# Mathematical Statistics II

STA2212H S LEC9101

Week 9

March 11 2025

THE CONVERSATION

Academic rigour, journalistic flair

Microplastics are tiny bits of plastic that show up in the environment. Svetlozar Hristov/iStock via Getty Images Plus

## What's that microplastic? Advances in machine learning are making identifying plastics in the environment more reliable

Published: March 6, 2025 8.35am EST

**Ambuj Tewari**

Professor of Statistics, University of Michigan

THE CONVERSATION

Academic rigour, journalistic flair

Microplastics are tiny bits of plastic that show up in the environment. Svetlozar Hristov/iStock via Getty Images Plus

# What's that microplastic? Advances in machine learning are making identifying plastics in the environment more reliable

Published: March 6, 2025 8.35am EST

**Ambuj Tewari**

Professor of Statistics, University of Michigan

Microplastics – the tiny particles of plastic shed when litter breaks down – are everywhere, from the deep sea to Mount Everest, and many researchers worry that they could harm human health.

I am a machine learning researcher. With a team of scientists, I have developed a tool to make identification of microplastics using their unique chemical fingerprint more reliable. We hope that this work will help us learn about the types of microplastics floating through the air in our study area,

# Today

1. Recap Mar 4 choosing test stats, hypothesis/signficance testing, multiple testing
2. Nonparametric tests, goodness-of-fit
3. Introduction to causal inference
4. Reviewing project guidelines
5. Conformal prediction

Upcoming seminar

Department Seminar Thursday March 6 11.00 – 12.00
Hydro Building, Room 9014
Bayesian modelling in neuroimaging
Michele Guindani, UCLA

# Project Guidelines

link

**Project Guidelines** STA 2212S: Mathematical Statistics II 2025

Presentation on April 1, 2025.
Report submission due April 16, 2025.

**Part 1: Presentation [10 points]**

On the last day of class (April 1), you will present your final project. This includes:

- Emailing a .pdf version of your team's slide deck pdf to `nancym.reid@utoronto.ca` by 09.00 April 1. You are responsible for the slides corresponding to your sections of the write-up. Please email **one** complete version for each team.
- Presenting the slides in no more than 10 minutes; each team member to present for no more than 5 minutes.

# Recap

$$X_1, \ldots, X_n \sim f(\boldsymbol{x}; \theta), \theta \in \Theta \subset \mathbb{R}^p$$

- testing $H_0 : \theta \in \Theta_0$

against some alternative

simple or composite H

Hyp. t.

- rejection region $\{\boldsymbol{x} : t(\boldsymbol{x}) > c_\alpha\}$

$$\mathrm{pr}_{H_0}\{t(\boldsymbol{X}) \geq c_\alpha\} \leq \alpha$$

$t \geq c_\alpha$ "reject" $H_0$

$\leq c_\alpha$ don't

Signif.

- $p$-value:

$$\mathrm{pr}_{H_0}\{t(\boldsymbol{X}) \geq t(\boldsymbol{x}^{obs})\}$$

large values

- significance function ($\theta \in \mathbb{R}$)

$$p(\theta) = \mathrm{pr}_\theta\{t(\boldsymbol{X}) \geq t(\boldsymbol{x}^{obs})\}$$

$\theta \in \Theta$

# Recap: Choosing test statistics $t(\cdot)$

1. Optimal choice – Neyman-Pearson lemma — Might be UMP (HW 7)
2. Pragmatic choice – likelihood-based test statistics — score test, Wald, LRT
3. Pragmatic choice – nonparametric test statistics

(a) Need to know distribution of test statistic under $H_0$ — $c_\alpha$, p-value

(b) Test statistic should be large when $H_0$ is not true — in probability

(c) Test statistic should have maximum power to detect departures from $H_0$

# … power of sign test

334 7 · Estimation and Hypothesis Testing

UMP for 1-sided

$\bar{x}$

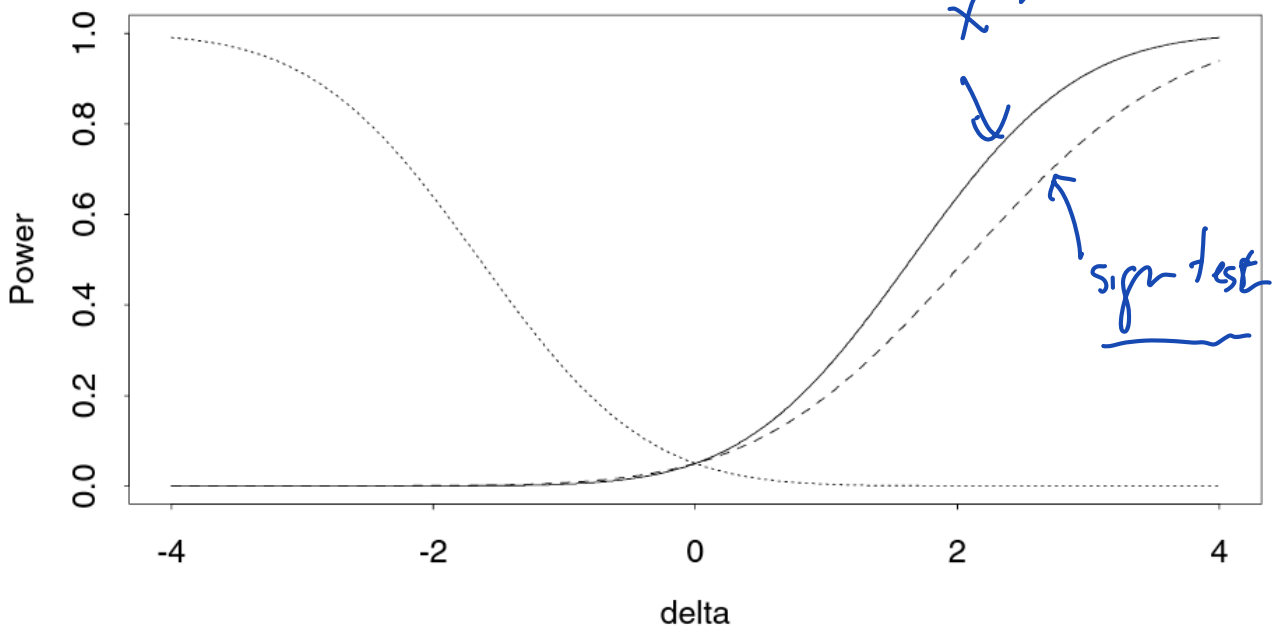



**Figure 7.6** Power functions for a test of whether the mean of a $N(\mu, \sigma^2)$ random sample of size $n$ equals $\mu_0$ against the alternative $\mu = \mu_1$, as a function of $\delta = n^{1/2}(\mu_1 - \mu_0)/\sigma$. The test size is $\alpha = 0.05$. The solid curve is the power function for a test of $\mu_1 > \mu_0$ based on $\overline{y}$, and the dashed line is the power function for the sign test. Both critical regions are of form $\overline{y} > t_\alpha$. The dotted curve is the power function for $\overline{y}$ when the critical region is $\overline{y} < t_\alpha$.

# Recap: Hypothesis tests and significance tests

- Hypothesis tests typically means:
  - $H_0$, $H_1$
  - critical/rejection region $R \subset \mathcal{X}$,
  - level $\alpha$, power $1 - \beta$
  - conclusion: "reject $H_0$ at level $\alpha$" or "do not reject $H_0$ at level $\alpha$"
  - planning: maximize power for some relevant alternative

minimize type II error

- Significance tests typically means:
  - $H_0$, ← $H_1$ vague or unspecified
  - test statistic $T$
  - observed value $t^{obs}$,
  - $p$-value $p^{obs} = \Pr(T \geq t^{obs}; H_0)$
  - alternative hypothesis often only implicit

large $T$ points to alternative

# Multiple testing

EH 15.2, AoS 10.7

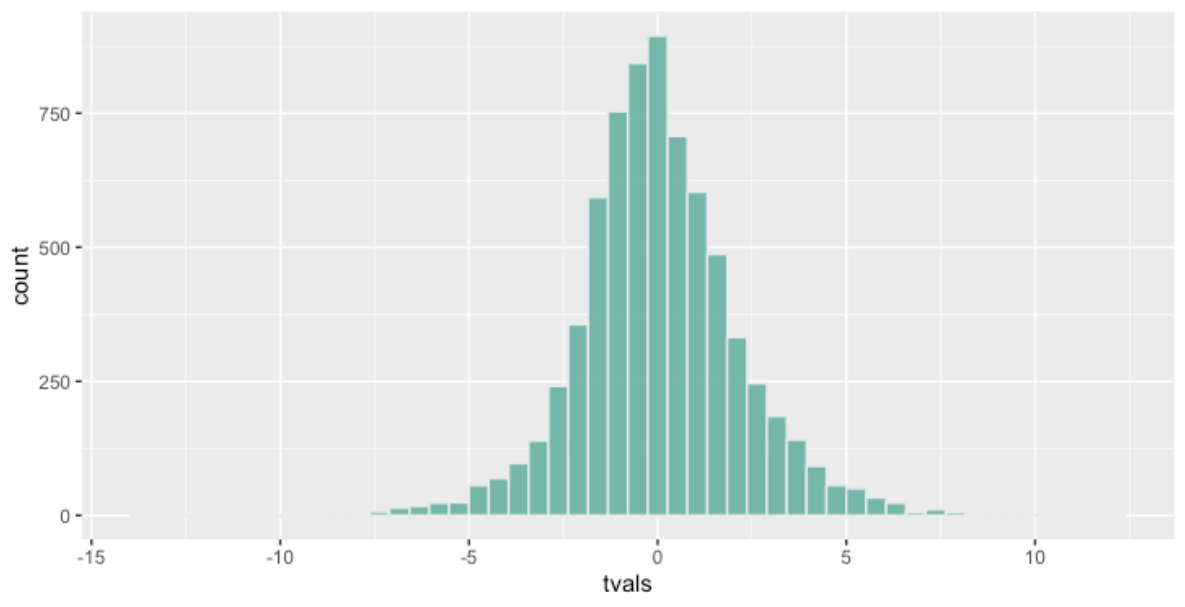
```
leukemia_big <- read.csv
  ("http://web.stanford.edu/~hastie/CASI_files/DATA/leukemia_big.csv")
  dim(leukemia_big)
  [1] 7128   72
```

- each row is a different gene; 47 AML responses and 25 ALL responses
- we could compute 7128 $t$-statistics for the mean difference between AML and ALL

```
tvals <- rep(0,7128)
for (i in 1:7128){
  leukemia_big[i,] %>% select(starts_with("ALL")) %>% as.numeric() -> x
  leukemia_big[i,] %>% select(starts_with("AML")) %>% as.numeric() -> y
  tvals[i] <- t.test(x,y,var.equal=T)$statistic
  }
```

# Multiple testing

EH 1.2, 15.2




$\underset{\sim}{X}, \underset{\sim}{Y}$ permute

45 ← Gp 1

27 ← Gp 2

t-stat. for each gene

$H_0: F_X(x) = F_Y(y)$ 7,128

```
summary(tvals)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-13.52611  -1.20672  -0.08406   0.02308   1.20886  12.26065
```

$H_1$: not $H_0$

# Benjamini-Hochberg

AoS 10.7; EH 15.2

Bonferroni use $\frac{\alpha}{m}$ cutoff for 'sign'

- order the $p$-values $p_{(1)}, \ldots, p_{(m)}$
- find $i_{max}$, the largest index for which

$$p_{(i)} \leq \frac{i}{m} q \quad \approx \alpha$$

- Let $BH_q$ be the rule that rejects $H_{0i}$ for $i \leq i_{max}$, not rejecting otherwise

- Theorem: If the $p$-values corresponding to valid null hypotheses are independent of each other, then

$$E \frac{\#\text{rejects incorrect}}{\#\text{reject's}} \qquad FDR(BH_q) = \pi_0 q \leq q, \qquad \text{where } \pi_0 = m_0/m$$

$\pi_0$ unknown but close to 1

- change the bound under dependence

$$p_{(i)} \leq \frac{i}{mC_m} q \qquad C_m = \sum_{i=1}^{m} \frac{1}{i}$$

# Example

$.05/10 = .005$

AoS Ex.10.28

Bonferroni

| index | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| pval | 0.00017 | 0.00448 | 0.00671 | 0.00907 | 0.01220 | 0.33626 | 0.3934 | 0.5388 | 0.5813 | 0.9862 |
| cut1 | 0.00500 | 0.01000 | 0.01500 | 0.02000 | 0.02500 | 0.03000 | 0.0350 | 0.0400 | 0.0450 | 0.0500 |
| cut2 | 0.00171 | 0.00341 | 0.00512 | 0.00683 | 0.00854 | 0.01024 | 0.0119 | 0.0137 | 0.0154 | 0.0171 |

B-H

dependence

simple composite

$H_0$: ✓

$H_1$: completely spec. dist ✓

p values are often easy to compute

$H_0$ composite

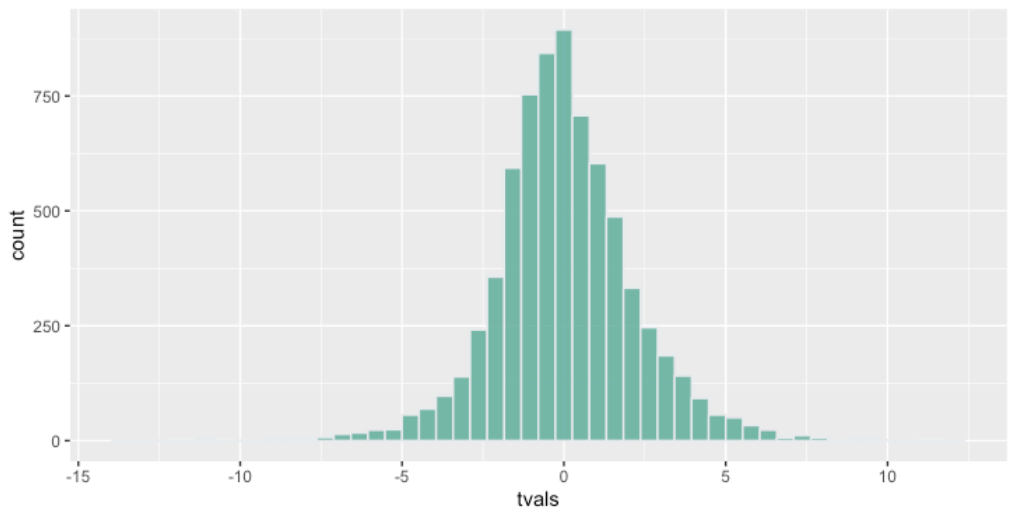
If $F_X = F_Y$

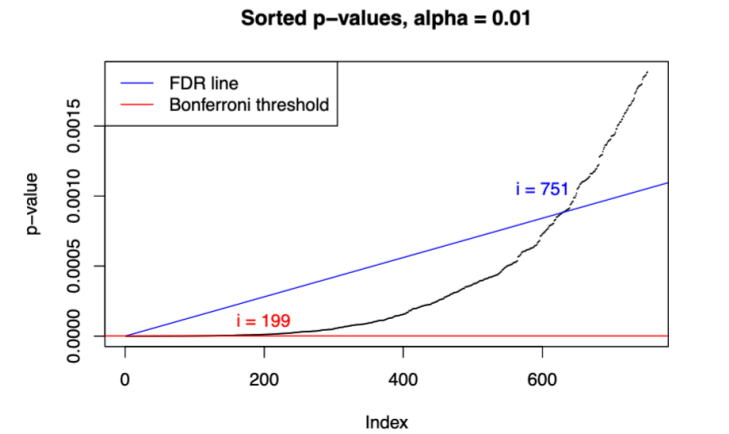
then all $\binom{72}{27}$ permutation of data have = prob.

convert comp. $H_0$ to simple (by cond'g)

# Multiple testing

EH 1.2, 15.2




The figure above shows sorted $p$-values of the $N = 7128$ $t$-tests. The red line corresponds to the threshold $\alpha/N$ from the Bonferroni method, and the blue line is the FDR line $(i/N)\alpha$. The

```
> summary(ttest)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-13.52611  -1.20672  -0.08406   0.02308   1.20886  12.26065
```

$X_1, \dots, X_{10000}$

$\hat{\beta}_5(\underline{X_5})$

"large" relative to est. s.e.

(Lasso) or $\hat{\beta}_5 > 0$, for $\lambda_{opt}$

# Multinomial goodness of fit tests

MS 9.2; MoS 10.8

composite

- $X_1, \ldots, X_n$ i.i.d.
- $H_0 : X_i \sim f(x; \theta)$; $H_1 : X_i$ arbitrary distribution
- Define $k$ sets $A_1, \ldots, A_k$ s.t.

$$\mathrm{pr}(X_i \in \cup_{j=1}^{k} A_j\} = 1$$

$\cup A_j = \mathcal{X}$ sample sp

- Define

$$Y_j = \sum_{i=1}^{n} 1\{X_i \in A_j\}$$

number of obs in category $j$

# Multinomial goodness of fit tests

MS 9.2; AoS 10.8

- $X_1, \ldots, X_n$ i.i.d.
- $H_0 : X_i \sim f(x; \theta)$; $H_1 : X_i$ arbitrary distribution
- Define $k$ sets $A_1, \ldots, A_k$ s.t.

$$\mathrm{pr}(X_i \in \cup_{j=1}^{k} A_j\} = 1$$

- Define

$$Y_j = \sum_{i=1}^{n} 1\{X_i \in A_j\}$$

number of obs in category $j$

- $Y = (Y_1, \ldots, Y_k) \sim Mult_k(n; p)$
- $\mathrm{pr}(Y_1 = y_1, \ldots, Y_k = y_k; p) =$
- $H_0 : p = p(\theta)$; $H_1 : p$ arbitrary

$p_j(\theta) = \int_{\{x \in A_j\}} f(x; \theta)\, dx$

$p_j = \Pr(Y_i \in A_j)$

$f(y; p) = \frac{n!}{y_1! \cdots y_k!} p_1^{y_1} \cdots p_k^{y_k}$, $\sum y_j = n$

$\sum p_j = 1$, $0 \le p_j \le 1$

# Multinomial goodness of fit tests

MS 9.2, AoS 10.8

- log-likelihood function $\ell(\underset{\sim}{p}) = \sum_{j=1}^{k} y_j \log p_j \quad j = 1, \dots, k$

- generalized likelihood ratio test $\max \ell(\underset{\sim}{p})$ over $H_0 \cup H_1$

  $\max \ell(\underset{\sim}{p})$ over $H_0$

$$\frac{\sup_{\underset{\sim}{p}} L(\underset{\sim}{p}; y)}{\sup_{\underset{\sim}{\theta}} L(\underset{\sim}{p}(\theta); y)} \quad \begin{matrix} \uparrow H_1 \\ \uparrow H_0 \end{matrix}$$

$$W = 2\left\{ \underbrace{\sup_{\underset{\sim}{p}} \ell(\underset{\sim}{p}; y)} - \sup_{\underset{\sim}{\theta}} \ell(\underset{\sim}{p}(\theta); y) \right\}$$

$\hat{\underset{\sim}{p}}$ under no constraint $= \frac{y}{n} \Rightarrow \hat{p}_j = \frac{y_j}{n}$

# Multinomial goodness of fit tests

MS 9.2, AoS 10.8

$\underline{y} \sim \text{Multi}_k(n, \underline{p}) = \frac{n!}{y_1! \cdots y_k!} \prod_{j=1}^{k} p_j^{y_j}$

- log-likelihood function

- generalized likelihood ratio test

$= \sum_{\text{bins}} \left( O \log \frac{O}{E} \right)$

- Theorem 9.1 (MS): Under $H_0$

$p = \dim(\theta)$

$$W = 2 \sum_{j=1}^{k} Y_j \log \left( \frac{Y_j}{np_j(\tilde{\theta})} \right) \xrightarrow{d} \chi^2_{k-1-p}$$

usual LR theory

$= 2 \sum Y_j \log \frac{Y_j}{n} - 2 \sum Y_j \log p_j(\tilde{\theta})$

$$L(p) = \prod_{j=1}^{k} p_j^{y_j} \qquad \sum p_j = 1 \;,\; \sum y_j = n$$

$$\ell(p) = \sum_{j=1}^{k} y_j \log p_j \;, \qquad \sum p_j = 1,\; \sum y_j = n$$

$$= \sum_{j=1}^{k-1} \{ y_j \log p_j \} + \underline{(n - y_1 - \cdots - y_{k-1})} \log(1 - p_1 - \cdots p_{k-1})$$

$$\frac{\partial \ell}{\partial p_j} = \frac{y_j}{p_j} - \frac{n - y_1 - \cdots - y_{k-1})}{(1 - p_1 - \cdots p_{k-1})} \;, \quad j = 1, \ldots, k-1$$

$$= \frac{y_j}{p_j} - \frac{y_k}{p_k} \qquad j = 1, \ldots, k-1$$

$$\Big|_{\hat{p}} = 0 \implies \cdots \quad \frac{y_j}{n_j} = \hat{p}_j$$

$$\ell(p(\theta)) = \sum_{j=1}^{k-1} y_j \log p_j(\theta) + \left(n - \sum^{k-1} y_j\right) \log\left\{1 - \sum^{k-1} p_j(\theta)\right\}$$

$$\frac{\partial \ell}{\partial \underset{\sim}{\theta}} : \quad \sum_{j=1}^{k} y_j \frac{p_j'(\underset{\sim}{\theta})}{p_j(\underset{\sim}{\theta})} \;; \quad \sum p_j(\theta) = 1$$

$$\sum y_j = n$$

$\theta_{k \times 1}$ vector $\quad \left.\frac{\partial \ell}{\partial \theta}\right|_{\hat{\theta}} = 0 \qquad \underset{\sim}{\tilde{\theta}}$ vector of length $k$

$\vdots$

# Multinomial goodness of fit tests

- log-likelihood function

$$\log \frac{Y_j}{np_j(\tilde{\theta})} = \log\Big(1 + \underbrace{\frac{Y_j}{np_j(\tilde{\theta})} - 1}_{\varepsilon}\Big)$$

$$\log(1+\varepsilon) \simeq \varepsilon - \frac{1}{2}\varepsilon^2 + \frac{1}{3}\varepsilon^3 - \frac{1}{4}\varepsilon^4 \ldots$$

- generalized likelihood ratio test

$p = \dim(\theta)$

- Theorem 9.1 (MS): Under $H_0$

$$\boxed{W = 2\sum_{j=1}^{k} Y_j \log\left(\frac{Y_j}{np_j(\tilde{\theta})}\right) \xrightarrow{d} \chi^2_{k-1-p}}$$

dist of $W$ $\approx$ dist of $Q$ for large $n$

- Theorem 92. (MS): Under $H_0$

$(O - E)^2$

$$\boxed{Q = \sum_{j=1}^{k} \frac{\{Y_j - np_j(\hat{\theta})\}^2}{np_j(\hat{\theta})} \xrightarrow{d} \chi^2_{k-1-p}}$$

$E$

# Multinomial goodness-of-fit tests

$(\hat{\theta}-\theta)^T I(\tilde{\theta})(\hat{\theta}-\theta)$

AoS 10.8; MS 9.2

? ?? "Wald"

$(p(\tilde{\theta})-p)\,\underline{I(p(\tilde{\theta}))}(\quad)$

outcome

Table 9.1 *Frequency of goals in First Division matches and "expected" frequency under Poisson model in Example 9.2*

| Goals | 0 | 1 | 2 | 3 | 4 | $\geq 5$ |
|---|---|---|---|---|---|---|
| Frequency | 252 | 344 | 180 | 104 | 28 | 16 |
| Expected | 248.9 | 326.5 | 214.1 | 93.6 | 30.7 | 10.2 |

test

| | |
|---|---|
| $n_1$ | $n_2$ |
| $n_3$ | $n_4$ |

$$p_0(\lambda) = 1 - \sum_{j=0}^{4} p_j(\lambda); \quad p_j(\lambda) = e^{-\lambda}\lambda^j/j!, \quad \tilde{\lambda} = 1.3118$$

$$Q = 11.09; \quad W = 10.87; \quad \mathrm{pr}(\chi^2_4 > [11.09, 10.87]) = [0.026, 0.028]$$

# Multinomial goodness-of-fit tests

SM Ex 4.38

136 4 · *Likelihood*

| | | Antigen 'B' | | |
|---|---|---|---|---|
| | | Absent | Present | Total |
| Antigen 'A' | Absent | 'O': 202 | 'B': 35 | 237 |
| | Present | 'A': 179 | 'AB': 6 | 185 |
| Total | | 381 | 41 | 422 |

**Table 4.3** Blood groups in England (Taylor and Prior, 1938). The upper part of the table shows a cross-classification of 422 persons by presence or absence of antigens 'A' and 'B', giving the groups 'A', 'B', 'AB', 'O' of the human blood group system. The lower part shows genotypes and corresponding probabilities under one- and two-locus models. See Example 4.38 for details.

| | Two-locus model | | One-locus model | |
|---|---|---|---|---|
| Group | Genotype | Probability | Genotype | Probability |
| 'A' | $(AA; bb), (Aa; bb)$ | $\alpha(1-\beta)$ | $(AA), (AO)$ | $\lambda_A^2 + 2\lambda_A\lambda_O$ |
| 'B' | $(aa; BB), (aa; Bb)$ | $(1-\alpha)\beta$ | $(BB), (BO)$ | $\lambda_B^2 + 2\lambda_B\lambda_O$ |
| 'AB' | $(AA; BB), (Aa; BB), (AA; Bb), (Aa; Bb)$ | $\alpha\beta$ | $(AB)$ | $2\lambda_A\lambda_B$ |
| 'O' | $(aa; bb)$ | $(1-\alpha)(1-\beta)$ | $(OO)$ | $\lambda_O^2$ |

log-linear models for cross-classified

$Q = 15.73$; $W = 17.66$ (two-locus)

$p < 10^{-5}$

$Q = 2.82$; $W = 3.17$ (single locus)

$p = 0.09; 0.07$

# Smooth goodness-of-fit tests

MS 9.3, SM p.327-9

- $X_1, \ldots, X_n$ *i.i.d.* $F(\cdot)$; $H_0 : F = F_0$ — simple e.g. $U(0,1)$ — comp e.g. $N(\mu, \sigma^2)$ ← $\hat{\mu}, \hat{\sigma}^2$ estimated

cumulative d.f.

- $\widehat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \leq t\}$

- three test statistics:
  1. $\sup_t |\widehat{F}_n(t) - F_0(t)|$ Kolmogorov-Smirnov
  2. $\int \{\widehat{F}_n(t) - F_0(t)\}^2 dF_0(t)$ Cramer-von Mises
  3. $\int \frac{\{\widehat{F}_n(t) - F_0(t)\}^2}{F_0(t)\{1 - F_0(t)\}} dF_0(t)$ Anderson-Darling

- SM Example 7.24 testing $N(\mu, \sigma^2)$ distribution
- SM Example 7.23; 6.14 testing $U(0, 1)$ distribution

# Smooth goodness-of-fit tests



- Special case $H_0 : F(t) = F_0(t) = t$ $\qquad X_i \sim U(0,1)$
- Recall

$$E_0\{\widehat{F}_n(t)\} = F_0(t) = t, \quad \mathrm{var}\{\widehat{F}_n(t)\} = t(1-t)/n$$

fixed $t$

- What about distribution of

$$\sup_t |\widehat{F}_n(t) - t| \qquad \int \{\widehat{F}_n(t) - t\}^2 dt \qquad \int \frac{\{\widehat{F}_n(t) - t\}^2}{F_0(t)\{1 - F_0(t)\}} dt$$

- need joint density of $\widehat{F}_n(t)\ \forall\ t$

# Smooth goodness-of-fit tests

MS 9.3

- Special case $H_0 : F(t) = F_0(t) = t$ $\qquad X_i \sim U(0,1)$
- Recall

$$E_0\{\widehat{F}_n(t)\} = F_0(t) = t, \quad \text{var}\{\widehat{F}_n(t)\} = t(1-t)/n$$

- What about distribution of

$$\sup_t |\widehat{F}_n(t) - t| \qquad \int \{\widehat{F}_n(t) - t\}^2 dt \qquad \int \frac{\{\widehat{F}_n(t) - t\}^2}{F_0(t)\{1 - F_0(t)\}} dt$$

- need joint density of $\widehat{F}_n(t)\ \forall\ t$

0 at $t=0$
1 at $t=1$

- define stochastic process $\boxed{B_n(t) = \sqrt{n}(\widehat{F}_n(t) - t)}$

$\{B_n(t) : 1 \geq t \geq 0\}$

$B_n(0) = 0 = B_n(1)$

- vector $(B_n(t_1), \ldots, B_n(t_k)) \xrightarrow{d} N_k(0, C)$, $\quad C_{ij} = \min(t_i, t_j) - t_i t_j$ $\qquad$ MS 9.3
- a Brownian bridge is a continuous function on $(0,1)$ with all finite-dimensional distributions as above

# Smooth goodness-of-fit tests

MS 9.3

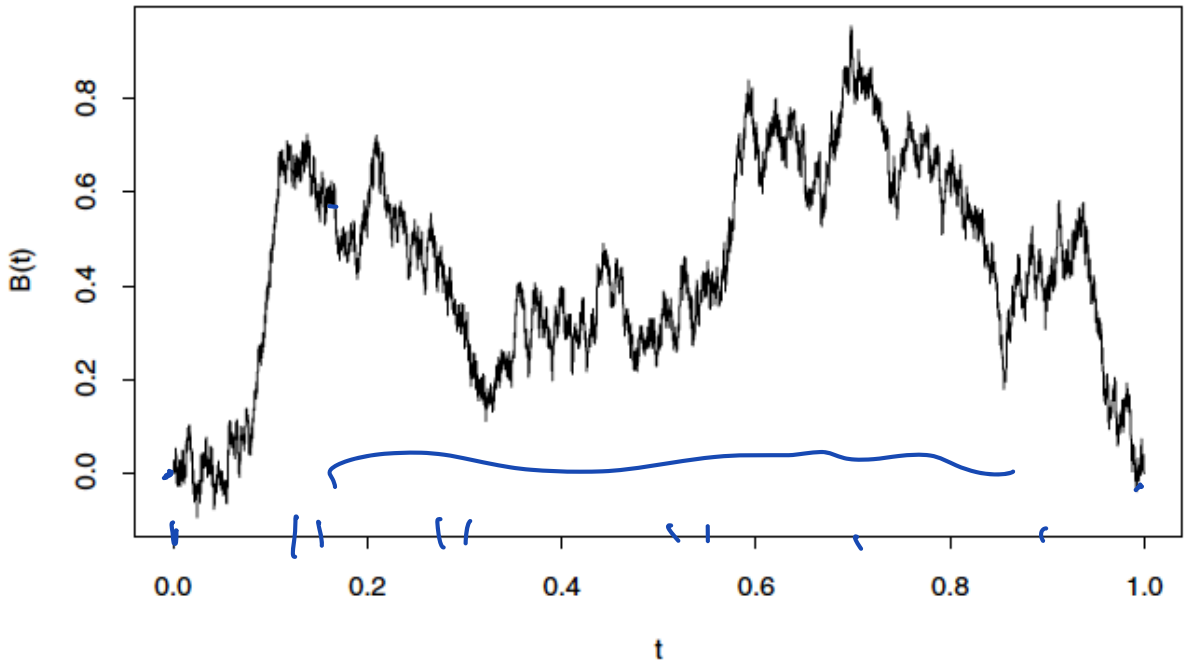



Figure 9.1 *A simulated realization of a Brownian bridge process.*

# Smooth goodness-of-fit tests



- Kolmogorov-Smirnov test

$$K_n = \sup_{0 \le t \le 1} |B_n(t)|$$

- Cramer-vonMises test

$$W_n^2 = \int_0^1 B_n^2(t)dt$$

- Anderson-Darling test

$$A_n^2 = \int_0^1 \frac{B_n^2(t)}{t(1-t)}dt$$

from $\hat{F}_n(t) - F_0(t)$
to $B_n(t)$

# Smooth goodness-of-fit tests



- Kolmogorov-Smirnov test

$$K_n = \sup_{0 \le t \le 1} |B_n(t)|$$

very conservative

- Cramer-vonMises test

$$W_n^2 = \int_0^1 B_n^2(t)dt$$

- Anderson-Darling test

$$A_n^2 = \int_0^1 \frac{B_n^2(t)}{t(1-t)} dt$$

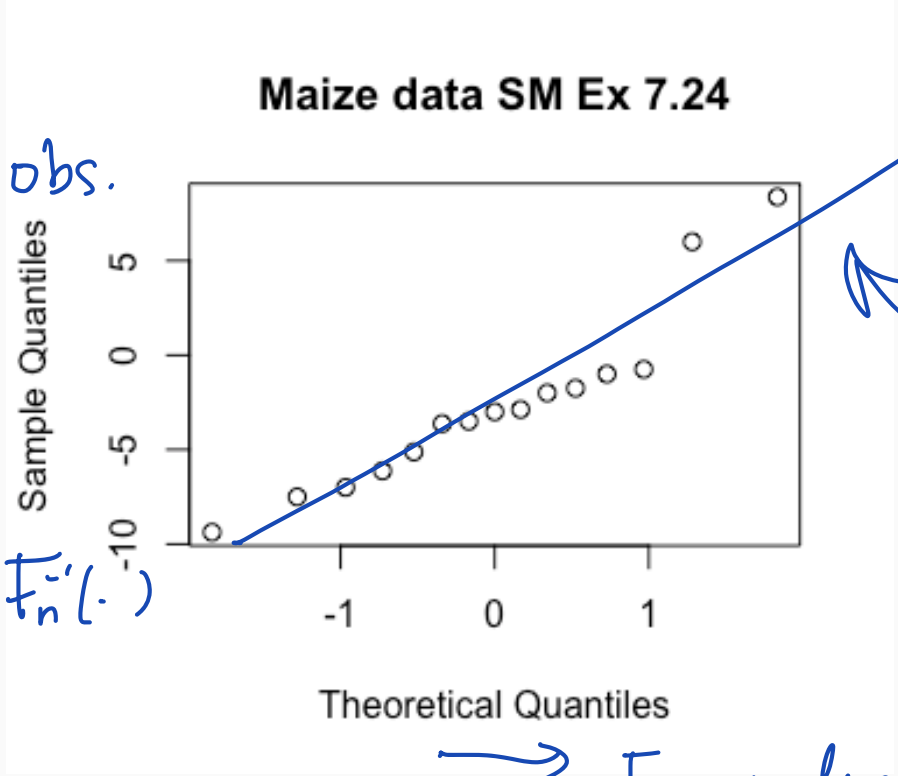
- limit theorems

$$K_n \xrightarrow{d} K, \qquad W_n^2 \xrightarrow{d} \sum_{j=1}^{\infty} \frac{Z_j^2}{j^2\pi^2}, \qquad A_n^2 \xrightarrow{d} \sum_{j=1}^{\infty} \frac{Z_j^2}{j(j+1)}$$

$$\mathrm{pr}(K > x) = 2\sum_{j=1}^{\infty}(-1)^{j+1}\exp(-2j^2x^2)$$

# Example

SM p.327-9




obs.

$F_n^{-1}(\cdot)$

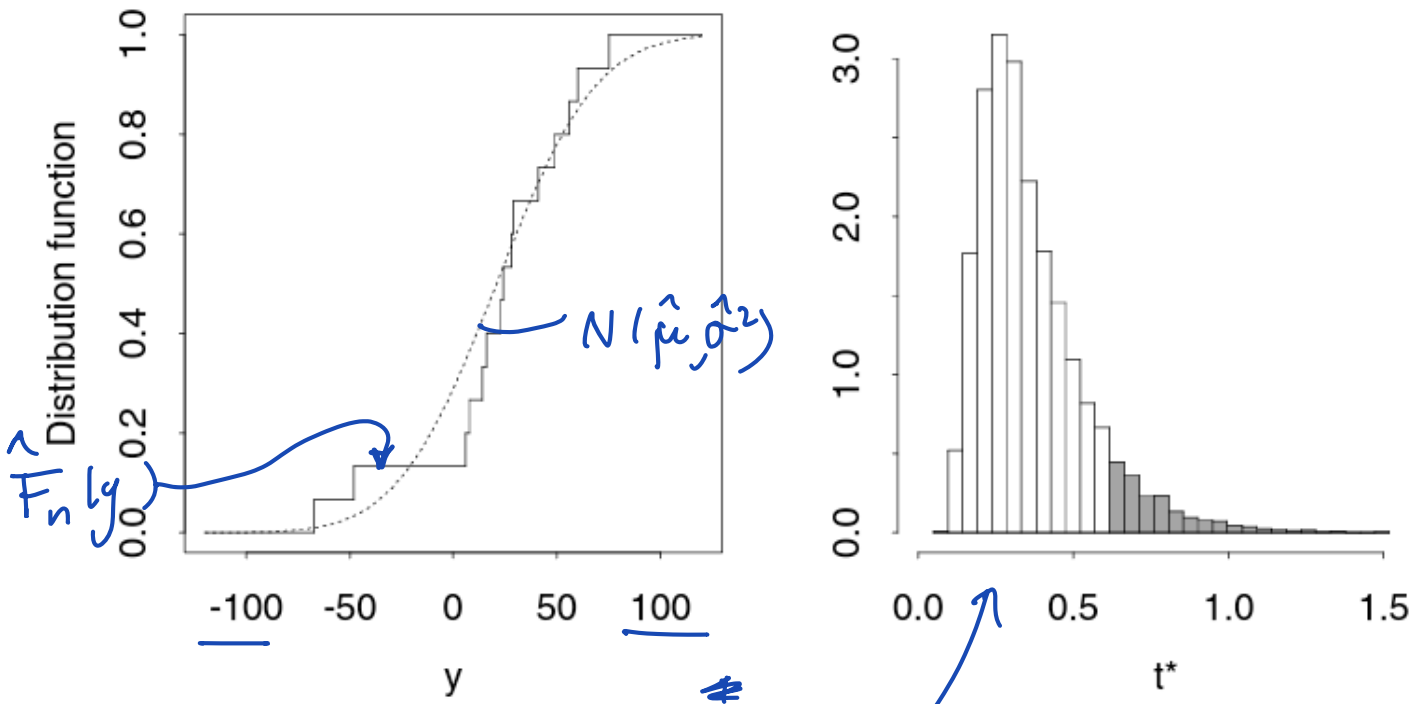
Exp$_N$ values of $X_{(i)}$

```
library(SMPracticals)
data(darwin)
cross <- seq(1,30,by=2)
self <- cross+1
diffs <- darwin[self,4]-darwin[cross,4]
qqnorm(diffs)
```

e.g. test st. ?

# Example: SM 7.24

**Figure 7.5** Analysis of maize data. Left: empirical distribution function for height differences, with fitted normal distribution (dots). Right: null density of Anderson–Darling statistic $T$ for normal samples of size $n = 15$ with location and scale estimated. The shaded part of the histogram shows values of $T^*$ in excess of the observed value $t_{\text{obs}}$.




$$\int_{-\infty}^{\infty} \frac{\{\hat{F}_n(t) - \hat{F}_0(t)\}^2}{\hat{F}_0(t)\{1 - \hat{F}_0(t)\}} \, d\hat{F}_0(t)$$

SM Example 7.24 testing $N(\mu, \sigma^2)$ distribution

# A note on Bayesian testing

- Relatively simple case: $\boldsymbol{X} \sim f(\boldsymbol{x}; \theta)$, $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$

simple

post. prob

$$\mathrm{pr}(H_0 \mid \boldsymbol{x}) = \frac{f(\boldsymbol{x} \mid H_0)\mathrm{pr}(H_0)}{f(\boldsymbol{x} \mid H_0)\mathrm{pr}(H_0) + f(\boldsymbol{x} \mid H_1)\mathrm{pr}(H_1)}$$

$$\mathrm{pr}(H_0 \mid \underline{x}) = \frac{L(\theta_0; \underline{x})\,\mathrm{pr}(H_0)}{L(\theta_0; \underline{x})\,\mathrm{pr}(H_0) + \underbrace{\int f(x; \theta)\pi(\theta)\,d\theta}_{f_{H_1}(x)\ \text{marg'l}} \cdot \mathrm{pr}(H_1)}$$

depends strongly on prior

some version of model selection

# A note on Bayesian testing

AoS 11.8

- Relatively simple case: $\boldsymbol{X} \sim f(\boldsymbol{x}; \theta), \quad H_0 : \theta = \theta_0, \quad H_1 : \theta \neq \theta_0$
-

$$\begin{aligned}
\mathrm{pr}(H_0 \mid \boldsymbol{x}) &= \frac{f(\boldsymbol{x} \mid H_0)\mathrm{pr}(H_0)}{f(\boldsymbol{x} \mid H_0)\mathrm{pr}(H_0) + f(\boldsymbol{x} \mid H_1)\mathrm{pr}(H_1)} \\
&= \frac{f(\boldsymbol{x} \mid \theta_0)\mathrm{pr}(H_0)}{f(\boldsymbol{x} \mid \theta_0)\mathrm{pr}(H_0) + \int f(\boldsymbol{x} \mid \theta)\pi(\theta)d\theta\,\mathrm{pr}(H_1)} \\
&= \frac{L_n(\theta_0)}{L_n(\theta_0) + \int L_n(\theta)\pi(\theta)d\theta}
\end{aligned}$$

$\mathrm{Pr}(H_0) = \mathrm{Pr}(H_1) = \frac{1}{2}$

- can't use improper priors; result is sensitive to the prior for $\theta$

$0 \leq \int L_n(\theta)\pi(\theta)d\theta \leq L_n(\hat{\theta})$

# Causality

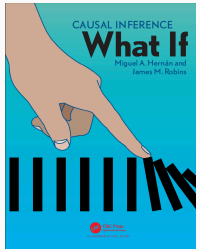
AoS 16,17; SM 9.1.2; Cox & Donnelly 9.2

- randomization; confounding; observational studies; experiments;
  "correlation is not causation", Simpson's 'paradox'

- counterfactuals; average treatment effect; conditional average treatment effect; ...

- graphical models; directed acyclic graphs; causal graphs; Markov assumptions...

- The Book

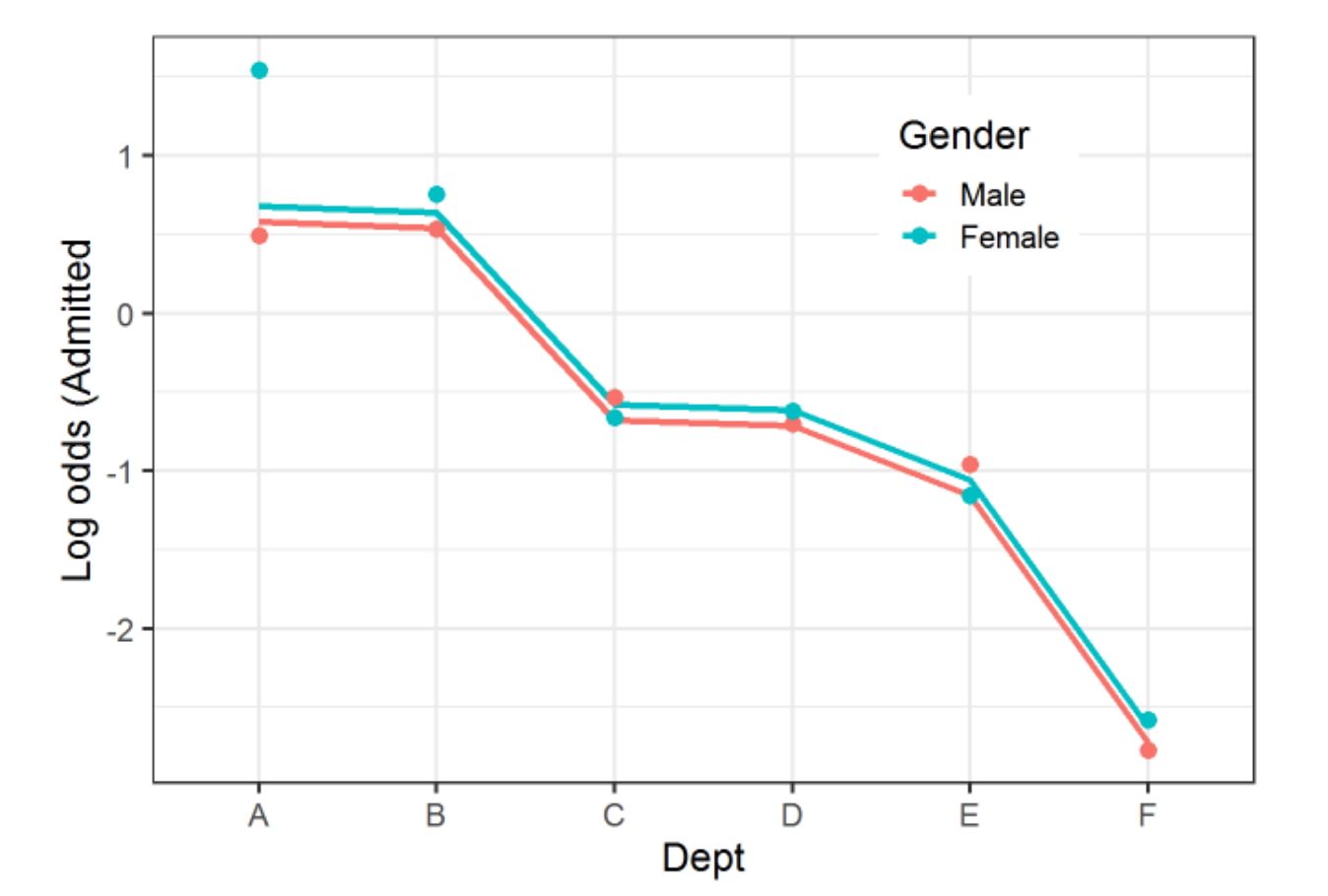


# Confounding variables

| | Men | | | Women | | |
|---|---|---|---|---|---|---|
| Major | Number of applicants | Number admitted | Percent admitted | Number of applicants | Number admitted | Percent admitted |
| A | 825 | 512 | 62 | 108 | 89 | 82 |
| B | 560 | 353 | 63 | 25 | 17 | 68 |
| C | 325 | 120 | 37 | 593 | 202 | 34 |
| D | 417 | 138 | 33 | 375 | 131 | 35 |
| E | 191 | 53 | 28 | 393 | 94 | 24 |
| F | 373 | 22 | 6 | 341 | 24 | 7 |
| Total | 2691 | 1198 | 44 | 1835 | 557 | 30 |

```
data(UCBAdmissions)
```

# ... Confounding variables



# ... Confounding variables

Radelet 1981

| race of defendant | death penalty imposed | death penalty not imposed | percentage |
|---|---|---|---|
| white | 19 | 141 | 11.88% |
| black | 17 | 149 | 10.24% |

# … Confounding variables

Radelet 1981

| race of defendant | death penalty imposed | death penalty not imposed | percentage |
|---|---|---|---|
| white | 19 | 141 | 11.88% |
| black | 17 | 149 | 10.24% |

| white victim | race of defendant | death penalty imposed | death penalty not imposed | percentage |
|---|---|---|---|---|
| | white | 19 | 132 | 12.58% |
| | black | 11 | 52 | 17.46% |

| black victim | race of defendant | death penalty imposed | death penalty not imposed | percentage |
|---|---|---|---|---|
| | white | 0 | 9 | 0% |
| | black | 6 | 97 | 5.83% |

| Age (years) | Smokers | Non-smokers |
|---|---|---|
| Overall | 139/582 (24) | 230/732 (31) |
| 18–24 | 2/55 (4) | 1/62 (2) |
| 25–34 | 3/124 (2) | 5/157 (3) |
| 35–44 | 14/109 (13) | 7/121 (6) |
| 45–54 | 27/130 (21) | 12/78 (15) |
| 55–64 | 51/115 (44) | 40/121 (33) |
| 65–74 | 29/36 (81) | 101/129 (78) |
| 75+ | 13/13 (100) | 64/64 (100) |

**Table 6.8** Twenty-year survival and smoking status for 1314 women (Appleton *et al.*, 1996). The smoker and non-smoker columns contain number dead/total (% dead).

# Causality and Counterfactuals

AoS Ch. 16; Cox & D 9.2

- $X$ – binary treatment indicator "treatment"
- $Y$ – binary outcome could be continuous
- "X causes Y" to be distinguished from "X is associated with Y"

# Causality and Counterfactuals

AoS Ch. 16; Cox & D 9.2

- $X$ – binary treatment indicator — "treatment"
- $Y$ – binary outcome — could be continuous
- "X causes Y" to be distinguished from "X is associated with Y"

- introduce potential outcomes $C_0, C_1$

$$Y = \begin{cases} C_0 & \text{if } X = 0 \\ C_1 & \text{if } X = 1 \end{cases}$$

- equivalently $Y = C_X$ or $Y = C_0(1 - X) + C_1 X$ — consistency equation

- causal treatment effect $\theta = E(C_1) - E(C_0)$ — want to estimate this

- association $\alpha = E(Y \mid X = 1) - E(Y \mid X = 0)$ — have data to estimate $\alpha$

# Counterfactual: Examples

AoS Ch.16; HR Ch.1

## Potential outcomes $C_0$, $C_1$

| $X$ | $Y$ | $C_0$ | $C_1$ |
|---|---|---|---|
| 0 | 4 | 4 | * |
| 0 | 7 | 7 | * |
| 0 | 2 | 2 | * |
| 0 | 8 | 8 | * |
| 1 | 3 | * | 3 |
| 1 | 5 | * | 5 |
| 1 | 8 | * | 8 |
| 1 | 9 | * | 9 |

treatment $X$, response $Y$

## Potential outcomes $Y^0$, $Y^1$

Table 2.1

| | $A$ | $Y$ | $Y^0$ | $Y^1$ |
|---|---|---|---|---|
| Rheia | 0 | 0 | 0 | ? |
| Kronos | 0 | 1 | 1 | ? |
| Demeter | 0 | 0 | 0 | ? |
| Hades | 0 | 0 | 0 | ? |
| Hestia | 1 | 0 | ? | 0 |
| Poseidon | 1 | 0 | ? | 0 |
| Hera | 1 | 0 | ? | 0 |
| Zeus | 1 | 1 | ? | 1 |
| Artemis | 0 | 1 | 1 | ? |
| Apollo | 0 | 1 | 1 | ? |
| Leto | 0 | 0 | 0 | ? |
| Ares | 1 | 1 | ? | 1 |
| Athena | 1 | 1 | ? | 1 |
| Hephaestus | 1 | 1 | ? | 1 |
| Aphrodite | 1 | 1 | ? | 1 |
| Cyclope | 1 | 1 | ? | 1 |
| Persephone | 1 | 1 | ? | 1 |
| Hermes | 1 | 0 | ? | 0 |
| Hebe | 1 | 0 | ? | 0 |
| Dionysus | 1 | 0 | ? | 0 |

# Causal Effect and Association

## Potential outcomes

Table 1.1

| | $Y^{a=0}$ | $Y^{a=1}$ |
|---|---|---|
| Rheia | 0 | 1 |
| Kronos | 1 | 0 |
| Demeter | 0 | 0 |
| Hades | 0 | 0 |
| Hestia | 0 | 0 |
| Poseidon | 1 | 0 |
| Hera | 0 | 0 |
| Zeus | 0 | 1 |
| Artemis | 1 | 1 |
| Apollo | 1 | 0 |
| Leto | 0 | 1 |
| Ares | 1 | 1 |
| Athena | 1 | 1 |
| Hephaestus | 0 | 1 |
| Aphrodite | 0 | 1 |
| Cyclope | 0 | 1 |
| Persephone | 1 | 1 |
| Hermes | 1 | 0 |
| Hebe | 1 | 0 |
| Dionysus | 1 | 0 |

## Observed outcomes

Table 1.2

| | $A$ | $Y$ |
|---|---|---|
| Rheia | 0 | 0 |
| Kronos | 0 | 1 |
| Demeter | 0 | 0 |
| Hades | 0 | 0 |
| Hestia | 1 | 0 |
| Poseidon | 1 | 0 |
| Hera | 1 | 0 |
| Zeus | 1 | 1 |
| Artemis | 0 | 1 |
| Apollo | 0 | 1 |
| Leto | 0 | 0 |
| Ares | 1 | 1 |
| Athena | 1 | 1 |
| Hephaestus | 1 | 1 |
| Aphrodite | 1 | 1 |
| Cyclope | 1 | 1 |
| Persephone | 1 | 1 |
| Hermes | 1 | 0 |
| Hebe | 1 | 0 |
| Dionysus | 1 | 0 |

# Causal treatment effect

AoS Eq. (16.2)

$\theta = \mathrm{E}(C_1) - \mathrm{E}(C_0)$

risk difference; ratio; odds

$\alpha = \mathrm{E}(Y \mid X = 1) - \mathrm{E}(Y \mid X = 0)$

If $X$ is is independent of $(C_0, C_1)$, $\theta = \alpha$

If $X$ is randomly assigned, then $X \perp (C_0, C_1)$

# Example 16.2

| $X$ | $Y$ | $C_0$ | $C_1$ |
|---|---|---|---|
| 0 | 0 | 0 | $0^*$ |
| 0 | 0 | 0 | $0^*$ |
| 0 | 0 | 0 | $0^*$ |
| 0 | 0 | 0 | $0^*$ |
| 1 | 1 | $1^*$ | 1 |
| 1 | 1 | $1^*$ | 1 |
| 1 | 1 | $1^*$ | 1 |
| 1 | 1 | $1^*$ | 1 |

$\theta = 0; \qquad \alpha = 1$

$(C_0, C_1)$ not independent of $X$

| $X$ | $Y$ | $C_0$ | $C_1$ |
|---|---|---|---|
| 0 | 0 | 0 | $0^*$ |
| 1 | 0 | 0 | $0^*$ |
| 1 | 0 | 0 | $0^*$ |
| 1 | 0 | 0 | $0^*$ |
| 1 | 1 | $1^*$ | 1 |
| 1 | 1 | $1^*$ | 1 |
| 1 | 1 | $1^*$ | 1 |
| 1 | 1 | $1^*$ | 1 |

$\theta = 0, \quad \alpha = 4/7 < 1$

thought experiment

## Potential outcomes

Table 1.1

| | $Y^{a=0}$ | $Y^{a=1}$ |
|---|---|---|
| Rheia | 0 | 1 |
| Kronos | 1 | 0 |
| Demeter | 0 | 0 |
| Hades | 0 | 0 |
| Hestia | 0 | 0 |
| Poseidon | 1 | 0 |
| Hera | 0 | 0 |
| Zeus | 0 | 1 |
| Artemis | 1 | 1 |
| Apollo | 1 | 0 |
| Leto | 0 | 1 |
| Ares | 1 | 1 |
| Athena | 1 | 1 |
| Hephaestus | 0 | 1 |
| Aphrodite | 0 | 1 |
| Cyclope | 0 | 1 |
| Persephone | 1 | 1 |
| Hermes | 1 | 0 |
| Hebe | 1 | 0 |
| Dionysus | 1 | 0 |

## Observed outcomes

Table 1.2

| | $A$ | $Y$ |
|---|---|---|
| Rheia | 0 | 0 |
| Kronos | 0 | 1 |
| Demeter | 0 | 0 |
| Hades | 0 | 0 |
| Hestia | 1 | 0 |
| Poseidon | 1 | 0 |
| Hera | 1 | 0 |
| Zeus | 1 | 1 |
| Artemis | 0 | 1 |
| Apollo | 0 | 1 |
| Leto | 0 | 0 |
| Ares | 1 | 1 |
| Athena | 1 | 1 |
| Hephaestus | 1 | 1 |
| Aphrodite | 1 | 1 |
| Cyclope | 1 | 1 |
| Persephone | 1 | 1 |
| Hermes | 1 | 0 |
| Hebe | 1 | 0 |
| Dionysus | 1 | 0 |

# Three types of causal statements

SM §9.1.2

1. A well-understood evidence-based mechanism, or set of mechanisms, that links a cause to its effect

2. two phenomena are linked by a stable association, whose direction is established and which cannot be explained by mutual dependence on some other allowable variable

3. observed association may be linked to causal effect via counterfactuals if $(C_0, C_0) \perp X$

not usually testable

# Conditional and marginal effects

AoS §16.3

- typically have additional explanatory variables (covariates) $Z$
- causal effect of treatment when $Z = z$

$$\theta_z = \mathrm{E}(C_1 \mid Z = z) - \mathrm{E}(C_0 \mid Z = z)$$

- marginal causal effect

$$\theta = \mathrm{E}_Z\{\mathrm{E}(C_1 \mid Z) - \mathrm{E}(C_0 \mid Z)\}$$

# Example



Table 2.2

| | $L$ | $A$ | $Y$ |
|---|---|---|---|
| Rheia | 0 | 0 | 0 |
| Kronos | 0 | 0 | 1 |
| Demeter | 0 | 0 | 0 |
| Hades | 0 | 0 | 0 |
| Hestia | 0 | 1 | 0 |
| Poseidon | 0 | 1 | 0 |
| Hera | 0 | 1 | 0 |
| Zeus | 0 | 1 | 1 |
| Artemis | 1 | 0 | 1 |
| Apollo | 1 | 0 | 1 |
| Leto | 1 | 0 | 0 |
| Ares | 1 | 1 | 1 |
| Athena | 1 | 1 | 1 |
| Hephaestus | 1 | 1 | 1 |
| Aphrodite | 1 | 1 | 1 |
| Cyclope | 1 | 1 | 1 |
| Persephone | 1 | 1 | 1 |
| Hermes | 1 | 1 | 0 |
| Hebe | 1 | 1 | 0 |
| Dionysus | 1 | 1 | 0 |

$\theta_{L=0}$

$\theta_{L=1}$

$L = 1$ critical condition

$L = 0$ stable condition

conditional randomization

# Causal regression function

AoS §16.2

- continuous "treatment" variable $X \in \mathbb{R}$
- counterfactual outcome $(C_0, C_1) \to$ counterfactual function $C(x)$
- observed response $Y = C(X)$ consistency
- causal regression function $\theta(x) = \mathrm{E}\{C(x)\}$
- association regression function $r(x) = \mathrm{E}(Y \mid X)$

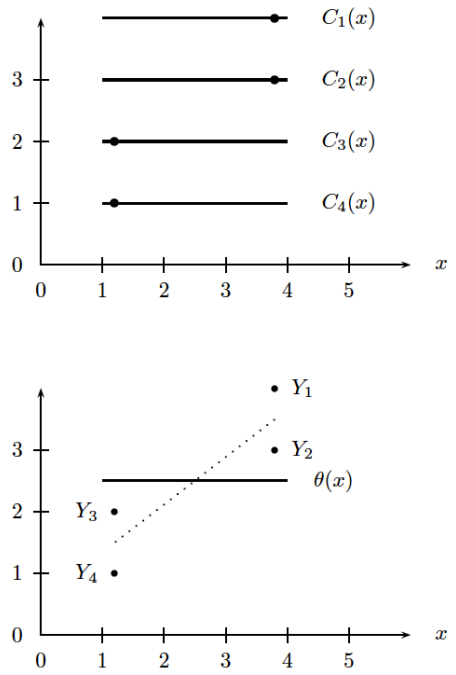



FIGURE 16.2. The top plot shows the counterfactual function $C(x)$ for four subjects. The dots represent their $X$ values. Since $C_i(x)$ is constant over $x$ for all $i$, there is no causal effect. Changing the dose will not change anyone's outcome. The lower plot shows the causal regression function $\theta(x) = (C_1(x) + C_2(x) + C_3(x) + C_4(x))/4$. The four dots represent the observed data points $Y_1 = C_1(X_1)$, $Y_2 = C_2(X_2)$, $Y_3 = C_3(X_3)$, $Y_4 = C_4(X_4)$. The dotted line represents the regression $r(x) = \mathbb{E}(Y|X = x)$. There is no causal effect since $C_i(x)$ is constant for all $i$. But there is an association since the regression curve $r(x)$ is not constant.

# No unmeasured confounding

AoS §16.3

- in observational studies treatment is not randomly assigned $\implies \theta(x) \neq r(x)$
- group subjects based on additional confounding variables
- No unmeasured confounding:

$$\{C(x); x \in \mathcal{X}\} \perp X \mid Z$$

- under the assumption of no unmeasured confounding, the causal regression function

typo in (16.7)

$$\theta(x) = \int \mathrm{E}(Y \mid X = x, Z = z) dF_Z(z)$$

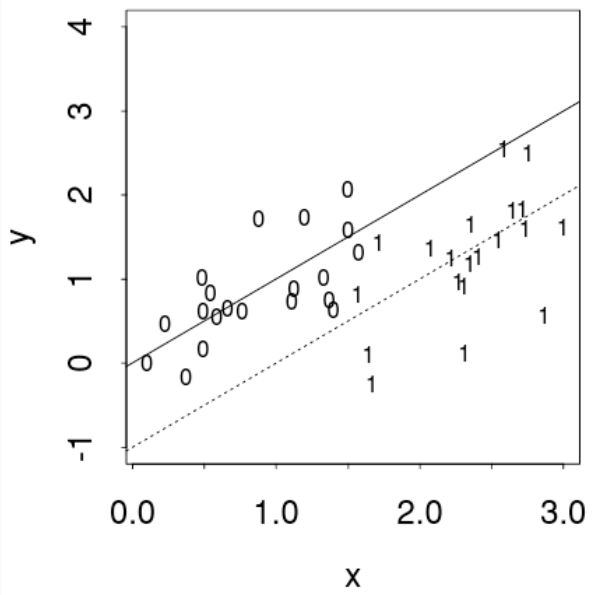
can be estimated by the association function

$$\hat{\theta}(x) = \frac{1}{n} \sum_{i=1}^{n} \hat{r}(x, Z_i) \qquad = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 \bar{Z}_n$$

causal reg function $\equiv$ adjusted treatment effect

# No unmeasured confounding

SM §9.1.1



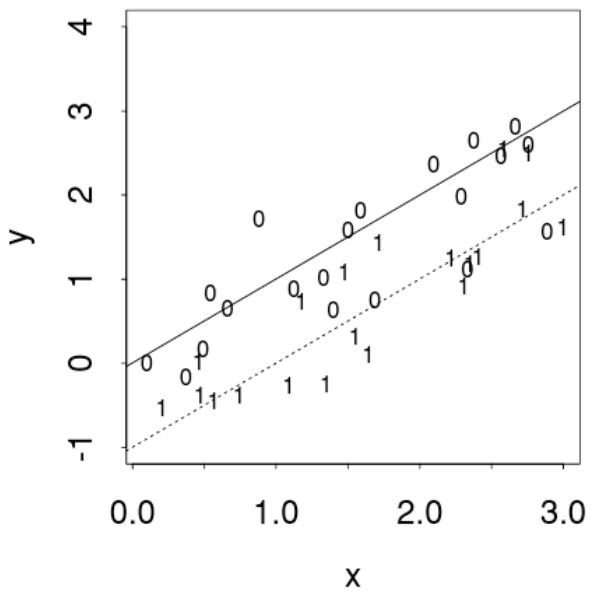




**Figure 9.2** Simulated results from experiments to compare the effect of a treatment $T$ on a response $Y$ that varies with a covariate $X$. The lines show the mean response for $T = 0$ (solid) and $T = 1$ (dots). Left: the effect of $T$ is confounded with dependence on $X$. Right: the experiment is balanced, with random allocation of $T$ dependent on $X$.

# No unmeasured confounding

SM §9.1.1

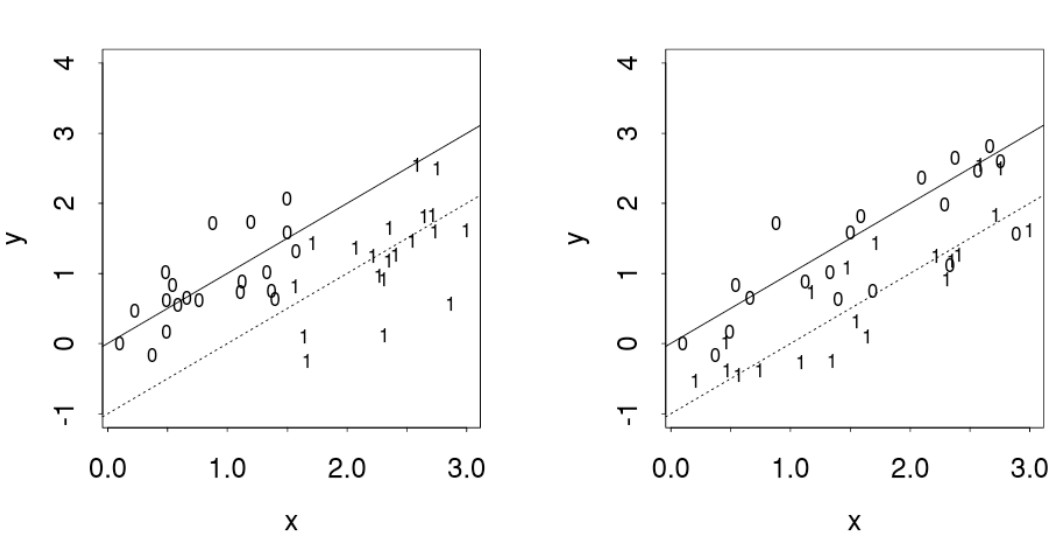



**Figure 9.2** Simulated results from experiments to compare the effect of a treatment $T$ on a response $Y$ that varies with a covariate $X$. The lines show the mean response for $T = 0$ (solid) and $T = 1$ (dots). Left: the effect of $T$ is confounded with dependence on $X$. Right: the experiment is balanced, with random allocation of $T$ dependent on $X$.

$C_1(x) - C_0(x) \equiv 1$

Left: $\bar{y}_1 - \bar{y}_0 = 0.2 \pm 0.3$

Right: $\bar{y}_1 - \bar{y}_0 = -1.2 \pm 0.3$

adjust for covariate: $y = \beta_0 + \beta_1 x + \delta t + \epsilon$

Left: $\hat{\delta} = -0.7 \pm 0.3$ Right: $\hat{\delta} = -1.25 \pm 0.16$

right randomized within pairs; matched on $x$

# Causality and observational data

CD 9.2.4

"Bradford-Hill guidelines" Evidence that an observed association is causal is strengthened if:

- the association is strong
- the association is found consistently over a number of independent studies
- the association is specific to the outcome studied
- the observation of a potential cause occurs earlier in time than the outcome
- there is a dose-response relationship
- there is subject-matter theory that makes a causal effect plausible
- the association is based on a suitable natural experiment

see also AoS §16.3

260 16. Causal Inference

| | $Y=1$ | $Y=0$ | $Y=1$ | $Y=0$ |
|---|---|---|---|---|
| $X=1$ | .1500 | .2250 | .1000 | .0250 |
| $X=0$ | .0375 | .0875 | .2625 | .1125 |
| | $Z=1$ (men) | | $Z=0$ (women) | |

confusion of causal effect with association

The marginal distribution for $(X, Y)$ is

| | $Y=1$ | $Y=0$ | |
|---|---|---|---|
| $X=1$ | .25 | .25 | .50 |
| $X=0$ | .30 | .20 | .50 |
| | .55 | .45 | 1 |

From these tables we find that,

$$\begin{aligned}
\mathbb{P}(Y=1|X=1) - \mathbb{P}(Y=1|X=0) &= -0.1 \\
\mathbb{P}(Y=1|X=1, Z=1) - \mathbb{P}(Y=1|X=0, Z=1) &= 0.1 \\
\mathbb{P}(Y=1|X=1, Z=0) - \mathbb{P}(Y=1|X=0, Z=0) &= 0.1.
\end{aligned}$$

To summarize, we *seem* to have the following information:

| Mathematical Statement | English Statement? |
|---|---|

# Directed graphs

AoS 17; HR 6; SM 6.2

- graphs can be useful for clarifying dependence relations among random variables

SM Markov random fields

- a Directed Acyclic Graph has random variables on the vertices and edges joining random variables

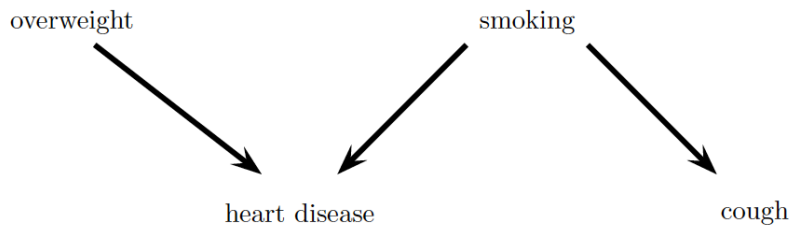



FIGURE 17.2. DAG for Example 17.4.

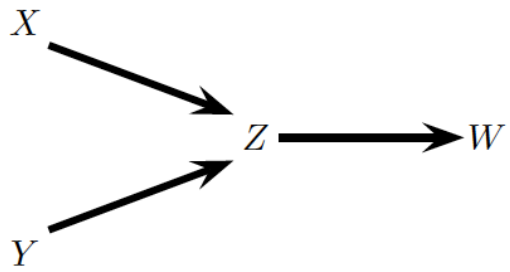



FIGURE 17.3. Another DAG.

# Directed graphs and causality

AoS 17.8

- variables at parent nodes are potential causes for responses at child nodes
- directed graphs often helpful adjunct to modelling with baseline variables, intermediate responses, and outcome variables of interest
- much hard to study the full joint distribution than the usual supervised learning approaches
- DAGs can be used to represent confounders

276 17. Directed Graphs and Conditional Independence

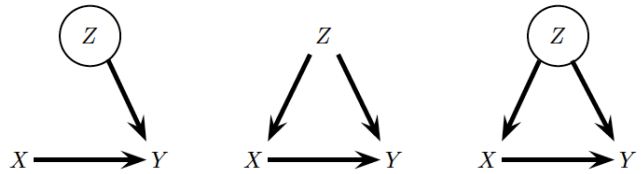



FIGURE 17.11. Randomized study; Observational study with measured confounders; Observational study with unmeasured confounders. The circled variables are unobserved.

276 17. Directed Graphs and Conditional Independence

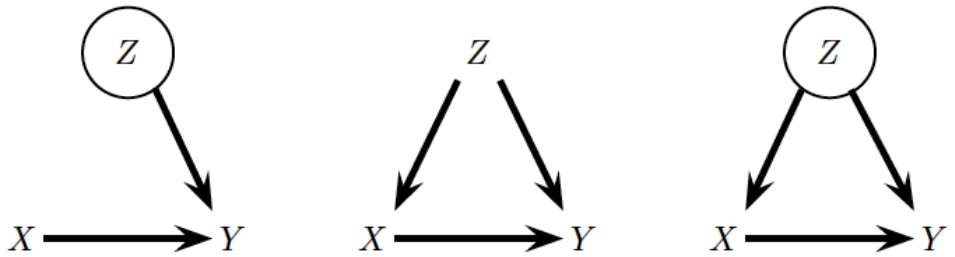



FIGURE 17.11. Randomized study; Observational study with measured confounders; Observational study with unmeasured confounders. The circled variables are unobserved.

randomized study observational study $\mathrm{E}(Y \mid x) = \int \mathrm{E}(Y \mid X, Z = z) dF_Z(z)$

unobserved confounder: $\theta \neq \alpha$