Bootstrap Methods and their Application

Anthony Davison

A short course based on the book
'Bootstrap Methods and their Application',
by A. C. Davison and D. V. Hinkley

Summary

- Bootstrap: simulation methods for frequentist inference.
- Useful when
  - standard assumptions invalid (n small, data not normal, $\ldots$);
  - standard problem has non-standard twist;
  - complex problem has no (reliable) theory;
  - or (almost) anywhere else.
- Aim to describe
  - basic ideas;
  - confidence intervals;
  - tests;
  - some approaches for regression

- Motivation
- Basic notions
- Confidence intervals
- Several samples
- Variance estimation
- Tests
- Regression

AIDS data

- UK AIDS diagnoses 1988–1992.
- Reporting delay up to several years!
- Problem: predict state of epidemic at end 1992, with realistic statement of uncertainty.
- Simple model: number of reports in row j and column k Poisson, mean

$$\mu_{jk} = \exp(\alpha_j + \beta_k).$$

- Unreported diagnoses in period j Poisson, mean

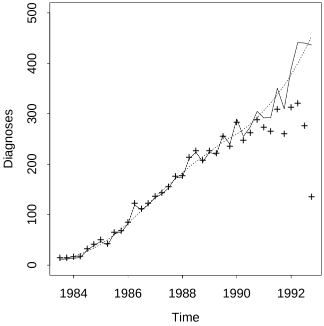
$$\sum_{k \text{ unobs}} \mu_{jk} = \exp(\alpha_j) \sum_{k \text{ unobs}} \exp(\beta_k).$$

- Estimate total unreported diagnoses from period j by replacing α_j and β_k by MLEs.
  - How reliable are these predictions?
  - How sensible is the Poisson model?

Diagnosis period		Reporting-delay interval (quarters):									Total reports to end
Year	Quarter	0†	1	2	3	4	5	6	⋯	≥14	of 1992
1988	1	31	80	16	9	3	2	8	⋯	6	174
	2	26	99	27	9	8	11	3	⋯	3	211
	3	31	95	35	13	18	4	6	⋯	3	224
	4	36	77	20	26	11	3	8	⋯	2	205
1989	1	32	92	32	10	12	19	12	⋯	2	224
	2	15	92	14	27	22	21	12	⋯	1	219
	3	34	104	29	31	18	8	6	⋯		253
	4	38	101	34	18	9	15	6	⋯		233
1990	1	31	124	47	24	11	15	8	⋯		281
	2	32	132	36	10	9	7	6	⋯		245
	3	49	107	51	17	15	8	9	⋯		260
	4	44	153	41	16	11	6	5	⋯		285
1991	1	41	137	29	33	7	11	6	⋯		271
	2	56	124	39	14	12	7	10			263
	3	53	175	35	17	13	11				306
	4	63	135	24	23	12					258
1992	1	71	161	48	25						310
	2	95	178	39							318
	3	76	181								273
	4	67									133

AIDS data

- Data (+), fits of simple model (solid), complex model (dots)
- Variance formulae could be derived — painful! but useful?
- Effects of overdispersion, complex model, ...?



Goal

Find reliable assessment of uncertainty when

- estimator complex
- data complex
- sample size small
- model non-standard

- Motivation
- **Basic notions**
- Confidence intervals
- Several samples
- Variance estimation
- Tests
- Regression

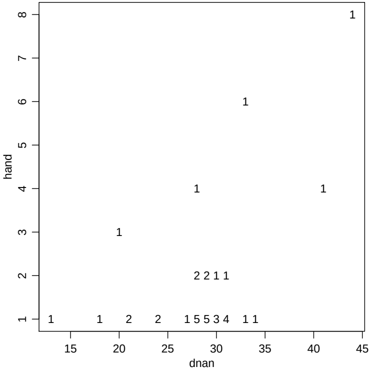
Handedness data

Table: Data from a study of handedness; `hand` is an integer measure of handedness, and `dnan` a genetic measure. Data due to Dr Gordon Claridge, University of Oxford.

	dnan	hand		dnan	hand		dnan	hand		dnan	hand
1	13	1	11	28	1	21	29	2	31	31	1
2	18	1	12	28	2	22	29	1	32	31	2
3	20	3	13	28	1	23	29	1	33	33	6
4	21	1	14	28	4	24	30	1	34	33	1
5	21	1	15	28	1	25	30	1	35	34	1
6	24	1	16	28	1	26	30	2	36	41	4
7	24	1	17	29	1	27	30	1	37	44	8
8	27	1	18	29	1	28	31	1			
9	28	1	19	29	1	29	31	1			
10	28	2	20	29	2	30	31	1			

Handedness data

Figure: Scatter plot of handedness data. The numbers show the multiplicities of the observations.



Handedness data

- Is there dependence between `dnan` and `hand` for these $n = 37$ individuals?
- Sample product-moment correlation coefficient is $\widehat{\theta} = 0.509$.
- Standard confidence interval (based on bivariate normal population) gives 95% CI $(0.221,\ 0.715)$.
- Data not bivariate normal!
- What is the status of the interval? Can we do better?

Frequentist inference

- Estimator $\widehat{\theta}$ for unknown parameter θ.
- Statistical model: data $y_1, \ldots, y_n \overset{\text{iid}}{\sim} F$, unknown
- Handedness data
  - $y = (\texttt{dnan}, \texttt{hand})$
  - F puts probability mass on subset of $\mathbb{R}^2$
  - $\widehat{\theta}$ is correlation coefficient
- Key issue: what is variability of $\widehat{\theta}$ when samples are repeatedly taken from F?
- Imagine F known — could answer question by
  - analytical (mathematical) calculation
  - simulation

Simulation with F known

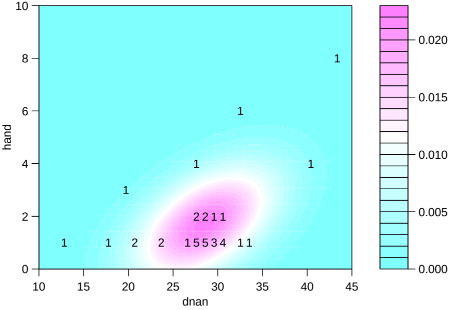
- For $r = 1, \ldots, R$:
  - generate random sample $y_1^*, \ldots, y_n^* \overset{\text{iid}}{\sim} F$;
  - compute $\widehat{\theta}_r$ using $y_1^*, \ldots, y_n^*$;
- Output after R iterations:

$$\widehat{\theta}_1^*, \widehat{\theta}_2^*, \ldots, \widehat{\theta}_R^*$$

- Use $\widehat{\theta}_1^*, \widehat{\theta}_2^*, \ldots, \widehat{\theta}_R^*$ to estimate sampling distribution of $\widehat{\theta}$ (histogram, density estimate, ...)
- If $R \to \infty$, then get perfect match to theoretical calculation (if available): Monte Carlo error disappears completely
- In practice R is finite, so some error remains

Handedness data: Fitted bivariate normal model

Figure: Contours of bivariate normal distribution fitted to handedness data; parameter estimates are $\widehat{\mu}_1 = 28.5$, $\widehat{\mu}_2 = 1.7$, $\widehat{\sigma}_1 = 5.4$, $\widehat{\sigma}_2 = 1.5$, $\widehat{\rho} = 0.509$. The data are also shown.



Handedness data: Parametric bootstrap samples

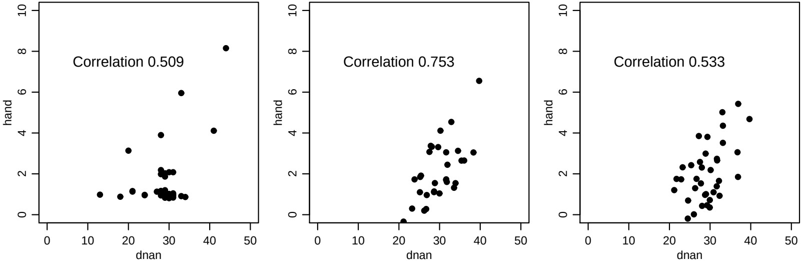
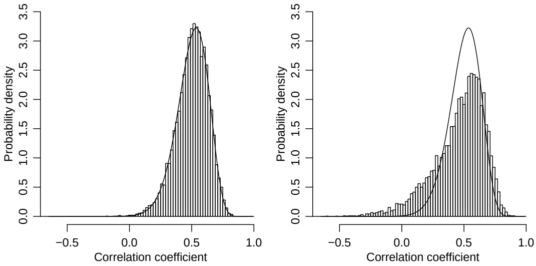


Figure: Left: original data, with jittered vertical values. Centre and right: two samples generated from the fitted bivariate normal distribution.

Handedness data: Correlation coefficient

Figure: Bootstrap distributions with $R = 10000$. Left: simulation from fitted bivariate normal distribution. Right: simulation from the data by bootstrap resampling. The lines show the theoretical probability density function of the correlation coefficient under sampling from a fitted bivariate normal distribution.



F unknown

- Replace unknown F by estimate $\widehat{F}$ obtained
  - parametrically — e.g. maximum likelihood or robust fit of distribution $F(y) = F(y; \psi)$ (normal, exponential, bivariate normal, ...)
  - nonparametrically — using empirical distribution function (EDF) of original data $y_1, \ldots, y_n$, which puts mass $1/n$ on each of the y_j
- Algorithm: For $r = 1, \ldots, R$:
  - generate random sample $y_1^*, \ldots, y_n^* \overset{\text{iid}}{\sim} \widehat{F}$;
  - compute $\widehat{\theta}_r$ using $y_1^*, \ldots, y_n^*$;
- Output after R iterations:

$$\widehat{\theta}_1^*, \widehat{\theta}_2^*, \ldots, \widehat{\theta}_R^*$$

Nonparametric bootstrap

- Bootstrap (re)sample $y_1^*, \ldots, y_n^* \stackrel{\text{iid}}{\sim} \widehat{F}$, where $\widehat{F}$ is EDF of $y_1, \ldots, y_n$
  - Repetitions will occur!
- Compute bootstrap $\widehat{\theta}^*$ using $y_1^*, \ldots, y_n^*$.
- For handedness data take $n = 37$ pairs $y^* = (\texttt{dnan}, \texttt{hand})^*$ with equal probabilities $1/37$ and replacement from original pairs $(\texttt{dnan}, \texttt{hand})$
- Repeat this R times, to get $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_R^*$
- See picture
- Results quite different from parametric simulation — why?

Handedness data: Bootstrap samples

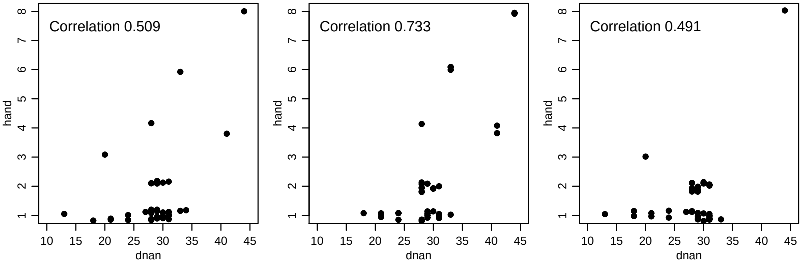


Figure: Left: original data, with jittered vertical values. Centre and right: two bootstrap samples, with jittered vertical values.

Using the $\widehat{\theta}^*$

- *Bootstrap replicates* $\widehat{\theta}^*_r$ used to estimate properties of $\widehat{\theta}$.
- Write $\theta = \theta(F)$ to emphasize dependence on F
- Bias of $\widehat{\theta}$ as estimator of θ is

$$\beta(F) = \mathrm{E}(\widehat{\theta} \mid y_1, \ldots, y_n \stackrel{\text{iid}}{\sim} F) - \theta(F)$$

estimated by replacing unknown F by known estimate $\widehat{F}$:

$$\beta(\widehat{F}) = \mathrm{E}(\widehat{\theta} \mid y_1, \ldots, y_n \stackrel{\text{iid}}{\sim} \widehat{F}) - \theta(\widehat{F})$$

- Replace theoretical expectation E() by empirical average:

$$\beta(\widehat{F}) \approx b = \overline{\widehat{\theta}^*} - \widehat{\theta} = R^{-1} \sum_{r=1}^{R} \widehat{\theta}^*_r - \widehat{\theta}$$

- Estimate variance $\nu(F) = \mathrm{var}(\widehat{\theta} \mid F)$ by

$$v = \frac{1}{R-1} \sum_{r=1}^{R} \left(\widehat{\theta}_r^* - \overline{\widehat{\theta}^*} \right)^2$$

- Estimate quantiles of $\widehat{\theta}$ by taking empirical quantiles of

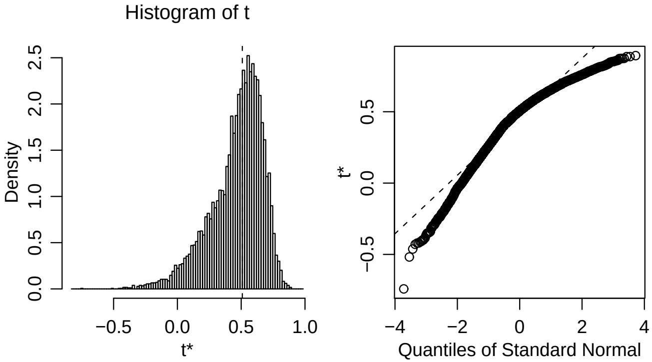
$$\widehat{\theta}_1^*, \ldots, \widehat{\theta}_R^*$$

- For handedness data, 10,000 replicates shown earlier give

$$b = -0.046, \quad v = 0.043 = 0.205^2$$

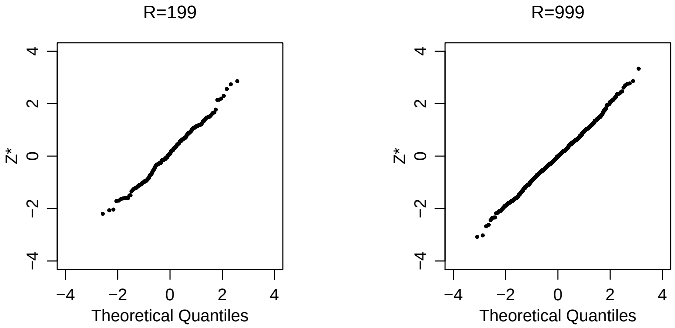
Handedness data

Figure: Summaries of the $\widehat{\theta}^*$. Left: histogram, with vertical line showing $\widehat{\theta}$. Right: normal Q–Q plot of $\widehat{\theta}^*$.



How many bootstraps?

- Must estimate moments and quantiles of $\widehat{\theta}$ and derived quantities. Nowadays often feasible to take $R \geq 5000$
- Need $R \geq 100$ to estimate bias, variance, etc.
- Need $R \gg 100$, prefer $R \geq 1000$ to estimate quantiles needed for 95% confidence intervals



Key points

- *Estimator is algorithm*
  - applied to original data $y_1, \ldots, y_n$ gives original $\widehat{\theta}$
  - applied to simulated data $y_1^*, \ldots, y_n^*$ gives $\widehat{\theta}^*$
  - $\widehat{\theta}$ can be of (almost) any complexity
  - for more sophisticated ideas (later) to work, $\widehat{\theta}$ must often be smooth function of data
- *Sample is used to estimate F*
  - $\widehat{F} \approx F$ — heroic assumption
- *Simulation replaces theoretical calculation*
  - removes need for mathematical skill
  - does not remove need for thought
  - check code *very* carefully — garbage in, garbage out!
- *Two sources of error*
  - statistical ($\widehat{F} \neq F$) — reduce by thought
  - simulation ($R \neq \infty$) — reduce by taking R large (enough)

- Motivation
- Basic notions
- Confidence intervals
- Several samples
- Variance estimation
- Tests
- Regression

Normal confidence intervals

- If $\widehat{\theta}$ approximately normal, then $\widehat{\theta} \dot{\sim} N(\theta+\beta, \nu)$, where $\widehat{\theta}$ has bias $\beta = \beta(F)$ and variance $\nu = \nu(F)$
- If β, ν known, $(1-2\alpha)$ confidence interval for θ would be (**D1**)

$$\widehat{\theta} - \beta \pm z_\alpha \nu^{1/2},$$

where $\Phi(z_\alpha) = \alpha.$

- Replace β, ν by estimates:

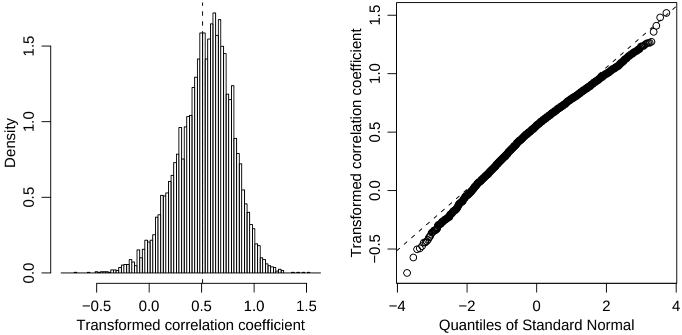
$$\begin{aligned} \beta(F) &\doteq \beta(\widehat{F}) \doteq b = \overline{\widehat{\theta}^*} - \widehat{\theta} \\ \nu(F) &\doteq \nu(\widehat{F}) \doteq v = (R-1)^{-1} \sum_r (\widehat{\theta}^*_r - \overline{\widehat{\theta}^*})^2, \end{aligned}$$

giving $(1-2\alpha)$ interval $\widehat{\theta} - b \pm z_\alpha v^{1/2}$.

- Handedness data: $R = 10,000$, $b = -0.046$, $v = 0.205^2$, $\alpha = 0.025$, $z_\alpha = -1.96$, so 95% CI is $(0.147, 0.963)$

Normal confidence intervals

- Normal approximation reliable? Transformation needed?
- Here are plots for $\widehat{\psi} = \frac{1}{2}\log\{(1+\widehat{\theta})/(1-\widehat{\theta})\}$:



Normal confidence intervals

- Correlation coefficient: try Fisher's z transformation:

$$\widehat{\psi} = \psi(\widehat{\theta}) = \tfrac{1}{2}\log\{(1+\widehat{\theta})/(1-\widehat{\theta})\}$$

for which compute

$$b_\psi = R^{-1}\sum_{r=1}^{R}\widehat{\psi}_r^* - \widehat{\psi}, \quad v_\psi = \frac{1}{R-1}\sum_{r=1}^{R}\left(\widehat{\psi}_r^* - \overline{\widehat{\psi}^*}\right)^2,$$

- $(1-2\alpha)$ confidence interval for θ is

$$\psi^{-1}\left\{\widehat{\psi} - b_\psi - z_{1-\alpha}v_\psi^{1/2}\right\}, \quad \psi^{-1}\left\{\widehat{\psi} - b_\psi - z_\alpha v_\psi^{1/2}\right\}$$

- For handedness data, get $(0.074, 0.804)$
- But how do we choose a transformation in general?

Pivots

- Hope properties of $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_R^*$ mimic effect of sampling from original model.
- Amounts to faith in 'substitution principle': may replace unknown F with known $\widehat{F}$ — false in general, but often more nearly true for pivots.
- *Pivot* is combination of data and parameter whose distribution is independent of underlying model.
- Canonical example: $Y_1, \ldots, Y_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$. Then

$$Z = \frac{\overline{Y} - \mu}{(S^2/n)^{1/2}} \quad \sim \quad t_{n-1},$$

  for all μ, σ^2 — so independent of the underlying distribution, provided this is normal
- Exact pivot generally unavailable in nonparametric case.

Studentized statistic

- Idea: generalize Student t statistic to bootstrap setting
- Requires variance V for $\widehat{\theta}$ computed from $y_1, \ldots, y_n$
- Analogue of Student t statistic:

$$Z = \frac{\widehat{\theta} - \theta}{V^{1/2}}$$

- If the quantiles z_α of Z known, then

$$\Pr\left(z_\alpha \le Z \le z_{1-\alpha}\right) = \Pr\left(z_\alpha \le \frac{\widehat{\theta} - \theta}{V^{1/2}} \le z_{1-\alpha}\right) = 1 - 2\alpha$$

(z_α no longer denotes a normal quantile!) implies that

$$\Pr\left(\widehat{\theta} - V^{1/2} z_{1-\alpha} \le \theta \le \widehat{\theta} - V^{1/2} z_\alpha\right) = 1 - 2\alpha$$

so $(1 - 2\alpha)$ confidence interval is $(\widehat{\theta} - V^{1/2} z_{1-\alpha}, \widehat{\theta} - V^{1/2} z_\alpha)$

- Bootstrap sample gives $(\widehat{\theta}^*, V^*)$ and hence

$$Z^* = \frac{\widehat{\theta}^* - \widehat{\theta}}{V^{*1/2}}$$

- R bootstrap copies of $(\widehat{\theta}, V)$:

$$(\widehat{\theta}_1^*, V_1^*), \quad (\widehat{\theta}_2^*, V_2^*), \quad \ldots, \quad (\widehat{\theta}_R^*, V_R^*)$$

  and corresponding

$$z_1^* = \frac{\widehat{\theta}_1^* - \widehat{\theta}}{V_1^{*1/2}}, \quad z_2^* = \frac{\widehat{\theta}_2^* - \widehat{\theta}}{V_2^{*1/2}}, \quad \ldots, \quad z_R^* = \frac{\widehat{\theta}_R^* - \widehat{\theta}}{V_R^{*1/2}}.$$

- Use $z_1^*, \ldots, z_R^*$ to estimate distribution of Z — for example, order statistics $z_{(1)}^* < \cdots < z_{(R)}^*$ used to estimate quantiles
- Get $(1 - 2\alpha)$ confidence interval

$$\widehat{\theta} - V^{1/2} z^*_{((1-\alpha)(R+1))}, \quad \widehat{\theta} - V^{1/2} z^*_{(\alpha(R+1))}$$

Why Studentize?

- Studentize, so $Z \xrightarrow{D} N(0,1)$ as $n \to \infty$. Edgeworth series:

$$\Pr(Z \le z \mid F) = \Phi(z) + n^{-1/2}a(z)\phi(z) + O(n^{-1});$$

  $a(\cdot)$ even quadratic polynomial.

- Corresponding expansion for Z^* is

$$\Pr(Z^* \le z \mid \widehat{F}) = \Phi(z) + n^{-1/2}\widehat{a}(z)\phi(z) + O_p(n^{-1}).$$

  Typically $\widehat{a}(z) = a(z) + O_p(n^{-1/2})$, so

$$\Pr(Z^* \le z \mid \widehat{F}) - \Pr(Z \le z \mid F) = O_p(n^{-1}).$$

- If don't studentize, $Z = (\widehat{\theta} - \theta) \xrightarrow{D} N(0, \nu)$. Then

  $$\Pr(Z \le z \mid F) = \Phi\left(\frac{z}{\nu^{1/2}}\right) + n^{-1/2} a'\left(\frac{z}{\nu^{1/2}}\right)\phi\left(\frac{z}{\nu^{1/2}}\right) + O(n^{-1})$$

  and

  $$\Pr(Z^* \le z \mid \widehat{F}) = \Phi\left(\frac{z}{\widehat{\nu}^{1/2}}\right) + n^{-1/2} \widehat{a}'\left(\frac{z}{\widehat{\nu}^{1/2}}\right)\phi\left(\frac{z}{\widehat{\nu}^{1/2}}\right) + O_p(n^{-1}).$$

  Typically $\widehat{\nu} = \nu + O_p(n^{-1/2})$, giving

  $$\Pr(Z^* \le z \mid \widehat{F}) - \Pr(Z \le z \mid F) = O_p(n^{-1/2}).$$

- Thus use of Studentized Z reduces error from $O_p(n^{-1/2})$ to $O_p(n^{-1})$ — better than using large-sample asymptotics, for which error is usually $O_p(n^{-1/2})$.

Other confidence intervals

- Problem for studentized intervals: must obtain V, intervals not scale-invariant
- Simpler approaches:
  - Basic bootstrap interval: treat $\widehat{\theta}-\theta$ as pivot, get

$$\widehat{\theta}-(\widehat{\theta}^*_{((R+1)(1-\alpha))}-\widehat{\theta}), \quad \widehat{\theta}-(\widehat{\theta}^*_{((R+1)\alpha)}-\widehat{\theta}).$$

  - Percentile interval: use empirical quantiles of $\widehat{\theta}^*_1,\ldots,\widehat{\theta}^*_R$:

$$\widehat{\theta}^*_{((R+1)\alpha)}, \quad \widehat{\theta}^*_{((R+1)(1-\alpha))}.$$

- Improved percentile intervals (BC_a, ABC, ...)
  - Replace percentile interval with

$$\widehat{\theta}^*_{((R+1)\alpha')}, \quad \widehat{\theta}^*_{((R+1)(1-\alpha''))},$$

    where α', α'' chosen to improve properties.
  - Scale-invariant.

Handedness data

- 95% confidence intervals for correlation coefficient θ, $R = 10,000$:

Normal	0.147	0.963
Percentile	−0.047	0.758
Basic	0.262	1.043
BC_a ($\alpha' = 0.0485, \alpha'' = 0.0085$)	0.053	0.792
Student	0.030	1.206
Basic (transformed)	0.131	0.824
Student (transformed)	−0.277	0.868

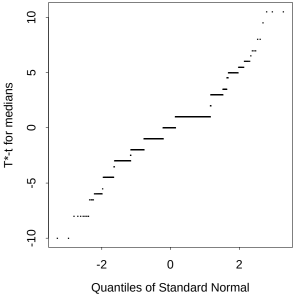
- Transformation is essential here!

General comparison

- Bootstrap confidence intervals usually too short — leads to under-coverage
- Normal and basic intervals depend on scale.
- Percentile interval often too short but is scale-invariant.
- Studentized intervals give best coverage overall, but
  - depend on scale, can be sensitive to V
  - length can be very variable
  - best on transformed scale, where V is approximately constant
- Improved percentile intervals have same error in principle as studentized intervals, but often shorter — so lower coverage

Caution

- Edgeworth theory OK for smooth statistics — beware rough statistics: must check output.
- Bootstrap of median theoretically OK, but very sensitive to sample values in practice.
- Role for smoothing?



Key points

- Numerous procedures available for ‘automatic’ construction of confidence intervals
- Computer does the work
- Need $R \geq 1000$ in most cases
- Generally such intervals are a bit too short
- Must examine output to check if assumptions (e.g. smoothness of statistic) satisfied
- May need variance estimate V — see later

- Motivation
- Basic notions
- Confidence intervals
- **Several samples**
- Variance estimation
- Tests
- Regression

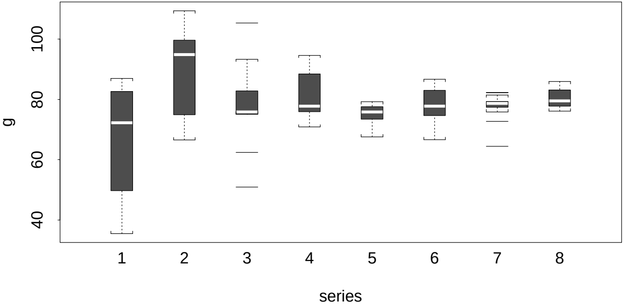
Gravity data

Table: Measurements of the acceleration due to gravity, g, given as deviations from 980,000 $\times 10^{-3}$ cms^{-2}, in units of cms$^{-2} \times 10^{-3}$.

Series							
1	2	3	4	5	6	7	8
76	87	105	95	76	78	82	84
82	95	83	90	76	78	79	86
83	98	76	76	78	78	81	85
54	100	75	76	79	86	79	82
35	109	51	87	72	87	77	77
46	109	76	79	68	81	79	76
87	100	93	77	75	73	79	77
68	81	75	71	78	67	78	80
	75	62			75	79	83
	68				82	82	81
	67				83	76	78
						73	78
						64	78

Gravity data

Figure: Gravity series boxplots, showing a reduction in variance, a shift in location, and possible outliers.



Gravity data

- Eight series of measurements of gravitational acceleration g made May 1934 – July 1935 in Washington DC
- Data are deviations from 9.8 m/s^2 in units of 10^{-3} cm/s^2
- Goal: Estimate g and provide confidence interval
- Weighted combination of series averages and its variance estimate

$$\widehat{\theta} = \frac{\sum_{i=1}^{8} \overline{y}_i \times n_i/s_i^2}{\sum_{i=1}^{8} n_i/s_i^2}, \quad V = \left(\sum_{i=1}^{8} n_i/s_i^2\right)^{-1},$$

giving

$$\widehat{\theta} = 78.54, \quad V = 0.59^2$$

and 95% confidence interval of $\widehat{\theta} \pm 1.96 V^{1/2} = (77.5, 79.8)$

Gravity data: Bootstrap

- Apply stratified (re)sampling to series, taking each series as a separate stratum. Compute $\widehat{\theta}^*$, V^* for simulated data
- Confidence interval based on

$$Z^* = \frac{\widehat{\theta}^* - \widehat{\theta}}{V^{*1/2}},$$

whose distribution is approximated by simulations

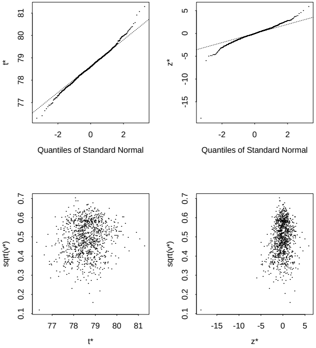
$$z_1^* = \frac{\widehat{\theta}_1^* - \widehat{\theta}}{V_1^{1/2}}, \ldots, z_R^* = \frac{\widehat{\theta}_R^* - \widehat{\theta}}{V_R^{1/2}},$$

giving

$$(\widehat{\theta} - V^{1/2} z^*_{((R+1)(1-\alpha))}, \widehat{\theta} - V^{1/2} z^*_{((R+1)\alpha)})$$

- For 95% limits set $\alpha = 0.025$, so with $R = 999$ use $z^*_{(25)}, z^*_{(975)}$, giving interval $(77.1, 80.3)$.

Figure: Summary plots for 999 nonparametric bootstrap simulations. Top: normal probability plots of t^* and $z^* = (t^* - t)/v^{*1/2}$. Line on the top left has intercept t and slope $v^{1/2}$, line on the top right has intercept zero and unit slope. Bottom: the smallest t_r^* also has the smallest v^*, leading to an outlying value of z^*.



Key points

- For several independent samples, implement bootstrap by stratified sampling independently from each
- Same basic ideas apply for confidence intervals

- Motivation
- Basic notions
- Confidence intervals
- Several samples
- Variance estimation
- Tests
- Regression

Variance estimation

- Variance estimate V needed for certain types of confidence interval (esp. studentized bootstrap)
- Ways to compute this:
  - double bootstrap
  - delta method
  - nonparametric delta method
  - jackknife

Double bootstrap

- Bootstrap sample $y_1^*, \ldots, y_n^*$ and corresponding estimate $\widehat{\theta}^*$
- Take Q second-level bootstrap samples $y_1^{**}, \ldots, y_n^{**}$ from $y_1^*, \ldots, y_n^*$, giving corresponding bootstrap estimates $\widehat{\theta}_1^{**}, \ldots, \widehat{\theta}_Q^{**}$,
- Compute variance estimate V as sample variance of $\widehat{\theta}_1^{**}, \ldots, \widehat{\theta}_Q^{**}$
- Requires total $R(Q+1)$ resamples, so could be expensive
- Often reasonable to take $Q \doteq 50$ for variance estimation, so need $O(50 \times 1000)$ resamples — nowadays not infeasible

Delta method

- Computation of variance formulae for functions of averages and other estimators
- Suppose $\widehat{\psi} = g(\widehat{\theta})$ estimates $\psi = g(\theta)$, and $\widehat{\theta} \dot{\sim} N(\theta, \sigma^2/n)$
- Then provided $g'(\theta) \neq 0$, have (**D2**)

$$\begin{aligned} \mathrm{E}(\widehat{\psi}) &= g(\theta) + O(n^{-1}) \\ \mathrm{var}(\widehat{\psi}) &= \sigma^2 g'(\theta)^2/n + O(n^{-3/2}) \end{aligned}$$

- Then $\mathrm{var}(\widehat{\psi}) \doteq \widehat{\sigma}^2 g'(\widehat{\theta})^2/n = V$
- Example (**D3**): $\widehat{\theta} = \overline{Y}$, $\widehat{\psi} = \log \widehat{\theta}$
- Variance stabilisation (**D4**): if $\mathrm{var}(\widehat{\theta}) \doteq S(\theta)^2/n$, find transformation h such that $\mathrm{var}\{h(\widehat{\theta})\} \doteq$constant
- Extends to multivariate estimators, and to $\widehat{\psi} = g(\widehat{\theta}_1, \ldots, \widehat{\theta}_d)$

Nonparametric delta method

- Write parameter $\theta = t(F)$ as functional of distribution F
- General approximation:

$$V \doteq V_L = \frac{1}{n^2} \sum_{j=1}^{n} L(Y_j; F)^2.$$

- $L(y; F)$ is *influence function value* for θ for observation at y when distribution is F:

$$L(y; F) = \lim_{\varepsilon \to 0} \frac{t\left\{(1-\varepsilon)F + \varepsilon H_y\right\} - t(F)}{\varepsilon},$$

where H_y puts unit mass at y. Close link to robustness.
- Empirical versions of $L(y; F)$ and V_L are

$$l_j = L(y_j; \widehat{F}), \quad v_L = n^{-2} \sum l_j^2,$$

usually obtained by analytic/numerical differentation.

Computation of l_j

- Write $\widehat{\theta}$ in weighted form, differentiate with respect to ε
- Sample average:

$$\widehat{\theta} = \overline{y} = \frac{1}{n}\sum y_j = \sum w_j y_j \Big|_{w_j \equiv 1/n}$$

Change weights:

$$w_j \mapsto \varepsilon + (1-\varepsilon)\tfrac{1}{n}, \quad w_i \mapsto (1-\varepsilon)\tfrac{1}{n}, \quad i \neq j$$

so (**D5**)

$$\overline{y} \mapsto \overline{y}_\varepsilon = \varepsilon y_j + (1-\varepsilon)\overline{y} = \varepsilon(y_j - \overline{y}) + \overline{y},$$

giving $l_j = y_j - \overline{y}$ and $v_L = \frac{1}{n^2}\sum(y_j - \overline{y})^2 = \frac{n-1}{n} n^{-1} s^2$

- Interpretation: l_j is standardized change in $\overline{y}$ when increase mass on y_j

Nonparametric delta method: Ratio

- Population $F(u, x)$ with $y = (u, x)$ and

$$\theta = t(F) = \int x\, dF(u, x) \Big/ \int u\, dF(u, x),$$

sample version is

$$\widehat{\theta} = t(\widehat{F}) = \int x\, d\widehat{F}(u, x) \Big/ \int u\, d\widehat{F}(u, x) = \overline{x}/\overline{u}$$

- Then using chain rule of differentiation (**D6**),

$$l_j = (x_j - \widehat{\theta} u_j)/\overline{u},$$

giving

$$v_L = \frac{1}{n^2} \sum \left(\frac{x_j - \widehat{\theta} u_j}{\overline{u}} \right)^2$$

Handedness data: Correlation coefficient

- Correlation coefficient may be written as a function of averages $\overline{xu} = n^{-1} \sum x_j u_j$ etc.:

$$\widehat{\theta} = \frac{\overline{xu} - \overline{x}\,\overline{u}}{\left\{(\overline{x^2} - \overline{x}^2)(\overline{u^2} - \overline{u}^2)\right\}^{1/2}},$$

  from which empirical influence values l_j can be derived
- In this example (and for others involving only averages), nonparametric delta method is equivalent to delta method
- Get

$$v_L = 0.029$$

  for comparison with $v = 0.043$ obtained by bootstrapping.
- v_L typically underestimates $\mathrm{var}(\widehat{\theta})$ — as here!

Delta methods: Comments

- Can be applied to many complex statistics
- Delta method variances often underestimate true variances:

$$v_L < \mathrm{var}(\widehat{\theta})$$

- Can be applied automatically (numerical differentation) if algorithm for $\widehat{\theta}$ written in weighted form, e.g.

$$\overline{x}_w = \sum w_j x_j, \quad w_j \equiv 1/n \text{ for } \overline{x}$$

and vary weights successively for $j = 1, \ldots, n$, setting

$$w_j = w_i + \varepsilon, \quad i \neq j, \quad \sum w_i = 1$$

for $\varepsilon = 1/(100n)$ and using the definition as derivative

Jackknife

- Approximation to empirical influence values given by

$$l_j \approx l_{\text{jack},j} = (n-1)(\widehat{\theta} - \widehat{\theta}_{-j}),$$

where $\widehat{\theta}_{-j}$ is value of $\widehat{\theta}$ computed from sample

$$y_1, \ldots, y_{j-1}, \quad y_{j+1}, \ldots, y_n$$

- Jackknife bias and variance estimates are

$$b_{\text{jack}} = -\frac{1}{n}\sum l_{\text{jack},j}, \quad v_{\text{jack}} = \frac{1}{n(n-1)}\left(\sum l_{\text{jack},j}^2 - n b_{\text{jack}}^2\right)$$

- Requires $n+1$ calculations of $\widehat{\theta}$
- Corresponds to numerical differentiation of $\widehat{\theta}$, with $\varepsilon = -1/(n-1)$

Key points

- Several methods available for estimation of variances
- Needed for some types of confidence interval
- Most general method is double bootstrap: can be expensive
- Delta methods rely on linear expansion, can be applied numerically or analytically
- Jackknife gives approximation to delta method, can fail for rough statistics

- Motivation
- Basic notions
- Confidence intervals
- Several samples
- Variance estimation
- Tests
- Regression

Ingredients

- Ingredients for testing problems:
  - data $y_1, \ldots, y_n$;
  - model M_0 to be tested;
  - test statistic $t = t(y_1, \ldots, y_n)$, with large values giving evidence against M_0, and observed value t_{obs}
- P-value, $p_{\text{obs}} = \Pr(T \geq t_{\text{obs}} \mid M_0)$ measures evidence against M_0 — small p_{obs} indicates evidence against M_0.
- Difficulties:
  - p_{obs} may depend upon 'nuisance' parameters, those of M_0;
  - p_{obs} often hard to calculate.

Examples

- Balsam-fir seedlings in 5×5 quadrats — Poisson sample?

0	1	2	3	4	3	4	2	2	1
0	2	0	2	4	2	3	3	4	2
1	1	1	1	4	1	5	2	2	3
4	1	2	5	2	0	3	2	1	1
3	1	4	3	1	0	0	2	7	0

- Two-way layout: row-column independence?

1	2	2	1	1	0	1
2	0	0	2	3	0	0
0	1	1	1	2	7	3
1	1	2	0	0	0	1
0	1	1	1	1	0	0

Estimation of p_{obs}

- Estimate p_{obs} by simulation from fitted null hypothesis model $\widehat{M}_0$.
- Algorithm: for $r = 1, \ldots, R$:
  - simulate data set $y_1^*, \ldots, y_n^*$ from $\widehat{M}_0$;
  - calculate test statistic t_r^* from $y_1^*, \ldots, y_n^*$.
- Calculate simulation estimate

$$\widehat{p} = \frac{1 + \#\{t_r^* \geq t_{\text{obs}}\}}{1 + R}$$

of

$$\widehat{p}_{\text{obs}} = \Pr(T \geq t_{\text{obs}} \mid \widehat{M}_0).$$

- Simulation and statstical errors:

$$\widehat{p} \approx \widehat{p}_{\text{obs}} \approx p_{\text{obs}}$$

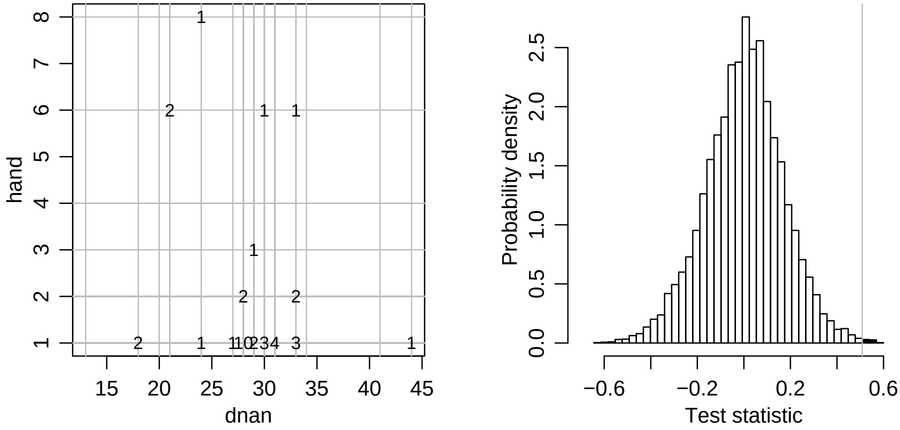
Handedness data: Test of independence

- Are `dnan` and `hand` positively associated?
- Take $T = \widehat{\theta}$ (correlation coefficient), with $t_{\text{obs}} = 0.509$; this is large in case of positive association (one-sided test)
- Null hypothesis of independence: $F(u, x) = F_1(u)F_2(x)$
- Take bootstrap samples independently from $\widehat{F}_1 \equiv (\texttt{dnan}_1, \dots, \texttt{dnan}_n)$ and from $\widehat{F}_2 \equiv (\texttt{hand}_1, \dots, \texttt{hand}_n)$, then put them together to get bootstrap data $(\texttt{dnan}_1^*, \texttt{hand}_1^*), \dots, (\texttt{dnan}_n^*, \texttt{hand}_n^*)$.
- With $R = 9,999$ get 18 values of $\widehat{\theta}^* \geq \widehat{\theta}$, so

  $$\widehat{p} = \frac{1 + 18}{1 + 9999} = 0.0019 :$$

  `hand` and `dnan` seem to be positively associated
- To test positive or negative association (two-sided test), take $T = |\widehat{\theta}|$: gives $\widehat{p} = 0.004$.

Handedness data: Bootstrap from $\widehat{M}_0$



Choice of R

- Take R big enough to get small standard error for $\widehat{p}$, typically ≥ 100, using binomial calculation:

$$\begin{aligned} \operatorname{var}(\widehat{p}) &= \operatorname{var}\left(\frac{1+\#\{t_r^* \geq t_{\mathrm{obs}}\}}{1+R}\right) \\ &\doteq \frac{1}{R^2}\, R p_{\mathrm{obs}}(1-p_{\mathrm{obs}}) = \frac{p_{\mathrm{obs}}(1-p_{\mathrm{obs}})}{R} \end{aligned}$$

  so if $p_{\mathrm{obs}} \doteq 0.05$ need $R \geq 1900$ for 10% relative error (**D7**)
- Can choose R sequentially: e.g. if $\widehat{p} \doteq 0.06$ and $R = 99$, can augment R enough to diminish standard error.
- Taking R too small lowers power of test.

Duality with confidence interval

- Often unclear how to impose null hypothesis on sampling scheme
- General approach based on duality between confidence interval $\mathcal{I}_{1-\alpha} = (\theta_\alpha, \infty)$ and test of null hypothesis $\theta = \theta_0$
- Reject null hypothesis at level α in favour of alternative $\theta > \theta_0$, if $\theta_0 < \theta_\alpha$
- Handedness data: $\theta_0 = 0 \notin \mathcal{I}_{0.95}$, but $\theta_0 = 0 \in \mathcal{I}_{0.99}$, so estimated significance level $0.01 < \widehat{p} < 0.05$: weaker evidence than before
- Extends to tests of $\theta = \theta_0$ against other alternatives:
  - if $\theta_0 \notin \mathcal{I}^{1-\alpha} = (-\infty, \theta^\alpha)$, have evidence that $\theta < \theta_0$
  - if $\theta_0 \notin \mathcal{I}_{1-2\alpha} = (\theta_\alpha, \theta^\alpha)$, have evidence that $\theta \neq \theta_0$

Pivot tests

- Equivalent to use of confidence intervals
- Idea: use (approximate) pivot such as $Z = (\widehat{\theta} - \theta)/V^{1/2}$ as statistic to test $\theta = \theta_0$
- Observed value of pivot is $z_{\text{obs}} = (\widehat{\theta} - \theta_0)/V^{1/2}$
- Significance level is

$$
\begin{aligned}
\Pr\left(\frac{\widehat{\theta} - \theta}{V^{1/2}} \geq z_{\text{obs}} \mid M_0\right) &= \Pr(Z \geq z_{\text{obs}} \mid M_0) \\
&= \Pr(Z \geq z_{\text{obs}} \mid F) \\
&\doteq \Pr(Z \geq z_{\text{obs}} \mid \widehat{F})
\end{aligned}
$$

- Compare observed z_{obs} with simulated distribution of $Z^* = (\widehat{\theta}^* - \widehat{\theta})/V^{*1/2}$, without needing to construct null hypothesis model $\widehat{M}_0$
- Use of (approximate) pivot is essential for success

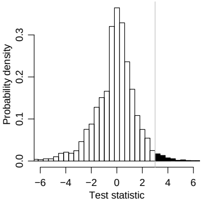
Example: Handedness data

- Test zero correlation ($\theta_0 = 0$), not independence; $\widehat{\theta} = 0.509$, $V = 0.170^2$:

$$z_{\text{obs}} = \frac{\widehat{\theta} - \theta_0}{V^{1/2}} = \frac{0.509 - 0}{0.170} = 2.99$$

- Observed significance level is

$$\widehat{p} = \frac{1 + \#\{z_r^* \geq z_{\text{obs}}\}}{1 + R} = \frac{1 + 215}{1 + 9999} = 0.0216$$



Exact tests

- Problem: bootstrap estimate is

$$\widehat{p}_{\text{obs}} = \Pr(T \geq t_{\text{obs}} \mid \widehat{M_0}) \neq \Pr(T \geq t \mid M_0) = p_{\text{obs}},$$

  so estimate the wrong thing
- In some cases can eliminate parameters from null hypothesis distribution by conditioning on sufficient statistic
- Then simulate from conditional distribution
- More generally, can use Metropolis–Hastings algorithm to simulate from conditional distribution (below)

Example: Fir data

- Data $Y_1, \ldots, Y_n \stackrel{\text{iid}}{\sim} \text{Pois}(\lambda)$, with λ unknown
- Poisson model has $\text{E}(Y) = \text{var}(Y) = \lambda$: base test of overdispersion on

$$T = \sum (Y_j - \overline{Y})^2 / \overline{Y} \mathrel{\dot\sim} \chi^2_{n-1};$$

  observed value is $t_{\text{obs}} = 55.15$
- Unconditional significance level:

$$\Pr(T \geq t_{\text{obs}} \mid \widehat{M}_0, \lambda)$$

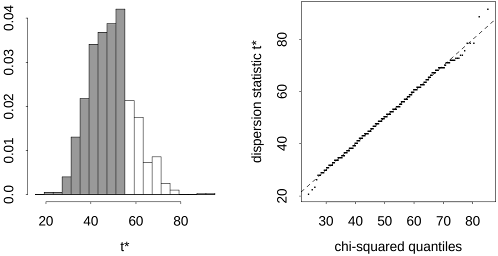
- Condition on value w of sufficient statistic $W = \sum Y_j$:

$$p_{\text{obs}} = \Pr(T \geq t_{\text{obs}} \mid \widehat{M}_0, W = w),$$

  independent of λ, owing to sufficiency of W
- Exact test: simulate from multinomial distribution of $Y_1, \ldots, Y_n$ given $W = \sum Y_j = 107$.

Example: Fir data

Figure: Simulation results for dispersion test. Left panel: $R = 999$ values of the dispersion statistic t^* obtained under multinomial sampling: the data value is $t_{\text{obs}} = 55.15$ and $\widehat{p} = 0.25$. Right panel: chi-squared plot of ordered values of t^*, dotted line shows χ^2_{49} approximation to null conditional distribution.



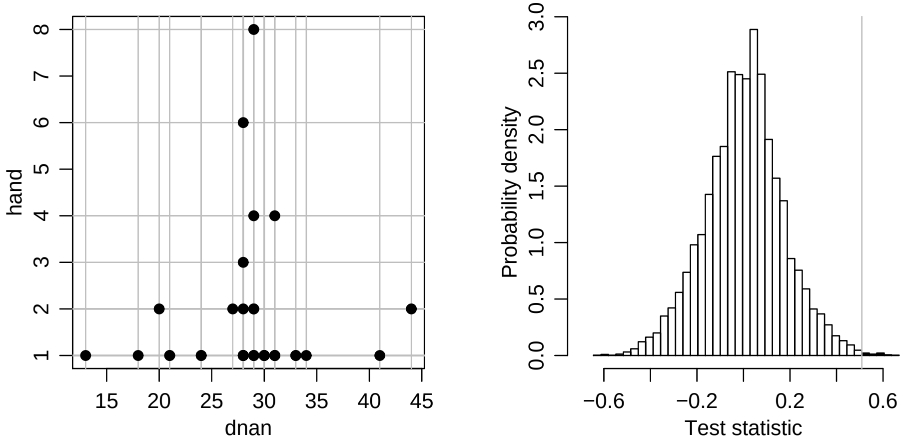
Handedness data: Permutation test

- Are `dnan` and `hand` related?
- Take $T = \widehat{\theta}$ (correlation coefficient) again
- Impose null hypothesis of independence: $F(u, x) = F_1(u)F_2(x)$, but condition so that marginal distributions $\widehat{F}_1$ and $\widehat{F}_2$ are held fixed under resampling plan — permutation test
- Take resamples of form

  $$(\mathtt{dnan}_1, \mathtt{hand}_{1^*}), \ldots, (\mathtt{dnan}_n, \mathtt{hand}_{n^*})$$

  where $(1^*, \ldots, n^*)$ is random permutation of $(1, \ldots, n)$
- Doing this with $R = 9,999$ gives one- and two-sided significance probabilities of 0.002, 0.003
- Typically values of $\widehat{p}$ very similar to those for corresponding bootstrap test

Handedness data: Permutation resample



Contingency table

1	2	2	1	1	0	1
2	0	0	2	3	0	0
0	1	1	1	2	7	3
1	1	2	0	0	0	1
0	1	1	1	1	0	0

- Are row and column classifications independent:

$$\Pr(\text{row } i,\ \text{column } j) = \Pr(\text{row } i) \times \Pr(\text{column } j)?$$

- Standard test statistic for independence is

$$T = \sum_{i,j} \frac{(y_{ij} - \widehat{y}_{ij})^2}{\widehat{y}_{ij}}, \quad \widehat{y}_{ij} = \frac{y_{i\cdot} y_{\cdot j}}{y_{\cdot\cdot}}$$

- Get $\Pr(\chi^2_{24} \geq 38.53) = 0.048$, but is $T \dot{\sim} \chi^2_{24}$?

Exact tests: Contingency table

- For exact test, need to simulate distribution of T conditional on sufficient statistics — row and column totals
- Algorithm (**D8**) for conditional simulation:
  1. choose two rows $j_1 < j_2$ and two columns $k_1 < k_2$ at random
  2. generate new values from hypergeometric distribution of $y_{j_1 k_1}$ conditional on margins of 2×2 table

$$\begin{array}{cc} y_{j_1 k_1} & y_{j_1 k_2} \\ y_{j_2 k_1} & y_{j_2 k_2} \end{array}$$

  3. compute test statistic T^* every $I = 100$ iterations, say
- Compare observed value $t_{\text{obs}} = 38.53$ with simulated T^* — get $\widehat{p} \doteq 0.08$

Key points

- Tests can be performed using resampling/simulation
- Must take account of null hypothesis, by
  - modifying sampling scheme to satisfy null hypothesis
  - inverting confidence interval (pivot test)
- Can use Monte Carlo simulation to get approximations to exact tests — simulate from null distribution of data, conditional on observed value of sufficient statistic
- Sometimes obtain permutation tests — very similar to bootstrap tests

- Motivation
- Basic notions
- Confidence intervals
- Several samples
- Variance estimation
- Tests
- Regression

Linear regression

- Independent data (x_1, y_1), ..., (x_n, y_n) with

$$y_j = x_j^{\mathrm{T}}\beta + \varepsilon_j, \quad \varepsilon_j \sim (0, \sigma^2)$$

- Least squares estimates $\widehat{\beta}$, leverages h_j, residuals

$$e_j = \frac{y_j - x_j^{\mathrm{T}}\widehat{\beta}}{(1-h_j)^{1/2}} \quad \dot\sim \quad (0, \sigma^2)$$

- Design matrix X is experimental ancillary — should be held fixed if possible, as

$$\mathrm{var}(\widehat{\beta}) = \sigma^2 (X^{\mathrm{T}}X)^{-1}$$

if model $y = X\beta + \varepsilon$ correct

Linear regression: Resampling schemes

- Two main resampling schemes
- Model-based resampling:

$$y_j^* = x_j^{\mathrm{T}} \widehat{\beta} + \varepsilon_j^*, \quad \varepsilon_j^* \sim \mathrm{EDF}(e_1 - \overline{e}, \ldots, e_n - \overline{e})$$

  - Fixes design but not robust to model failure
  - Assumes ε_j sampled from population
- Case resampling:

$$(x_j, y_j)^* \sim \mathrm{EDF}\{(x_1, y_1), \ldots, (x_n, y_n)\}$$

  - Varying design X but robust
  - Assumes (x_j, y_j) sampled from population
  - Usually design variation no problem; can prove awkward in designed experiments and when design sensitive.

Cement data

Table: Cement data: y is the heat (calories per gram of cement) evolved while samples of cement set. The covariates are percentages by weight of four constituents, tricalciumaluminate x_1, tricalcium silicate x_2, tetracalcium alumino ferrite x_3 and dicalcium silicate x_4.

	x_1	x_2	x_3	x_4	y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

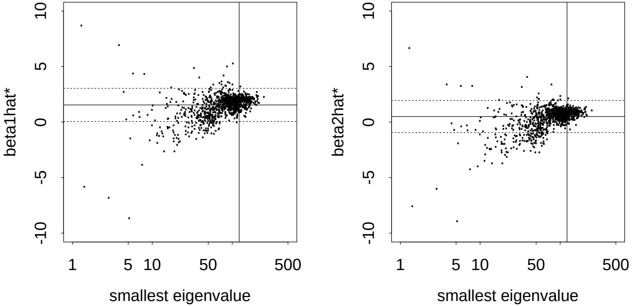
Cement data

- Fit linear model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

  and apply case resampling
- Covariates compositional: $x_1 + \cdots + x_4 \doteq 100\%$ so X almost collinear — smallest eigenvalue of $X^{\mathrm{T}}X$ is $l_5 = 0.0012$
- Plot of $\widehat{\beta}_1^*$ against smallest eigenvalue of $X^{*\mathrm{T}}X^*$ reveals that $\mathrm{var}^*(\widehat{\beta}_1^*)$ strongly variable
- Relevant subset for case resampling — post-stratification of output based on l_5^*?

Cement data


beta1hat*
10
5
0
-5
-10
1
5
10
50
500
smallest eigenvalue
beta2hat*
10
5
0
-5
-10
1
5
10
50
500
smallest eigenvalue

Cement data

Table: Standard errors of linear regression coefficients for cement data. Theoretical and error resampling assume homoscedasticity. Resampling results use $R = 999$ samples, but last two rows are based only on those samples with the middle 500 and the largest 800 values of ℓ_1^*.

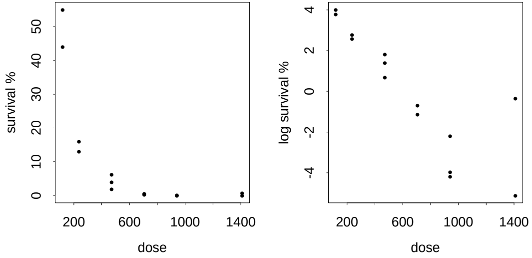
	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$
Normal theory	70.1	0.74	0.72
Model-based resampling, $R = 999$	66.3	0.70	0.69
Case resampling, all $R = 999$	108.5	1.13	1.12
Case resampling, largest 800	67.3	0.77	0.69

Survival data

dose x	117.5	235.0	470.0	705.0	940.0	1410
survival % y	44.000	16.000	4.000	0.500	0.110	0.700
	55.000	13.000	1.960	0.320	0.015	0.006
			6.120		0.019	

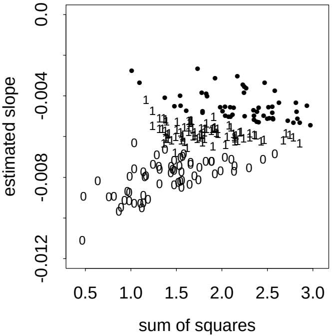
- Data on survival % for rats at different doses
- Linear model:

$$\log(\text{survival}) = \beta_0 + \beta_1 \text{dose}$$



Survival data

- Case resampling
- Replication of outlier: none (0), once (1), two or more (•).
- Model-based sampling including residual would lead to change in intercept but not slope.



Generalized linear model

- Response may be binomial, Poisson, gamma, normal, ...

$$y_j \quad \sim \quad \text{mean } \mu_j, \ \ \text{variance } \phi V(\mu_j),$$

where $g(\mu_j) = x_j^{\mathrm{T}}\beta$ is linear predictor; $g(\cdot)$ is link function.

- MLE $\widehat{\beta}$, fitted values $\widehat{\mu}_j$, Pearson residuals

$$r_{Pj} = \frac{y_j - \widehat{\mu}_j}{\{V(\widehat{\mu}_j)(1-h_j)\}^{1/2}} \quad \dot{\sim} \quad (0, \phi).$$

- Bootstrapped responses

$$y_j^* = \widehat{\mu}_j + V(\widehat{\mu}_j)^{1/2}\varepsilon_j^*$$

where $\varepsilon_j^* \ \sim \ \mathrm{EDF}(r_{P1} - \overline{r}_P, \ldots, r_{Pn} - \overline{r}_P)$. However
  - possible that $y_j^* \notin \{0, 1, 2, \ldots,\}$
  - r_{Pj} not exchangeable, so may need stratified resampling

AIDS data

- Log-linear model: number of reports in row j and column k follows Poisson distribution with mean

$$\mu_{jk} = \exp(\alpha_j + \beta_k)$$

- Log link function

$$g(\mu_{jk}) = \log \mu_{jk} = \alpha_j + \beta_k$$

and variance function

$$\operatorname{var}(Y_{jk}) = \phi \times V(\mu_{jk}) = 1 \times \mu_{jk}$$

- Pearson residuals:

$$r_{jk} = \frac{Y_{jk} - \widehat{\mu}_{jk}}{\{\widehat{\mu}_{jk}(1 - h_{jk})\}^{1/2}}$$

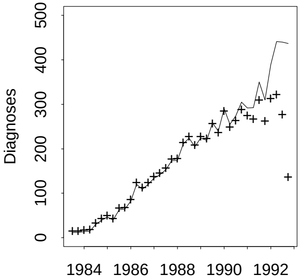
- Model-based simulation:

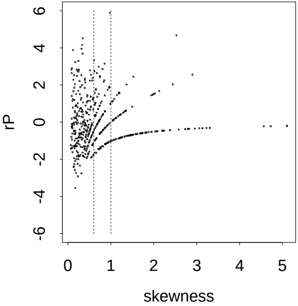
$$Y^*_{jk} = \widehat{\mu}_{jk} + \widehat{\mu}_{jk}^{1/2} \varepsilon^*_{jk}$$

Diagnosis period		Reporting-delay interval (quarters):									Total reports to end
Year	Quarter	0†	1	2	3	4	5	6	···	≥14	of 1992
1988	1	31	80	16	9	3	2	8	···	6	174
	2	26	99	27	9	8	11	3	···	3	211
	3	31	95	35	13	18	4	6	···	3	224
	4	36	77	20	26	11	3	8	···	2	205
1989	1	32	92	32	10	12	19	12	···	2	224
	2	15	92	14	27	22	21	12	···	1	219
	3	34	104	29	31	18	8	6	···		253
	4	38	101	34	18	9	15	6	···		233
1990	1	31	124	47	24	11	15	8	···		281
	2	32	132	36	10	9	7	6	···		245
	3	49	107	51	17	15	8	9	···		260
	4	44	153	41	16	11	6	5	···		285
1991	1	41	137	29	33	7	11	6	···		271
	2	56	124	39	14	12	7	10			263
	3	53	175	35	17	13	11				306
	4	63	135	24	23	12					258
1992	1	71	161	48	25						310
	2	95	178	39							318
	3	76	181								273
	4	67									133

AIDS data

- Poisson two-way model deviance 716.5 on 413 df — indicates strong overdispersion: $\phi > 1$, so Poisson model implausible
- Residuals highly inhomogeneous — exchangeability doubtful





AIDS data: Prediction intervals

- To estimate prediction error:
  - simulate complete table y^*_{jk};
  - estimate parameters from incomplete y^*_{jk}
  - get estimated row totals and 'truth'

$$\widehat{\mu}^*_{+,j} = e^{\widehat{\alpha}^*_j} \sum_{k \text{ unobs}} e^{\widehat{\beta}^*_k}, \quad y^*_{+,j} = \sum_{k \text{ unobs}} y^*_{jk}.$$

  - Prediction error

$$\frac{y^*_{+,j} - \widehat{\mu}^*_{+,j}}{\widehat{\mu}^{*1/2}_{+,j}}$$

  studentized so more nearly pivotal.
- Form prediction intervals from R replicates.

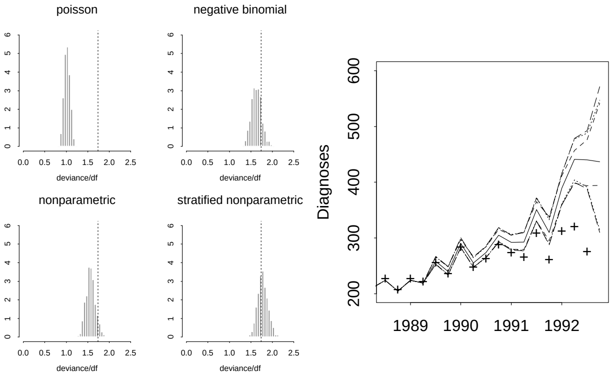
AIDS data: Resampling plans

- Resampling schemes:
  - parametric simulation, fitted Poisson model
  - parametric simulation, fitted negative binomial model
  - nonparametric resampling of r_P
  - stratified nonparametric resampling of r_P
- Stratification based on skewness of residuals, equivalent to stratifying original data by values of fitted means
- Take strata for which

$$\widehat{\mu}_{jk} < 1, \quad 1 \leq \widehat{\mu}_{jk} < 2, \quad \widehat{\mu}_{jk} \geq 2$$

AIDS data: Results

- Deviance/df ratios for the sampling schemes, $R = 999$.
- Poisson variation inadequate.
- 95% prediction limits.



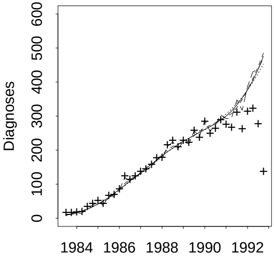
AIDS data: Semiparametric model

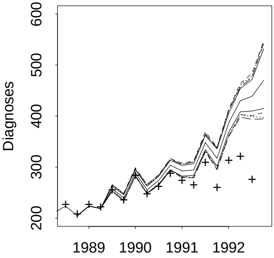
- More realistic: generalized additive model

$$\mu_{jk} = \exp\{\alpha(j) + \beta_k\},$$

where $\alpha(j)$ is locally-fitted smooth.

- Same resampling plans as before
- 95% intervals now generally narrower and shifted upwards





Key points

- Key assumption: independence of cases
- Two main resampling schemes for regression settings:
  - Model-based
  - Case resampling
- Intermediate schemes possible
- Can help to reduce dependence on assumptions needed for regression model
- These two basic approaches also used for more complex settings (time series, ...), where data are dependent

Summary

- Bootstrap: simulation methods for frequentist inference.
- Useful when
  - standard assumptions invalid (n small, data not normal, $\ldots$);
  - standard problem has non-standard twist;
  - complex problem has no (reliable) theory;
  - or (almost) anywhere else.
- Have described
  - basic ideas
  - confidence intervals
  - tests
  - some approaches for regression

Books

- Chernick (1999) *Bootstrap Methods: A Practicioner's Guide*. Wiley
- Davison and Hinkley (1997) *Bootstrap Methods and their Application*. Cambridge University Press
- Efron and Tibshirani (1993) *An Introduction to the Bootstrap*. Chapman & Hall
- Hall (1992) *The Bootstrap and Edgeworth Expansion*. Springer
- Lunneborg (2000) *Data Analysis by Resampling: Concepts and Applications*. Duxbury Press
- Manly (1997) *Randomisation, Bootstrap and Monte Carlo Methods in Biology*. Chapman & Hall
- Shao and Tu (1995) *The Jackknife and Bootstrap*. Springer