MAS472/6004 Computational Inference

Richard Wilkinson

r.d.wilkinson@sheffield.ac.uk

Computational Inference

Simple computational tools for solving hard statistical problems.

- Monte Carlo/simulation
- MC and simulation in frequentist inference
- Random number generation/ simulating from probability distributions
- Further Bayesian computation

Methods implemented via simple programs in R.

Chapter 1: Monte Carlo methods

Problem 1: estimating probabilities

A particular site is being considered for a wind farm. At that site, the log of the wind speed in m/s on day t is known to follow an $AR(2)$ process:

$$Y_t = 0.6Y_{t-1} + 0.4Y_{t-2} + \varepsilon_t, \qquad (1)$$

with $\varepsilon_t \sim N(0, 0.01)$.

If $Y_1 = Y_2 = 1.5$, what is the **probability** that the wind speed $\exp(Y_t)$ will be below 15 kmh for more than 10 days in a 100 day period?

Problem 2: estimating variances

Given a sample of 5 standard normal random variables $X_1, \ldots, X_5$, what is the **variance** of

$$\max_i\{X_i\} - \min_i\{X_i\}$$

Problem 3: Estimating percentiles

The concentration of pollutant at any point in region following release from point source can be describe by the model

$$C(x,y,z) = \frac{Q}{2\pi u_{10}\sigma_z\sigma_y} \exp\left[-\frac{1}{2}\left\{\frac{y^2}{\sigma_y^2} + \frac{(z-h)^2}{\sigma_z^2}\right\}\right], \quad (2)$$

C: air concentration of pollutant, Q: release rate, u_{10}: wind speed at $10m$ above ground, σ_y, σ_z: diffusion parameters in horizontal and vertical directions, h:release height, (x,y,z): coordinates along wind direction, cross wind and above ground.

Given $Q = 100$, $h = 50$m, but u, σ_z, σ_y uncertain. If

$\log u_{10} \sim N(2, .1)$ $\log \sigma_y^2 \sim N(10, 0.2)$ $\log \sigma_z^2 \sim N(5, 0.05)$

What is the **95th percentile** of $C(100, 100, 40)$?

Problem 4:Estimating expectations

A hospital ward has 8 beds

- The number of patients arriving each day is uniformly distributed between 0 and 5 inclusive.
- The length of stay for each patient is also uniformly distributed between 1 and 3 days inclusive.

If all 8 beds are free initially, what is the **expected** number of days before there are more patients than beds?

Problem 5: Optimal decisions

The Monty Hall Problem

On a game show you are given the choice of three doors.

- Behind one door is a car; behind the others, goats.

The rules of the game are

- After you have chosen a door, the game show host, Monty Hall, opens one of the two remaining doors to reveal a goat.
- You are now asked whether you want to stay with your first choice, or to switch to the other unopened door.

What is the **optimal strategy**? And what is the resulting probability of winning?

These 5 problems are all either hard or impossible to tackle analytically. However, the **Monte Carlo method**, can be used to obtain approximate answers to all of them.

> *Monte Carlo methods are a broad class of computational algorithms relying on repeated random sampling to obtain numerical results. They use randomness to solve problems that might be deterministic in principle.*

Some useful results

Monte Carlo is primarily used to calculate integrals. For example

- Expectation of a random variable $X \sim f(\cdot)$, or a function of it
$$\mathbb{E}g(X) = \int g(x)f(x)\mathrm{dx}$$
- Variance
$$\mathbb{V}\mathrm{ar}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2. \quad (3)$$
- Probability $\mathbb{P}(X < a)$ is the expectation of $\mathbb{I}_{X<a}$, the indicator function which is 1 if $X < a$ and otherwise is 0. Then
$$\begin{aligned}\mathbb{P}(X < a) &= 1 \times \mathbb{P}(X < a) + 0 \times \mathbb{P}(X \geq a) \\ &= \mathbb{E}\{\mathbb{I}(X < a)\} = \int \mathbb{I}_{X<a} f(X)\mathrm{dx}\end{aligned}$$

Monte Carlo Integration - I

Suppose we are interested in the integral
$$I = \mathbb{E}(g(X)) = \int g(x)f(x)\mathrm{dx}$$
Let $X_1, X_2, \ldots, X_n$ be independent random variables with pdf $f(x)$. Then a Monte Carlo approximation to I is
$$\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i). \quad (4)$$

Example:

Monte Carlo Integration - II

Some properties of $\hat{I}$.

(1) $\hat{I}_n$ is an unbiased estimator of I. **Proof:**

Monte Carlo Integration - III

(2) $\hat{I}_n$ converges to I as $n \to \infty$.

Proof:

Monte Carlo Integration - IV

The SLLN tells us $\hat{I}_n$ converges, but not how fast. It doesn't tell us how large n must be to achieve a certain error.

(3)

$$\mathbb{E}[(\hat{I}_n - I)^2] = \frac{\sigma^2}{n}$$

where $\sigma^2 = \mathbb{V}\text{ar}(g(X))$. Thus the 'root mean square error' (RMSE) of $\hat{I}_n$ is

$$\text{RMSE}(\hat{I}_n) = \frac{\sigma}{\sqrt{n}} = O(n^{-1/2}).$$

Thus, our estimate is more accurate as $n \to \infty$, and is less accurate when σ^2 is large. σ^2 will usually be unknown, but we can estimate it:

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(g(X_i) - \hat{I}_n)^2$$

We call $\hat{\sigma}$ the *Monte Carlo standard error*.

Monte Carlo Integration - V

We write[1]

$$\text{RMSE}(\hat{I}_n) = O(n^{-1/2})$$

to emphasise the rate of convergence of the error with n.

To get 1 digit more accuracy requires a 100-fold increase in n. A 3-digit improvement would require us to multiply n by 10^6.

Consequently Monte Carlo is not usually suited for problems where we need a very high accuracy. Although the error rate is low (the RMSE decreases slowly with n), it has the nice properties that the RMSE

- does not depend on $d = \dim(x)$
- does not depend on the smoothness of f

Consequently Monte Carlo is very competitive in high dimensional problems that are not smooth.

Monte Carlo Integration - VI

In addition to the rate of convergence, the central limit theorem tells us the asymptotic[2] distribution of $\hat{I}_n$

(4)

$$\frac{\sqrt{n}(\hat{I}_n - I)}{\sigma} \to N(0,1) \text{ in distribution as } n \to \infty$$

Informally, $\hat{I}_n$ is approximately $N(I, \frac{\sigma^2}{n})$ for large n.

This allows us to calculate confidence intervals for I.

See the R code on MOLE.

- If we require $E\{f(X)\}$, random observations from distribution of $f(X)$ can be generated by generating $X_1, \ldots, X_n$ from distribution of X, and then evaluating $f(X_1), \ldots, f(X_n)$.
- Preceding results can be applied when estimating variances or probabilities of events.
- Percentiles estimated by taking the sample percentile from the generated sample of values $X_1, \ldots, X_n$.
- We expect the estimate to be more accurate as n increases. Determining a percentile is equivalent to inverting a CDF. If wish to know the 95th percentile, we must find ν such that

$$P(X \leq \nu) = 0.95, \tag{5}$$

Monte Carlo solutions to the example problems

Question 1

Define E: the event that in 100 days the wind speed is below 15kmh for more than 10 days.
To estimate $\mathbb{P}(E)$, generate lots of individual time series, and count proportion of series in which E occurs

1. Generate ith realisation of the time series process:
   For $t = 3, 4, \ldots, 100$:
   - Set $Y_t \leftarrow 0.6Y_{t-1} + 0.4Y_{t-2} + N(0, 0.01)$
2. Count number of elements of $\{Y_1, \ldots, Y_{100}\}$ less than $\log 4.167$:
   - Set $X_i \leftarrow \sum_{t=1}^{100} I\{Y_t < \log 4.167\}$
3. Determine if event E has occured for time series i:
   - Set $E_i \leftarrow I\{X_i > 10\}$
4. Estimate $\mathbb{P}(E)$ by $\frac{1}{N}\sum_{i=1}^{N} E_i$

Question 2

Define Z to be the difference between max and min of 5 standard normal random variables. Estimate the variance

Question 3

Transformation of a random variable:
Given random variables $X_1, \ldots, X_d$ we want to know the distribution of $Y = f(X_1, \ldots, X_d)$.

- The Monte Carlo method can be used
  - Sample unknown inputs from their distributions,
  - evaluate the function to obtain output value from its distribution.
- Given suitably large sample, 95th percentile from distribution of $C(100, 100, 40)$ can be estimated by the 95th percentile from sample of simulated values of $C(100, 100, 40)$.

For $i = 1, 2, \ldots, N$:

1. Sample a set of input values:
   - Sample $u_{10,i}$ from $\log N(2, .1)$
   - Sample $\sigma^2_{y,i}$ from $\log N(10, 0.2)$
   - Sample $\sigma^2_{z,i}$ from $\log N(5, 0.05)$
2. Evaluate the model output C_i:
   - Set $C_i \leftarrow \frac{100}{2\pi u_{10,i}\sigma_{z,i}\sigma_{y,i}} \exp\left[-\frac{1}{2}\left\{\frac{40^2}{\sigma^2_{y,i}} + \frac{100}{\sigma^2_{z,i}}\right\}\right]$
3. Return the 95th percentile of $C_1, C_2, \ldots, C_N$.

Question 4

- Define W to be the number of days before the first patient arrives to find no available beds.
- The question has asked us to give $E(W)$.
- If we can generate $W_1, \ldots, W_n$ from the distribution of W, we can then estimate $E(W)$ by $\bar{W}$.

See the R code on MOLE for a way to simulate this process.

The Monty Hall Problem

- Simulate N separate games by randomly letting x take values in $\{1, 2, 3\}$ with equal probability. x represents which door the car is behind.
- Simulate the contestant randomly picking a door by choosing a value y in $\{1, 2, 3\}$ (it doesn't matter how we do this, we can always choose 1 if you like, the results are the same).
- Now the game show host will open the door which hasn't been picked that contains a goat. For each of the N games, record the success of the two strategies
  1. stick with choice y
  2. change to the unopened door.
- Calculate the success rate for each strategy.

Example 1

Consider the probability p that a standard normal random variable will lie in the interval $[0, 1]$. This can be written as an integral

$$p = \int_0^1 \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx. \tag{6}$$

Two methods for estimating/evaluating this probability are

1. numerical integration/quadrature, e.g., trapezium rule, Simpson's rule etc
2. given a sample of standard normal random variables $Z_1, \ldots, Z_n$, look at the proportion of Z_is occurring in the interval $[0, 1]$.

Example 1: An alternative method

1. Y is a RV with $f(Y)$ any function of Y. To generate a random value from the distribution of $f(Y)$, generate a random Y from the distribution of Y, and then evaluate $f(Y)$.
2. Providing $\mathbb{E}\{f(Y)\}$ exists, given a sample $f(Y_1), \ldots, f(Y_n)$,

$$\frac{1}{n}\sum_{i=1}^{n} f(Y_i)$$

   is an unbiased estimator of $\mathbb{E}\{f(Y)\}$.
3. Let X be a random variable with a $U[0, 1]$ distribution. For an arbitrary function $f(X)$, what is the expectation of $f(X)$?

4 Now choose f to be the function $f(X) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{X^2}{2}\right)$. Then if $X \sim U[0,1]$

$$\mathbb{E}\{f(X)\} = \int_0^1 \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) 1\mathrm{dx} \tag{8}$$

Given a sample $f(X_1), \ldots, f(X_n)$ from the distribution of $f(X)$, we can estimate $E\{f(X)\}$ by the *unbiased* **Monte Carlo** estimator $\hat{p}$

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} f(X_i), \tag{9}$$

where X_i is drawn randomly from the $U[0,1]$ distribution.

Key idea

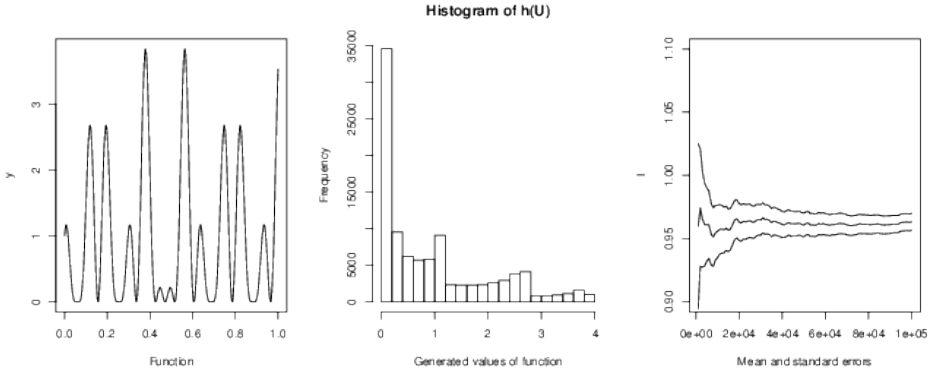
re-express the integral of interest (6) as an *expectation*.

Example 2

Consider the integral $\int_0^1 h(x)\mathrm{dx}$ where

$$h(x) = [\cos(50x) + \sin(20x)]^2$$

Generate $X_1, \ldots, X_n$ from $U[0,1]$ and estimate with $\hat{I}_n = \frac{1}{n} \sum h(X_i)$.



The general framework

$$R = \int f(x)dx \tag{10}$$

Let $g(x)$ be some density function that is easy to sample from. How do we re-write (10) as the expectation of a function of a random variable X with density function $g(x)$?

So we now have $R = E\{h(X)\}$, where X has the density function $g(x)$. If we now sample $X_1, \ldots, X_n$ from $g(x)$, then evaluate $h(X_1), \ldots, h(X_n)$,

$$\hat{R} = \frac{1}{n} \sum_{i=1}^{n} h(X_i) \tag{12}$$

is an unbiased estimator of R.

Example 3

Use Monte Carlo integration to estimate

$$R = \int_{-1}^{1} \exp(-x^2)dx. \tag{13}$$

We'll consider two different choices for $g(x)$.

1. A uniform density on $[-1,1]$: $g(x) = 0.5$ for $x \in [-1,1]$. We sample $X_1, \ldots, X_n$ from $U[-1,1]$, and estimate R by

$$\hat{R} = \frac{1}{n} \sum_{i=1}^{n} \frac{\exp(-X_i^2)}{g(X_i)} = \frac{1}{n} \sum_{i=1}^{n} 2\exp(-X_i^2). \tag{14}$$

2 A normal density function $N(0, 0.5)$.
Note: sampled value X from $g(x)$ not constrained to lie in $[-1, 1]$.
Re-write R as

$$R = \int_{-\infty}^{\infty} I\{-1 \le x \le 1\} \exp(-x^2) dx, \tag{15}$$

where $I\{\}$ denotes the indicator function.
We now sample $X_1, \ldots, X_n$ from $N(0, 0.5)$ and estimate R by

$$\hat{R} = \frac{1}{n} \sum_{i=1}^{n} \frac{I\{-1 \le X_i \le 1\} \exp(-X_i^2)}{g(X_i)} = \frac{1}{n} \sum_{i=1}^{n} \pi^{1/2} I\{-1 \le X_i \le 1 \tag{16}$$

Key idea

$g(x)$ needs to mimic $f(x)$ as closely as possible. Consider again $R = \int_{-1}^{1} \exp(-x^2) \mathrm{dx}$.

Two terrible choices of g:

1. A uniform density on $[0, 1]$: $g(x) = 1$ for $x \in [0, 1]$.

$$R = \int_{-\infty}^{\infty} I\{-1 \le x \le 1\} \exp(-x^2) dx, \tag{17}$$

For $x \in [-1, 0)$, we have $f(x) > 0$ and $g(x) = 0$. Must have $g(x) > 0$ for all x where $f(x) > 0$.

2. A normal density $N(0, 0.09)$.
In this case, we have $g(x) > 0$ for $x \in [-1, 1]$, but we when we sample x from g, we expect around 95% of the values to lie in the range $(-0.6, 6)$.
The Monte Carlo estimate of R is given by

$$\hat{R} = \frac{1}{n} \sum_{i=1}^{n} I\{-1 \le X_i \le 1\} \frac{\exp(-X_i^2)\sqrt{0.18\pi}}{\exp(-5.56X_i^2)}. \tag{18}$$

Convergence

- Provided $f(x) > 0 \Rightarrow g(x) > 0$, $\hat{R}$ will converge to R as $n \to \infty$.
- Use the central limit theorem to derive a confidence interval for $\hat{R}$:

$$\hat{R} \sim N\left(R, \frac{\sigma^2}{n}\right), \tag{19}$$

where we estimate σ^2 by

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left\{h(X_i) - \hat{R}\right\}^2 \tag{20}$$

- We can then report the confidence interval as

$$\hat{R} \pm Z_{1-\alpha/2} \sqrt{\hat{\sigma}^2/n}, \tag{21}$$

- Estimates of σ^2 in the example: $U[-1, 1] : 0.16$, $N(0, 0.5) : 0.42$, $N(0, 0.09) : 6.81$.

Comparison of Monte Carlo with numerical integration

Mid-ordinate rule

Consider finding $I = \int_0^1 f(x)$ dx. There are many different numerical integration schemes we might use.
For example, the mid-ordinate rule is one of the simplest methods, and approximates I by a sum

$$\tilde{I}_n = \frac{1}{n} \sum_{i=1}^{n} f(x_i),$$

The points $x_i = (i - \frac{1}{2})h$ are equally spaced at intervals of $h = 1/n$.

Comparison of MC with numerical integration II

Mid-ordinate rule error analysis

For smooth 1-d functions the error rates for quadrature rules can be much better than Monte Carlo

For example, if $f : [0, 1] \to \mathbb{R}$ and $f''(x)$ is continuous, then

$$|I - \tilde{I}_n| \leq \frac{1}{24n^2} \max_{0 \leq x \leq 1} |f''(x)|$$

So

$$\text{RMSE}(\tilde{I}) = O(n^{-2})$$

i.e., it is a second order method. Other rules achieve higher errror rates. For example, Simpson's rule is a fourth order method.

This is much faster than Monte Carlo: to get an extra digit of accuracy we only need multiply n by a factor of $\sqrt{10} = 3.2$

Comparison of MC with numerical integration III

Curse of dimensionality

Classial quadrature methods work well for smooth 1d problems. But for d-dimensional integrals we have a problem. Suppose

$$I = \int_0^1 \int_0^1 \dots \int_0^1 f(x_1, \dots, x_d) \mathrm{d}x_1 \dots \mathrm{d}x_d$$

We can use the same N point 1-d quadrature rules on each of the d integrals.

This uses $n = N^d$ evaluations of f. The 1d mid-ordinate rule has error $O(N^{-2})$, so the d-dimensional mid-ordinate rule has error

$$|I - \tilde{I}| = O(N^{-2}) = O(n^{-2/d})$$

For $d = 4$ this is the same as Monte Carlo. For larger d it is worse.

In addition, we require f to be smooth ($f''(x)$ to be continuous) for the method to work well.

Monte Carlo has the same $O(n^{-1/2})$ error rate regardless of $\dim(x)$ or $f''(x)$

Chapter II
Simulation methods in for classical statistics

Classical statistical theory contains many methods for testing hypotheses in numerous different situations.

Derivation of these tests can be difficult or impossible in some cases and often relies on asymptotic results or approximations.

If the test we wish to perform is non-standard then deriving a suitable test procedure may not be possible (or we may have forgotten the correct test!).

In this Chapter we consider what can be done using simulation methods.

2.1 Monte Carlo tests

Recap of hypothesis testing framework

Suppose that we have a null hypothesis H_0 represented by a completely specified model and that we wish to test this hypothesis using data $X_1, \ldots, X_n$. We proceed as follows

1. Assume H_0 is true.
2. Find a test statistic $T(X_1, \ldots, X_n)$ for which large values indicate departure from H_0.
3. Calculate the theoretical sampling distribution of T under H_0.
4. The observed value $T_{obs} = T(x_1, \ldots, x_n)$ of the test statistic is compared with the distribution of T under H_0. Either
   - (Neyman-Pearson) reject H_0 if $T_{obs} > c$. Here c is chosen so that $\mathbb{P}(T \geq c|H_0) = \alpha$ where α is the **size** of the test, i.e., $\mathbb{P}(\text{reject } H_0|H_0 \text{ true}) = \alpha$.
   - (Fisher) compute the p-value $p = \mathbb{P}(T \geq T_{obs}|H_0)$ and report it. This represents the strength of evidence against H_0.

Example 1: normal parametric test

Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. Suppose that $\sigma^2 = 1$ is known. Consider the null hypothesis

$$H_0 : \mu = 0.$$

1.
2.
3.
4.

It may not be possible to derive the sampling distribution of T under H_0.

- T is not some fairly simple function,
- or if $X_1, \ldots, X_n$ are not independent samples from the population of interest (dependent data are common in real problems).

Moreover, in deriving the distribution of T, we assume that n is large, equal variances, normality etc. If these assumptions don't hold then our distribution for T will be incorrect.

Monte Carlo Tests

We may not know the distribution of T under H_0, but often it is possible to simulate from the model to produce sample data sets

$$\{X_1^{(i)}, \ldots, X_n^{(i)}\}$$

for $i = 1, \ldots, m-1$.

From these we can calculate $m-1$ sample values of the statistic under H_0,

$$\{T_1, \ldots, T_{m-1}\}$$

We can then estimate the distribution of T under H_0 from this sample and can estimate the critical value c or the p-value by a Monte Carlo approximation, i.e., estimate $\mathbb{P}(T > T_{obs}|H_0)$ by

$$\frac{1}{m-1}\sum_{i=1}^{m-1} \mathbb{I}_{T_i > t_{obs}}.$$

Monte Carlo Testing Algorithm

1. Generate $m-1$ sample test statistics $t_1, \ldots, t_{m-1}$ according to H_0.
2. For a test of size α, define $k = m\alpha$. If t_{obs} is one of the k^{th} largest values in $\{T_1, \ldots, T_{m-1}, t_{obs}\}$ then reject H_0.

i.e. reject H_0 if $t_{obs} > T_{(m-k)}$

where $T_{(1)}, \ldots, T_{(m)}$ are the order statistics of $T_1, \ldots, T_{m-1}, t_{obs}$.

Example: normal parametric test revisited

In this simple case we know the distribution of T under H_0, but it is informative to consider the Monte Carlo test.

1. Generate 1000 samples of size n from a $N(0, \sigma^2)$ distribution and calculate T.

```
t.sample <- c()
for(i in 1:999){
    temp <- rnorm(n=n, mean=0, sd=sigma)
    t.sample[i] <- mean(temp)
}
z.sample <- t.sample*sqrt(n)/sigma
z <- c(z.sample, 2)
```

add observation $Z_{obs} = 2$ to simulated data

For $\alpha = 0.05$, we find the 95th percentile of the sampling distribution

```
c<-quantile(z, 0.95)
```

Then we compare c with the observed value of 2. I found $c = 1.67$ so we would reject H_0 at the 5% level.

If instead, we wanted to estimate the p-value $\mathbb{P}(T \geq T_{obs}|H_0)$ we could estimate it using the R command

```
sum(z>2)/1000
```

For my implementation I found a p-value of 0.028 which again suggests we should reject H_0 at the 5% level.

Note that this is a random test: if we repeat it multiple times we will get a slightly different answer each time.

Example 2: Chi-squared tests

Exam grades are to be compared between 16 boys and 19 girls in a single class. The data are

	A	B	C	D
boys	3	4	5	4
girls	8	8	3	0

The null hypothesis is that there is no difference between boys and girls in exam performance.

In other words, a girl and boy chosen at random have the same probability of obtaining any particular grade.

To apply the standard chi-squared test in this case we would calculate the table of expected values under H_0 and then calculate the test statistics

Calculating for the data we find $T_{obs} = 7.907$, which has a p-value of 0.048.

However, as a rule of thumb, to use the χ^2 test, the expected number of counts in each cell should be at least 5. In this case, 4 of the 8 values are less than 5, which means the assumptions used to show that T has a χ^2 distribution with 3 degrees of freedom do not hold.

Consider using a Monte Carlo test to perform the test.

1. Under H_0, probabilities of obtaining each grade are given by the estimates

	A	B	C	D
probability	$\frac{11}{35}$	$\frac{12}{35}$	$\frac{8}{35}$	$\frac{4}{35}$

2. We then generate a new set of results for boys and girls; the boys' results are sampled from a Multinomial(16, $\frac{11}{35}$, $\frac{12}{35}$, $\frac{8}{35}$, $\frac{4}{35}$) distribution, and the girls' from a Multinomial(19, $\frac{11}{35}$, $\frac{12}{35}$, $\frac{8}{35}$, $\frac{4}{35}$). An example is shown below:

	A	B	C	D
boys	3	5	6	2
girls	4	5	7	3

3. Calculate T for these data, which for this simulated dataset is 5.323.

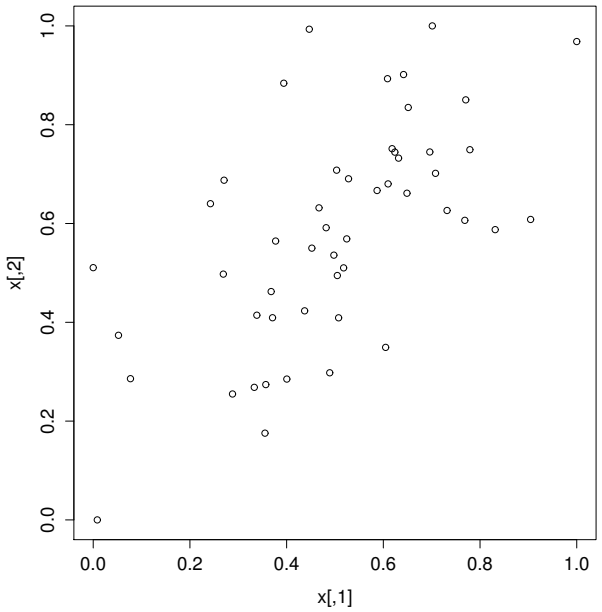
We then repeat $m - 1$ times to get $T_1, \ldots, T_{m-1}$

We rank $T_1, \ldots, T_{m-1}$ together with the observed value of the test statistic T_{obs}. In R we generate 99 test statistics and find 75 to be less than T_{obs}, and 24 to be greater. In this case the null hypothesis is not rejected at the 5% level.

Notice that this is a different conclusion to that reached using the χ^2 test.

In this case, the Monte Carlo test should preferred as we are working with the true distribution of the test statistic and not an approximation.

In general, when conducting hypothesis tests we do not have to be so reliant on distributional approximations, and **we should always consider the option of working with exact distributions**.



Example 3: Testing for randomness in spatial patterns

H_0: The spatial locations of each subject are randomly distributed over the unit square: both coordinates have $U[0, 1]$ distributions.

Various possibilities for test statistic. We will consider *nearest-neighbour* of each subject.

Let d_i denote distance from subject i to next closest subject, and define the test statistic T to be

$$T = \left(\sum_{i=1}^{50} d_i \right)^{-1} . \tag{2}$$

If locations are clustered, nearest neighbours will be small $\Rightarrow T$ will be large.

Under H_0, don't know theoretical sampling distribution of T. Straightforward to simulate values of T under H_0, so can estimate the critical values (such as the 95th percentile) we need for the hypothesis test.

1. Generate locations (x, y) of each subject by sampling x and y independently from $U[0, 1]$.
2. For each subject, find the closest observation and measure the distance to it to obtain the nearest-neighbour distance for that observation
3. Take the reciprocal of the sum of the 50 nearest-neighbour distances to get T_i.

Given a sample $T_1, \ldots, T_{m-1}$, we then rank $T_1, \ldots, T_{m-1}$ and the observed T_{obs} in order to give $T_{(1)}, \ldots, T_{(m)}$. For a test of size 5%, if T_{obs} is one of the $0.05 \times m$ largest values, then H_0 is rejected.

p-values

We can estimate the p-value $\mathbb{P}(T \geq T_{obs}|H_0)$ of a Monte Carlo test by looking at the number of observations greater than T_{obs}

$$\hat{p} = \frac{1}{m}\left(\sum_{i=1}^{m-1} \mathbb{I}_{T_i \geq T_{obs}} + 1\right)$$

Exercise: If $p = \mathbb{P}(T \geq T_{obs}|H_0)$ show that $\hat{p} \geq \frac{1}{m}$ and

$$\sum_{i=1}^{m-1} \mathbb{I}_{T_i \geq T_{obs}} \sim Bin(m-1, p)$$

so that the estimate $\hat{p}$ has expectation

$$\mathbb{E}(\hat{p}) = p + \frac{1-p}{m}$$

and is therefore a biased estimator of p. Note that for large m the bias is small.

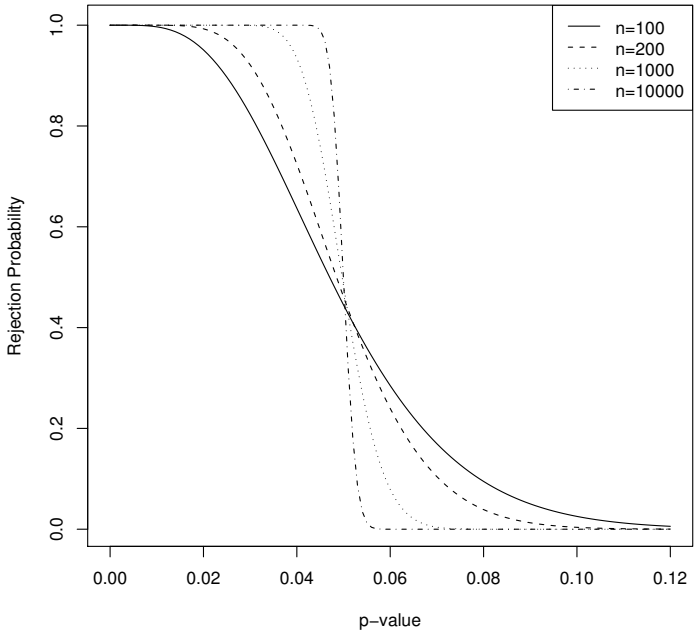
How large should m be?

We need to choose m sufficiently large so that the random sample $T_1, \ldots, T_{m-1}$ allows us to estimate the critical region to a sufficient degree of accuracy.

The Monte Carlo test has a random critical point and so 'blurs' the critical region.

- We reject H_0 if T_{obs} is one of the $k-1$ largest values in $\{T_1, \ldots, T_{m-1}\}$, where $m = k\alpha$.
- If p is the true p-value then we reject H_0 with probability

$$\begin{aligned} R(p) &= \sum_{r=0}^{k-1} \binom{m-1}{r} p^r (1-p)^{m-r-1} \\ &= \mathbb{P}(\text{Bin}(m-1, p) \leq k-1) \end{aligned}$$



$R(p)$ can be interpreted as the proportion of times the Monte Carlo test will reject H_0 when we observe T_{obs}.

For p-values smaller than 0.05 we want $R(p)$ to be large and for p-values greater than 0.05 we want $R(p)$ to be small. We choose m to make this so.

We conclude from the figure that a sample size of $m = 100$ is usually acceptable as long as the results aren't interpreted too rigidly.

Of course, this is only an issue if generating test statistics requires substantial computational effort. If it is trivial to generate sample test statistics (which it is in all but the most complex of cases), then a much large value of m can be used.

2.2 Randomisation Tests

Monte Carlo tests allowed us to do hypothesis tests when the null hypothesis specified a complete distribution for the data, e.g., $H_0 : X_i \sim N(0, 1)$.

We now consider a second technique known as **randomisation tests** for deriving the sampling distribution of the test statistic, where no distributional assumptions about the data are required.

The general scenario under consideration is that of an investigation into whether or not a particular treatment/covariate/factor has an effect on some response.

Our aim is to test this without fully specifying a distribution for the data.

Example 1: Cholesterol data

A small study was conducted to investigate the effect of diet on cholesterol levels. Volunteers were randomly allocated to one of two diets, and cholesterol levels were recorded at the end of the trial period

Diet A	233	291	312	250	246	197	268	224
Diet B	185	263	246	224	212	188	250	148

The interest is in whether or not there is a significant difference between the mean cholesterol levels for the two groups. The null hypothesis is

H_0 : mean cholesterol levels with the diets are equal

A standard classical analysis of this data might be to assume

$$X_i^{(j)} \sim N(\mu_j, \sigma^2)$$

for $i = 1, \ldots, 8$ and $j = 1, 2$ with σ^2 an unknown common variance.

The standard test is then a two sample t-test, based on the statistic

$$T = \frac{\bar{X}^{(1)} - \bar{X}^{(2)}}{\sqrt{s^2/8 + s^2/8}}, \tag{3}$$

where s^2 is the pooled estimate of variance.

Then under H_0 (and assuming normality of the data!), the test statistic T has a t-distribution with 14 degrees of freedom.

For this data, the observed test statistic T_{obs} is 2.0034 with a p-value of 0.0649 for a two-sided test.

But what if we want to analyse the data **without assuming normality**? E.g., because the sample sizes are small

Randomization tests can be used to find a distribution for T without making any distributional assumptions about the data.

If H_0 is true, then any difference in the two sample means would be solely due to how the 16 individuals were assigned to the two groups. So if H_0 is true, what is the probability of observing a sizeable difference between the two group means?

It must be equal to the probability of assigning the individuals to the two groups in such a way that the imbalance occurs, as long as the individuals were assigned to the two groups at random in the actual study. This is the principle idea behind randomisation tests.

Randomisation Test

1. Suppose the 16 individuals in the study have been labelled

Diet A	1	2	3	4	5	6	7	8
Diet B	9	10	11	12	13	14	15	16

2. Randomly re-assign the 16 individuals to the two groups.
3. Re-calculate the test-statistic for this permuted data
4. Repeat 2 and 3 to obtain B sampled test-statistics, denoted $T_1, \ldots, T_B$.
5. For a two-sided test, the estimated p-value of the observed test statistic T_{obs} is

$$\frac{1}{B}\sum_{i=1}^{B} \mathbb{I}_{|T_i| \geq |T_{obs}|}$$

Using 10000 random permutations gave a p-value of 0.063.

Equivalent test statistics

The significance level of T_{obs} is determined using

$$\hat{p} = \frac{1}{B}\sum_{i=1}^{B} I\{|T_i| \geq |T_{obs}|\}.$$

Notice that multiplying T_{obs} and all T_is by some constant would have no effect on significance level; ordering would be preserved. An **equivalent test statistic** is one that preserves ordering and hence does not change the p-value. In the example, an equivalent test statistic would be

$$T = \bar{X}_1 - \bar{X}_2. \tag{4}$$

ie no need to compute the denominator in Equation (3).

Exact randomisation tests

Could consider systematically *every* possible permutation, rather than a random sample of permutations to determine the significance level.

- This is known as an **exact** randomisation test or **permutation** test
- Can be computationally demanding/impracticable if number of possible permutations is large.
- A large sample of random permutations should be sufficient.

Outliers

In parametric tests, outlying observations in the data can cause problems.

- In the comparison of means problem, an outlier can increase the difference $\bar{X}^{(1)} - \bar{X}^{(2)}$ and will inflate the within group variance.
- Consequently the true significance of the test statistic may be underestimated.

In a randomisation test, you are comparing the relative size of the observed test statistic to its value under alternative random permutations.

Hence, the outlier will not have the same effect.

Example 2

This is illustrated in some data from a study reported in Ezinga (1976) for two treatments A and B:

A	0.33	0.27	0.44	0.28	0.45	0.55	0.44	0.76	0.59	0.01
B	0.28	0.80	3.72	1.16	1.00	0.63	1.14	0.33	0.26	0.63

The sample group means are $\bar{X}_A = 0.412$ and $\bar{X}_B = 0.995$, and the observed test statistic for a two sample t-test is $T = 1.78$.

For a two-tailed test this gives a p-value of 0.11, so not significant at the 5% level. Using a randomisation test, T is now significant at the 5% level with a p-value of about 0.03.

Exercise: Check this conclusion in R

Example 3: Analysis of Variance

Randomisation tests are applicable in many different contexts. Analysis of variance is another example. Below are responses measured on four treatment groups:

Group A	-0.10	-1.10	0.74	-3.80	
Group B	0.94	-0.30	0.67	0.86	1.19
Group C	-0.25	0.84	0.04	0.25	
Group D	0.99	0.08	0.98	0.75	0.53

Test the null hypothesis

$$H_0 : \text{all four groups have equal means}$$

Qn: What classical hypothesis test would you use?

Alternatively, a randomisation test could be applied:

1. Randomly re-assign the observations to the four treatments, keeping the numbers in each treatment the same.
2. Evaluate the test statistic
$$F = \frac{(\sum_{i=1}^{4} n_i(\bar{x}_i - \bar{x})^2)/3}{(\sum_{i=1}^{4}\sum_{j=1}^{n_i}(x_{i,j} - \bar{x}_i)^2)/14}$$
for the permuated data.
3. Repeat steps 1 and 2 B times to obtain sampled test statistics $F_1, \ldots, F_B$.
4. Estimate the significance level of F_{obs} by
$$\frac{1}{B}\sum_{i=1}^{B} \mathbb{I}_{F_i \geq F_{obs}}.$$

Based on a sample size $B = 10000$, the estimated p-value for F_{obs} was 0.03, suggesting slightly stronger evidence against the null hypothesis (compared with the parametric test).

Example 4: One-sample randomisation tests

Randomisation tests can be used for one-sample problems, but under stricter assumptions. This is demonstrated with the following example:

Given observations

$$\{10.61, 9.46, 7.02, 11.68, 9.58, 11.96, 11.28, 7.63, 6.42, 8.85\}$$

drawn from some population with mean μ, test the null hypothesis

$$H_0 : \mu = 10.$$

It is not immediately obvious what can be permuted here. However, supposing the two following assumptions hold:

- Each observation has been sampled randomly from its population
- The population distribution is symmetric about its mean.

Now suppose H_0 is true, and consider randomly sampling a value X from the population, and then evalutating $Y = X - 10$. If the population distribution is symmetric about 10, then Y must have an equal probability of being positive or negative.

In this example, subtracting 10 from each observation and taking the resulting mean gives a sample mean of -0.551. We will use the absolute value of this sample mean as the test statistic, so $T_{obs} = 0.551$ (for a two-sided test).

$$T = \left| \frac{1}{B} \sum_{i=1}^{B} Y_i \right|$$

If H_0 is true, and both assumptions hold, then the observed sample mean could simply be due to an imbalance of positive and negative Y values. This can be tested as follows:

Fisher's Randomisation test

1. Subtract hypothesised population mean from each observations:

$$\{0.61, -0.54, -2.98, 1.68, -0.42, 1.96, 1.28, -2.37, -3.58, -1.15\}$$

2. Calculate the observed test statistics: $T_{obs} = 0.551$
3. With probability 0.5 for each observation, change the sign of $X - \mu$. E.g.

$$\{-0.61, -0.54, -2.98, -1.68, 0.42, 1.96, -1.28, -2.37, -3.58, -.15\}$$

4. Re-calculate the test-statistic for the new simulated observations: $T = 0.951$.
5. Repeate 3 and 4 to obtain B sampled test-statistics $T_1, \ldots, T_B$.
6. Estimate the significance of T_{obs} by $\frac{1}{B} \sum_{i=1}^{B} \mathbb{I}_{|T_i| \geq |T_{obs}|}$

With $B = 10000$, the estimated significance of T_{obs} is 0.4021.

Using a conventional t-test, the significance of T_{obs} is 0.3982, so there is close agreement between the two methods in this example.

Example 5

Two treatments A and B, unknown population means μ_A and μ_B.

treatment A	130	119	119	168	130
treatment B	154	115	169	137	186

Consider $H_0 : \mu_B - \mu_A = 20$.

How would we test this using a randomisation test?

Cannot just permute data and evaluate difference between the sample means, as population means not equal under H_0.

Suppose we were to add 20 to each observation in group A. Under H_0, what is the expectation of {20+ a response in group A}?

If H_0 is true, then this expectation is $20 + \mu_A = \mu_B$. Adding 20 to each response in group A:

treatment A+20	150	139	139	188	150
treatment B	154	115	169	137	186

Under H_0 groups have equal population means. We can now use randomisation test in the usual way.

Summary of randomisation tests

Some argue that randomisation tests should always be used, as samples of data are never truly randomly drawn from the population of interest; some members of the population are always going to be more accessible than others.

On the other side, there is no theory to show that the results of a randomisation test can be generalised to the whole population; evidence against the null hypothesis is obtained for the observed sample only.

Consequently, in either case, a 'non-statistical' judgement has to made; that the sample can be treated as effectively random for a conventional test, or that the results can be generalised to the population for a randomisation test.

Two advantages of randomisation tests are that they can be used for any test statistic (i.e. in cases when it is not possible to analytically derive the distribution of the test statistic), and that we don't have to assume a particular distribution for the data.

Note that in most of the examples, almost identical results were obtained using the two methods. In this case, the randomisation test could be seen as a means of supporting the results from the parametric test.

The requirement for the randomisation test to be valid is that the subjects are assigned randomly to each treatment. If random allocation is not explicitly part of the experimental procedure then there needs to be the belief that the actual allocation was as likely to occur as any other.

2.3 Bootstrapping

The bootstrap is a method for assessing properties of a statistical estimator in a *non-parametric* framework. That is, we do not assume that the data are obtained from any parametric distribution (eg. normal, exponential etc).

The bootstrap is usually used to assess the variance of a statistical estimator but it is not exclusively used for this purpose.

The name comes from the story 'The Surprising Adventures of Baron Munchausen', where the main character pulls himself out of a swamp, by pulling on his own bootstraps.

The idea behind bootstrapping is that we can use the data muliple times to generate 'new' data sets to assess the properties of parameters.

Recap: CDFs

2.3.1 The Empirical Distribution Function

Define

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{X_i \leq x}$$

to be the *empirical distribution function (edf)* for data $\{X_1, \ldots, X_n\}$.

- $\hat{F}$ takes values in $\{0, \frac{1}{n}, \ldots, \frac{n}{n}\}$
- to sample from $\hat{F}$ we sample **WITH REPLACEMENT** from $\{X_1, \ldots, X_n\}$.

Note that $\hat{F}$ is a random quantity. We consider the edf to be an estimator for F. If X_i are all from distribution F then the following results hold.

Properties of the EDF - I

1. $\hat{F}(x)$ is an unbiased estimator of $F(x)$.

$$\mathbb{E}\hat{F}(x) = F(x)$$

Proof

Properties of the EDF - II

2. $\hat{F}(x) \to F(x)$ as $n \to \infty$ with probability 1.

Proof

Properties of the EDF - III

3.

$$\frac{\sqrt{n}(\hat{F}(x) - F(x))}{\sqrt{F(x)(1-F(x))}} \to N(0,1) \text{ in distribution}$$

as $n \to \infty$.

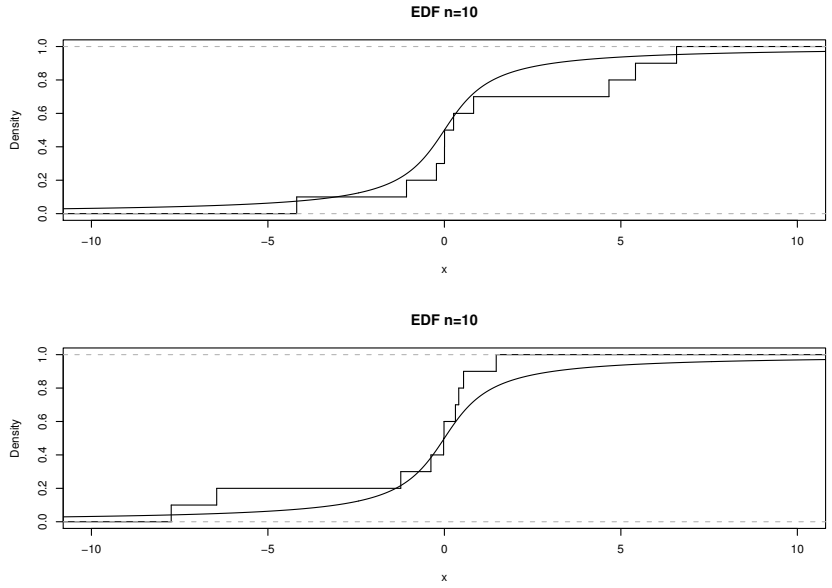
Proof:

Properties of the EDF - IV

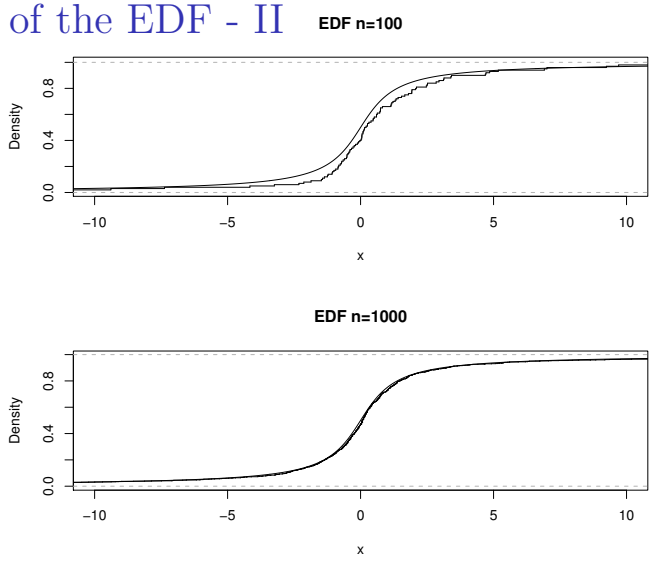
4 If X_i are an independent identically distributed sequence (so it doesn't matter if we change the order), then knowledge of $\hat{F}$ is equivalent to knowledge of $\{X_1, \ldots, X_n\}$.

Example of the EDF

Suppose $X_i \sim$ Cauchy. Then we can examine the edf for increasing values of n.



Example of the EDF - II



Notice that we've repeated for $n = 10$ and that the edf is different each time. Notice also that the edf becomes more accurate as n gets larger.

Parameters, statistics and properties

Bootstrapping texts can sometimes be confusing because of the language usage. We say

- θ is a parameter if it is a property of the underlying population, i.e., $\theta = \theta(F)$.
- $\hat{\theta}$ is a statistic which estimates θ if $\hat{\theta}$ is a function of the sample $X_1, \ldots, X_n$ - this is equivalent to being a function of the empirical distribution function

$$\hat{\theta} = \theta(\hat{F}) \equiv \theta(X)$$

- We then talk about properties of $\hat{\theta}$ such as its bias, its expectation, its standard error etc. For bootstrapping applications these properties are usually sampling properties, that is, if we repeatedly collected similar samples, what properties would $\hat{\theta}$ have?

Difficulties arise when we note that properties are also parameters of the statistic $\hat{\theta}$ and that we estimate them with statistics of the statistics.

Plug-in Principle

For example, suppose we have a sample of size n, $\{X_1, \ldots, X_n\}$ say, from unknown density F.

Suppose that interest lies in some parameter θ of the distribution F which we write $\theta = \theta(F)$ where we consider θ to be a functional of the distribution F.

We estimate θ by $\hat{\theta}$ where $\hat{\theta}$ is a function of the observations $\{X_1, \ldots, X_n\}$. Usually we have that $\hat{\theta} = \theta(\hat{F})$, that is, if we apply the functional $\theta(\cdot)$ to the edf $\hat{F}$ we get the statistic $\hat{\theta}$.

The parameter θ and the statistic $\hat{\theta}$ are both found by using the functional $\theta(\cdot)$. For the parameter we have $\theta = \theta(F)$, and for the statistic we have $\hat{\theta} = \theta(\hat{F})$.

This is what we call the *plug-in principle*. To estimate parameter $\theta = \theta(F)$ when we don't know F, we plug-in the empirical distribution function $\hat{F}$ to find the estimator $\hat{\theta} = \theta(\hat{F})$.

Examples of parameters and the plug-in principle

1 **Population mean**

$$\theta = \theta(F) = \mathbb{E}_F X = \int x \mathrm{dF(x)} = \int \mathrm{xf(x)dx}$$

Use the plug-in principle

Here $\delta(x)$ is the Dirac delta function which is defined by its behaviour under integration

$$\int_A \delta(x-a)\mathrm{dx} = \begin{cases} 1 \text{ if } a \in A \\ 0 \text{ if } a \notin A \end{cases}$$

The delta function $\delta(x-a)$ is the derivative of the indicator function $\mathbb{I}_{x \leq a}$.

Examples of parameters and the plug-in principle

2 **Population variance**

$$\theta = \theta(F) = \mathbb{V}\text{ar}_F(X) = \int (x - \mathbb{E}_F(X))^2 \text{dF(x)}$$

Examples of parameters and the plug-in principle

3 Probability

$$\theta = \mathbb{P}_F(X > c) = \int_c^\infty \text{dF(x)}$$

2.3.2 Estimating sampling properties with the bootstrap

For our estimates to be of any value, it is necessary to know their properties, such as the bias or the standard error:

- The bias is defined as

$$\text{bias}(\hat{\theta}) = \mathbb{E}\hat{\theta} - \theta$$

- The standard error is

$$\text{se}(\theta) = \sqrt{[\mathbb{E}(\hat{\theta} - \theta)^2]}$$

If we believed that X_i were from a specific parametric model

$$\text{e.g. } F = \Phi \text{ so that } X_i \sim N(\mu, \sigma^2)$$

then we could calculate the bias and standard error analytically. If these calculations were difficult or impossible (for example, if $\theta =$ trimmed mean) then we can use simulations from F to estimate the standard error and bias of the statistics.

What if we don't have a parametric model for F?

The bootstrap can be used to estimate the sampling distribution in this case.

The idea is that instead of sampling from the population of interest, i.e. from $F(\cdot)$, we instead sample with replacement from the sample $\{x_1, \ldots, x_n\}$, i.e. from $\hat{F}(\cdot)$.

Example 1: Heart-attack study

A controlled, randomized, double-blind study was carried out to investigate whether or not aspirin reduces the risk of heart attacks in healthy middle-aged men. Data from the study is

	heart attacks (fatal plus non-fatal)	subjects
aspirin	104	11037
placebo	189	11034

Heart-attack study -II

Define θ to be the true ratio of proportions of heart attacks in those with aspirin to those with a placebo, the relative risk.

From the data, the estimate of θ suggests that aspirin lowers the risk of a heart attack:

$$\hat{\theta} = \frac{104/11037}{189/11034} = 0.55$$

But how confident can we be?

Can we calculate a confidence interval for θ?

It is possible to derive a parametric confidence interval for θ from theory by assuming that the log relative risk is normally distributed. But what if we've forgotten how, or don't wish to assume normality?

Heart-attack study -III

Bootstrapping enables us to derive these intervals without assuming that the log relative risks are normally distributed:

1. Estimate the probability $\hat{p}_1$ of a patient with aspirin having a heart-attack:
$$\hat{p}_1 = \frac{104}{11037} = 0.00942$$
2. Estimate the probability $\hat{p}_2$ of a patient with a placebo having a heart-attack:
$$\hat{p}_2 = \frac{189}{11034} = 0.0171$$
3. Simulate data for a new experiment: sample r_1 from Binomial$(11037, 0.00942)$ and r_2 from Binomial$(11034, 0.0171)$. The new data is known as a bootstrap sample.
4. Obtain a new estimate of the ratio:
$$\hat{\theta}^*_s = \frac{r_1/11037}{r_2/11034}$$

Heart-attack study -IV

Steps 3 and 4 are then repeated a large number of times, to obtain a sample

$$\{\hat{\theta}^*_1, \ldots, \hat{\theta}_B\}$$

We can then use the 2.5th and 97.5th percentiles of this sample as a 95% confidence interval for θ.

With $B = 10000$, performing this procedure in R gave a 95% interval of $(0.43, 0.69)$ for θ.

We will now formally introduce the bootstrap and look at some examples in detail.

The bootstrap

The basic idea behind the bootstrap is to find properties of statistics $\hat{\theta}$ by resampling from $\hat{F}$ (rather than F).

- If we could generate from F, we could simulate sample data sets $\{X_1^{(i)}, \ldots, X_n^{(i)}\}$ for $i = 1, \ldots, B$ and find $\hat{\theta}^{(i)}$. We can then learn properties of $\hat{\theta}$ from $\{\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}\}$.

But usually we don't know F and so can't produce these samples. Instead we can bootstrap. This involves two ideas:

(i) Replace F by $\hat{F}$.

(ii) We sample from $\hat{F}$ and find the properties of $\hat{\theta}$ under $\hat{F}$.

The bootstrap

The bootstrap algorithm

1. Generate B bootstrap replicates from $\hat{F}$.

$$\mathbf{X}^{*(i)} = \{X_1^{*(i)}, \ldots, X_n^{*(i)}\} \text{ for } i = 1, \ldots, B$$

2. Calculate B bootstrap parameter estimates

$$\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$$

3. Calculate the property of interest for $\hat{\theta}$ from $\{\hat{\theta}_i^*\}_{i=1}^B$ e.g.

$$\text{se}_{boot}(\hat{\theta}) = \sqrt{\mathbb{E}_F(\hat{\theta} - \theta)^2} \approx \sqrt{\frac{1}{B-1}\sum(\hat{\theta}_i^* - \bar{\hat{\theta}})^2}$$

where $\bar{\hat{\theta}} = \frac{1}{B}\sum \hat{\theta}_i^*$

The bootstrap

We call iid samples of size n from $\hat{F}$ *bootstrap replicates*.

They can be generated by sampling with replacement from $\{x_1, \ldots, x_n\}$.

In R this can be achieved by using the command

```
sample(n=1, size=n, data=x, replace=T)
```


The bootstrap estimate of standard error

Suppose $\hat{\theta}(X)$ is some statistic based on $X = \{x_1, \ldots, x_n\}$ used for estimating parameter θ. The standard error of $\hat{\theta}(X)$ is

$$\text{se}(\hat{\theta}) = \sqrt{\mathbb{V}\text{ar}_F(\hat{\theta}(X))}$$

Here the variance is with respect to distribution F. The bootstrap estimate is found by

ii replacing F with $\hat{F}$.

$$\text{se}_F(\hat{\theta}) \overset{O_p(\frac{1}{\sqrt{n}})}{\approx} se_{\hat{F}}(\hat{\theta})$$

iiii Approximating $se_{\hat{F}}$ using simulation:

$$se_{\hat{F}}(\hat{\theta}) \overset{O_p(\frac{1}{\sqrt{B}})}{\approx} \left(\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}(X^{(b)}) - \bar{\hat{\theta}})^2\right)^{\frac{1}{2}} =: \text{se}_{boot}$$

where $X^{*(b)} = \{X_1^{*(b)}, \ldots, X_n^{*(b)}\}$ is a bootstrap sample from $\{x_1, \ldots, x_n\}$ ie, se^2_{boot} is the variance of $\hat{\theta}(X^*)$ when X^* are drawn from $\hat{F}$.

The bootstrap estimate of standard error - II

To make this more explicit, note that the first step is using the plug-in principle again.

If we consider the variance of $\hat{\theta}$ to be a functional of F -

$$\mathbb{V}\text{ar}(\hat{\theta})[F] = \mathbb{E}_F(\hat{\theta} - \mathbb{E}_F(\hat{\theta}))^2$$

then when we plug-in $\hat{F}$ we find

$$\mathbb{V}\text{ar}_{\hat{F}}(\hat{\theta}) = \mathbb{E}_{\hat{F}}(\hat{\theta} - \mathbb{E}_{\hat{F}}\hat{\theta})^2.$$

The bootstrap estimate of standard error - III

The second step is then to estimate se_{boot} by simulation by replacing $\mathbb{V}\text{ar}_{\hat{F}}(\hat{\theta}(X^*))$ with an estimate

$$\mathbb{V}\text{ar}_{\hat{F}}(\hat{\theta}(X^*)) \approx \mathbb{V}\text{ar}_{boot}(\hat{\theta}(X^*)) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}(X^{*(b)}) - \bar{\hat{\theta}})^2$$

where

$$\bar{\hat{\theta}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}(X^{*(b)})$$

and where

$$X^{*(b)} = \{X_1^{*(b)}, \ldots, X_n^{*(b)}\}$$

are B bootstrap replicates from $\hat{F}$.

Bootstrap estimate of bias

The bias of an estimator $\hat{\theta}$ of parameter θ is defined as

$$\text{bias} = \mathbb{E}_F(\hat{\theta}) - \theta$$

i.e., how the mean of the estimator of θ differs from the true value of θ.

An estimate is found by replacing F by $\hat{F}$.

$$\text{bias}_{\hat{F}} = \mathbb{E}_{\hat{F}}(\hat{\theta}) - \hat{\theta}$$

That is, the difference between the expected value and the estimated value.

This, again, is the plug-in principle.

Bootstrap estimate of bias - II

We can estimate $\mathbb{E}_{\hat{F}}(\hat{\theta})$ from bootstrap samples as

$$\mathbb{E}_{\hat{F}}(\theta) \approx \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^*_{(b)}$$

giving us the bootstrap estimator of the bias to be

$$\text{bias}_{\hat{F}}(\hat{\theta}) \approx \text{bias}_{boot}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^*_{(b)} - \hat{\theta}$$

So in general there are two approximation steps in the bootstrap procedure:

1. Replace F by $\hat{F}$.
2. Simulate from $\hat{F}$ to form the estimate of the property of interest.

The error in the first approximation scales with the number of data points (and so is fixed for any given problem). The error in the second approximation scales with B, the number of bootstrap replicates, and so can be controlled.

$$\text{se}_F(\hat{\theta}) \overset{O_p(\frac{1}{\sqrt{n}})}{\approx} se_{\hat{F}}(\hat{\theta}) \overset{O_p(\frac{1}{\sqrt{B}})}{\approx} \text{se}_{boot}(\hat{\theta})$$

Here $Y_n = O_p(x_n)$ means that Y_x/x_n is stochastically bounded, i.e., for any $\epsilon > 0$, there exists $M > 0$, such that for all n

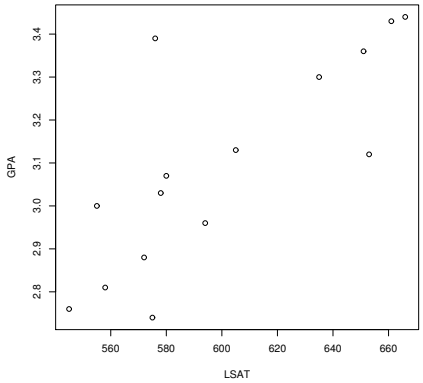
$$\mathbb{P}(|Y_n/x_n| > M) < \epsilon$$

Lawschool example

A sample of 15 law schools was taken, and two measurements were made for each school:

x_i : LSAT, average score for the class on a national law test
y_i : GPA, average undergraduate grade-point average for the class

We are interested in the correlation coefficient between these two quantities, which we estimate to be $\hat{\theta} = 0.776$.



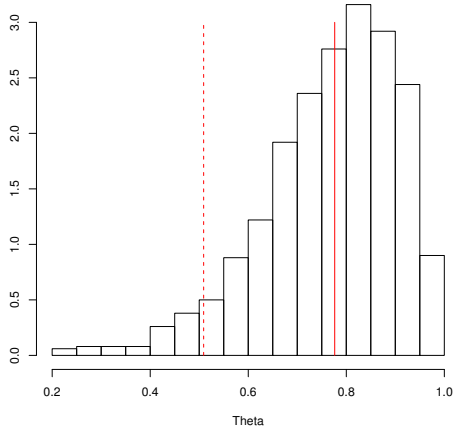
Lawschool example - II

But how accurate is this estimate of the correlation coefficient? We use the bootstrap to estimate the standard error of $\hat{\theta} = \text{cor}(LSAT, GPA)$.

1. Sample 15 data points with replacement from the observed data $z = \{(x_1, y_1), \ldots, (x_{15}, y_{15})\}$ to obtain new data z^*.
2. Evaluate the sample correlation coefficient $\hat{\theta}^*$ for the newly sampled data z^*.
3. Repeate steps 1 and 2 to obtain $\hat{\theta}^{*(1)}, \ldots, \hat{\theta}^{*(B)}$.
4. Estimate the standard error of the sample correlation coefficient by the sample standard deviation of $\hat{\theta}^{*(1)}, \ldots, \hat{\theta}^{*(B)}$.

Lawschool example - III

With $B = 1000$, I found the estimated standard error of $\hat{\theta}$ is 0.137. It can help to plot a histogram of the bootstrap replicates as this gives more information about the distribution of $\hat{\theta}$.



Bootstrap confidence interval

We have two methods of calculating CIs.

1. **Normal interval** Given an estimate of the standard error of $\hat{\theta}$, if we assume that the distribution of $\hat{\theta}$ is approximately normal, then an approximate 95% confidence interval is given by

$$\hat{\theta} \pm 1.96\text{se}(\hat{\theta}^*)$$

   - For the law dataset we find a 95% CI for $\text{cor}(LSAT, GPA)$ of $[0.51, 1.04] \equiv [0.51, 1.00]$.

   This interval is not accurate unless the distribution of bootstrap samples is approximately normal.

Bootstrap confidence interval - II

2. **Percentile confidence interval**
   For a 95% confidence interval, we need to find the two values l and u with

$$\mathbb{P}(\hat{\theta}^* > u) = 0.975$$
$$\mathbb{P}(\hat{\theta}^* < l) = 0.025$$

   ie, we need to identify 2.5th and 97.5th percentiles from the distribution of $\hat{\theta}^*$. We can find this from the 2.5th and 97.5th percentiles of the sample $\{\hat{\theta}^{*(1)}, \ldots, \hat{\theta}^{*(B)}\}$. We generally need a larger value of B to get accurate percentile estimates than that required to find an accurate estimate of the standard error.
   - For the law dataset we find a 95% CI of $[0.45, 0.96]$.

Hypothesis testing with the bootstrap

Example: mice survival times

Treatment	94	197	16	38	99	141	23		
Control	52	104	146	10	50	31	40	27	46

Is there a difference between two group means?

- Denote 7 treatment observations by $\mathbf{x} = \{x_1, \ldots, x_7\}$, and 9 control observations by $\mathbf{y} = \{y_1, \ldots, y_9\}$.
- Could perform two-sample t test, assuming normally distributed responses and equal variances in the two groups.
- Define μ_X: population treatment mean, μ_Y: population control mean. For one-sided test of $H_0 : \mu_X = \mu_Y$, observed $p-$value is 0.1405.

Bootstrap hypothesis test

- Alternative to assuming normality
- Denote F_X: distribution of treatment survival time, F_Y: distribution of control survival time.
- Write null hypothesis as $H_0 : F_X = F_Y = F$, with F the single common distribution of all the responses.
- estimate F by $\hat{F}$, empirical cdf of all 16 observations.

Bootstrap two-sample significance test

1. Sample 16 values with replacement from $\{x_1, \ldots, x_7, y_1, \ldots, y_9\}$.
2. Set $\{x_1^*, \ldots, x_7^*\}$ to be the first 7 sampled values, and $\{y_1^*, \ldots, y_9^*\}$ to be the remaining 9 sampled values.
3. Calculate the bootstrap test statistic

$$T^* = \frac{\bar{x}^* - \bar{y}^*}{\hat{\sigma^*}\sqrt{1/7 + 1/9}}$$

   for the sampled data.
4. Repeat steps 1 to 3 B times to obtain $T^*_{(1)}, \ldots, T^*_{(B)}$.
5. Estimate the significance of the observed T_{obs} by

$$\frac{1}{N}\sum_{i=1}^{N} I\{T^*_{(i)} \geq T_{obs}\} \qquad (5)$$

The Parametric bootstrap

Thus far we have been using the **non-parametric bootstrap**, ie,

- sample from $\hat{F}$ making no assumptions about the distribution of the data.

The **parametric bootstrap** can be used when we believe $F = F_\theta$, i.e. we have a parametric model for the data.

Then instead of sampling from $\hat{F}$, we sample from $F_{\hat{\theta}}$.

In the mice example, we would replace step 1. on slide 77 by

1a. Estimate population mean $\hat{\mu}$ and variance $\hat{\sigma}^2$, e.g.,

$$\hat{\mu} = \frac{1}{16}(\sum x_i + \sum y_i)$$

1b. Sample 16 values from a $N(\hat{\mu}, \hat{\sigma}^2)$ distribution.

Steps 2-5 remain unchanged.

The Bootstrap and Regression

A formal regression type model has the structure

$$y_i = f(x_i, \beta) + \epsilon_i$$

where f is a specified function acting on the covariates x_i with parameters β, and ϵ_i is a realisation from a specified error structure. With this model framework, there are two alternative ways to bootstrap the model.

1. Fit the regression model, form the empirical distribution of the residuals, generate bootstrap replications of the data by substituting these back into the model, and re-fit the model to obtain bootstrap distributions of β. This is called *model-based resampling*.
2. Bootstrap from the pairs (x_i, y_i), re-fit the model to each realization, form the bootstrap distribution of β.

Model-based resampling

We will fit a model of the form

$$GPA_i = \beta_0 + \beta_1 LSAT_i + \epsilon_i$$

to the law data. A least squares fit to these data gives $\hat{\beta}_0 = 0.3794$ and $\hat{\beta}_1 = 0.0045$, but how accurate are these values? We can perform the following set of steps to find the standard error of these estimates

Model-based resampling - II

1. Find the fitted residuals

$$\hat{\epsilon}_i = GPA_i - \hat{\beta}_0 - \hat{\beta}_1 LSAT_i$$

2. Sample $\epsilon_1^*, \ldots, \epsilon_{15}^*$ with replacement from $\{\hat{\epsilon}_1, \ldots, \hat{\epsilon}_{15}\}$
3.

$$\text{Set } GPA_i^* = \hat{\beta}_0 + \hat{\beta}_1 LSAT_i + \epsilon_i^*$$

4. Fit the least squares regression to $\{(LSAT_1, GPA_1^*), \ldots, (LSAT_{15}, GPA_{15}^*)\}$ to find estimates $\beta = (\beta_0^*, \beta_1^*)$.
5. Repeat steps 2 to 4 B times to find bootstrap replicates

$$\{(\beta_0^{*(1)}, \beta_1^{*(1)}), \ldots, (\beta_0^{*(B)}, \beta_1^{*(B)})\}$$

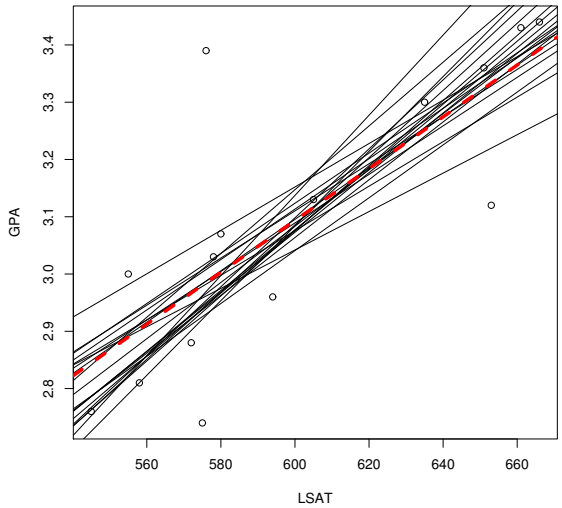
and use these replicates to estimate se$(\hat{\beta}_0)$ and se$(\hat{\beta}_1)$.

Model-based resampling - III

Using 1000 bootstrap replicates, I find the standard errors are

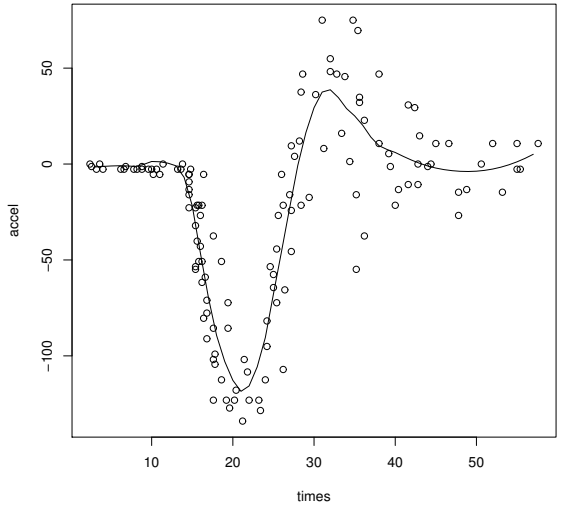
$$\text{se}(\hat{\beta}_0) = 0.586, \qquad \text{se}(\hat{\beta}_1) = 0.000973$$

and the plot shows a sample of 20 bootstrap regression lines.



Motorcycle example

We consider data providing measurements of acceleration against time for a simulated motorcycle accident. The data are shown in the figure.

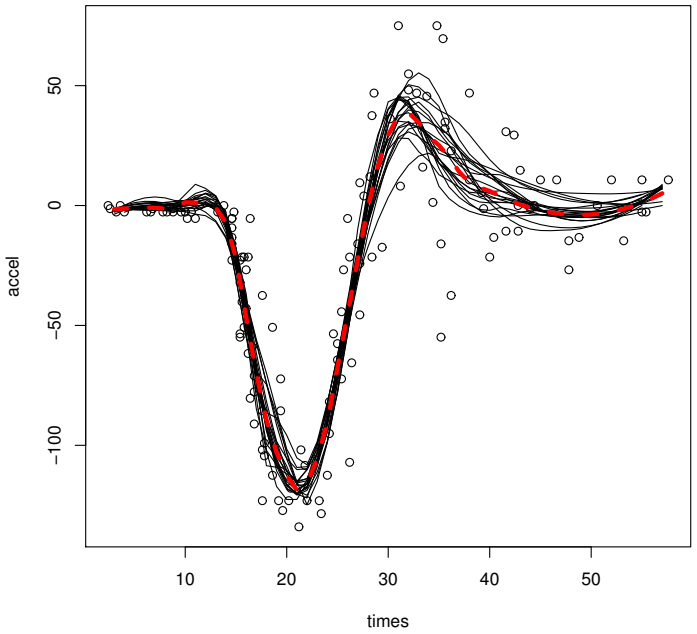


Motorcycle example

Clearly the relationship is nonlinear, and has structure that will not easily be modelled parametrically. We use the `loess()` command in R to fit a locally weighted least-squares regression line to the data (the details aren't important for this course, but for completeness sake we set `span`=1/3 which determines the proportion of the data to be included in the moving window which specifies which points are to be regressed upon.). The figure shows the best fit.

Because of the non-parametric structure of the model, classical approaches to the assessment of parameter precision are not available. We can get a sense of how accurate the parameters are by using a bootstrapping scheme (of the first type). This is achieved by simply bootstrapping the pairs (x,y) in the original plot and fitting `loess` curves to each simulated bootstrap series. A figure showing 20 bootstrap samples is shown below.

Motorcycle example



R code is available in `motorcycle.txt`.

Summary

1. Monte Carlo tests
   - Will work with any test statistic and hypothesis, but requires specification of distribution of data under null hypothesis
   - Only procedure out of three that produces 'completely new' data.
2. Randomization tests
   - Can generally only handle tests of no treatment effect between different treatment groups. One sample tests can be performed, but under stricter assumptions.
   - No distribution is required/assumed for the data, only that allocation of subjects to treatment groups is random.

3 Bootstrapping
   - Arguably the most widely applicable method of the three.
   - Main use is to construct confidence intervals
   - Dependent on the empirical cdf being a good approximation to the true distribution.
   - Accuracy ultimately depends on size of **original** sample.

2.4 Prediction errors and cross-validation

We fit models by minimizing some measure of error, eg, we fit a linear model by minimizing sum of squares

$$S(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^\top \beta)^2$$

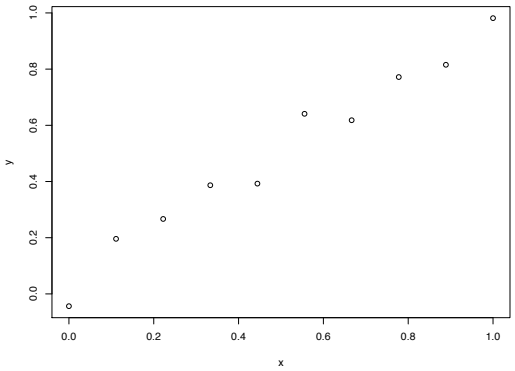
When choosing between competing models we might be tempted to take the model that achieves the lowest error rate on the training data.

However, the error we achieve on the training data is not the same as the error we expect when predicting new data.

We need to be careful when choosing between models not to over-fit and choose a model that is too complex.

Example: Over-fitting

Suppose we are given data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$



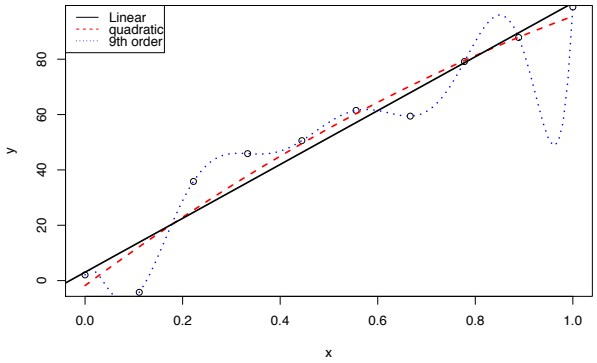
and we want to choose between

$$\mathcal{M}_1 : y = \beta_0 + \beta_1 x + \epsilon \qquad \mathcal{M}_2 : y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

$$\vdots$$

$$\mathcal{M}_d : y = \beta_0 + \beta_1 x + \ldots + \beta_d x^d + \epsilon$$

Example: Over-fitting



The plot shows the fitted curves for $\mathcal{M}_1, \mathcal{M}_2$ and $\mathcal{M}_9$. The residual sum of squares is 676 ($\mathcal{M}_1$), 590 ($\mathcal{M}_2$), and 0 ($\mathcal{M}_9$).

> *With four parameters I can fit an elephant, and with five I can make him wiggle his trunk. John von Neumann*

Example: Over-fitting

$\mathcal{M}_9$ is a perfect fit to the training data - the residual sum of squares is 0.

- With n data points, we can always find a polynomial of degree $n - 1$ that fits perfectly.

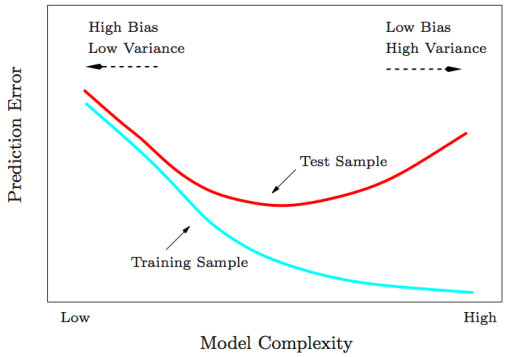
But $\mathcal{M}_9$ is over-fit - it is modelling the noise not the signal and would fail to accurately predict new data.

We know in general that fitting high order polynomials to regression data is not a sensible thing to do, but how can we demonstrate this?

- Some methods adjust the training error to account for the model complexity, e.g., AIC, BIC, C_p statistic.

Alternatively, in data-rich environments, we can simply split the data into a training set, and a test set. We fit the model on the training-set, and then test its predictive accuracy on the test set.

Training vs test set performance



Making a model more complex will **always** result in a better fit to the training data. But there is a **bias-variance** trade-off

- bias occurs from errors in the model structure, ie, from models that are too simplistic
- variance occurs from needing to estimate parameters - for complex models with many parameters, fitting can be sensitive to small fluctuations in the training set leading us to fit the noise rather than the signal.

Cross-validation

Cross-validation is means of efficiently assessing **predictive accuracy**, and extends the idea of having a test and training datasets.

Leave-one-out cross-validation (LOO-CV) For $i = 1, \ldots, n$

1. Fit the model to the reduced data set (or training set),

$$\{(x_1, y_1), \ldots, (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \ldots, (x_n, y_n)\}$$

2. Obtain from the fitted model the predicted value $\hat{y}_i$ at x_i.
3. Compute the squared error $\epsilon_i = (\hat{y}_i - y_i)^2$

An average squared prediction error can then be reported as $\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2$, or the root-mean-square (rms) prediction error as $\sqrt{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2}$.

All predictions are on held-out data (test data) and so this gives us a measure of a model's predictive skill.

k-fold cross-validation

Note that this is not the expected prediction error of the actual model (as we have only fit to $n-1$ data points), though it should be close if n is sufficiently large (so that the fit to $n-1$ points is very similar to the fit to n points).

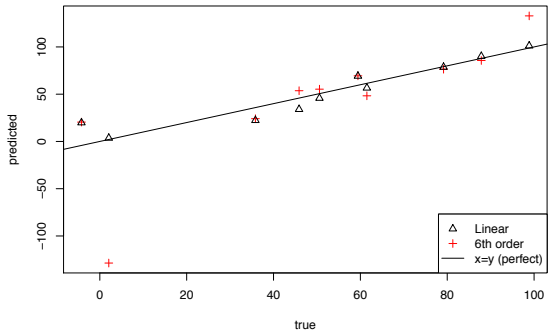
This approach left out one point at a time, and is called **leave-one-out cross-validation**.

K-fold cross-validation splits the data into K chunks of approximately equal size. Then for $k = 1, \ldots, K$,

- Delete chunk k from the data
- Fit the model to the rest of the data
- Use the fitted model to predict the data in chunk k and compute the prediction error.

Setting $K = n$ gives leave-one-out cross-validation!

If we do LOO-CV for our example, then we can plot the predicted value against the true observed value for different models. A perfect model would have predictions on the line $y = x$.



We can see that linear regression is much better in terms of predictive performance than the 6th degree polynomial.

The mean square prediction error for the straight line is 1065.8 whereas for the 8th order polynomial it is 1936.

What K should we use in K-fold cross-validation?

There is a variance-bias trade-off here too!

- The variance of our estimate of the predictive error grows as K gets larger
  - For large K, e.g. LOO-CV with $K = n$, the data doesn't typically get *shaken up* enough. In LOO-CV each fold only differs by two data points, and so estimates from each fold are highly correlated. Hence our estimate of the average prediction error has a high variance (ie is unreliable).
  - For small K, the folds are very different, and so the error estimates are less correlated, and we get a stable estimate.
- The bias of our estimate of the predictive error shrinks as K gets larger
  - Since each training set contains only $\frac{K-1}{K}n$ data points, rather than n, the estimate of the prediction error is usually biased upwards (ie is too large)
  - The bias is minimized for $K = n$, but this has high variance.

$K = 5$ or $K = 10$ are both usually considered good choices but it can vary between applications.

The `cvTools` package in R can be used to do cross-validation.

Chapter III
Simulating random variables

- Inference techniques used so far have been based on **simulation**
- We now consider how to simulate X from $f_X(x)$.
- In semester 1 used MCMC - but simpler methods are needed in order to do MCMC.
- Starting point: generate U from $U[0,1]$ distribution
- Then consider transformation $g(U)$ to obtain a random draw from $f_X(x)$.

How could we generate $U[0,1]$ r.v.s with coin tosses?

Sampling from $U(0,1)$

Need to simulate independent random variables uniformly distributed on $[0,1]$.

Definition: A sequence of pseudo-random numbers $\{u_i\}$ is a deterministic sequence of numbers in $[0,1]$ having the same statistical properties as a similar sequence of random numbers. Ripley 1987.

The sequence $\{u_i\}$ is reproducible provided u_1 is known.

A good sequence would be "unpredictable to the uninitiated".

Congruential generators (D.H. Lehmer, 1949)

The general form of a congruential generator is

$$N_i = (aN_{i-1} + c) \bmod M,$$
$$U_i = N_i/M, \text{ where integers } a, c \in [0, M-1]$$

If $c = 0$, it is called a *multiplicative congruential generator* (otherwise, *mixed*).
These numbers are restricted to the M possible values

$$0, \quad \frac{1}{M}, \quad \frac{2}{M}, \quad \ldots, \quad \frac{M-1}{M}.$$

Clearly, they are *rational* numbers, but if M is large they will practically cover the reals in $[0,1]$.

N_1: the **seed**. Can be re-set so you can reproduce same set of uniform random numbers. In R, use `set.seed(i)`, where i an integer.

As soon as some N_i repeats, say, $N_i = N_{i+T}$, then the whole subsequence repeats, i.e. $N_{i+t} = N_{i+T+t}$, $t = 1, 2, \ldots$.
The least such T is called the *period*.
A good generator will have a long period.

The period cannot be longer than M and also depends on a and c.
Several useful Theorems exist concerning periods of congruential generators. For example, for $c > 0$, $T = M$ if and only if

1. c and M have no common factors (except 1),
2. $1 = a \pmod p$ for every prime number that divides M,
3. $1 = a \pmod 4$ if 4 divides M.

Usually M is chosen to make the modulus operation efficient, and then a and c are chosen to make the period as long as possible. Ripley suggests $c = 0$ or $c = 1$ is usually a good choice.

The NAG Fortran Library G05CAF

$$M = 2^{59} \qquad a = 13^{13} \qquad c = 0$$

Another recommended one is

$$M = 2^{32} \qquad a = 69069 \qquad c = 1.$$

so that

$$N_i = (69069 N_{i-1} + 1) \text{mod } 2^{32}$$

and

$$U_i = 2^{-32} N_i$$

Lattice structure

Notice that for a congruential generator

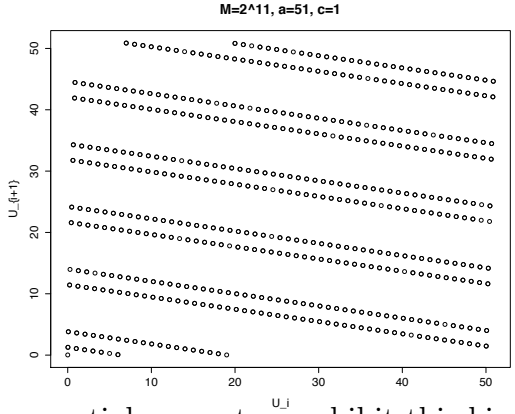
$$N_i - aN_{i-1} = c - bM,$$

where $b > 0$ is an integer. Therefore,

$$U_i - aU_{i-1} = \frac{c}{M} - b.$$

The LHS lies in $(-a, 1)$ since $U_i \in [0, 1)$.
Therefore, b can take at most $a + 1$ distinct values.

If we plot points (U_{i-1}, U_i), all the points will lie on at most $a + 1$ parallel lines.



All linear congruential generators exhibit this kind of lattice structure, not just for pairs (U_{i-1}, U_i), but also for triples (U_{i-2}, U_{i-1}, U_i), and in higher dimensions.

A good generator is expected to have *fine lattice structure*, that is, points $(U_{i-k+1}, \ldots, U_{i-1}, U_i) \in [0, 1)^k$ must lie on many hyperplanes in $\mathbb{R}^k$ for all small k ($k \ll M$).

RANDU - lattice structure

$M = 2^{31}$, $a = 2^{16} + 3 = 65539$, and $c = 0$.

Once very popular, RANDU has eventually been found out to be a rather poor generator.

RANDU - lattice structure II

Let $U_i = N_i/m$ then for this generator

$$U_{i+2} - 6U_{i+1} + 9U_i = k \text{ an integer.}$$

Since $0 \leq U_i < 1$

$$-6 < U_{i+2} - 6U_{i+1} + 9U_i < 10.$$

Therefore $k = -5, -4, \ldots, -1, 0, +1, \ldots, 9$.
Hence k can take on 15 integer values only, and subsequently (U_{i-2}, U_{i-1}, U_i) must lie on at most 15 parallel planes.

This is an example of *coarse lattice structure*, unsatisfactory coverage of $[0, 1)^3$.

Generation from non-$U(0, 1)$

We have a sequence $U_1, U_2, U_3, \ldots$ of independent uniform random numbers in $[0, 1]$.

We want $X_1, X_2, \ldots$ distributed independently and identically from some specified distribution.
The answer is to transform the $U_1, U_2, \ldots$ sequence into $X_1, X_2, \ldots$ sequence.
The idea is to find a function $g(U_1, U_2, U_3, \ldots)$ that has the required distribution.
There are always many ways of doing this. A good algorithm should be quick because millions of random numbers may be required.

3.2 The inversion method

Let X be any continuous random variable and define $Y = F_X(X)$, where F_X is the distribution function of X: $F_X(x) = P(X \leq x)$.
Claim: $Y \sim U[0, 1]$.
Proof $Y \in [0, 1]$ and the distribution function of Y is

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P\big(F_X(X) \leq y\big) \\ &= P\big(X \leq F_X^{-1}(y)\big) = F_X\big(F_X^{-1}(y)\big) = y \end{aligned}$$

which is the distribution function of a uniform random variable on $[0, 1]$.

So whatever the distribution of X, $Y = F_X(X)$ is uniformly distributed on $[0, 1]$. The inversion method turns this backwards. Let $U = F_X(X)$, then $X = F_X^{-1}(U)$.

- So to generate $X \sim F_X$ take a single uniform variable U, and set $X = F_X^{-1}(U)$.

Example: exponential distribution

Let $X \sim Exp(1/\lambda)$ (mean λ), i.e.

$$f(x) = \lambda^{-1}\mathrm{e}^{-x/\lambda} \quad (x \geq 0)$$

$$F(x) = \int_0^x \lambda^{-1}\mathrm{e}^{-z/\lambda}\,\mathrm{dz} = [-\mathrm{e}^{-\mathrm{z}/\lambda}]_0^{\mathrm{x}} = 1 - \mathrm{e}^{-\mathrm{x}/\lambda}.$$

Set $U = 1 - \mathrm{e}^{-X/\lambda}$ and solve for X

$$X = -\lambda \ln(1-U).$$

Note that $1-U$ is uniformly distributed on $[0,1]$, so we might as well use

$$X = -\lambda \ln U.$$

Question: What are the limitations of the inversion method?

Discrete distributions

The inversion method works for discrete random variables in the following sense.
Let X be discretely distributed with possible values x_i having probabilities p_i. So

$$P(X = x_i) = p_i, \qquad \sum_{i=1}^{k} p_i = 1.$$

Then $F_X(x) = \sum_{x_i \leq x} p_i$ is a step function.
Inversion gives $X = x_i$ if $\sum_{x_j < x_i} p_j < U \leq \sum_{x_j \leq x_i} p_j$ which clearly gives the right probability values.

- Think of this as splitting $[0,1]$ into intervals of length p_i. The interval in which U falls is the value of X.

Question: What problems might we face using this method? Eg Consider a Poisson(100) distribution.

Discrete distributions - example

Let $X \sim \mathrm{Bin}(4, 0.3)$. The probabilities are

$$P(X=0) = .2401, \quad P(X=1) = .4116, \quad P(X=2) = .2646$$
$$P(X=3) = .0756, \quad P(X=4) = .0081.$$

The algorithm says
$X = 0$ if $0 \leq U \leq .2401$,
$X = 1$ if $.2401 < U \leq .6517$,
$X = 2$ if $.6517 < U \leq .9163$,
$X = 3$ if $.9163 < U \leq .9919$,
$X = 4$ if $.9919 < U \leq 1$.

Carrying out the binomial algorithm means the following. Let $U \sim U(0,1)$.

1. Test $U \leq .2401$. If true, return $X = 0$.
2. If false, test $U \leq .6517$. If true, return $X = 1$.
3. If false, test $U \leq .9163$. If true, return $X = 2$.
4. If false, test $U \leq .9919$. If true, return $X = 3$.
5. If false, return $X = 4$.

Discrete distributions - example

Consider the speed of this. The expected number of steps (which roughly equates to speed) is

$$1 \times .2401 + 2 \times .4116 + 3 \times .2646 + 4 \times .0756 + 4 \times .0081$$

$$= 1 + E(X) - 0.0081 = 2.1919$$

To speed things up we can rearrange the order so that the later steps are less likely.

1. Test $U \leq .4116$. If true return $X = 1$.
2. If false, test $U \leq .6762$. If true return $X = 2$.
3. If false, test $U \leq .9163$. If true return $X = 0$.
4. and 5. as before.

Expected number of steps:
$1 \times .4116 + 2 \times .2646 + 3 \times .2401 + 4 \times (0.0956 + 0.0081) = 1.9959$.
Approximate 10% speed increase.

3.3 Other Transformations

(a) If $U \sim U(0,1)$ set $V = (b-a)U + a$ then $V \sim U(a,b)$ where $a < b$.

(b) If Y_i are iid exponential with parameter λ then

$$X = \sum_{i=1}^{n} Y_i = -\frac{1}{\lambda} \sum_{i=1}^{n} \log U_i = -\frac{1}{\lambda} \log \left(\prod_{i=1}^{n} U_i \right)$$

has a $Ga(n, \lambda)$ distribution.

(c) If $X_1 \sim Ga(p,1)$, $X_2 \sim Ga(q,1)$, X_1 and X_2 independent then $Y = X_1/(X_1 + X_2) \sim Be(p,q)$.

(d) Composition: if

$$f = \sum_{i=1}^{r} p_i f_i$$

where $\sum p_i = 1$ and each f_i is a density, then we can sample from f by first sampling I from the discrete distribution $p = \{p_1, \ldots, p_r\}$ and then taking a sample from f_I.

The Box-Müller algorithm for the normal distribution

We cannot generate a normal random variable by inversion, because F_X is not known in closed form (nor its inverse).
The Box–Müller method (1958). Let $U_1, U_2 \sim U[0,1]$. Calculate

$$X_1 = \sqrt{-2 \ln U_1} \cos(2\pi U_2),$$
$$X_2 = \sqrt{-2 \ln U_1} \sin(2\pi U_2).$$

Then X_1 and X_2 are independent $N(0,1)$ variables.
The method is not particularly fast, but is easy to program and quite memorable.

3.4 Rejection Algorithm

Fundamental Theorem of Simulation:
Simulating

$$X \sim f(x)$$

is equivalent to simulating

$$(X, U) \sim U\{(x,u) : 0 < u < f(x)\}.$$

Note that $f(x,u) = \mathbb{I}_{0<u<f(x)}$ so that

$$\int f(x,u)\mathrm{du} = \int_0^{\mathrm{f(x)}} \mathrm{du} = \mathrm{f(x)}$$

as required.

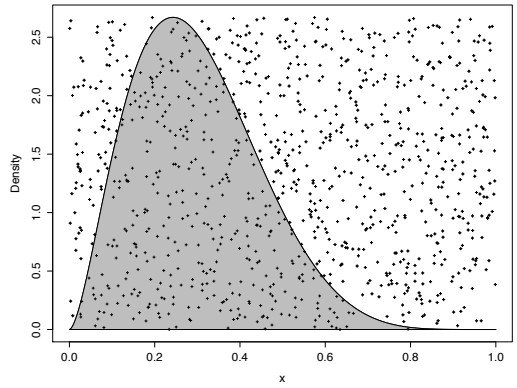
Hence, f is the marginal density of the joint distribution $(X, U) \sim U\{(x,u) : 0 < u < f(x)\}$.

Rejection Algorithm Explained

The problem with this result is that simulating uniformly from the set

$$\{(x,u) : 0 < u < f(x)\}$$

may not be possible. A solution is to simulate the pair (X, U) in a bigger set, where simulation is easier, and then take the pair if the constraint is satisfied.



Rejection: Uniform bounding box

Suppose that $f(x)$ is zero outside the interval $[a, b]$ (so that $\int_a^b f(x)\mathrm{dx} = 1$) and that f is bounded above by m.

- Simulate the pair $(Y, U) \sim U[a, b] \times [0, m]$ ($Y \sim U[a, b]$, $U \sim U[0, m]$ independently).
- Accept the pair if the constraint $0 < U < f(Y)$ is satisfied.

This results in the correct distribution for the accepted Y value, call it X.

$$\begin{aligned}\mathbb{P}(X \le x) &= \mathbb{P}(Y \le x | U < f(Y)) \\ &= \frac{\int_a^x \int_0^{f(y)} \mathrm{dudy}}{\int_a^b \int_0^{f(y)} \mathrm{dudy}} \\ &= \int_a^x f(y)\mathrm{dy}.\end{aligned}$$

Note: we can use the rejection algorithm even if we only know f upto a normalising constant (as is often the case in Bayesian statistics - see chapter 4).

Example: Sampling from a beta distribution

Consider sampling from $X \sim \mathrm{Beta}(\alpha, \beta)$ for $\alpha, \beta > 1$ which has pdf

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \quad 0 < x < 1.$$

We note

$$f(x) \propto f_1(x) = x^{\alpha-1}(1-x)^{\beta-1} \quad 0 < x < 1$$

and that $M = \sup_{0<x<1} x^{\alpha-1}(1-x)^{\beta-1}$ occurs at $x = \dfrac{\alpha - 1}{\alpha + \beta - 2}$ (mode) and hence

$$M = \frac{(\alpha-1)^{\alpha-1}(\beta-1)^{\beta-1}}{(\alpha+\beta-2)^{\alpha+\beta-2}}.$$

The rejection algorithm is

1. Generate $Y \sim U(0, 1)$ and $U \sim U(0, M)$.
2. If $U \le f_1(Y) = Y^{\alpha-1}(1-Y)^{\beta-1}$ then let $X = Y$ (accept) else go to 1 (reject).

Generalising the Rejection Idea

If the support of f is not finite, then bounding it within a rectangle will not work. Instead of using a box to bound the density $f(x)$ (ie requiring $f(x) < m$ for some constant m) we can use a function $m(x)$ such that $f(x) \le m(x)$ for all x.

Suppose the larger bounding set is

$$\mathcal{L} = \{(y, u) : 0 < u < m(y)\}$$

then all we require is that simulation of a uniform from $\mathcal{L}$ is feasible. Note

- The closer m is to f the more efficient our algorithm.
- Because $m(x) \ge f(x)$, m cannot be a probability density. We write

$$m(x) = Mg(x) \text{ where } \int m(x)\mathrm{dx} = \int \mathrm{Mg(x)dx} = \mathrm{M}$$

for some density g.

Generalising the Rejection Idea II

This suggests a more general implementation of the fundamental theorem:

Corollary: Let $X \sim f(x)$ and let $g(x)$ be a density function that satisfies $f(x) \le Mg(x)$ for some constant $M \ge 1$. Then, to simulate $X \sim f$, it is sufficient to generate

$$Y \sim g \qquad \text{and} \qquad U|Y = y \sim U(0, Mg(y))$$

and set $X = Y$ if $U \le f(Y)$.

Proof:

$$\begin{aligned}\mathbb{P}(X \in A) &= \mathbb{P}(Y \in A | U \le f(Y)) \\ &= \frac{\int_A \int_0^{f(y)} \frac{\mathrm{du}}{Mg(y)} g(y)\mathrm{dy}}{\int \int_0^{f(y)} \frac{\mathrm{du}}{Mg(y)} g(y)\mathrm{dy}} \\ &= \int_A f(y)\mathrm{dy}\end{aligned}$$

The Rejection Algorithm

The rejection algorithm is usually stated in a slightly modified form:

Rejection Algorithm

If g is such that f/g is bounded, so there exists M such that $Mg(x) \geq f(x)$ for all x then

1. Generate Y from density g, and U from $U(0,1)$.
2. If $U \leq f(Y)/Mg(Y)$ set $X = Y$. Otherwise, return to step 1.

produces simulations from f

We keep sampling new Y and U until the condition is satisfied.

Exercise: Convince yourself that these two descriptions of the rejection algorithm are the same.

Example: Sampling from a beta distribution revisited

Use rejection to sample from $X \sim \text{Beta}(\alpha, \beta)$. Let $g(y) = \alpha y^{\alpha-1}$, $0 < y < 1$, then

$$\frac{f_1(x)}{g(x)} = \frac{(1-x)^{\beta-1}}{\alpha} \quad \text{is bounded if and only if } \beta \geq 1$$

Then $M = \sup_x \left\{ \frac{f_1(x)}{g(x)} \right\} = \frac{1}{\alpha}$ occurs at $x = 0$.

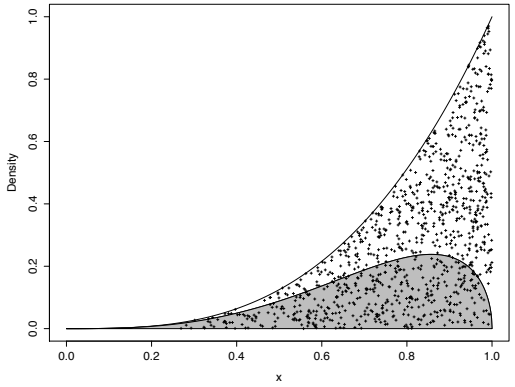
1. Simulate Y with pdf $g(y) = \alpha y^{\alpha-1}$, $0 < y < 1$ and $U \sim U(0,1)$.
2. If $U \leq \dfrac{f_1(Y)}{Mg(Y)} = \dfrac{(1-Y)^{\beta-1}}{\left(\frac{1}{\alpha}\right)\alpha} = (1-Y)^{\beta-1}$ then set $X = Y$ else go to 1.

How to simulate Y with pdf $g(y) = \alpha y^{\alpha-1}$?

- We note that the cdf of Y is $G(y) = y^\alpha$, $0 < y < 1$.
- Therefore we can use inversion. Let $Z \sim U(0,1)$ then solve $Z = G(Y) = Y^\alpha$ and so $Y = Z^{\frac{1}{\alpha}}$.

Full algorithm is:

1. Generate $U \sim U(0,1)$ and $Z \sim U(0,1)$. Let $Y = Z^{\frac{1}{\alpha}}$.
2. If $U \leq (1-Y)^{\beta-1}$ then set $X = Y$ else go to 1.



Efficiency of the rejection method

Each time we generate a (Y, U) pair,

$$\text{Prob(Reject)} = P\big(U \geq f(Y)/Mg(Y)\big) = 1 - \frac{1}{M}, \quad \text{Prob(Accept)} = \frac{1}{M}.$$

The number of tries until we accept Y is a geometric random variable with expectation M.

Note that M here must be calculated with the normalised density f, i.e., $M = \sup \frac{f(x)}{g(x)}$.
If we used an unnormalised density $f_1(x)$, where $\int f_1(x)\mathrm{d}x = c$, so that $f(x) = \frac{1}{c} f_1(x)$, then if we used

$$M = \sup \frac{f_1(x)}{g(x)}$$

the acceptance rate is

$$\mathbb{P}(\text{Accept}) = \frac{c}{M}$$

For maximum efficiency, we want M as small as possible, i.e. $\sup f(x)/g(x)$ as small as possible. This means finding a g that

(a) we can sample from efficiently, and

(b) mimics f as closely as possible.

There are many good generators based on rejection from a well-chosen g.

Rejection Example III

Let θ have von Mises distribution with pdf

$$f(\theta) = \frac{\exp(k\cos\theta)}{2\pi I(k)} \quad 0 < \theta < 2\pi \quad (k \geq 0)$$

where $I(k)$ is the normalising constant.

Let $f_1(\theta) = \frac{1}{2\pi}\exp(k\cos\theta)$, $0 < \theta < 2\pi$.

$Y \sim U(0, 2\pi)$ so that $g(y) = \frac{1}{2\pi}$, $0 < y < 2\pi$.

Then

$$M = \sup_\theta \left\{ \frac{f_1(\theta)}{g(\theta)} \right\} = \sup_\theta \{\exp(k\cos\theta)\} = \exp k.$$

Let $U \sim U(0, 1)$.

If

$$U \leq \frac{f_1(Y)}{Mg(Y)} = \frac{\exp(k\cos Y)}{2\pi \cdot \frac{1}{2\pi} \cdot \exp k} = \exp\left(k(\cos Y - 1)\right)$$

we accept $\theta = Y$ otherwise reject.

Truncated distributions

Suppose we wish to sample X from the following distribution:

$$f_X(x) \propto \begin{cases} g_X(x) & \text{for } x \in A \\ 0 & \text{otherwise} \end{cases}$$

where $g_X(x)$ is a known density that we can sample from, e.g. $g_X(x)$ is the $N(0, 1)$ density, and $A = [0, \infty)$.

$$f_X(x) = \begin{cases} k\, g_X(x) & \text{for } x \in A \\ 0 & \text{otherwise} \end{cases}$$

where k is a normalising constant, given by

$$k^{-1} = \int_A g_X(x)dx$$

$$f_X(x) \propto \begin{cases} g_X(x) & \text{for } x \in A \\ 0 & \text{otherwise} \end{cases}$$

Consider using rejection method to sample X from $f_X(x)$. We sample Y from the full (non-truncated) density $g_X(x)$.

$$\frac{f_X(x)}{g_X(x)} = \begin{cases} k & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

So $M = \sup_x \frac{f_X(x)}{g_X(x)} = k$.
Rejection algorithm: sample u from $U[0,1]$ and y from $g_Y(y)$, and accept $X = y$ if $u \leq \frac{f_X(y)}{M\, g_Y(y)}$.
But since

$$\frac{f_X(x)}{M\, g_X(x)} = \begin{cases} \frac{f_X(x)}{k\, g_X(x)} = 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

we will always have $u \leq \frac{f_X(y)}{M\, g_Y(y)}$ if $y \in A$, and $u \geq \frac{f_X(y)}{M\, g_Y(y)}$ if $y \notin A$.
So we don't need to sample u. Can just do

1. generate y from $g_Y(y)$
2. if $y \in A$, accept $X = y$
3. otherwise, return to step 1.

As usual, acceptance probability will be high if M is small, i.e $\int_A g_Y(y)\, dy$ is near 1. So if the truncated region is large, rejection sampling will be inefficient.

3.5 Multivariate generators

Now suppose we want to generate a random vector $\boldsymbol{X} = (X_1, \ldots, X_p)$ from density $f(\boldsymbol{x})$. We can note the following simple points.

1. If the elements of $\boldsymbol{X}$ are to be independent, i.e.

   $$f(\boldsymbol{x}) = f_1(x_1) f_2(x_2) \ldots f_p(x_p),$$

   then we can separately generate X_1 from f_1, X_2 from $f_2, \ldots, X_p$ from f_p using different uniforms.
2. Inversion no longer works as the theorem can't be generalised.
3. Rejection *does* work. If we can generate from $g(\boldsymbol{x})$ (and g may be a product of independent components) and find $M \geq \sup_{\boldsymbol{x}} \frac{f(\boldsymbol{x})}{g(\boldsymbol{x})}$ and otherwise reject.

Sequential methods

We can obviously write

$$f(\boldsymbol{x}) = f_1(x_1) f_2(x_2|x_1) f_3(x_3|x_1, x_2) \ldots .$$

So we can first generate X_1 from f_1. Then for that given value of X_1, generate X_2 from f_2, and so on.

Example

Suppose we wish to sample $\{x_1, x_2\}$ from the density function

$$f(\theta, \phi) \propto x_2^{-\frac{1}{2}} x_2^{-(\alpha+1)} \mathrm{e}^{-\frac{2\beta + \lambda(x_1-\mu)^2}{2x_2}}$$

Firstly, consider the marginal distribution of x_1

$$f(x_1|x_2) \propto \mathrm{e}^{-\frac{\lambda(x_1-\mu)^2}{2x_2}}$$

as we can ignore factors not depending on x_1.

Thus we can recognise that

$$f(x_1|x_2) \sim N(\mu, \frac{x_2}{\lambda})$$

Next consider the marginal of x_2

$$\begin{aligned} f(x_2) &\propto \int f(x_1, x_2)\mathrm{d}x_1 \\ &\propto x_2^{-\frac{1}{2}} x_2^{-(\alpha+1)} \mathrm{e}^{-\frac{\beta}{x_2}} \left(\frac{x_2}{\lambda}\right)^{\frac{1}{2}} \\ &\propto x_2^{-(\alpha+1)} \mathrm{e}^{-\frac{\beta}{x_2}} \end{aligned}$$

where the term on the right in rd is the missing constant from the $N(\mu, \frac{x_2}{\lambda})$ distribution.

We can recognise this as an inverse gamma distribution $x_2 \sim \Gamma^{-1}(\alpha, \beta)$.

So to simulate random variables from f we can first simulate x_2 from an inverse-Gamma distribution (e.g. by rejection sampling) and then simulate $x_1 \sim N(\mu, \frac{x_2}{\lambda})$ using, e.g., Box-Muller.

Multivariate normal distributions

How can we generate $\mathbf{X}$ from a $N(\mathbf{m}, V)$ distribution, for some non-diagonal matrix V?

We know how to generate iid $N(0, 1)$ rvs from the Box-Muller algorithm, so perhaps we can take a sequence of independent standard normal random variables $Z_1, Z_2, \ldots$ and transform these in some way?

One technique involves the use of the **Cholesky square root** of the matrix V. For any (symmetric, square) positive definite matrix V, we can find a square root U (called the Cholesky decomposition), such that $U^T U = V$.

To find the Cholesky square root of a matrix V in R, type `chol(V)`.

Multivariate normal distributions II

$$\text{Set } \mathbf{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} \quad \text{where } Z_i \sim N(0,1) \quad \text{and} \quad n = \dim \mathbf{X}.$$

Consider

$$\mathbf{Y} = \mathbf{m} + U^T \mathbf{Z}.$$

Then $\mathbf{Y}$ must have a multivariate normally distribution (why?), and

$$\begin{aligned} \mathbb{E}(\mathbf{m} + U^T\mathbf{Z}) &= \mathbf{m}, \\ \mathbb{V}\mathrm{ar}(\mathbf{m} + U^T\mathbf{Z}) &= U^T I_n U = V, \end{aligned}$$

(with I_n the $n \times n$ identity matrix $= \mathbb{V}\mathrm{ar}\mathbf{Z}$).
Hence to generate $\mathbf{X}$, we generate independent standard normal random variables $\mathbf{Z}$, and then transform them by $\mathbf{m} + U^T\mathbf{Z}$ to obtain $\mathbf{X}$.

3.6 Importance sampling

In order to estimate an integral of the form $\int h(x)f(x)\mathrm{d}x$ we find that it is sometimes better to generate values not from the distribution $f(x)$, but instead from some other distribution $g(x)$ and to then account for this by using a weighting. This is the idea behind importance sampling.

To introduce the idea we consider a simple example.

Example of Monte Carlo/Importance Sampling

Let X be Cauchy $f(x) = \dfrac{1}{\pi(1+x^2)}$, $-\infty < x < \infty$.

Let $\theta = P(X > 2) = I = \displaystyle\int_2^\infty \frac{1}{\pi(1+x^2)}\,\mathrm{dx}$ $(= 0.1476)$.

Use Monte Carlo Methods to estimate θ.

(i) Generate n Cauchy variates, $X_1, \ldots, X_n$.
Let Y_1 be the number that are greater than 2,
$Y_1 = \sum \mathbb{I}_{X_i > 2}$. Then $Y_1 \sim B(n, \theta)$ so that

$$E(Y_1) = n\theta, \quad V(Y_1) = n\theta(1-\theta)$$

$$\hat{\theta}_1 = \frac{Y_1}{n}$$

$$E(\hat{\theta}_1) = \frac{E(Y_1)}{n} = \frac{n\theta}{n} = \theta$$

and

$$V(\hat{\theta}_1) = \frac{V(Y_1)}{n^2} = \frac{n\theta(1-\theta)}{n^2} = \frac{\theta(1-\theta)}{n} = \frac{0.126}{n}.$$

Example of Monte Carlo/Importance Sampling - II

(ii) Note that $\theta = \frac{1}{2}P(|X| > 2)$ - we want to use this to reduce the variance of our estimator $\hat{\theta}$.
Generate n Cauchy variates.
Let Y_2 be the number that are greater than 2 in modulus then $Y_2 \sim B(n, 2\theta)$
and $\hat{\theta}_2 = \frac{1}{2}\dfrac{Y_2}{n}$

$$\implies E(\hat{\theta}_2) = \tfrac{1}{2}\frac{E(Y_2)}{n} = \tfrac{1}{2} \cdot \frac{n2\theta}{n} = \theta$$

and

$$V(\hat{\theta}_2) = \frac{V(Y_2)}{2^2 n^2} = \frac{n2\theta(1-2\theta)}{2^2 n^2} = \frac{\theta(1-2\theta)}{2n} = \frac{0.052}{n}.$$

Example of Monte Carlo/Importance Sampling - III

(iii) The relative inefficiency of these methods is due to generation of values outside the domain of interest $[2, \infty)$.
Alternatively note we can write

$$\theta = \frac{1}{2} - \int_0^2 \frac{1}{\pi(1+x^2)}\,\mathrm{dx}.$$

This integral can be considered the expectation of $h(X) = \frac{2}{\pi(1+x^2)}$ where $X \sim U[0,2]$ as the density of $U[0,2]$ is $g(x) = 1/2$.
An alternative method of evaluation of θ is therefore

$$\hat{\theta}_3 = \frac{1}{2} - \frac{1}{n}\sum_{i=1}^{n} h(U_i)$$

where $U_i \sim U[0,2]$.

Example of Monte Carlo/Importance Sampling - IV

We can see that

$$\mathbb{E}(\hat{\theta}_3) = \frac{1}{2} - \frac{1}{n}\sum_{i=1}^{n}\int_0^2 \frac{2}{\pi(1+x^2)}\mathrm{dx} = \frac{1}{2} - \mathbb{P}(0 < \mathrm{X} < 2)$$

where $X \sim$ Cauchy, so that it too is an unbiased estimator.

The variance of $\hat{\theta}_3$ is $\mathbb{V}\mathrm{ar}(h(U))/n$ and we can see that

$$\begin{aligned}
\mathbb{E}h(U) &= \int_0^2 h(x)\frac{1}{2}\mathrm{dx} = 0.5 - 0.1475 = 0.3525 \\
\mathbb{E}h(U)^2 &= \int_0^2 h(x)^2\frac{1}{2}\mathrm{dx} = \int_0^2 \frac{2}{\pi^2(1+\mathrm{x}^2)^2}\mathrm{dx} \\
&= \frac{1}{\pi^2}\left[\frac{x}{x^2+1} + \tan^{-1}(x)\right]_0^2 = 0.1527
\end{aligned}$$

Hence $\mathbb{V}\mathrm{ar}(h(x)) = 0.1527 - 0.3525^2 = 0.02851$ and thus

$$\mathbb{V}\mathrm{ar}(\hat{\theta}_3) = \frac{0.02851}{n}$$

Example of Monte Carlo/Importance Sampling - V

(iv) Finally, note that another possibility is to note that if $y = \frac{1}{x}$

$$\theta = \int_{+2}^{\infty} \frac{1}{\pi(1+x^2)}\,\mathrm{dx} = \int_0^{\frac{1}{2}} \frac{\mathrm{y}^{-2}\,\mathrm{dy}}{\pi(1+\mathrm{y}^{-2})} = \int_0^{\frac{1}{2}} \mathrm{h(y)}\,\mathrm{dy}.$$

This can be seen as the expectation of $h(X) = \frac{X^{-2}}{2\pi(1+X^{-2})}$ where $X \sim U[0, \frac{1}{2}]$. We can estimate this as

$$\hat{\theta}_4 = \frac{1}{n}\sum_{i=1}^{n} h(U_i)$$

where $U_1, \ldots, U_n \sim U[0, 1/2]$.
Again, we have $\mathbb{E}\hat{\theta}_4 = \theta$ and now

$$\mathbb{E}h(U)^2 = \int_0^{1/2} h(x)^2 \cdot 2\mathrm{dx} = \frac{1}{4\pi^2}\left[\frac{\mathrm{x}}{\mathrm{x}^2+1} + \tan^{-1}(\mathrm{x})\right]_0^{1/2} = 0.02188$$

$$\text{Hence } \mathbb{V}\mathrm{ar}(\hat{\theta}_4) = \frac{0.02188 - 0.1476^2}{n} = \frac{0.0000955}{n}$$

Summary of Example

We found 4 unbiased estimators of θ, each with a different variance.

$$\mathbb{V}\mathrm{ar}(\hat{\theta}_1) = \frac{0.126}{n} \qquad \mathbb{V}\mathrm{ar}(\hat{\theta}_2) = \frac{0.052}{n}$$

$$\mathbb{V}\mathrm{ar}(\hat{\theta}_3) = \frac{0.02851}{n} \qquad \mathbb{V}\mathrm{ar}(\hat{\theta}_4) = \frac{0.0000955}{n}$$

The best estimator is the one with the smallest variance, namely $\hat{\theta}_4$.
Compared with $\hat{\theta}_1$, the evaluation of $\hat{\theta}_4$ requires $\sqrt{(0.126/0.0000955)} \approx 36$ times fewer simulations to achieve the same precision.
By carefully considering our simulation method we can hope to get more accurate estimates.
Estimate $\hat{\theta}_2$ and $\hat{\theta}_4$ are both types of importance sampling.

Importance Sampling

Consider calculating the integral

$$I = \mathbb{E}_f h(X) = \int h(\boldsymbol{x}) f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

Importance sampling
Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be independently and identically distributed random variables with common density $g(\boldsymbol{x})$.
Define $w(\boldsymbol{x}) = f(\boldsymbol{x})/g(\boldsymbol{x})$, so that

$$\mathbb{E}_g\{h(\boldsymbol{X}_i)w(\boldsymbol{X}_i)\} = \int h(\boldsymbol{x})w(\boldsymbol{x})g(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int \mathrm{h}(\boldsymbol{x})\mathrm{f}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \mathrm{I}.$$

Therefore

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} w(\boldsymbol{X}_i)h(\boldsymbol{X}_i) \qquad (1)$$

is an unbiased estimator of I.

Some comments:

- $g(\boldsymbol{x})$ is called the importance function, and $w(\boldsymbol{X}_i)$ are called the importance weights.
- The sum (1) will converge for the same reasons the Monte Carlo sum does.
- Notice that this sum is valid for any choice of the distribution g, as long as $\mathrm{supp}(f) \subseteq \mathrm{supp}(g)$.
- This is a very general representation that expresses the fact that a given integral is not intrinsically associated with a given distribution.
- Because very little restriction is put on the choice g, we can choose a distribution which is easy to sample from, and one which gives nice properties for the sum.

Cauchy example revisited

We can now understand the estimator $\hat{\theta}_4$ in the Cauchy example. Recall that we want to estimate

$$\mathbb{E}\mathbb{I}_{X>2} = \int h(x)f(x)\mathrm{dx}$$

where $h(x) = \mathbb{I}_{x>2}$ and $f(x) = \frac{1}{\pi(1+x^2)}$.
Noticing that for large x, $f(x)$ is similar to the density

$$g(x) = 2/x^2 \text{ for } x > 2.$$

suggests $g()$ might be a good importance density. We can sample from g by letting $X_i = 1/U_i$ where $U_i \sim U[0, \frac{1}{2}]$ (inversion method). Thus our estimator is

$$\begin{aligned}\hat{\theta} &= \frac{1}{n}\sum_{i=1}^{n} h(x_i)\frac{f(x_i)}{g(x_i)} = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i^2}{2\pi(1+x_i^2)} \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{u_i^{-2}}{2\pi(1+u_i^{-2})} = \hat{\theta}_4\end{aligned}$$

The variance of the estimator

Since the $\boldsymbol{X}_i$s are iid, $\quad \mathbb{V}\mathrm{ar}(\hat{I}) = \frac{\sigma^2}{n}, \quad$ where

$$\begin{aligned}\sigma^2 &= \mathbb{V}\mathrm{ar}_g\{h(\boldsymbol{X})w(\boldsymbol{X})\} = \mathbb{E}\{h(\boldsymbol{X})^2 w(\boldsymbol{X})^2\} - \mathbb{E}\{h(\boldsymbol{X})w(\boldsymbol{X})\}^2 \\ &= \int h(\boldsymbol{x})^2 w(\boldsymbol{x})^2 g(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x} - \mathrm{I}^2 \\ &= \int \frac{h(\boldsymbol{x})^2 f(\boldsymbol{x})^2}{g(\boldsymbol{x})}\ \mathrm{d}\boldsymbol{x} - \mathrm{I}^2 \quad \text{since} \quad \mathrm{g}(\boldsymbol{x}) = \frac{\mathrm{f}(\boldsymbol{x})}{\mathrm{w}(\boldsymbol{x})}.\end{aligned}$$

We do not of course know σ^2 in practice, but we can see that $\hat{I}$ will be a better estimator if we can make $w(\boldsymbol{X})$ less variable. Our objective, therefore, is to find a distribution $g(\boldsymbol{x})$ that we know how to obtain independent samples from, and which mimics $h(\boldsymbol{x})f(\boldsymbol{x})$ as closely as possible.

Optimal choice of g

Theorem The choice of $g = g^* = \frac{|h(x)|f(x)}{\int |h(z)|f(z)\mathrm{dz}}$ minimises the variance of the estimator (1).
Proof We've seen that it is sufficient to minimise

$$\int \frac{h^2(\boldsymbol{x})f^2(\boldsymbol{x})}{g(\boldsymbol{x})}\ \mathrm{d}\boldsymbol{x} = \mathbb{E}_{\mathrm{g}}\left(\frac{\mathrm{h}^2(\mathrm{X})\mathrm{f}^2(\mathrm{X})}{\mathrm{g}^2(\mathrm{X})}\right)$$

and using Jensen's inequality we can see that

$$\begin{aligned}\mathbb{E}_g\left(\frac{h^2(X)f^2(X)}{g^2(X)}\right) &\geq \left(\mathbb{E}_g\left[\frac{|h(X)|f(X)}{g(X)}\right]\right)^2 \\ &= \left(\int |h(x)|f(x)\mathrm{dx}\right)^2\end{aligned}$$

and that this lower bound is achieved by choosing $g = g^*$.
NB: We won't be able to calculate g^*! But the theorem suggests that choosing g to look like hf will be a good choice.

Unnormalised densites

Suppose we only know f upto a normalising constant, i.e., we know

$$f(x) = \frac{f_1(x)}{c} \qquad \text{where } c = \int f_1(x)\mathrm{dx}$$

We can still use importance sampling

Importance sampling with unnormalised densites
Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be independently and identically distributed random variables with common density $g(\boldsymbol{x})$. Define $\tilde{w}(\boldsymbol{x}) = f_1(\boldsymbol{x})/g(\boldsymbol{x})$. Estimate I by

$$\hat{I} = \frac{\sum_{i=1}^n \tilde{w}(\boldsymbol{X}_i)h(\boldsymbol{X}_i)}{\sum_{i=1}^n \tilde{w}(\boldsymbol{X}_i)}$$

Alternatively, we can write this as

$$\hat{I} = \sum_{i=1}^{n} w_i h(\boldsymbol{X}_i) \qquad \text{where} \qquad w_i = \frac{\tilde{w}(X_i)}{\sum \tilde{w}(X_i)}$$

$\frac{1}{n}\sum \tilde{w}(\boldsymbol{X}_i)$ is an unbiased estimator of c as

$$\mathbb{E}_g \tilde{w}(X) = \int \frac{f_1(x)}{g(x)} g(x) \mathrm{d}x = \int f_1(x) \mathrm{d}x = c.$$

When we use unnormalised densities, $\hat{I}$ is a biased estimator of I, however it is possible to prove that we still have $\hat{I} \to I$ almost surely as $n \to \infty$.

This will be important when we use importance sampling to estimate Bayesian quantities.

Effective sample size

How variable the weights are tells us how efficient our choice of g is.

In the best case, where $g = f$, then $\tilde{w}(X) = 1$ so that $w_i = \frac{1}{n}$, which is the case in plain Monte Carlo. In this case $\mathbb{V}\mathrm{ar}(w(X)) = 0$.

If f and g are very different, then the weights will be very variable, and we can find that one or two particles (X_i) dominate the sum.

We often calculate the **effective sample size**

$$ESS = \frac{1}{\sum w_i^2}$$

- In the best case, $w_i = \frac{1}{n}$ and ESS$= n$ - so we have an effective sample size equal to the true sample size.
- The worst case is when one of the $w_i = 1$ and all the others are equal to zero. Then ESS$= 1$, i.e., we effectively have only a single sample.

We want to choose g so that the ESS is large.

3.7 Variance reduction techniques

Antithetic variables

The method of antithetic variables uses two correlated estimators and combines them to get an estimator with a lower variance (i.e. a better estimator).
Suppose we have two different estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ of θ,

- with the same mean and variance
- but which are negatively correlated

Define $\hat{\theta}_3 = \frac{1}{2}(\hat{\theta}_1 + \hat{\theta}_2)$. Then

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(\hat{\theta}_3) &= \frac{1}{4}(\mathbb{V}\mathrm{ar}(\hat{\theta}_1) + \mathbb{V}\mathrm{ar}(\hat{\theta}_2) + 2\mathbb{C}\mathrm{ov}(\hat{\theta}_1, \hat{\theta}_2)) \\
&= \frac{1}{2}(\mathbb{V}\mathrm{ar}(\hat{\theta}_1) + \mathbb{C}\mathrm{ov}(\hat{\theta}_1, \hat{\theta}_2)) \\
&< \frac{1}{2}\mathbb{V}\mathrm{ar}(\hat{\theta}_1)
\end{aligned}$$

This is twice the cost of computing $\hat{\theta}_1$ but the variance is more than halved!

Antithetic variables - II

We need to find two estimators which are negatively correlated. This can be done as follows:

- If $U \sim U[0,1]$ then $1 - U \sim U[0,1]$ also.
- If F is the distribution function of X then $X_1 = F^{-1}(U)$ and $X_2 = F^{-1}(1-U)$ are both distributed according to F
- and $\mathbb{C}\mathrm{ov}(X_1, X_2) < 0$.

Proof (non-examinable):
Let $h(u) = F^{-1}(u)$. Then $h(u)$ is a non-decreasing function.
We need to show

$$\mathbb{E}h(U)h(1-U) \leq (\mathbb{E}h(U))^2$$

Let $Q = \mathbb{E}h(U)$. The since h is non-decreasing on $[0,1]$

$$h(0) \leq Q \leq h(1)$$

Let $f(y) = \int_0^y h(1-x)dx - Qy$ on $[0,1]$
Then $f(0) = f(1) = 0$ and

$$f'(y) = h(1-y) - Q$$

is also a non-increasing function.
Since $f'(0) = h(1) - Q \geq 0$ and $f'(1) = h(0) - Q \leq 0$ we must have

$$f(u) \geq 0 \text{ on } [0,1]$$

Therefore

$$0 \leq \int_0^1 f(y)h'(y)dy = [fh]_0^1 - \int f'h(y)dy$$
$$= -\int_0^1 f'(y)h(y)dy$$

Therefore

$$\int_0^1 f'(y)h(y) = \int_0^1 h(y)(h(1-y)-Q)dy = \int_0^1 h(y)h(1-y)dy - Q^2 \leq 0$$

Hence $\int_0^1 h(y)h(1-y)dy \leq Q^2$ as required.

Cauchy Example Revisited

Above we used

$$\hat{\theta}_3 = \frac{1}{2} - \frac{2}{n}\sum_{i=1}^{n}\left[\frac{1}{\pi(1+u_i^2)}\right]$$

as an estimator of $\mathbb{P}(X > 2)$ where $X \sim Cauchy$.
An estimator with a smaller variance can be found using antithetic variables

$$\frac{1}{2}\left(\frac{1}{2} - \frac{2}{n}\sum_{i=1}^{n}\left[\frac{1}{\pi(1+u_i^2)}\right] + \frac{1}{2} - \frac{2}{n}\sum_{i=1}^{n}\left[\frac{1}{\pi(1+(2-u_i)^2)}\right]\right)$$

which gives

$$\hat{\theta}_{antithetic} = \frac{1}{2} - \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{\pi(1+u_i^2)} + \frac{1}{\pi(1+(2-u_i)^2)}\right]$$

The for $n = 10$ we find the variance of $\hat{\theta}_3$ is 2.7×10^{-4} whereas the variance of $\hat{\theta}_{antithetic}$ is 5.5×10^{-6} - a substantial improvement.

3.8 Bayesian inference

Unnormalised densities frequently occur when we are doing Bayesian inference.

Suppose we are interested in some posterior expectation, for example, the posterior mean:

$$I = \mathbb{E}(\theta|x) = \int \theta f(\theta|x)\mathrm{d}\theta$$

where

$$f(\theta|x) = \frac{f(\theta)f(x|\theta)}{f(x)} \qquad \text{by Bayes theorem.}$$

The denominator $f(x) = \int f(\theta)f(x|\theta)\mathrm{d}x$ is often intractable and unknown, and so we instead work with the unnormalised density

$$f_1(\theta|x) = f(\theta)f(x|\theta) = \text{ prior } \times \text{ likelihood}$$

Rejection sampling for Bayesian inference

You may have seen in MAS364 (or Autumn of MAS6004) how to sample from a posterior distribution using MCMC. We can also use rejection sampling, or estimate posterior expectations using importance sampling.

So to sample posterior samples of θ from $f(\theta|x)$, using proposal density g (assuming $f_1(\theta|x)/g(\theta) \leq M$ for all θ), we can do

1. Simulate $\theta \sim g(\cdot)$
2. Accept θ with probability

$$\frac{f(\theta)f(x|\theta)}{Mg(\theta)}$$

otherwise reject θ.

If we use $g(\theta) = f(\theta)$, ie, use the prior as the proposal, then this reduces to accept θ with probability $\frac{f(x|\theta)}{M}$, but this is usually inefficient (ie, M is large, so the acceptance rate $1/M$ is small).

Importance sampling for Bayesian inference

Suppose we wish to estimate the posterior expectation

$$\mathbb{E}(r(\theta)|\mathbf{x}) = \int r(\theta) f(\theta|\mathbf{x}) \mathrm{d}\theta$$

We could use importance sampling, using the prior distribution as the importance distribution, ie, $g = f$.
If we do not know $f(\mathbf{x})$ then we can use the following importance sampling approach:

- Simulate $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ from the prior $f(\boldsymbol{\theta})$
- Set $\tilde{w}_i = f(\mathbf{x}|\boldsymbol{\theta})$
- Set $w_i = \tilde{w}_i / \sum \tilde{w}_i$ and estimate $\mathbb{E}(r(\boldsymbol{\theta})|\mathbf{x})$ by

$$\sum_{i=1}^{n} w_i r(\boldsymbol{\theta}_i)$$

This is inefficient if the prior is very different to the posterior as we will spend too much time sampling $\boldsymbol{\theta}_i$ where the likelihood is very small, and so the weights $w(\boldsymbol{\theta}_i)$ will also be very small.

If this is the case, then the effective sample size will be small.

Choice of g and the normal approximation

A more efficient alternative to using the prior distribution for g, is to build a normal approximation to the posterior and use this as g

Let $h(\boldsymbol{\theta}) = \log f(\boldsymbol{\theta}|\mathbf{x})$. Now define $\mathbf{m}$ to be posterior mode of $\boldsymbol{\theta}$, so $\mathbf{m}$ maximises both $f(\boldsymbol{\theta}|\mathbf{x})$ and $h(\boldsymbol{\theta})$.

We may need to use numerical optimisation to find $\mathbf{m}$, e.g. using the `optim` command in R.
We can then use a Taylor expansion of $h(\boldsymbol{\theta})$ around $\mathbf{m}$

$$h(\boldsymbol{\theta}) = h(\mathbf{m}) + (\boldsymbol{\theta} - \mathbf{m})^T \mathbf{h}'(\mathbf{m}) + \frac{1}{2}(\boldsymbol{\theta} - \mathbf{m})^T M (\boldsymbol{\theta} - \mathbf{m}) + \ldots$$

to build a Gaussian approximation to the posterior (known as the Laplace approximation).
Here, $h'(\mathbf{m})$ the vector of first derivatives of $h(\boldsymbol{\theta})$, and M the matrix of second derivatives of $h(\boldsymbol{\theta})$, both evaluated at $\boldsymbol{\theta} = \mathbf{m}$.

Since $\mathbf{m}$ maxmises $h(\mathbf{m})$ we have $h'(\mathbf{m}) = \mathbf{0}$. Hence

$$f(\boldsymbol{\theta}|\mathbf{x}) = \exp\{h(\boldsymbol{\theta})\} \simeq \exp\{h(\mathbf{m})\} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta} - \mathbf{m})^T V^{-1} (\boldsymbol{\theta} - \mathbf{m})\right\}, \tag{2}$$

where $-V^{-1} = M$.

Thus, our approximation of $f(\boldsymbol{\theta}|\mathbf{x})$ is a multivariate normal distribution, mean vector $\mathbf{m}$, variance matrix $-M^{-1}$. This will be a good approximation if posterior mass is concentrated around $\mathbf{m}$.

NB: We do not need $f(\mathbf{x})$ to obtain M, since

$$h(\boldsymbol{\theta}) = \log f(\boldsymbol{\theta}|\mathbf{x}) = \log f(\boldsymbol{\theta}) + \log f(\mathbf{x}|\boldsymbol{\theta}) - \log f(\mathbf{x}),$$

so $\log f(\mathbf{x})$ will disappear when we differentiate $h(\boldsymbol{\theta})$.

Assessing convergence

Suppose we wish to estimate $\mathbb{E}\{r(\boldsymbol{\theta})|\mathbf{x}\}$ for some $r(\boldsymbol{\theta})$. If $f(\mathbf{x})$ known, then

$$\hat{\mathbb{E}}\{r(\boldsymbol{\theta})|\mathbf{x}\} = \frac{1}{n} \sum_{i=1}^{n} r(\boldsymbol{\theta}_i) w(\boldsymbol{\theta}_i),$$

and can use central limit theorem to obtain a confidence interval for $\mathbb{E}\{r(\boldsymbol{\theta})|\mathbf{x}\}$, as in MC integration.
We can check our estimate by
1) Increasing the sample size n to check the stability of any estimate.
2) Increasing the standard deviation in the $g(\boldsymbol{\theta})$ density, to check stability to the choice of g, e.g., if we're using a normal approximation, we could multiply V by 4 etc.

Example: leukaemia data

Patients suffering from leukaemia are given a drug, 6-mercaptopurine (6-MP), and the number of days x_i until freedom from symptoms is recorded of patient i:

$$6^*, 6, 6, 6, 7, 9^*, 10^*, 10, 11^*, 13, 16, 17^*,$$

$$19^*, 20^*, 22, 23, 25^*, 32^*, 32^*, 34^*, 35^*.$$

A * denotes censored observation.
Will suppose that time x to the event of interest follows a *Weibull* distribution:

$$f(x|\alpha, \beta) = \alpha\beta(\beta x)^{\alpha-1} \exp\{-(\beta x)^\alpha\}$$

for $x > 0$.
For censored observations, we have

$$P(x > t|\alpha, \beta) = \exp\{-(\beta t)^\alpha\}.$$

Example: leukaemia data

Likelihood

Define

- d: number of uncensored observations,
- $\sum_u \log x_i$: sum of logs of all uncensored observations.

Writing $\boldsymbol{\theta} = (\alpha, \beta)^T$, the log likelihood is then given by

$$\log f(\mathbf{x}|\boldsymbol{\theta}) = d\log\alpha + \alpha d\log\beta + (\alpha-1)\sum_u \log x_i - \beta^\alpha \sum_{i=1}^n x_i^\alpha.$$

Suppose our prior distributions for α and β are both exponential with

$$\begin{aligned} f(\alpha) &= 0.001\exp(-0.001\alpha), \\ f(\beta) &= 0.001\exp(-0.001\beta). \end{aligned}$$

Example: leukaemia data

Building an approximation to the posterior

1) **Obtain the posterior mode of $\boldsymbol{\theta}$**. Maximise log posteior, i.e.

$$h(\boldsymbol{\theta}) = d\log\alpha + \alpha d\log\beta + (\alpha-1)\sum_u \log x_i - \beta^\alpha \sum_{i=1}^n x_i^\alpha - 0.001\alpha - 0.001\beta + K$$

for some constant K.

In R, we can find the mode to be $\mathbf{m} = (1.354, 0.030)$ using the `optim` command.

2) **Derive the matrix of second derivatives of $h(\boldsymbol{\theta})$.**

$$M = \begin{pmatrix} \frac{\partial^2}{\partial\alpha^2}h(\boldsymbol{\theta}) & \frac{\partial^2}{\partial\alpha\partial\beta}h(\boldsymbol{\theta}) \\ \frac{\partial^2}{\partial\alpha\partial\beta}h(\boldsymbol{\theta}) & \frac{\partial^2}{\partial\beta^2}h(\boldsymbol{\theta}) \end{pmatrix},$$

evaluated at $\boldsymbol{\theta} = \mathbf{m}$.

$$\begin{aligned} \frac{\partial^2}{\partial\alpha^2}h(\boldsymbol{\theta}) &= -\frac{d}{\alpha^2} - \sum(\beta x_i)^\alpha(\log(\beta x_i))^2 \\ \frac{\partial^2}{\partial\beta^2}h(\boldsymbol{\theta}) &= \frac{1}{\beta^2}\left\{\beta^\alpha\alpha(1-\alpha)\sum_{i=1}^n x_i^\alpha - d\alpha\right\}, \\ \frac{\partial^2}{\partial\alpha\partial\beta}h(\boldsymbol{\theta}) &= \frac{1}{\beta}\left[d - \beta^\alpha\left\{\alpha\log\beta\sum_{i=1}^n x_i^\alpha + \sum_{i=1}^n x_i^\alpha + \alpha\sum_{i=1}^n x_i^\alpha\log x_i\right\}\right] \end{aligned}$$

$$M = \begin{pmatrix} -31.618 & 175.442 \\ 175.442 & -18806.085 \end{pmatrix}.$$

3) **Obtain the normal approximation to use as $g(\boldsymbol{\theta})$.**
$g(\boldsymbol{\theta})$: bivariate normal, mean $\mathbf{m}$, variance matrix $V = -M^{-1}$:

$$\boldsymbol{\theta} \sim N\left\{ \begin{pmatrix} 1.354 \\ 0.030 \end{pmatrix}, \begin{pmatrix} 0.0334 & 0.0003 \\ 0.0003 & 0.00006 \end{pmatrix} \right\}$$

4) **Sample $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ from $g(\boldsymbol{\theta})$ and compute the importance weights** $w(\boldsymbol{\theta}_1), \ldots, w(\boldsymbol{\theta}_n)$. The weights are given by

$$w(\boldsymbol{\theta}_i) = \frac{\tilde{w}(\boldsymbol{\theta}_i)}{\sum_{i=1}^{n} \tilde{w}(\boldsymbol{\theta}_i)}, \qquad \text{with} \qquad \tilde{w}(\boldsymbol{\theta}_i) = \frac{f(\boldsymbol{\theta}_i) f(\mathbf{x}|\boldsymbol{\theta}_i)}{g(\boldsymbol{\theta}_i)}$$

NB the Gaussian approximation may give us negative samples. Since $\alpha > 0$ and $\beta > 0$, we shoould simply discard negative $\boldsymbol{\theta}$ values, i.e., use a truncated normal density for $g(\theta)$.

Note that when we compute $w(\boldsymbol{\theta}_i)$, it is not necessary to rescale $g(\theta)$ so that it integrates to 1, as any normalising constant in $g(\theta)$ will cancel.

5) **Estimate the posterior mean of $\boldsymbol{\theta}$**
We compute the estimate

$$\hat{E}(\boldsymbol{\theta}|\mathbf{x}) = \sum_{i=1}^{n} \boldsymbol{\theta}_i w(\boldsymbol{\theta}_i).$$

In R, with $n = 100000$, this gives $\hat{E}(\boldsymbol{\theta}|\mathbf{x}) = (1.346, 0.031)^T$.
6) **Check for convergence**
We repeat steps 4 and 5 with more dispersion in $g(\boldsymbol{\theta})$:

$g(\boldsymbol{\theta})$	$\hat{E}(\boldsymbol{\theta}|\mathbf{x})$
$N(\mathbf{m}, V)$	$(1.346, 0.031)^T$
$N(\mathbf{m}, 4V)$	$(1.384, 0.031)^T$
$N(\mathbf{m}, 16V)$	$(1.380, 0.031)^T$

Finally, double the sample size (no effect observed).
For percentiles, we can do resampling in R.

See computer class 5 for more details and code to implement this approach.

Chapter 4
Likelihood-based inference

4.1 Likelihoods

Data $\mathbf{x} = \{x_1, \ldots, x_n\}$, joint distribution of $\mathbf{x}$ depends on unknown θ.
Likelihood is density (or probability if x_i is discrete) of the data x conditional on the parameter θ, i.e.

$$f(\mathbf{x}|\theta).$$

Function of θ *for fixed* $\mathbf{x}$, so denote the likelihood function by $L(\theta; \mathbf{x})$:

$$L(\theta; \mathbf{x}) = f(\mathbf{x}|\theta).$$

If $x_1, \ldots, x_n$ are independent, then
$f(\mathbf{x}|\theta) = f(x_1|\theta) \times f(x_2|\theta) \times \ldots \times f(x_n|\theta)$, and so

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta).$$

Used for point and interval estimation, and hypothesis testing.

Score statistics, Fisher information and the Cramer-Rao minimum variance bound

The score statistic is defined to be $\frac{\partial}{\partial\theta} l(\theta; \mathbf{x}) = \frac{\partial}{\partial\theta} \log f(\mathbf{x}|\theta)$.
$\mathbf{X}$: unobserved value of $\mathbf{x}$. Define the *random variable*

$$\frac{\partial}{\partial\theta} l(\theta; \mathbf{X}) = \frac{\partial}{\partial\theta} \log f(\mathbf{X}|\theta).$$

Transformation of a r.v. $\mathbf{X}$, where transformation is derivative, w.r.t. θ, of the log of the density of $\mathbf{X}$.

N.B. We treat $l(\theta; \mathbf{X})$ as a function of the random data $\mathbf{X}$, *evaluated at the true value of* θ, rather than a function of the parameter θ for fixed data $\mathbf{x}$.

$$\begin{aligned}
\frac{\partial}{\partial\theta} l(\theta; \mathbf{X}) &= \frac{\partial}{\partial\theta} \log L(\theta; \mathbf{X}) \\
&= \left\{ \frac{\partial}{\partial\theta} L(\theta; \mathbf{X}) \right\} \times \frac{1}{L(\theta; \mathbf{X})} = \left\{ \frac{\partial}{\partial\theta} f(\mathbf{X}|\theta) \right\} \times \frac{1}{f(\mathbf{X}|\theta)}.
\end{aligned}$$

$$\begin{aligned}
E\left\{ \frac{\partial}{\partial\theta} l(\theta; \mathbf{X}) \right\} &= \int \left\{ \frac{\partial}{\partial\theta} l(\theta; \mathbf{x}) \right\} f(\mathbf{x}|\theta) d\mathbf{x} \\
&= \int \left\{ \frac{\partial}{\partial\theta} f(\mathbf{x}|\theta) \right\} \times \frac{1}{f(\mathbf{x}|\theta)} f(\mathbf{x}|\theta) d\mathbf{x} \\
&= \frac{\partial}{\partial\theta} \int f(\mathbf{x}|\theta) d\mathbf{x} \\
&= \frac{\partial}{\partial\theta} 1 = 0.
\end{aligned}$$

Expected value of the derivative of the log-likelihood at the true value of θ is 0.
Consider example of $X \sim exp(rate = \theta)$. Then
$l(\theta; X) = \log\theta - \theta X$ and

$$\frac{\partial}{\partial\theta} l(\theta; \mathbf{X}) = \frac{1}{\theta} - X,$$

so

$$\begin{aligned} E\left\{\frac{\partial}{\partial\theta} l(\theta; \mathbf{X})\right\} &= \int \left(\frac{1}{\theta} - x\right) \theta \exp(-\theta x) dx \\ &= \frac{1}{\theta}\int \theta \exp(-\theta x) dx - \int x\theta \exp(-\theta x) dx \\ &= \frac{1}{\theta} - \frac{1}{\theta} = 0. \end{aligned}$$

However, the expected value of the derivative of the log-likelihood evaluated at the *wrong* value of θ, say θ^*, is not 0.
For example,

$$\frac{\partial}{\partial\theta} l(\theta; \mathbf{X})\Big|_{\theta=\theta^*} = \frac{1}{\theta^*} - X,$$

with

$$\begin{aligned} E\left\{\frac{\partial}{\partial\theta} l(\theta; \mathbf{X})\Big|_{\theta=\theta^*}\right\} &= \int \left(\frac{1}{\theta^*} - x\right) \theta \exp(-\theta x) dx \\ &= \frac{1}{\theta^*} - \frac{1}{\theta}, \end{aligned}$$

which is non-zero for $\theta^* \neq \theta$.

To derive an expression for the variance of $\frac{\partial}{\partial\theta} l(\theta; \mathbf{X})$, we note that

$$\begin{aligned} & 0 = \int \frac{\partial}{\partial\theta} l(\theta; \mathbf{x}) f(\mathbf{x}|\theta) d\mathbf{x} \\ \Rightarrow\quad & 0 = \frac{\partial}{\partial\theta}\int \frac{\partial}{\partial\theta} l(\theta; \mathbf{x}) f(\mathbf{x}|\theta) d\mathbf{x} \\ \Rightarrow\quad & 0 = \int \left\{\frac{\partial^2}{\partial\theta^2} l(\theta; \mathbf{x})\right\} f(\mathbf{x}|\theta) d\mathbf{x} + \int \left\{\frac{\partial}{\partial\theta} l(\theta; \mathbf{x})\right\}\left\{\frac{\partial}{\partial\theta} f(\mathbf{x}|\theta)\right\} d\mathbf{x} \\ \Rightarrow\quad & 0 = \int \left\{\frac{\partial^2}{\partial\theta^2} l(\theta; \mathbf{x})\right\} f(\mathbf{x}|\theta) d\mathbf{x} \\ & \quad + \int \left\{\frac{\partial}{\partial\theta} l(\theta; \mathbf{x})\right\}\left\{\frac{\partial}{\partial\theta} l(\theta; \mathbf{x})\right\} f(\mathbf{x}|\theta) d\mathbf{x} \\ \Rightarrow\quad & E\left[\left\{\frac{\partial}{\partial\theta} l(\theta; \mathbf{X})\right\}^2\right] = -E\left\{\frac{\partial^2}{\partial\theta^2} l(\theta; \mathbf{X})\right\}. \end{aligned}$$

$$E\left[\left\{\frac{\partial}{\partial\theta} l(\theta; \mathbf{X})\right\}^2\right] = -E\left\{\frac{\partial^2}{\partial\theta^2} l(\theta; \mathbf{X})\right\}$$

Since $E\{\frac{\partial}{\partial\theta} l(\theta; \mathbf{X})\} = 0$, we have

$$Var\left\{\frac{\partial}{\partial\theta} l(\theta; \mathbf{X})\right\} = -E\left\{\frac{\partial^2}{\partial\theta^2} l(\theta; \mathbf{X})\right\}.$$

The term $-E\left\{\frac{\partial^2}{\partial\theta^2} l(\theta; \mathbf{X})\right\}$ is known as the **Fisher information** which we will denote by $\mathcal{I}_E(\theta)$:

$$\mathcal{I}_E(\theta) \equiv -E\left\{\frac{\partial^2}{\partial\theta^2} l(\theta; \mathbf{X})\right\}.$$

Fisher information: measure of amount of information a sample size of n contains about θ. For independent observations $X_1, \ldots, X_n$,

$$l(\theta; \mathbf{X}) = \sum_{i=1}^{n} \log f(X_i|\theta),$$

$$\mathcal{I}_E(\theta) = -nE\left\{\frac{\partial^2}{\partial\theta^2} l(\theta; X_i)\right\},$$

hence Fisher information is proportional to sample size.
• Example. Consider $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ with σ^2 known. Then

$$\begin{aligned} -E\left\{\frac{\partial^2}{\partial\theta^2} l(\theta; \mathbf{X})\right\} &= -E\left\{\frac{\partial^2}{\partial\theta^2} \frac{-1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \theta)^2\right\} \\ &= \frac{n}{\sigma^2}, \end{aligned}$$

Fisher information is n/σ^2. As σ^2 decreases, the observations more likely to be close to θ, so data more informative about θ.

Fisher information can be used to give a bound on the variance of an estimator.
Let $T(\mathbf{X})$ be an unbiased estimator, with $X_1, \ldots, X_n$ independent. Then it is possible to prove that

$$Var(T) \geq \frac{1}{\mathcal{I}_E(\theta)}.$$

This is known as the **Cramer-Rao minimum variance bound**.

Asymptotic normality

For large n, the distribution of the m.l.e $\hat{\theta}$ is approximately normal, with

$$\hat{\theta} \sim N\{\theta, \mathcal{I}_E(\theta)^{-1}\}.$$

Thus for large n, the m.l.e. $\hat{\theta}$ is *approximately* unbiased, and achieves the Cramer-Rao minimum variance bound.

In the multivariate case with $\theta = (\theta_1, \ldots, \theta_d)$ we have

$$\mathcal{I}_E(\theta) = \begin{pmatrix} e_{1,1}(\theta) & \cdots & e_{1,d}(\theta) \\ \vdots & & \vdots \\ e_{d,1}(\theta) & \cdots & e_{d,d}(\theta) \end{pmatrix},$$

with

$$e_{i,j}(\theta) = E\left\{-\frac{\partial^2}{\partial\theta_i\, \partial\theta_j} l(\theta)\right\}.$$

So for large n, the distribution of the m.l.e of θ is approximately multivariate normal:

$$\hat{\theta} \sim N_d(\theta, \mathcal{I}_E(\theta)^{-1}),$$

Example: normally distributed data

Consider $X_1, \ldots, X_n$ with $X_i \sim N(\theta_1, \theta_2)$, with both θ_1 and θ_2 unknown. We write $\theta = (\theta_1, \theta_2)^T$.

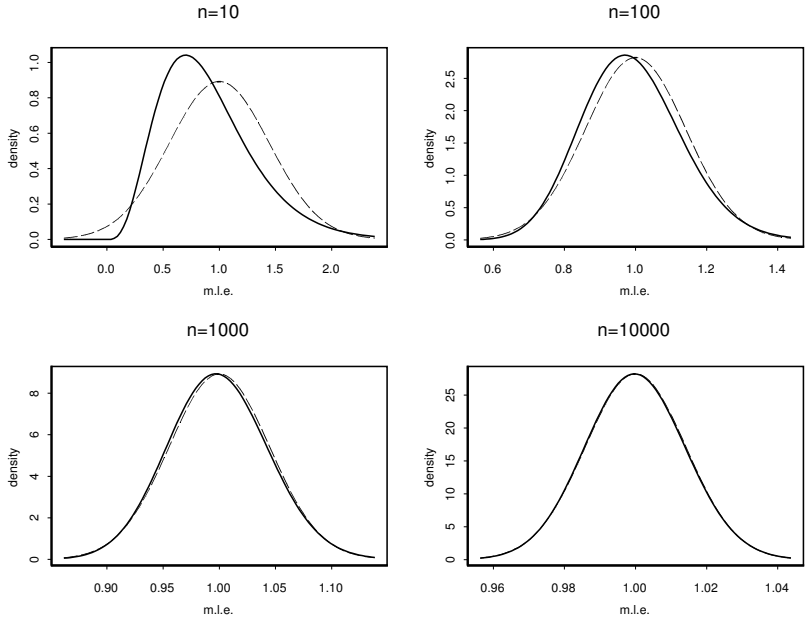
$$l(\theta; \mathbf{x}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\theta_2 - \frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i - \theta_1)^2,$$

$$\begin{aligned}\hat{\theta}_1 &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ \hat{\theta}_2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\end{aligned}$$

$$\mathcal{I}_E(\theta) = \begin{pmatrix} \frac{n}{\theta_2} & 0 \\ 0 & \frac{n}{2\theta_2^2} \end{pmatrix}.$$

For large n, the approximate distribution of $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)^T$ is

$$\begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} \sim N\left\{ \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \begin{pmatrix} \frac{\theta_2}{n} & 0 \\ 0 & \frac{2\theta_2^2}{n} \end{pmatrix} \right\}$$



Confidence intervals based on asymptotic normality

Suppose we want a $100(1-\alpha)\%$ confidence interval for any particular element of θ, say θ_j. For suitably large n, we have

$$\hat{\theta}_j \sim N(\theta_j, \gamma_{j,j}),$$

where $\gamma_{j,j}$ is the $\{j,j\}$ element of $\mathcal{I}_E(\theta)^{-1}$.
This then gives us an approximate interval as

$$(\hat{\theta}_j - z_{1-\frac{\alpha}{2}}\sqrt{\gamma_{j,j}}, \hat{\theta}_j + z_{1-\frac{\alpha}{2}}\sqrt{\gamma_{j,j}}),$$

θ unknown, so approximate $\mathcal{I}_E(\theta)$ by observed information matrix

$$\mathcal{I}_O(\theta) = \begin{pmatrix} -\frac{\partial^2}{\partial\theta_1^2} l(\theta) & \cdots & -\frac{\partial^2}{\partial\theta_1\partial\theta_d} l(\theta) \\ \vdots & & \vdots \\ -\frac{\partial^2}{\partial\theta_d\partial\theta_1}(\theta) & \cdots & -\frac{\partial^2}{\partial\theta_d^2} l(\theta) \end{pmatrix},$$

evaluated at $\theta = \hat{\theta}$.

$\tilde{\gamma}_{i,j}$: the i,jth element of the inverse of $\mathcal{I}_O(\theta)$, we use

$$(\hat{\theta}_j - z_{1-\frac{\alpha}{2}}\sqrt{\tilde{\gamma}_{j,j}}, \hat{\theta}_j + z_{1-\frac{\alpha}{2}}\sqrt{\tilde{\gamma}_{j,j}}),$$

as an approximate confidence interval. Since we know that $\hat{\theta} \to \theta$ as $n \to \infty$, with probability 1, we would expect $\mathcal{I}_O(\theta)$ to be similar to $\mathcal{I}_E(\theta)$ for large sample sizes.

4.2 Profile Likelihood

- RV X, density function f, parameters $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_d\}$
- Given $\mathbf{x} = (x_1, \ldots, x_n)$, only want inferences about *subset* of $\boldsymbol{\theta}$.
- Partition $\boldsymbol{\theta}$ into $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ with $\boldsymbol{\theta}_1$ the parameters of direct interest.
- $\boldsymbol{\theta}_2$, the parameters not of direct interest are known as **nuisance parameters**.

- Example: $X \sim N(\mu, \sigma^2)$ with both μ and σ^2 unknown, though we may only be interested in the mean parameter μ.
- Can use asymptotic distribution of m.l.e. to derive confidence intervals for individual parameters.
- Will now consider an alternative form of likelihood function which in some cases can produce more accurate confidence intervals.

Partitioning $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, **profile** log-likelihood function for $\boldsymbol{\theta}_1$ is

$$l_p(\boldsymbol{\theta}_1; \mathbf{x}) = \max_{\boldsymbol{\theta}_2} l(\boldsymbol{\theta}). \tag{1}$$

To get the profile log-likelihood function for θ_1:

1. Treat $\boldsymbol{\theta}_1$ as a constant in $l(\boldsymbol{\theta}; \mathbf{x})$.
2. Find the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_2$ in terms of the data $\mathbf{x}$ *and* $\boldsymbol{\theta}_1$.
3. Plug in this expression for $\hat{\boldsymbol{\theta}}_2$ into the full log-likelihood $l(\boldsymbol{\theta}; \mathbf{x})$ to get the profile log-likelihood $l_p(\boldsymbol{\theta}_1; \mathbf{x})$.

- Writing $\boldsymbol{\theta} = (\theta_i, \boldsymbol{\theta}_{-i})$, plotting $l_p(\theta_i)$ gives us profile of log-likelihood surface viewed from θ_i axis.
- If $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)$ maximises $l(\boldsymbol{\theta})$, then $\hat{\boldsymbol{\theta}}_1$ maximises $l_p(\boldsymbol{\theta}_1)$ and $\hat{\boldsymbol{\theta}}_2$ maximises $l_p(\boldsymbol{\theta}_2)$.
- Useful exploratory tool; allows you to plot a likelihood $l_p(\theta_i)$ for a single parameter θ_i.
- Can be used to derive more accurate confidence intervals.

Example 1

$X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ i.i.d.

$$l(\mu, \sigma^2; \mathbf{x}) = -n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2. \tag{2}$$

Fixing μ, the MLE of σ^2 is $\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$.Substituting this back into the full log-likelihood $l(\mu, \sigma^2; \mathbf{x})$, we get

$$l_p(\mu; \mathbf{x}) = -\frac{n}{2} \log\{\frac{1}{n} \sum (x_i - \mu)^2\} - \frac{n}{2}. \tag{3}$$

Fixing σ^2, the MLE of μ is $\bar{x}$.The profile log-likelihood for σ^2 is

$$l_p(\sigma^2; \mathbf{x}) = -n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \bar{x})^2. \tag{4}$$

Inference using the deviance function

- Can construct CI for $\boldsymbol{\theta}$ based on asymptotic normality of MLE. Alternative approach: use **deviance function**.
- For arbitrary $\boldsymbol{\theta}^*$,

$$D(\boldsymbol{\theta}^*) = 2\{l(\hat{\boldsymbol{\theta}}; \mathbf{x}) - l(\boldsymbol{\theta}^*; \mathbf{x})\}. \tag{5}$$

  $\hat{\boldsymbol{\theta}}$ maximises log-likelihood, so $D(\boldsymbol{\theta}^*) \geq 0$.
- If $D(\boldsymbol{\theta}^*)$ is small, then $l(\boldsymbol{\theta}^*)$ must be close to $l(\hat{\boldsymbol{\theta}})$, which suggests that $\boldsymbol{\theta}^*$ is a plausible estimate for the true unknown value of $\boldsymbol{\theta}$.
- A confidence interval (or region if $\boldsymbol{\theta}$ is a vector) could then be of the form

$$C = \{\boldsymbol{\theta}^* : D(\boldsymbol{\theta}^*) \leq c\}, \tag{6}$$

  for some suitable value of c.

- With data $x_1, \ldots, x_n$, for sufficiently large n, it can be shown that at the true value of $\boldsymbol{\theta}$, $D(\boldsymbol{\theta}) \sim \chi^2_d$, where d is the dimensionality of $\boldsymbol{\theta}$.
- An approximate $(1 - \alpha)$ confidence region for $\boldsymbol{\theta}$ is then given by

$$C_\alpha = \{\boldsymbol{\theta}^* : D(\boldsymbol{\theta}^*) \leq c_\alpha\}, \tag{7}$$

  with c_α the $(1 - \alpha)$ percentage point of the χ^2_d distribution.
- Usually more accurate than asymptotic normality approximation, may require greater computational effort.

Profile likelihood and the deviance function

- $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, with $\boldsymbol{\theta}_1$ a k-dimensional subset of $\boldsymbol{\theta}$. **Profile deviance**:

$$D_p(\boldsymbol{\theta}_1^*) = 2\{l(\hat{\boldsymbol{\theta}}; \mathbf{x}) - l_p(\boldsymbol{\theta}_1^*; \mathbf{x})\}, \tag{8}$$

  with $\hat{\boldsymbol{\theta}}$ the maximum likelihood estimator of $\boldsymbol{\theta}$.
- Based on a sample of size n, with n sufficiently large,

$$D_p(\boldsymbol{\theta}_1) \sim \chi^2_k. \tag{9}$$

- Can obtain a confidence interval for any element θ_i as

$$C_\alpha = \{\theta_i^* : D_p(\theta_i^*) \leq c_\alpha\}, \tag{10}$$

  again, with c_α the $(1-\alpha)$ percentage point of the χ^2_1 distribution.
- This will often be more accurate than the interval

$$\hat{\theta}_i \pm z_{\frac{\alpha}{2}} \sqrt{\psi_{i,i}} \tag{11}$$

  stated earlier.

Example: leukaemia data

- Leukaemia patients given drug, 6-mercaptopurine (6-MP), and the number of days t_i until freedom from symptoms is recorded:

  $6^*, 6, 6, 6, 7, 9^*, 10^*, 10, 11^*, 13, 16, 17^*, 19^*, 20^*, 22, 23, 25^*, 32^*, 32^*,$

  A * denotes an observation censored at that time.
- Weibull model:

$$f_T(t) = \alpha\beta(\beta t)^{\alpha-1} \exp\{-(\beta t)^\alpha\} \tag{12}$$

  for $t > 0$. $\alpha = 1$ gives exponential distribution.
- For censored data

$$P(T > t) = \exp\{-(\beta t)^\alpha\}. \tag{13}$$

  d: no. of uncensored observations, $\sum_u \log t_i$: sum of all logs of the uncensored observations.

$$l(\alpha, \beta; \mathbf{x}) = d\log\alpha + \alpha d\log\beta + (\alpha-1)\sum_u \log t_i - \beta^\alpha \sum_{i=1}^n t_i^\alpha. \tag{14}$$

Treat α as fixed, and find MLE of β as function of data and α.

$$\hat{\beta} = \left(\frac{d}{\sum_{i=1}^n t_i^\alpha}\right)^{\frac{1}{\alpha}}. \tag{15}$$

The profile log-likelihood of α is then given by

$$\begin{aligned} l_p(\alpha) &= l(\alpha, \hat{\beta}) \\ &= d\log\alpha + \alpha d\log\left(\frac{d}{\sum_{i=1}^n t_i^\alpha}\right)^{\frac{1}{\alpha}} + (\alpha-1)\sum_u \log t_i - d \end{aligned}$$

- Finding the full MLE $(\hat{\alpha}, \hat{\beta})$ cannot be done analytically, so numerical methods have to be used.
- To construct the confidence interval, onlyneed $\hat{\alpha}$ that maximises $l_p(\hat{\alpha})$, as $l_p(\hat{\alpha}) = l(\hat{\alpha}, \hat{\beta})$.
- For a 95% confidence interval, the 95th percentage point of the χ^2_1 distribution is 3.841.The confidence interval is then given by

$$\begin{aligned} C_{0.05} &= \{\alpha^* : D_p(\alpha^*) \leq 3.841\} && (16) \\ &= [\alpha^* : 2\{l_p(\hat{\alpha}) - l_p(\alpha^*)\} \leq 3.841] && (17) \\ &= \{\alpha^* : l_p(\alpha^*) > l_p(\hat{\alpha}) - 3.841/2\}. && (18) \end{aligned}$$

- Numerically, we estimate the MLE $\hat{\alpha}$ to be 1.35, with $l_p(\hat{\beta}) = -41.66$.
- From the graph, we can then read off the 95% confidence interval for α as (0.73,2.2).
- This contains the value 1, so the simpler exponential distribution is plausible for this dataset.

Example: machine component failure

- Level of corrosion w in a machine component recorded and component tested until a failure is observed, at time t.
- Denote each observation by (w_i, t_i), where w_i is the level of corrosion, and t_i is the failure time.
- Possible model: $T \sim Exponential(\lambda)$ distribution, with λ a function of the corrosion level w:

$$\lambda = \alpha w^{\beta}. \tag{19}$$

  w treated as fixed, i.e. model distribution of the failure time conditional on the corrosion.
- $\beta = 0$ implies same expected time to failure, α^{-1} for all components, regardless of the corrosion level w.

The density of a single observation (w, t) is given by

$$f_T(t) = \alpha w^{\beta} \exp\{-\alpha w^{\beta} t\}. \tag{20}$$

$$l(\alpha, \beta; \mathbf{x}) = n \log \alpha + \beta \sum_{i=1}^{n} \log w_i - \alpha \sum_{i=1}^{n} w_i^{\beta} t_i. \tag{21}$$

We can derive an expression for the profile log-likelihood of β: Treating β as fixed, we obtain the MLE of α as

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^{n} w_i^{\beta} t_i}. \tag{22}$$

We then substitute this expression for α in the full log-likelihood $l(\alpha, \beta)$ to get the profile log-likelihood for β:

$$l_p(\beta; \mathbf{x}) = n \log \left(\frac{n}{\sum_{i=1}^{n} w_i^{\beta} t_i} \right) + \beta \sum_{i=1}^{n} \log w_i - n. \tag{23}$$

- Numerically, estimate $\hat{\beta} = 0.473$, with $l_p(\hat{\beta}; \mathbf{x}) = -20.01$.
- From graph, read off 95% confidence interval for β as (0.11,0.95).
- Doesn't contain zero, and so there is clear evidence that $\beta \neq 0$
- For comparison, compute confidence interval for β using normal approximation.
- Observed information matrix is given by

$$\begin{pmatrix} -\frac{\partial^2 l}{\partial \alpha^2} & -\frac{\partial^2 l}{\partial \alpha \partial \beta} \\ -\frac{\partial^2 l}{\partial \alpha \partial \beta} & -\frac{\partial^2 l}{\partial \beta^2} \end{pmatrix} = \begin{pmatrix} n\alpha^{-2} & \sum w_i^{\beta} t_i \log w_i \\ \sum w_i^{\beta} t_i \log w_i & \alpha \sum w_i^{\beta} t_i (\log w_i)^2 \end{pmatrix} \tag{24}$$

- Obtain $\hat{\alpha}$ by substituting $\beta = 0.473$ into formula, gives $\hat{\alpha} = 1.099$.
- Substitute $\alpha = 1.099$, $\beta = 0.473$ into observed information matrix, invert to get

$$V = \begin{pmatrix} 0.0534 & -0.0241 \\ -0.0241 & 0.0442 \end{pmatrix}. \tag{25}$$

- CI for β using asymptotic normality is

$$\hat{\beta} \pm 1.96 \times 0.0442^{0.5}, \tag{26}$$

  which gives (0.0611,0.8849).