12:00	**Jenny Bryan,** University of British Columbia
12:00-2:00	Lunch
2:00-3:00	**RamnathVaidyanathan**, McGill
3:00-3:30	Tea
3:30-4:30	**Mariah Hamel**, Plotly Inc.

Recap: overdispersion etc.

- saturated model: $y_i \sim Bin(n_i, p_i), \quad \tilde{p}_i = y_i/n_i,$

  $$\ell(\tilde{p}) = \Sigma\{y_i \log(y_i/n_i) + (n_i - y_i)\log(1 - y_i/n_i)\}$$

- what's the saturated model for linear regression? what is the maximized log-likelihood for this model?
- with binomial data, large-ish n_i, residual deviance compares regression model to saturated model
- if it's too large, we have the wrong model
- lack of independence among individual Bernoullis; a few outliers; wrong predictors ELM p. 43,4
- estimate $\tilde{\phi} = X^2/(n-p)$ ELM p. 45
- inflate variance $\hat{\beta} \dot{\sim} N(\beta, \tilde{\phi}(X^{\mathrm{T}}WX))$ instead of $N(\beta, X^{\mathrm{T}}WX)$

... overdispersion

```
> summary(bmod)

Call:
glm(formula = cbind(survive, total - survive) ~ location + period,
    family = binomial, data = troutegg)
...
period8      -2.3256     0.2429  -9.573  < 2e-16 ***
period11     -2.4500     0.2341 -10.466  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1021.469  on 19  degrees of freedom
Residual deviance:   64.495  on 12  degrees of freedom
AIC: 157.03

> summary(bmod2)

Call:
glm(formula = cbind(survive, total - survive) ~ location + period,
    family = quasibinomial, data = troutegg)

period8      -2.3256     0.5609  -4.146 0.001356 **
period11     -2.4500     0.5405  -4.533 0.000686 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 5.330358)
```

... overdispersion

SM §10.6, p.512

- $Y \mid \epsilon \sim \text{Bin}(m, \epsilon p)$
- $\text{E}(\epsilon) = 1, \quad \text{var}(\epsilon) = \xi$
- $\text{E}(Y) = \text{E}\{\text{E}(Y \mid \epsilon)\} = \text{E}(mp\epsilon) = mp$
- $\text{var}(Y) = \text{var}\{\text{E}(Y \mid \epsilon)\} + \text{E}\{\text{var}(Y \mid \epsilon)\}$
- $\text{var}(Y) = m\{p(1-p) + \xi p^2(m-1)\}$ ntbc
- variance is larger than $mp(1-p)$ see also ELM p.44
- can't be detected if $m = 1$ m plays the role of n_i

Matched case-control studies

ELM §2.12

- Cases $Y = 1$; Controls $Y = 0$ retrospective c-c study
- on the logit scale, we can estimate the effect of x on $\Pr(Y = 1 \mid x)$ Jan 14
- even though we have over-sampled the cases

- in a matched case-control study, we choose controls with same covariates
- then we do not model the effects of those covariates on response cannot
- if effect of covariate is more complex than $\beta_j x_j$, we avoid specifying the functional form
- we might indirectly adjust for effects that are hard to ascertain
- e.g. match on place of residence could help control for ‘environmental effects’ ELM p.48
- matched case control data not representative of population

... matched case-control studies

- suppose we have $1 : M$ matching one case, M matched controls
- for person i in matched set j, we have
  $y_{ij}, x_{ij}, \quad i = 0, 1, \ldots, M$
- model:

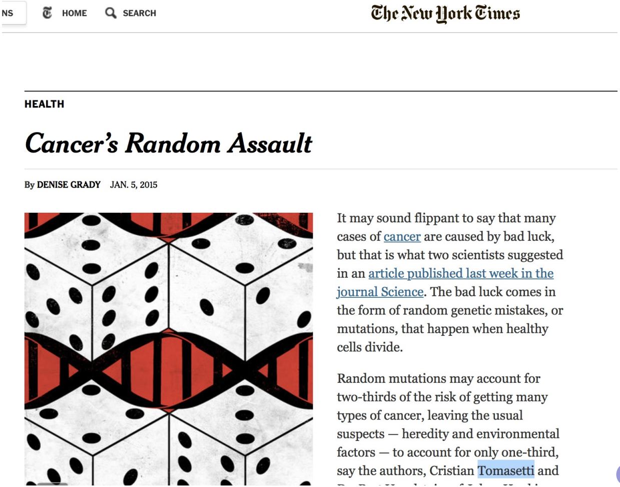
$$\log \frac{p_j(x_{ij})}{1 - p_i(x_{ij})} = \alpha_j + x_{ij}^{\mathrm{T}}\beta$$

- different intercept for each matched set confounding variables
- same effect of covariates across patients and sets β
- data: in matched set j, we have 1 case (person 0) and M controls (persons $1, \ldots, M$)
-

$$\begin{aligned} \Pr(y_{0j} = 1 \mid \Sigma_{i=1}^{M} y_{ij} = 1) &= \frac{\Pr(y_{0j} = 1, y_{1j} = 0, \ldots, y_{Mj} = 0)}{\Pr(y_{1j} = 0, \ldots, y_{Mj} = 0)} \\ &= \frac{\exp(x_{0j}^{\mathrm{T}}\beta)}{\Sigma_{i=0}^{M} \exp(x_{ij}^{\mathrm{T}}\beta)} \end{aligned}$$

... matched case-control studies

In the News



HEALTH

Cancer's Random Assault

By DENISE GRADY JAN. 5, 2015

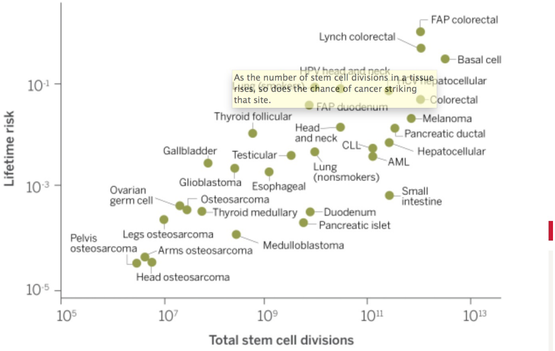
It may sound flippant to say that many cases of cancer are caused by bad luck, but that is what two scientists suggested in an article published last week in the journal Science. The bad luck comes in the form of random genetic mistakes, or mutations, that happen when healthy cells divide.

Random mutations may account for two-thirds of the risk of getting many types of cancer, leaving the usual suspects — heredity and environmental factors — to account for only one-third, say the authors, Cristian Tomasetti and

Link

... in the news

Science News: “The bad luck of cancer” (also published online as “Simple math explains why you may or may not get cancer”).
Science: “Variation in cancer risk among tissues can be explained by the number of stem cell divisions”.
Economist: “Chancing your arm: a recent study does not show that two-thirds of cancer cases are due to bad luck”.

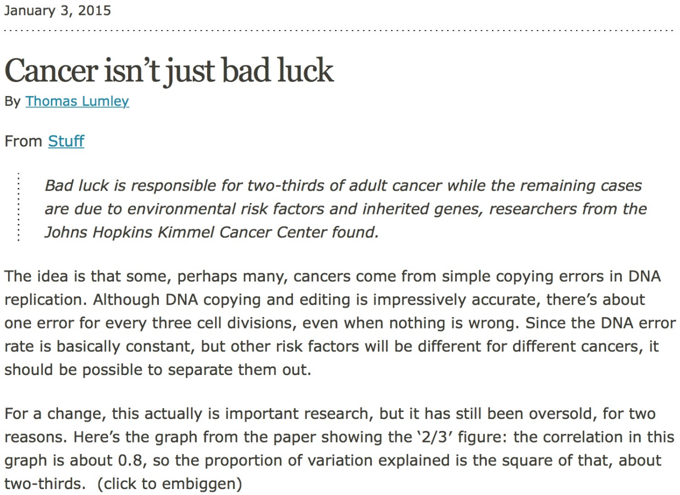


As the number of stem cell divisions in a tissue rises, so does the chance of cancer striking that site.

DATA: TOMASETTI ET AL./SCIENCE

Cancer in the News

statschat.org.nz



January 3, 2015

Cancer isn't just bad luck

By Thomas Lumley

From Stuff

> *Bad luck is responsible for two-thirds of adult cancer while the remaining cases are due to environmental risk factors and inherited genes, researchers from the Johns Hopkins Kimmel Cancer Center found.*

The idea is that some, perhaps many, cancers come from simple copying errors in DNA replication. Although DNA copying and editing is impressively accurate, there's about one error for every three cell divisions, even when nothing is wrong. Since the DNA error rate is basically constant, but other risk factors will be different for different cancers, it should be possible to separate them out.

For a change, this actually is important research, but it has still been oversold, for two reasons. Here's the graph from the paper showing the '2/3' figure: the correlation in this graph is about 0.8, so the proportion of variation explained is the square of that, about two-thirds. (click to embiggen)

... cancer

- For a change, this actually is important research, but it has still been oversold, ...
- there are labels such as “Lung (smokers)” and “Lung (non-smokers)”, so it’s not as simple as ‘bad luck’. Some risk factors have been taken into account. It’s not obvious whether this makes the correlation higher or lower.
- the proportion of variation explained isn’t a proportion of cancer risk
- Using a log scale for incidence is absolutely right when showing the biological relationship, but you can’t read proportions of incidence explained off that graph
- Using the log scale gives a lot more weight to the very rare cancers in the lower left corner, which turn out not to have important modifiable risk factors. Using an untransformed y-axis gives equal weight to all cancers, which is what you want from a medical or public health point of view.

... cancer

- Using the log scale gives a lot more weight to the very rare cancers in the lower left corner, which turn out not to have important modifiable risk factors. Using an untransformed y-axis gives equal weight to all cancers, which is what you want from a medical or public health point of view.
- Except, even that isn't quite right. If you look at my two graphs it's clear that the correlation will be driven by the top three points. Two of those are familial colorectal cancers, and the incidence quoted is the incidence in people with the relevant mutations; the third is basal cell carcinoma, which barely counts as cancer from a medical or public health viewpoint If we leave out the familial cancers and basal cell carcinoma, the proportion explained drops to about 10%.
- If we leave out put back basal cell carcinoma as well, something statistically interesting happens. The correlation shoots back up again, but only because it's being driven by a single point. A more honest correlation estimate, predicting each point based on the other points and not based on itself, is much lower.
- So, in summary: the "two-thirds of cancers explained" is Just Wrong.

... cancer

- Statsguy
- The problem is that it applies only to explaining the variation in cancer risk from one tissue to another. It tells us nothing about how much of the risk within a given tissue is due to modifiable factors.
- Plumbum
- Imagine a hypothetical world in which cancer occurs during stem cell division with some significant probability only if a given environmental factor is present, and that environmental factor is present equally in all tissue types. In this world cancer incidence across tissue types is perfectly correlated with the number of stem cell divisions, but nevertheless all cancer is caused by the environmental factor.