Anchoring and Bias

In the absence of hard data, a person's estimate of an unknown parameter, such as

- the risk of being in an earthquake,
- the literacy rate in South Africa, or
- the proportion of unwashed students in a IB lecture,

can be distorted by providing a reference point or **anchor**.

For example, in one study subjects were asked to give the percent of African countries in the United Nations. In each case, a number between 0 and 100 was assigned as an initial value and the subject was asked if this was too high or two low, and what adjustment needed to be made. Despite the fact that each of the subjects knew that the initial value had been determined randomly (by spinning a wheel in the subject's presence), there still tended to be a bias toward the initial value.

A knowledge of the psychology of human judgment can be relevant in interpreting the results of a study. One must watch out for **bias** introduced by the way a question is phrased.

Suppose a jury is shown a videotape of a car collision. The following two questions can elicit quite different answers.

- How fast was the car going when it hit the wall?
- How fast was the car going when it crashed into the wall?

A survey method for a sensitive question

How can we get accurate answers to a sensitive question which respondents might be reluctant to answer truthfully?

Examples:

(a) "Have you ever used illegal drugs?"

(b) "Have you ever used a sick day leave when you weren't really sick?"

(c) "Have you not bathed or showered in the last 24 hours?"

Method 1: The Innocuous Question

Let Q_s be the sensitive question and Q_i be an innocuous question which has a known probability of yielding a YES response. For example,

$Q_s =$ "Have you not bathed in the last 24 hours?"
$Q_i =$ "Flip a coin. Did you get a head?"

The respondent answers Q_s with probability θ and Q_i with probability $(1-\theta)$. The key point is that the respondent determines which question she answers by using some probability device which is under her control. For example: She rolls a die. If the result is $\{1, 2, 3, \text{ or } 4\}$ answer Q_s; if it is $\{5 \text{ or } 6\}$ answer Q_i. Since only the respondent knows which question she is answering, there should be no stigma attached to a YES or N0 response.

If the known probability of a YES to Q_i is α, we find that

$$\mathbb{P}(\text{YES}) = \theta p + (1-\theta)\alpha$$

$$p = \frac{\mathbb{P}(\text{YES}) - (1-\theta)\alpha}{\theta}$$

and hence if the number of YESs in a sample of size n is X,

$$\hat{p} = \frac{X/n - (1-\theta)\alpha}{\theta}$$

For the experiment in Lecture 1, (1997 class), $\theta = 251/365 = 0.688$, $\alpha = 1$.

If p were 0.20 we would have $\mathbb{P}(\text{YES}) = .450$.

We observed $X/n = 89/194 = 0.46$, so $\hat{p} = \mathbf{0.21}$.

Method 2: Warner (1965)

Let Q_s be the sensitive question and $\bar{Q}_s$ be its complement. For example,

$Q_s =$ "Have you not bathed in the last 24 hours?"
$\bar{Q}_s =$ "Have you bathed in the last 24 hours?"

With some (known) probability θ a subject answers Q_s, otherwise (with probability $1-\theta$) he or she answers $\bar{Q}_s$.

Let $p =$ proportion in the population for which the true response to Q_s is YES. So the chance of getting a YES response is given by

$$\mathbb{P}(\text{YES}) = \theta p + (1-\theta)(1-p)$$

We solve easily for p to give

$$p = \frac{\mathbb{P}(\text{YES}) - (1-\theta)}{2\theta - 1}$$

If the number of YES answers in a sample of size n is Y, we can estimate p with

$$\tilde{p} = \frac{Y/n - (1-\theta)}{2\theta - 1}$$

For $\theta = 251/365$, $p = 0.20$, $\mathbb{P}(\text{YES}) = 0.387$.

Comparison of variances

Both $\hat{p}$ and $\tilde{p}$ are unbiased estimators of p. We might ask, which method is more efficient in the sense of having smaller variance.

Determine this for the case $\theta = 251/365$, $p = 0.20$, $\alpha = 1$.

Recall that

$$\hat{p} = \frac{X/n - (1-\theta)\alpha}{\theta} \quad \text{and} \quad \tilde{p} = \frac{Y/n - (1-\theta)}{2\theta - 1}$$

and $\mathrm{var}(X/n) = (1/n)\mathbb{P}(\mathrm{YES})(1 - \mathbb{P}(\mathrm{YES}))$. This gives

$$\begin{aligned}
\frac{\mathrm{var}\,\tilde{p}}{\mathrm{var}\,\hat{p}} &= \frac{\mathrm{var}(Y/n)/(2\theta-1)^2}{\mathrm{var}(X/n)/\theta^2} \\
&= \frac{(.387)(.613)n^{-1}/(137/365)^2}{(.450)(.550)n^{-1}/(251/365)^2} \\
&= 3.22
\end{aligned}$$

Thus the first scheme is more than 3 times as efficient as the second.

How many words did Shakespeare know?

Shakespeare's known works comprise 884,647 words. He wrote 31,534 different words, of which 14,376 appear only once, 4,343 twice, etc. How many words did he know but not use?

Let us suppose that he knew S words. Suppose word i occurs in such a way that the number of its occurrences in a sample of any 884,647v words of Shakepeare is distributed as a Poisson RV with parameter $\lambda_i v$, $v > 0$.

Let N_x be the number of words which occur x times in a random sample of 884,647 words of Shakespeare. As an observed value of N_x we have n_x, e.g., $n_1 =$14,376. Then

$$\eta_x := \mathbb{E} N_x = \mathbb{E}\left[\sum_{i=1}^{S} 1\{\text{word } i \text{ is used } x \text{ times}\}\right] = \sum_{i=1}^{S} \frac{\lambda_i^x}{x!} e^{-\lambda_i}$$

Suppose we want to make an estimate of $\Delta(t)$, the expected number of distinct words that will occur in a sample of 884,647$(1+t)$ words, $t > 0$, but which are not include amongst the first 884,647 such words. Now

$$\Delta(t) = \mathbb{E}\left[\sum_{i=1}^{S} 1\{\text{word } i \text{ is in large sample but not in subset}\}\right]$$
$$= \sum_{i=1}^{S} e^{-\lambda_i}\left(1 - e^{-\lambda_i t}\right) .$$

Thus far we have

$$\eta_x = \mathbb{E}N_x = \sum_{i=1}^{S} \frac{\lambda_i^x}{x!} e^{-\lambda_i} \quad \text{and} \quad \Delta(t) = \sum_{i=1}^{S} e^{-\lambda_i} \left(1 - e^{-\lambda_i t}\right)$$

Using the fact that

$$1 - e^{-\lambda_i t} = \lambda_i t - \frac{\lambda_i^2}{2!} t^2 + \cdots$$

and substituting, this gives

$$\Delta(t) = \eta_1 t - \eta_2 t^2 + \eta_3 t^3 - \cdots .$$

Recall that $\eta_x = \mathbb{E}N_x$. Thus an unbiased estimator of $\Delta(t)$ is

$$\widehat{\Delta(t)} = N_1 t - N_2 t^2 + N_3 t^3 - \cdots$$

and we have $n_1 t - n_2 t^2 + n_3 t^3 - \cdots$ as an estimate of $\Delta(t)$.

For the Shakespeare data, and $t = 1$, this gives $\widehat{\Delta(1)}$ =11,430.

Thus if the exisiting known works of Shakepeare were twice as large as they actually are then we would expect to see about 11,430 new words in addition to the 31,534 we have already seen. The expected error of this estimate is less than 150.

What about $\widehat{\Delta(\infty)}$, i.e., an estimate of the total number of words which Shakespeare knew? Unfortunately, the above estimator does not converge as $t \to \infty$.

However, other methods, too complicated to explain here, suggest that Shakepeare knew at least 35,000 words which he did not use.

A Confidence Interval for Remaining Life

J. Richard Gott, Princeton astrophysicist, has written an article in *Nature* explaining how to obtain 95% confidence intervals for the remaining lifetime of about anything you wish.

The idea is that if we are at a randomly chosen point of the lifetime, then with probability $.95$ we are somewhere between $1/40$ and $39/40$ through the total life.

The Rule of 39

Suppose a restauranter wants to obtain a 95% confidence interval for the remaining time that a party will remain at their table. He notes how long they have been there already, say T.

Assuming that the point at which the data is obtained is equally likely to be at any point during the meal, the probability is $1/20$ that the diners are between $1/40$th and $39/40$th of the way through their total stay. In this case their remaining stay is at least $T/39$ and no more than $39T$.

Thus the probability that the interval $[T/39, 39T]$ contains the true length of the remaining stay is at least $.95$.

In practice, of course there should be better ways to estimate this quantity, taking account of, for example, what course of the meal the party is on, or how long a typical party tends to take over a meal.

Confidence Interval for Remaining Life

But now consider an example where we can't possibly know 'what course of the meal the party is on' or 'how long a typical party tends to take over a meal'.

Limits on Human Existence

Suppose we want to estimate how much longer the human race will remain in existence. We estimate that the human race has been around for about $200,000$ years.

If we're between $1/40$th and $39/40$th of the way through the lifetime of the human race, then we have no more than $39 \times 200,000 = 7.8$ million years left to go. Similarly, we have at least $1/39 \times 200,000 = 5128$ years to go. Thus a $> 95\%$ confidence interval for the remaining lifetime of the human race is

$$[5,000,\ 8,000,000] \text{ years.}$$

A Confidence Interval for Remaining Life Limits on Human Existence — is this sense or nonsense?

Having developed this beautiful rule of thumb, Gott goes on to demonstrate, by example, the dangers of taking this kind of thing too seriously. He concludes that the space program, now 32 years old, will (with 95% confidence) end before another 1200 years are up, surely too short a time for us to colonize the galaxy and thus escape the $8,000,000$ year deadline derived above.

Fortunately for the human race, Gott's whole theory is only about a year old, and thus can be expected to last somewhere between another 39 years and another 9 days. Thus it will most likely die long before it has a chance to doom the space program, and thereby the whole human race.

Reference: *Formula projects limits seen on human existence.* The New York Times, 1 June 1993, Sec. C Page 1. Macolm W. Browne

Utility

Is playing lotteries rational? To investigate this question we need a notation for a lottery. Denote 'the consumer wins prize x with probability p and prize y with probability $1-p$' by

$$p \circ x \oplus (1-p) \circ y\,.$$

The set of lotteries is denoted $\mathcal{L}$. Let $\sim$ denote indifference between two lotteries. Rational preferences between lotteries should obey some reasonable assumptions, namely for all prizes x, y and $p, q \in [0,1]$

- $1 \circ x \oplus (1-1) \circ y \sim x$,
- $p \circ x \oplus (1-p) \circ y \sim (1-p) \circ y \oplus p \circ x$,
- $q \circ [p \circ x \oplus (1-p) \circ y] \oplus (1-q) \circ y$
  $\sim (qp) \circ x \oplus (1-qp) \circ y$.

These are enough to show that there must exist a function $u(\cdot)$, mapping lotteries to $\mathbb{R}$ such that

$$\begin{aligned} & p \circ x \oplus (1-p) \circ y \;\succ\; q \circ w \oplus (1-q) \circ z \\ \iff & u\big(p \circ x \oplus (1-p) \circ y\big) > u\big(q \circ w \oplus (1-q) \circ z\big) \end{aligned}$$

That is, we can determine a consumer's preference amongst two lotteries just by comparing the values of the **utility** $u(\cdot)$.

Note that u is not unique, since if $u(\cdot)$ works, so does $au(\cdot) + b$, $a > 0$.

The expected utility property

It is reasonable to guess that there is a $u(\cdot)$ which satisfies the **expected utility property** (EUP):

$$u\big(p \circ x \oplus (1-p) \circ y\big) = pu(x) + (1-p)u(y)\,.$$

But this has to be proved. The EUP is not as obvious as it first appears. We might think that the utility of a lottery could be worth more than simply the expected value of the utility from the prizes. I.e., we might think the consumer is the sort of person who likes lotteries. That would help 'explain' why people like to play lotteries. In fact, the EUP obliterates such an explanation and we must conclude that playing lotteries is irrational, or that some of the 'obvious' assumptions about $\sim$ used to prove the EUP do not apply to some consumers.

To prove the EUP, let us assume there is a best prize b and worse prize w. Define $u(b) = 1$, $u(w) = 0$. For an arbitrary lottery z, define $u(z) = p_z$ where p_z is chosen so that

$$p_z \circ b \oplus (1-p_z) \circ w \sim z.$$

Then $p \circ x \oplus (1-p) \circ y$

$$\sim p \circ \big(p_x \circ b \oplus (1-p_x) \circ w\big) \oplus (1-p) \circ \big(p_y \circ b \oplus (1-p_y) \circ w\big)$$

$$\sim \big[pp_x + (1-p)p_y\big] \circ b \oplus \big[1 - pp_x - (1-p)p_y\big] \circ w$$

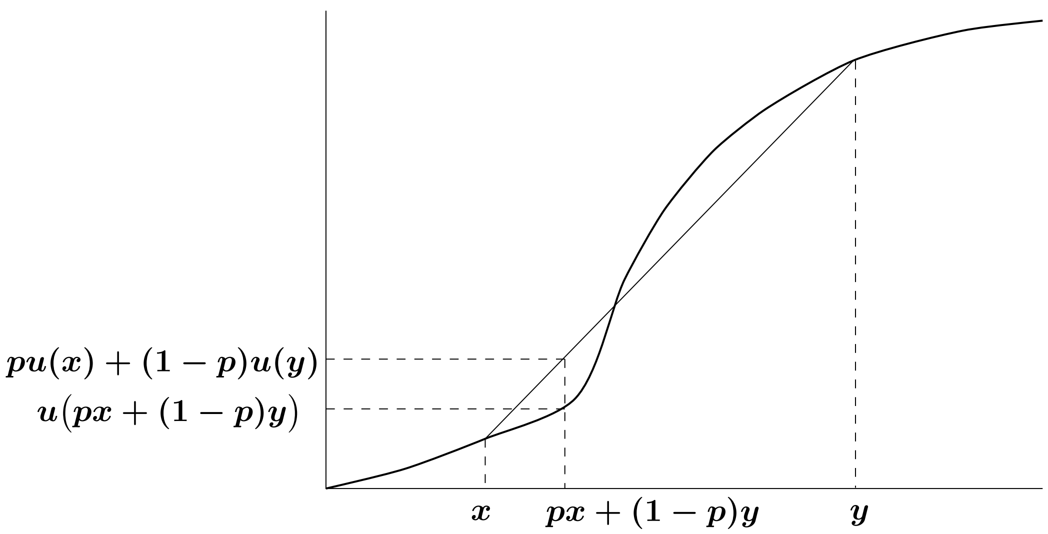
$$\sim \big[pu(x) + (1-p)u(y)\big] \circ b \oplus \big[1 - pu(x) - (1-p)u(y)\big] \circ w$$

Hence this u has the EUP.

A utility function

Playing a lottery can only be rational if the utility of the expected value of the lottery, $= \boldsymbol{u(px + (1-q)y)}$ is less than the utility of the lottery,
$= \boldsymbol{u(p \circ x \oplus (1-p) \circ y) = pu(x) + (1-p)u(y)}$.

This occurs for the lottery illustrated in the following diagram.



However, most people have a concave utility function. If so, and if the assumption leading to the EUP are valid, then playing lotteries is irrational.

Allias Paradox

Which of these choices do you find most attractive?

A. $\pounds 1$ million guaranteed

B. 89% chance of $\pounds 1$ million
10% chance of $\pounds 2.5$ million
1% chance of nothing

Now consider these choices:

C. 89% chance of nothing
11% chance of $\pounds 1$ million

D. 90% chance of nothing
10% chance of $\pounds 2.5$ million

Most people prefer A to B and D to C. However, this is inconsistent. Preference of A to B means

$$u(1) > .89u(1) + .10u(2.5) + .01u(0)$$

Whereas preference of D to C means

$$\begin{aligned} .89u(0) + .11u(1) &< .90u(0) + .10u(2.5) \\ .11u(1) &< .01u(0) + .10u(2.5) \\ u(1) &< .89u(1) + .10u(2.5) + .01u(0) \end{aligned}$$

Ellsberg Paradox

A bag contains 300 balls, of which 100 are red and 200 are either blue and green. One is drawn at random.

Which of gamble do you find most attractive?

Gamble A. You get £ 1000 if the ball is red.

Gamble B. You get £ 1000 if the ball is blue.

A bag contains 300 balls, of which 100 are red and 200 are either blue and green. One is drawn at random.

Which gamble do you find most attractive?

Gamble C. You get £ 1000 if the ball is not red.

Gamble D. You get £ 1000 if the ball is not blue.

Most people prefer A to B and C to D. However, this is inconsistent. Preference of A to B means we think $\mathbb{P}(\text{red}) > \mathbb{P}(\text{blue})$ and preference of C to D means we think $\mathbb{P}(\text{not red}) > \mathbb{P}(\text{not blue})$. But these cannot both be true since,

$$\mathbb{P}(\text{not red}) = 1 - \mathbb{P}(\text{red}) \text{ and } \mathbb{P}(\text{not blue}) = 1 - \mathbb{P}(\text{blue})$$

Some psychologists think the Allias and Ellsberg paradoxes require new models to describe people's behaviour. Other think these paradoxes are just 'optical illusions'.

An Estimation Game

Players 1 and 2 are to play the following game.

Player 1

- thinks of any real number, say $\boldsymbol{\theta}$;
- adds an error to $\boldsymbol{\theta}$ that is equally likely to be $+\mathbf{10}$ or $-\mathbf{10}$ (i.e., chosen by his tossing a fair coin);
- tells the result to Player 2.

Player 2 learns $\boldsymbol{x}$.

He now knows that $\boldsymbol{\theta}$ is either $\boldsymbol{x} - \mathbf{10}$ or $\boldsymbol{x} + \mathbf{10}$.

Player 2 must try to guess (estimate) $\boldsymbol{\theta}$.

With what probability can Player 2 guess correctly?

Solution to the Estimation Game

It is surprising, but Player 2 can guess θ correctly with a probability > 0.5. How does he do that?

Player 2 should privately sample a real number, say y, from a distribution on $(-\infty, \infty)$, say from $N(0, 1)$.

If $x < y$ he should guess $\theta = x + 10$.

If $x > y$ he should guess $\theta = x - 10$.

To see why this works, consider 3 cases.

1. If $y > \theta + 10$, then $x < y$ and Player 2 guesses $\theta = x + 10$. He is correct half the time, i.e., when Player 1 added the error -10 rather than $+10$.
2. Similarly, if $y < \theta - 10$, then $x > y$ and Player 2 is also correct with probablity 0.5.
3. However, if y is in the interval $(\theta - 10, \theta + 10)$ then Player 2's guess is always correct. E.g., if Player 1 subtracted 10 then $x = \theta - 10 < y$ and Player 1 correctly guesses $\theta = x + 10$.

As long as there is a positive probability that $y \in (\theta - 10, \theta + 10)$ — which is ensured by sampling y from $N(0, 1)$ — the total probability that Player 2 is correct is therefore > 0.5.

This is significant compared to χ_1^2 whose 5% point is 3.84.

A Love Story

"You haven't told me yet," said Lady Nuttal, "what it is your fiance does for a living?"

"He's a statistician," replied Lamia, with an annoying sense of being on the defensive.

Lady Nuttal was obviously taken aback. It had not occurred to her that statisticians entered into normal social relationships. The species, she would have surmised, was perpetuated in some collateral manner, like mules.

"But Aunt Sara, it's a very interesting profession," said Lamia warmly.

"I don't doubt it," said her aunt, who obviously doubted it very much. "To express anything important in mere figures is so plainly impossible that there must be endless scope for well-paid advice on how to do it. But don't you think that life with a statistician would be rather, shall we say, humdrum?"

Lamia was silent. She felt reluctant to discuss the surprising depth of emotional possibility which she had discovered below Edward's numerical veneer.

"It's not the figures themselves," she said finally. "it's what you do with them that matters."

(K.A.C. Manderville, The Undoing of Lamia Gurdleneck)

Benford's law

Benford's law states that the distribution of the leading digit in data sets is typically not equi-distributed but rather given by the distribution $p(k) = \log_{10}(k+1) - \log_{10}(k)$ for $k = 1, 2, \ldots, 9$. (The leading digit of $.0034$ is 3, of 243 is 2 etc.). Numerous explanations for this have been given but perhaps the most persuasive is that Benford's distribution is the unique distribution for the leading digits that is not changed by a change of units, i.e., multiplying the data by a constant c.

Here are 56 physical constants:

Speed of light	299792458 m s(-1)
Gravitation constant	6.67259e-11 m3 kg(-1) s(-2)
Planck constant	6.6260755e-34 J s
Planck constant/2pi	1.0545726691251e-34 J s
Planck mass	2.17671e-08 kg
Planck length	1.61605e-35 m
Planck time	5.39056e-44 s
Elementary charge	1.60217733e-19 C
Vacuum permeability	1.25663706143592e-06 N A(-2)
Vacuum permitivity	8.85418781762039e-12 F m(-1)
Magnetic flux quantum	2.06783461e-15 Wb
Quantized Hall resistance	25812.8056 Ohm
Electron mass	9.1093897e-31 kg
Muon mass	1.8835327e-28 kg
Proton mass	1.6726231e-27 kg
Neutron mass	1.6749286e-27 kg
Deuteron mass	3.343586e-27 kg
Proton/Electron mass	1836.152701 1
Fine-structure constant	0.00729735308 1
Rydberg constant	10973731.534 m(-1)
Bohr radius	5.29177249e-11 m
Hartree energy	4.3597482e-18 J
Bohr magneton	9.2740154e-24 J T(-1)
Nuclear magneton	5.0507866e-27 J T(-1)
Electron g-factor	2.002319304386 1
Electron radius	2.81794092e-15 m
Von-Klitzing constant	25812.807 Ohm
Proton mass	1.0072764666 amu
Electron mass	0.0005485799111 amu

Astronomical Unit (AU)	1.4959789e11 m
Parsec	206264.806 AU
Light Year (Ly)	9.46053e15 m
Sidereal Year	3.155815 sec
Mass of the Sun	1.989e30 kg
Radius of the Sun	6.96e5 km
Luminosity of the Sun	3.90e26 W
Solar Constant	1370 W/m2
Boltzmann constant	1.380658e-23 J K(-1)
Avogadro constant	6.0221367e+23 mol(-1)
Faraday constant	96485.309 C mol(-1)
Gas constant	8.31451 J mol(-1) K(-1)
Stefan-Boltzmann const.	5.67051e-08 W m(-2) K(-4)
Molar volume	22.441 l/mol
1st radiation constant	3.7417749e-16 W m2
2nd radiation constant	0.01438769 m K
Electron volt	1.60217733e-19 J
Atomic mass unit	1.6605402e-27 kg
Standard acceleration	9.80665 m s(-2)
Standard athmosphere	101325 Pa
Thermodynamic calorie	4.184 J
Inch	0.0254 m
Foot	0.3048 m
Yard	0.9144 m
Mile	1609.344 m
Ounce	0.02834952 kg
Pound	0.45359232 kg

UNIVERSAL CONSTANTS	
Speed of light	299792458 m s(-1)
Gravitation constant	6.67259e-11 m3 kg(-1) s(-2)
Planck constant	6.6260755e-34 J s
Planck constant/2pi	1.0545726691251e-34 J s
Planck mass	2.17671e-08 kg
Planck length	1.61605e-35 m
ASTRONOMICAL CONSTANTS	
Astronomical Unit (AU)	1.4959789e11 m
Parsec	206264.806 AU
Light Year (Ly)	9.46053e15 m
Sidereal Year	3.155815 sec
Mass of the Sun	1.989e30 kg
Radius of the Sun	6.96e5 km
...	

The number of these $\mathbf{56}$ constants whose leading digit is $\mathbf{1, 2, \ldots, 9}$ are respectively $\mathbf{19, 11, 5, 3, 5, 4, 1, 2, 6}$. So the numbers in the sets $\mathbf{\{1\}}$,$\mathbf{\{2, 3\}}$,$\mathbf{\{4, 5, 6, 7\}}$ are 19, 16, 13.

Under Benford's distribution we would expect these to be equal, i.e., $\mathbf{16}$. The Pearson's chi-squared statistic is therefore

$$\boldsymbol{T = \sum_i \frac{(o_i - e_i)^2}{e_i}}$$

$$\boldsymbol{= \frac{(19-16)^2 + (16-16)^2 + (13-16)^2}{16} = 1.125}$$

The $\mathbf{90\%}$ point of the $\boldsymbol{\chi_2^2}$ is $\mathbf{4.61}$. So this statistic reveals no significant departure from Benford's distribution.

Jane Austen and her imitator†

When Jane Austen died she left the novel Sanditon only partially finished, but a summary of its remainder. A highly literate admirer finished the novel attempting to emulate Jane Austen's style. Here are counts of some common words.

	by Austen			by Imitator
word	Sense and Sensibility	Emma	Sanditon	Sanditon
a	147	186	101	83
an	25	26	11	29
this	32	39	15	15
that	94	105	37	22
with	59	74	28	43
without	18	10	10	4
Total	375	440	202	196

† This example is taken from pages 485-489 of Rice's book, where he quotes Morton's 1978 analysis in his book *Literary Detection*.

A test of homogeneity

We might first of all check that the first three columns have the same distribution (expected counts shown in parentheses).

	by Austen		
word	Sense and Sensibility	Emma	Sanditon
a	147 (160.0)	186 (187.8)	101 (86.2)
an	25 (22.9)	26 (26.8)	11 (12.3)
this	32 (31.7)	39 (37.2)	15 (17.1)
that	94 (87.0)	105 (102.1)	37 (46.9)
with	59 (59.4)	74 (69.7)	28 (32.0)
without	18 (14.0)	10 (16.4)	10 (7.5)
Total	375	440	202

The usual statistic

$$\sum_{ij} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = \mathbf{12.27}$$

and this is to be compared to χ^2_{10}, whose $\mathbf{10\%}$ point is $\mathbf{15.99}$. So the data are consistent with authorship by the same person.

A second test of homogeneity

Now we compare Austen and the imitator, pooling all the data for Austin in one column.

word	Austen		Imitator	
a	434	(433.5)	83	(83.52)
an	62	(76.3)	29	(14.73)
this	86	(84.7)	15	(16.31)
that	236	(216.3)	22	(41.79)
with	161	(171.0)	43	(33.00)
without	38	(35.2)	4	(6.85)
Total	1017		196	

The statistic is

$$\sum_{ij} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = 32.81$$

which is highly significant compared to the χ^2_5, whose 0.5% point is 16.75. The imitator was not succesful in imitating this aspect of Austen's work.

Their main differences are in frequencies of use of the words *an* and *that*.

A child's puzzle

Place the 16 court cards Ace, King, Queen and Jack, of Spades, Hearts, Diamonds and Clubs, in a 4×4 array so that no row or column contains more than one card of the same value or the same suit.

Solution:

A$\spadesuit$	K$\heartsuit$	Q$\diamondsuit$	J$\clubsuit$
K$\diamondsuit$	A$\clubsuit$	J$\spadesuit$	Q$\heartsuit$
Q$\clubsuit$	J$\diamondsuit$	A$\heartsuit$	K$\spadesuit$
J$\heartsuit$	Q$\spadesuit$	K$\clubsuit$	A$\diamondsuit$

Latin Squares

A simpler puzzle is to simply place the letters A,B,C,D in a 4×4 array so that no letter appears more than once in any row or column. Such an array is called a **Latin square**.

A	B	C	D
B	A	D	C
C	D	A	B
D	C	B	A

Euler wrote about these in 1782.

There are many Latin squares. Not counting those that are equivalent by permutation of rows and columns the number of distinct $n \times n$ Latin squares is

n	no. squares
2	1
3	1
4	4
5	56
6	9,408
7	16,942,080
$\vdots$	

Experimental design

The aim of a good experimental design is to get the most information from the least amount of experimental effort. Suppose we want to compare four possible methods of caring for apple trees: A,B,C,D.

We have resources to do 16 experiments, which we do by dividing a square plot of land into 16 blocks, and then using one method (treatment) in each block. At the end of the season we will compare the yields under different methods. There are various ways we could allocate the treatments.

(i)

A	A	B	B
A	A	B	B
A	A	B	B
C	C	D	D

(ii)

A	A	B	B
A	A	B	B
C	C	D	D
C	C	D	D

(iii)

A	B	C	D
B	A	D	C
C	D	A	B
D	C	B	A

If we use (i) then we won't learn as much about treatments C and D as we do about A and B.

In (ii) each treatment is used the same number of times. But suppose there is a strong prevailing wind from the east, and the soil in the north is less fertile than in the south. Then this makes it hard to compare B and C on an equal basis.

It is (iii) that looks best. Each treatment appears exactly once in each row and column. Each treatment will be equally confounded with wind and soil effects.

Eliminating nuisance parameters

Suppose the yield in plot (i, j) can be modelled as

$$y_{ij} = \mu_i + \lambda_j + \theta_{ij} + \epsilon_{ij},$$

where μ_i is an effect that applies to experiments in the ith row; λ_j is an effect that applies to experiments in the jth column; θ_{ij}, $\theta_{ij} \in \{\theta_A, \theta_B, \theta_C, \theta_D\}$, is a effect due to the treatment used in that plot; and ϵ_{ij} are IID $N(0, \sigma^2)$ errors.

Here μ_i and λ_j are **nuisance parameters**. It is only $\theta_A - \theta_B$, $\theta_A - \theta_C$, etc. which really interest us.

To estimate $\theta_A - \theta_B$, say, we would take

$$\hat{\theta}_A - \hat{\theta}_B = \sum_{ij} a_{ij} y_{ij}$$

where the matrix (a_{ij}) is chosen such that we have an unbiased estimator, i.e.,

$$\sum_{ij} a_{ij}(\mu_i + \lambda_j + \theta_{ij}) = \theta_A - \theta_B.$$

The variance of this estimator is $\sigma^2 \sum_{ij} a_{ij}^2$.

It can be shown that the maximum of the variances of $\hat{\theta}_A - \hat{\theta}_B$, $\hat{\theta}_A - \hat{\theta}_C$, etc., is minimized by the 'orthogonal design' obtained with the Latin square

A	B	C	D
B	A	D	C
C	D	A	B
D	C	B	A

Eliminating even more nuisance parameters

Suppose we hire 4 workers, Aphrodite, Boreas, Ganymede and Dionysius, to pick the apple trees. The model is now

$$y_{ij} = \mu_i + \lambda_j + \pi_{ij} + \theta_{ij} + \epsilon_{ij},$$

where $\pi_{ij} \in \{\pi_\alpha, \pi_\beta, \pi_\gamma, \pi_\delta\}$ is an effect due to the picker of plot (i, j). Now we use a **Graeco-Latin square** design:

A α	B β	C γ	D δ
B γ	A δ	D α	C β
C δ	D γ	A β	B α
D β	C α	B δ	A γ

e.g., plot (2,1) gets treatment B and is picked by Ganymede.

$n \times n$ Graeco-Latin squares exist for all n except $n = 2$ and 6. This was conjectured by Euler, and proved in 1900. (Euler also thought there was none for $n = 10, 14, 18, \ldots$)

e.g., $n = 3$

A α	B β	C γ
B γ	C α	A β
C β	A γ	B α

Existence of Graeco-Latin squares

Euler conjectured that no $n \times n$ Graeco-Latin square exists for $n = 6, 10, 14 \ldots$.

This was proved true for $n = 6$ in 1900, and false for $n = 10, 14, \ldots$ by Bose, Shrikhande and Parker in 1959.

Here is a 10×10 Graeco-Latin square which disproves Euler's conjecture.

4 6	**5** 7	**6** 8	**7** 0	**8** 1	**0** 2	**1** 3	**2** 4	**3** 5	**9** 9
7 1	**9** 4	**3** 7	**6** 5	**1** 2	**4** 0	**2** 9	**0** 6	**8** 8	**5** 3
9 3	**2** 6	**5** 4	**0** 1	**3** 8	**1** 9	**8** 5	**7** 7	**6** 0	**4** 2
1 5	**4** 3	**8** 0	**2** 7	**0** 9	**7** 4	**6** 6	**5** 8	**9** 2	**3** 1
3 2	**7** 8	**1** 6	**8** 9	**6** 3	**5** 5	**4** 7	**9** 1	**0** 4	**2** 0
6 7	**0** 5	**7** 9	**5** 2	**4** 4	**3** 6	**9** 0	**8** 3	**2** 1	**1** 8
8 4	**6** 9	**4** 1	**3** 3	**2** 5	**9** 8	**7** 2	**1** 0	**5** 6	**0** 7
5 9	**3** 0	**2** 2	**1** 4	**9** 7	**6** 1	**0** 8	**4** 5	**7** 3	**8** 6
2 8	**1** 1	**0** 3	**9** 6	**5** 0	**8** 7	**3** 4	**6** 2	**4** 9	**7** 5
0 0	**8** 2	**9** 5	**4** 8	**7** 6	**2** 3	**5** 1	**3** 9	**1** 7	**6** 4

How far can we go?

Suppose the workers can only pick one plot a day. They pick Monday–Thursday. The weather might be different on these days, or the workers might become more or less efficient as the week progresses. Clearly we would now like a design based on what is known as a 'complete hyper-square'. E.g., plot (2,1) gets treatment B and is picked by Ganymede on Thursday.

A α 1	B β 2	C γ 3	D δ 4
B γ 4	A δ 3	D α 2	C β 1
C δ 2	D γ 1	A β 4	B α 3
D β 3	C α 4	B δ 1	A γ 2

In this square each pair of 'symbols', one drawn from each of two out of five sets (row, column, Roman letter, Greek letter, or number), appears together once and only once.

That's all we can do. We can eliminate at most n nuisance parameters using a $n \times n$ square.

It is a theorem that complete hyper-squares exist for all n that are a power of a prime number.

This is all quite amusing, but in practice other experimental designs are often better than Latin squares. For one thing, it is not usually the case that all 'factors' have the same number of 'levels' (4 in these examples).

Stein's Paradox

Suppose X_1 is a sample from $N(\theta_1, \sigma^2)$ and on the basis of this sample we want to estimate θ_1. Clearly $\hat{\theta} = X_1$ minimizes $\mathbb{E}(\hat{\theta}_1 - \theta_1)^2$, to a value of σ^2.

Now suppose $X_1, \ldots, X_k$ are independent samples, $X_i \sim N(\theta_i, \sigma^2)$, and we want to estimate $\theta_1, \ldots, \theta_k$, so as to minimize

$$\mathbb{E}(\hat{\theta}_1 - \theta_1)^2 + \mathbb{E}(\hat{\theta}_2 - \theta_2)^2 + \cdots + \mathbb{E}(\hat{\theta}_k - \theta_k)^2 .$$

Here $\theta_1, \ldots, \theta_k$ have nothing at all to do with one another.

E.g., we might have

$\theta_1 =$ mean IQ of Cambridge students;

$\theta_2 =$ mean diameter of craters on the moon;

$\theta_3 =$ mean weight of New Zealand sheep, etc.

You might think we should take $\hat{\theta}_i = X_i$.

In fact, for $k > 2$ we do better with

$$\boxed{\hat{\theta}_i = X_i + \frac{(k-2)(\bar{X} - X_i)\sigma^2}{\sum_j (X_j - \bar{X})^2}, \quad i = 1, \ldots, k}$$

It is paradoxical that we do better to take account of data other than simply X_i when forming our estimate of θ_i!

Stein's lemma

To evaluate the performance of the Stein estimator we first need **Stein's lemma**:

$$\mathbb{E}[(X-\theta)g(X)] = \sigma^2 \mathbb{E}g'(X).$$

The proof of this lemma is by integration by parts:

$$\mathbb{E}[(X-\theta)g(X)]$$

$$= \int_{-\infty}^{\infty} (x-\theta)g(x)\frac{e^{-(x-\theta)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}dx$$

$$= -\sigma^2 g(x)\frac{e^{-(x-\theta)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}\Bigg|_{-\infty}^{\infty}$$

$$+ \sigma^2 \int_{-\infty}^{\infty} g'(x)\frac{e^{-(x-\theta)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}dx$$

$$= \sigma^2 \mathbb{E}g'(X)\,.$$

Evaluation of Stein's estimator

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{k}(\hat{\theta}_i-\theta_i)^2\right]\\
&=\sum_{i=1}^{k}\mathbb{E}\left[X_i+\frac{(k-2)(\bar{X}-X_i)\sigma^2}{\sum_j(X_j-\bar{X})^2}-\theta_i\right]^2\\
&=\sum_{i=1}^{k}\left\{\mathbb{E}\left[\bar{X}_i-\theta_i\right]^2+2\mathbb{E}\left[(\bar{X}_i-\theta_i)\frac{(k-2)\sigma^2(\bar{X}-X_i)}{\sum_j(X_j-\bar{X})^2}\right]\right.\\
&\qquad\left.+\mathbb{E}\left[\frac{(k-2)(\bar{X}-X_i)\sigma^2}{\sum_j(X_j-\bar{X})^2}\right]^2\right\}\\
&=k\sigma^2+2(k-2)\sigma^4\sum_{i=1}^{k}\mathbb{E}\left[\frac{\partial}{\partial X_i}\frac{(\bar{X}-\bar{X}_i)}{\sum_j(X_j-\bar{X})^2}\right]\\
&\qquad+(k-2)^2\sigma^4\mathbb{E}\left[\frac{1}{\sum_j(X_j-\bar{X})^2}\right]\\
&=k\sigma^2-(k-2)^2\sigma^4\mathbb{E}\left[\frac{1}{\sum_j(X_j-\bar{X})^2}\right]\\
&<k\sigma^2=\mathbb{E}\left[\sum_{i=1}^{k}(X_i-\theta_i)^2\right].
\end{aligned}
$$

As noted above, it is remarkable that we should gain by taking account of the values of $X_2, \ldots, X_k$ when estimating θ_1, since θ_1 might have nothing to do with $\theta_2, \ldots, \theta_k$.

Some intuition about the Stein estimator

The Stein estimator is

$$\hat{\theta}_i = X_i + \frac{(k-2)(\bar{X} - X_i)\sigma^2}{\sum_j (X_j - \bar{X})^2}$$

A small value of the denominator, say $S = \sum_j (X_j - \bar{X})^2$, would suggest that we should not reject the hypothesis $H_0 : \theta_1 = \cdots = \theta_k$. If H_0 is true then we would minimize $\sum_i \mathbb{E}(\hat{\theta}_i - \theta)^2$ to σ^2 $(< k\sigma^2)$ by taking $\hat{\theta}_i = \bar{X}$.

The Stein estimator shrinks $\hat{\theta}_i$ towards $\bar{X}$ precisely when S is small.

It is also interesting to compare this to the estimation game in which Player 1 thinks of any real number, say θ; adds an error to θ that is equally likely to be $+10$ or -10 (i.e., chosen by his tossing a fair coin) and tells the result to Player 2.

Player 2 learns x and knows that θ is either $x \pm 10$. He has a better than 50% chance of guessing θ correctly if he samples a real number, say y, from a distribution on $(-\infty, \infty)$, say from $N(0,1)$. If $x < y$ he should guess $\theta = x + 10$. If $x > y$ he should guess $\theta = x - 10$.

If Player 2 plays this game against two players, 1 and $1'$, he could use Player 1's x as his y' for guessing the θ' of Player $1'$, and the x' of Player $1'$ as his y for guessing the θ of Player 1.

Factor analysis

Suppose n subjects take a psychological test of p questions, say $p = 70$. The ith candidate's answers are $x_{i1}, \dots, x_{ip}$.

The total variation in the data is

$$S = \sum_{j=1}^{p} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

Factor analysis tries to 'explain' this variation via a smaller number of '**factors**', $z_1, \dots, z_m$, (say $m = 4$), such that

(a) $z_j = \beta_1^j x_1 + \cdots + \beta_p^j x_p$, with $||\beta^j|| = 1$,
(subject i **scores** $z_{ij} = \beta_1^j x_{i1} + \cdots + \beta_p^j x_{ip}$ on factor j);

(b) z_j and z_k are uncorrelated, in the sense that for $j \neq k$

$$\sum_{i=1}^{n} (z_{ij} - \bar{z}_j)(z_{ik} - \bar{z}_k) = 0, \text{ and}$$

(c) the variation in the factor scores (which is always $\leq S$)

$$S' = \sum_{j=1}^{m} \sum_{i=1}^{n} (z_{ij} - \bar{z}_j)^2$$

is almost as large as S.

E.g., subjects are separated almost as much by their scores on 4 uncorrelated factors as they were by their answers to 70 questions. (We will have $S' = S$ once $m = p$.)

The Myers–Briggs personality typing inventory

In this test subjects answer 70 questions of the sort:

1. *When the phone rings do you:*
   (a) hasten to get to it first, or (b) hope someone else will answer?

2. ...

(You can take an on-line test which is similar to the Myers–Briggs at the web site `http://sunsite.unc.edu/personality/keirsey.html`.)

These answers are converted to scores on 4 binary factors. To get a sense of how you might score on these factors, consider:

Do you prefer to draw energy from
- the outside world of people, activities or things (**E**)?
- the internal world of ideas, emotions, or impressions (**I**)?

Do you prefer to take in information
- through the five senses and noticing what is actual (**S**)?
- through a "sixth sense" and noticing what might be (**N**)?

Do you prefer to organize and structure information to decide
- in a logical, objective way (**T**)?
- in a personal, value-oriented way (**F**)?

Do you have a preference for living
- a planned and organized life (**J**)?
- a spontaneous and flexible life (**P**)?

Note your type, e.g., **ENTP**.

The four factors

How a person is energized:

Extroversion (**E**)- Preference for drawing energy from the outside world of people, activities or things.

Introversion (**I**)- Preference for drawing energy from one's internal world of ideas, emotions, or impressions.

What a person pays attention to:

Sensing (**S**)- Preference for taking in information through the five senses and noticing what is actual.

Intuition (**N**)- Preference for taking in information through a "sixth sense" and noticing what might be.

How a person decides:

Thinking (**T**)- Preference for organizing and structuring information to decide in a logical, objective way.

Feeling (**F**)- Preference for organizing and structuring information to decide in a personal, value-oriented way.

Life style a person adopts:

Judgement (**J**)- Preference for living a planned and organized life.

Perception (**P**)- Preference for living a spontaneous and flexible life.

The 16 personality types

ENFJ : “Pedagogue”. Outstanding leader of groups. Can be aggressive at helping others to be the best that they can be. 5% of the total population.

INFJ : “Author”. Strong drive and enjoyment to help others. Complex personality. 1% of the total population.

ENFP : “Journalist”. Uncanny sense of the motivations of others. Life is an exciting drama. 5% of the total population.

INFP : “Questor”. High capacity for caring. Calm and pleasant face to the world. High sense of honor derived from internal values. 1% of the total population.

ENTJ : “Field Marshall”. The basic driving force and need is to lead. Tend to seek a position of responsibility and enjoys being an executive. 5% of the total population.

INTJ : “Scientist”. Most self-confident and pragmatic of all the types. Decisions come very easily. A builder of systems and the applier of theoretical models. 1% of the total population.

ENTP : “Inventor”. Enthusiastic interest in everything and always sensitive to possibilities. Non-conformist and innovative. 5% of the total population.

INTP : “Architect”. Greatest precision in thought and language. Can readily discern contradictions and inconsistencies. The world exists primarily to be understood. 1% of the total population.

The 16 personality types continued

ESTJ : “Administrator”. Much in touch with the external environment. Very responsible. Pillar of strength. 13% of the total population.

ISTJ : “Trustee”. Decisiveness in practical affairs. Guardian of time-honored institutions. Dependable. 6% of the total population.

ESFJ : “Seller”. Most sociable of all types. Nurturer of harmony. Outstanding host or hostesses. 13% of the total population.

ISFJ : “Conservator”. Desires to be of service and to minister to individual needs - very loyal. 6% of the total population.

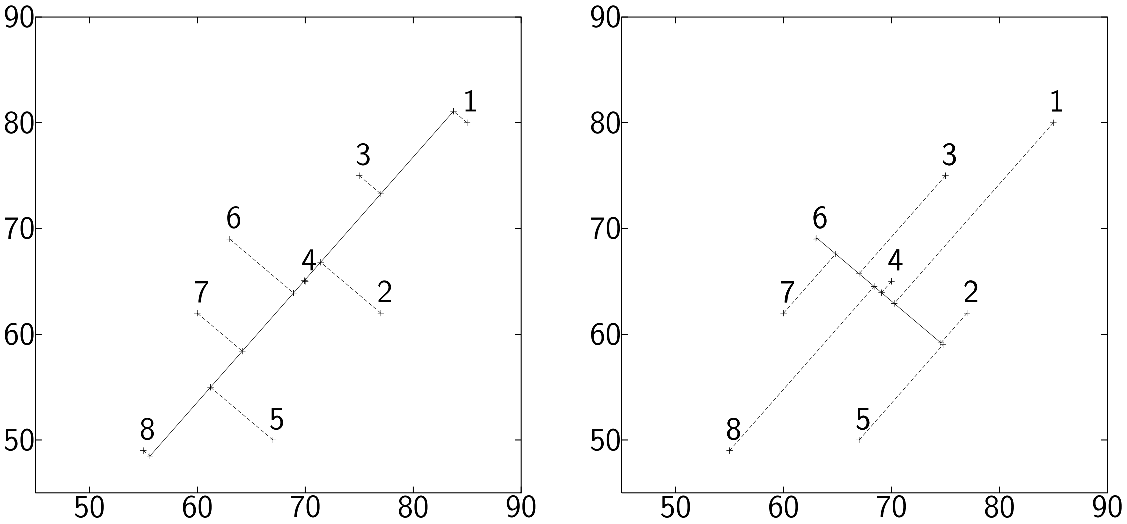
ESTP : “Promotor”. Action! When present, things begin to happen. Fiercely competitive. Entrepreneur. Often uses shock effect to get attention. Negotiator par excellence. 13% of the total population.

ESFP : “Entertainer”. Radiates attractive warmth and optimism. Smooth, witty, charming, clever. Fun to be with. Very generous. 13% of the total population.

ISTP : “Artisan”. Impulsive action. Life should be of impulse rather than of purpose. Action is an end to itself. Fearless, craves excitement, master of tools. 5% of the total population.

ISFP : “Artist”. Interested in the fine arts. Expression primarily through action or art form. The senses are keener than in other types. 5% of the total population.

Factor scores



$$\text{IQ factor} = \mathbf{.653}(\text{math score}) + \mathbf{.757}(\text{verbal score})$$
$$\text{mathmo factor} = \mathbf{.757}(\text{math score}) - \mathbf{.653}(\text{verbal score})$$

$$\text{math score} = \mathbf{.653}(\text{IQ factor}) + \mathbf{.757}(\text{mathmo factor})$$
$$\text{verbal score} = \mathbf{.757}(\text{IQ factor}) - \mathbf{.653}(\text{mathmo factor})$$

student	math score	verbal score	IQ factor	mathmo factor
1	85	80	116.1	12.1
2	77	62	97.2	17.8
3	75	75	105.8	7.8
4	70	65	94.9	10.5
5	67	50	81.6	18.1
6	63	69	93.4	2.6
7	60	62	86.1	4.9
8	55	49	73.0	9.6