Today

- HW 1: due February 4, 11.59 pm.
- Aspects of Design — CD Chapter 2
- Continue with Chapter 2 of ELM
- In the News:

Recap: data on proportions

- data: $y_i, x_i, n_i, \quad i = 1, \ldots n$
- possible model: $Y_i \sim Bin(n_i, p_i)$
- regression: $\text{logit}(p_i) = x_i^{\mathrm{T}}\beta$ more generally $p_i = g(x_i^{\mathrm{T}}\beta)$
- inference: $\hat{\beta} \dot{\sim} N\{\beta, j^{-1}(\hat{\beta})\}$
- residual deviance: $2\{\ell(\tilde{p}; y) - \ell(\hat{p}; y)\} \dot{\sim} \chi^2_{n-q}$
- $\tilde{p}_i = y_i/n_i; \quad \hat{p}_i = p_i(\hat{\beta}) = g(x_i^{\mathrm{T}}\hat{\beta})$
- change in deviance: $2\{\ell(\hat{p}_A; y) - \ell(\hat{p}_B; y)\} \dot{\sim} \chi^2_\nu$

STA 2201S: Applied Statistics II Spring 2015

Homework 1

Due February 4, 11.59 pm on Blackboard. On the course web page you can find the assingment under "Course Materials".

- Homework Questions
- Reference paper for Q1

January 7

- Slides
- iPad slides annotated
- Buzzfeed article
- Information about `knitr` and `Sweave`
- R code to reproduce analysis in slides
- Report of the Presidential Commision on the Space Shuttle Challenger Accident. The oring data is about 1/3 of the way down this page.

Course Information

Text

Extending the Linear Model with R by J. Faraway.

Recommended

Statistical Models by A.C. Davison.
Principles of Applied Statistics by D.R. Cox and C.A. Donnelly

Example 10.18

aggregated data presented in textbook

Table 10.8 Data on nodal involvement (Brown, 1980).

m	r	age	stage	grade	xray	acid
6	5	0	1	1	1	1
6	1	0	0	0	0	1
4	0	1	1	1	0	0
4	2	1	1	0	0	1
4	0	0	0	0	0	0
3	2	0	1	1	0	1
3	1	1	1	0	0	0
3	0	1	0	0	0	1
3	0	1	0	0	0	0
2	0	1	0	0	1	0
2	1	0	1	0	0	1
2	1	0	0	1	0	0
1	1	1	1	1	1	1
1	1	1	1	0	1	1
1	1	1	0	1	1	1
1	1	1	0	0	1	1
1	0	1	0	1	0	0
1	1	0	1	1	1	0
1	0	0	1	1	0	0
1	1	0	1	0	1	0
1	1	0	0	1	0	1
1	0	0	0	0	1	1

... example 10.18

binary data presented in R library

```
> library(SMPracticals); data(nodal)
> head(nodal)
  m r aged stage grade xray acid
1 1 1    0     1     1    1    1
2 1 1    0     1     1    1    1
3 1 1    0     1     1    1    1
4 1 1    0     1     1    1    1
5 1 1    0     1     1    1    1
6 1 0    0     1     1    1    1
```

- all covariates 0/1
- several patients have the same value of the covariates

  covariate classes: ELM
- these can be added up to make a binomial observation

... example 10.18

```
 > nodal2[1:4,]
  m r aged stage grade xray acid
1 6 5    0     1     1    1    1
2 6 1    0     0     0    0    1
3 4 0    1     1     1    0    0
4 4 2    1     1     0    0    1


> ex1018binom = glm(cbind(r,m-r) ~ ., data = nodal2, family = binomial)
> summary(ex1018binom)  # stuff omitted
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.0794     0.9868  -3.121  0.00180 **
aged          -0.2917     0.7540  -0.387  0.69881
stage         1.3729     0.7838   1.752  0.07986 .
grade         0.8720     0.8156   1.069  0.28500
xray          1.8008     0.8104   2.222  0.02628 *
acid          1.6839     0.7915   2.128  0.03337 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(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: 41.693

Number of Fisher Scoring iterations: 5
```

... example 10.18: variable selection

```
> step(ex1018binom)

Coefficients:
(Intercept)        stage         xray         acid
     -3.052        1.645        1.912        1.638

Degrees of Freedom: 22 Total (i.e. Null);  19 Residual
Null Deviance:     40.71
Residual Deviance: 19.64  AIC: 39.26
```

– we can drop `aged` 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

... example 10.18: 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
-

```
> update(ex1018binom, . ~ . - aged - stage)

Call:  glm(formula = cbind(r, m - r) ~ grade + xray + acid, family = binomial,
    data = nodal2)

Coefficients:
(Intercept)        grade         xray         acid
     -2.734        1.420        1.750        1.797

Degrees of Freedom: 22 Total (i.e. Null);  19 Residual
Null Deviance:     40.71
Residual Deviance: 21.28  AIC: 40.9

> deviance(ex1018binom)
[1] 18.06869
> pchisq(21.28-18.07,df=2,lower=F)
[1] 0.2008896
```

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*

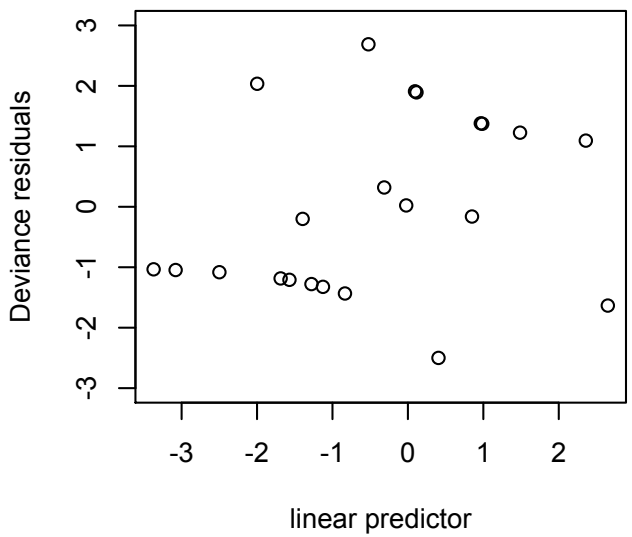
- for binomial data $\ell(\beta; y) = \Sigma[y_i x_i^{\mathrm{T}} \beta - n_i \log\{1 + \exp(x_i^{\mathrm{T}} \beta)\}]$

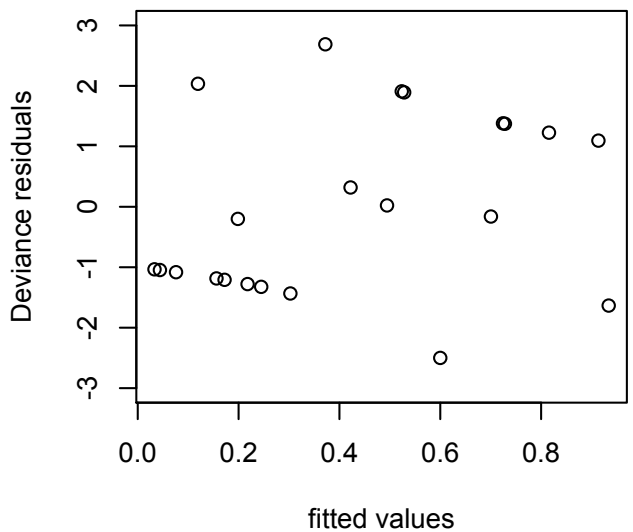
... example 10.18: residuals

```
> summary(ex1018binom)

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
```





Binomial model residuals

- Residual Deviance is log-likelihood ratio statistic for the fitted model compared to the saturated model
- saturated model maximized at $\tilde{p}_i = y_i/n_i$

$$\ell(\tilde{p}) = \sum_{i=1}^{n} \{y_i \log(y_i/n_i) + (n_i - y_i) \log(1 - y_i/n_i)\}$$

- fitted model maximized at $\hat{\beta}$

$$\ell(\hat{\beta}) = \sum_{i=1}^{n} \{y_i \log p_i(\hat{\beta}) + (n_i - y_i) \log(1 - p_i(\hat{\beta}))\}$$

- twice the difference:

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

- see p.29, after (2.1), where $n_i p_i(\hat{\beta}) = \hat{y}_i$

Deviance residuals

```
> summary(ex1018binom)

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
```

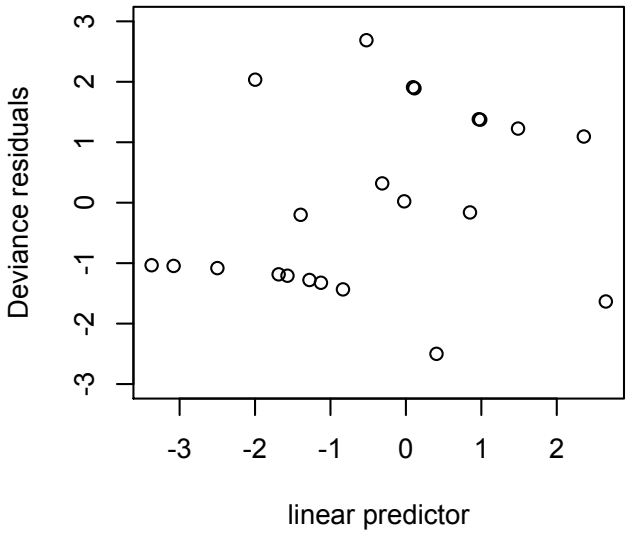
Deviance:
$2\sum_{i=1}^{n}[y_i \log\{y_i/n_i p_i(\hat{\beta})\} + (n_i - y_i)\log\{(n_i - y_i)/(n_i - n_i p_i(\hat{\beta}))\}]$

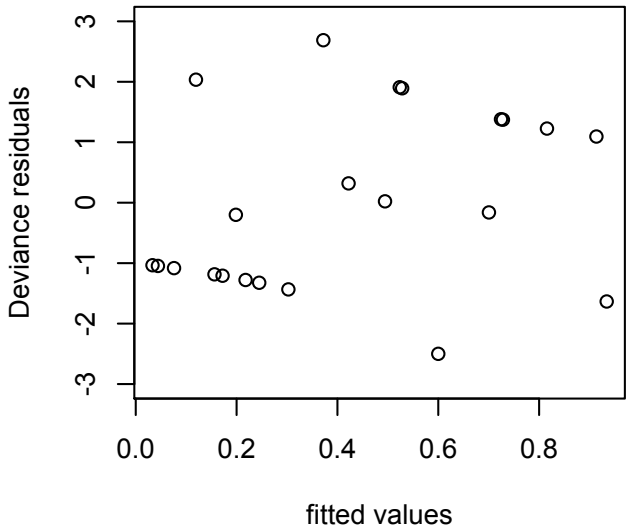
approximately distributed as χ^2_{n-q} $n_i \to \infty$

Deviance residuals:

$$\begin{aligned} r_{Di} &= \pm\sqrt{(2[y_i \log\{y_i/n_i\hat{p}_i\} + (n_i - y_i)\log\{(n_i - y_i)/(n_i - n_i\hat{p}_i)\}])} \\ &\dot{\sim} N(0,1) \end{aligned}$$

... example 10.18: residuals





Binary Data

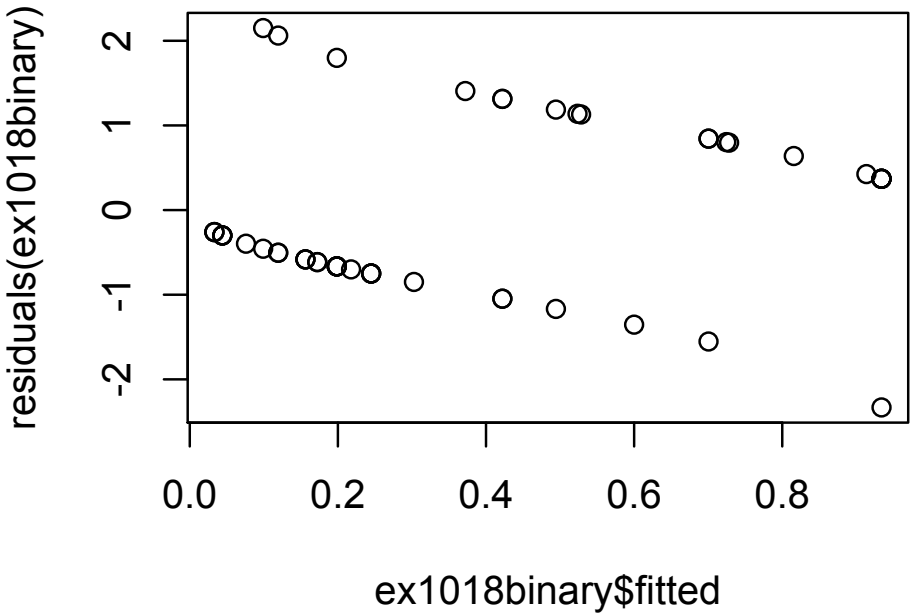
- if all $n_i = 1$, data is binary — as distinguished from binomial
- example: Example 10.18; data as in `SMPracticals`

```
> head(nodal)
  m r aged stage grade xray acid
1 1 1    0     1     1    1    1
2 1 1    0     1     1    1    1
3 1 1    0     1     1    1    1
4 1 1    0     1     1    1    1
5 1 1    0     1     1    1    1
6 1 0    0     1     1    1    1

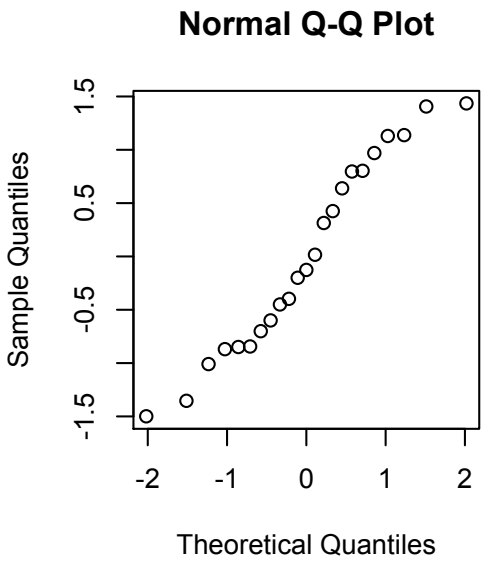
> ex1018binary <- glm(cbind(r,m-r) ~ ., data = nodal, family = binomial)
> summary(ex1018binary)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.0794     0.9868  -3.121   0.0018 **
aged1        -0.2917     0.7540  -0.387   0.6988
stage1        1.3729     0.7838   1.752   0.0799 .
grade1        0.8720     0.8156   1.069   0.2850
xray1         1.8008     0.8104   2.222   0.0263 *
acid1         1.6839     0.7915   2.128   0.0334 *
---
...
    Null deviance: 70.252  on 52  degrees of freedom
Residual deviance: 47.611  on 47  degrees of freedom
AIC: 59.611
```

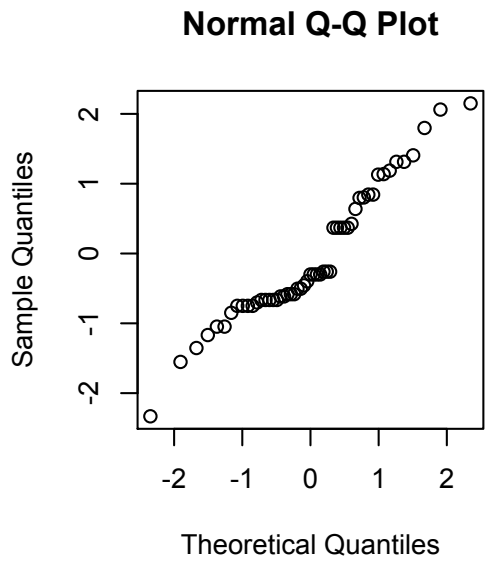
Binary Data

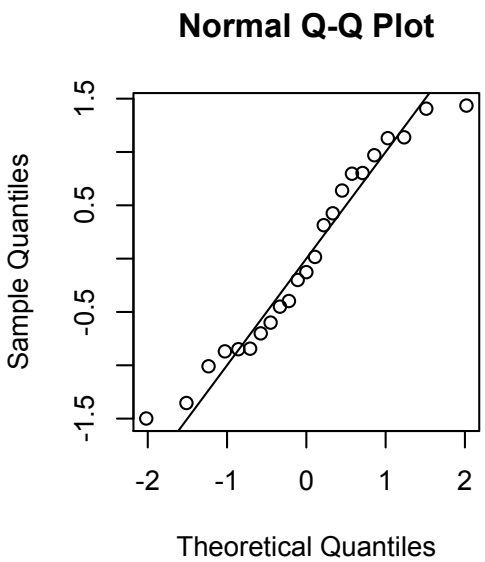
```
plot(ex1018binary$fitted.values,
residuals(ex1018binary))
```

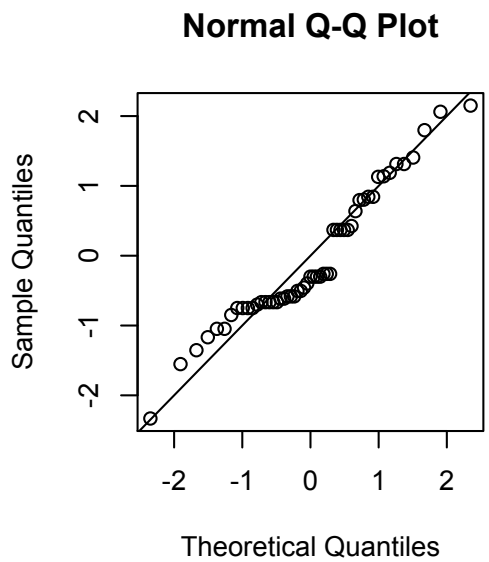


Cannot use residual deviance to measure goodness-of-fit


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-2
-1
0
1
2
-1.5
-0.5
0.5
1.5


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-2
-1
0
1
2


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-2
-1
0
1
2
-1.5
-0.5
0.5
1.5


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-2
-1
0
1
2

Logistic regression

- read §2.4 for one motivation of logistic regression model
- read §2.5 (and AS I) for interpretation of parameters in terms of log odds
- see Example `babyhood` in §2.5 for logistic regression with qualitative covariates
- what is the algebraic form of the model? how are the dummy covariates coded? what is x_i^{T}?
- note construction of confidence intervals for odds, exponentiating confidence intervals for change in log-odds (§2.5)

2×2 table

ELM §2.6

- special case: covariate takes values 0, 1
- $\Pr(Y_i = 1 \mid x_i = 0) = \dfrac{\exp(\beta_0)}{1 + \exp(\beta_0)}$
- $\Pr(Y_i = 1 \mid x_i = 1) = \dfrac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)}$
- $Y = 1$ is the event of interest – death, cure, heart attack, ...
- $x = 1$ is the factor of interest – treatment, smoking status, exposure, ...
- units with $Y = 1$ are cases (dead, sick, recovered, ...)
- units with $Y = 0$ controls (alive, well, not recovered ...)

Prospective and retrospective sampling

CD §3.6

Population

	$y = 0$	$y = 1$
$x = 0$	π_{00}	π_{01}
$x = 1$	π_{10}	π_{11}

Prospective study

	$y = 0$	$y = 1$
$x = 0$	$\pi_{00}/(\pi_{00} + \pi_{01})$	$\pi_{01}/(\pi_{00} + \pi_{01})$
$x = 1$	$\pi_{10}/(\pi_{10} + \pi_{11})$	$\pi_{11}/(\pi_{10} + \pi_{11})$

Retrospective study

	$y = 0$	$y = 1$
$x = 0$	$\pi_{00}/(\pi_{00} + \pi_{10})$	$\pi_{01}/(\pi_{01} + \pi_{11})$
$x = 1$	$\pi_{10}/(\pi_{00} + \pi_{10})$	$\pi_{11}/(\pi_{01} + \pi_{11})$

cross-product ratio in 2nd and 3rd table the same as that in 1st

... prospective and retrosp. sampling

ELM p.35, SM §10.4.2.2

$$p^*(x) = \frac{\pi_1 p(x)}{\pi_1 p(x) + \pi_0(1 - p(x))}$$

$$\text{logit } p^*(x) = \log \frac{\pi_1}{\pi_0} + \text{logit } p(x)$$

$$\pi_1 = \Pr(\text{included in study} \mid \text{disease})$$

$$\pi_0 = \Pr(\text{included in study} \mid \text{not disease})$$

$$p^*(x) = \Pr(\text{disease} \mid \text{included in study}, x)$$

$$p(x) = \Pr(\text{disease} \mid x)$$

prospective study: $\pi_1 = \pi_0$

retrospective study $\pi_1 > \pi_0$

but may not be known

Overdispersion

ELM §2.11, SM 10.6

- $Y_i \sim Bin(n_i, p_i) \Rightarrow \mathrm{E}(Y_i) = n_i p_i, \quad \mathrm{Var}(Y_i) = n_i p_i (1 - p_i)$
- variance is determined by the mean
- ```
  > bmod
  Degrees of Freedom: 19 Total (i.e. Null);  12 Residual
  Null Deviance:      1021
  Residual Deviance: 64.5  AIC: 157
  ```
- quasi-binomial: $\mathrm{E}(Y_i) = n_i p_i, \quad \mathrm{Var}(Y_i) = \phi n_i p_i (1 - p_i)$
- estimate ϕ? over-dispersion parameter

- usually use $X^2/(n-p)$, where

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

Design and Analysis CD Ch. 1

- choice of material/individuals to study – "units of analysis"
- "For studies of a new phenomenon it will usually be best to examine situations in which the phenomenon is likely to appear in the most striking form, even if this is in some sense artificial"
- statistical analysis needs to take account of the design (even if statistician enters the project at the analysis stage)
- need to be clear at the design stage about broad features of the statistical analysis – more publicly convincing **and** "reduces the possibility that the data cannot be satisfactorily analysed"
- "it is unrealistic and indeed potentially dangerous to follow an initial plan unswervingly ... it may be a crucial part of the analysis to clarify the research objectives"

Experimental and observational studies CD Ch. 1

- experiment is a study in which all key elements are under the control of the investigator
- in an observational study key elements cannot be manipulated by the investigator.
- “It often, however, aids the interpretation of an observation study to consider the question: what would have been done in a comparable experiment?”
- Example: hormone replacement therapy and heart disease
- observational study – strong and statistically significant reduction in heart disease among women taking hormone replacement therapy
- women’s health study (JAMA, 2002, p.321) – statistically significant increase in risk among women randomized to hormone replacement therapy

Principles of measurement CD Ch. 1

- “construct validity – measurements do actually record the features of concern”
- “record a number of different features sufficient to capture concisely the important aspects”
- reliable – i.e. reasonably reproducible
- “cost of the measurements is commensurate with their importance”
- “measurement process does not appreciably distort the system under study”

Types and phases of analysis CD Ch. 1

- “A general principle, sounding superficial but difficult to implement, is that analyses should be as simple as possible, but no simpler.”
- the method of analysis should be transparent
- main phases of analysis
  - data auditing and screening;
  - preliminary analysis;
  - formal analysis;
  - presentation of conclusions

Design of Studies

CD, Ch.2

- common objectives
- to avoid systematic error, that is distortion in the conclusions arising from sources that do not cancel out in the long run
- to reduce the non-systematic (random) error to a reasonable level by replication and other techniques
- to estimate realistically the likely uncertainty in the final conclusions
- to ensure that the scale of effort is appropriate

... design of studies

- we concentrate largely on the careful analysis of individual studies
- in most situations synthesis of information from different investigations is needed
- but even there the quality of individual studies remains important
- examples include overviews (such as the Cochrane reviews)
- in some areas new investigations can be set up and completed relatively quickly; design of individual studies may then be less important

... design of studies

- formulation of a plan of analysis
- establish and document that proposed data are capable of addressing the research questions of concern
- main configurations of answers likely to be obtained should be set out
- level of detail depends on the context
- even if pre-specified methods must be used, it is crucial not to limit analysis
- planned analysis may be technically inappropriate
- more controversially, data may suggest new research questions or replacement of objectives
- latter will require confirmatory studies

Unit of study and analysis

- smallest subdivision of experimental material that may be assigned to a treatment context: Expt
- Example: RCT – unit may be a patient, or a patient-month (in crossover trial)
- Example: public health intervention – unit is often a community/school/...
- split plot experiments have two classes of units of study and analysis
- in investigations that are not randomized, it may be helpful to consider what the primary unit of analysis would have been, had a randomized experiment been feasible
- the unit of analysis may not be the unit of interpretation – ecological bias systematic difference between impact of x at different levels of aggregation
- on the whole, limited detail is needed in examining the variation within the unit of study

Types of observational studies

- secondary analysis of data collected for another purpose
- estimation of a 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
- experiment: investigator has complete control over treatment assignment
- census
- meta-analysis: statistical assessment of a collection of studies on the same topic

Avoidance of systematic error

CD §2.4

- "distortion in the conclusions arising from irrelevant sources that do not cancel out in the long run"
- can arise through systematic aspects of, for example, a measuring process, or the spatial or temporal arrangement of units
- this can often be avoided by design, or adjustment in analysis
- can arise by the entry of personal judgement into some aspect of the data collection process
- this can often be avoided by randomization and blinding

In the News

NS HOME SEARCH The New York Times

HEALTH

Cancer's Random Assault

By DENISE GRADY JAN. 5, 2015

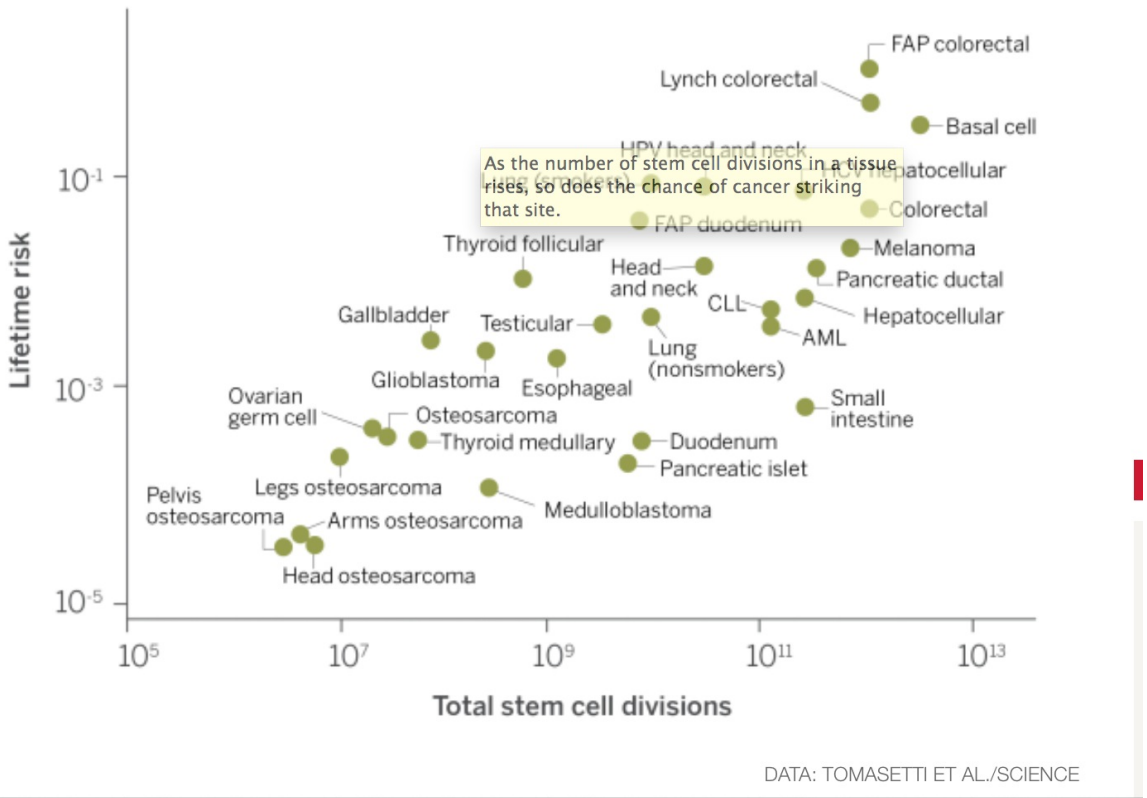
It may sound flippant to say that many cases of cancer are caused by bad luck, but that is what two scientists suggested in an article published last week in the journal Science. The bad luck comes in the form of random genetic mistakes, or mutations, that happen when healthy cells divide.

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
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



... in the news

"Best way for professors to get good student evaluations"? Slate, Dec.9

"What's in a name: exposing gender bias in student ratings of teaching", MacNeill et al., *Innovations in Higher Education* 2014 published online Dec. 4.