CHAPTER

13

Statistics and Randomness in the World

Probability goes from a model to data. Statistics goes from data back to the world.

IN THIS CHAPTER

13.1	Reading the Headlines	STEP I 1992
13.2	Counting Voles	STEP II 2000
13.3	The German Tank Problem	STEP III 1998
13.4	Mean or Median?	STEP II 2000
13.5	Asking Awkward Questions	STEP II 2023
13.6	Errors That Add Up	STEP II 1993
13.7	Manufacturing Correlation	STEP III 2010

Learning from Data

All models are wrong, but some are useful.

— George Box, statistician

In probability, we know how the dice are loaded and ask what will happen. In statistics, we see what happened and ask how the dice must be loaded. Probability reasons from causes to effects; statistics reasons back from effects to causes. Every estimate of a population, every opinion poll, every clinical trial and every scientific measurement is a statistical inference: a conclusion about the world drawn from limited, noisy data.

Statistics is a young subject compared with most of mathematics. Its modern form was created in the early twentieth century by Karl Pearson, Ronald Fisher, Jerzy Neyman and others, often driven by practical problems: brewing beer (the t-test was invented at the Guinness brewery), growing crops, testing medicines. Today, with the explosion of data, statistics and its close relative machine learning are among the most important applications of mathematics.

Why statistics matters

Decisions about medicines, public health, economic policy, climate and justice all rest on statistical evidence. Misreading it has real consequences: a patient frightened by a false positive (Section 12.2), a policy based on a chance fluctuation (Section 13.1), a computer system whose small errors pile up (Section 13.6). Understanding what data can and cannot tell you, and how uncertain any conclusion is, is part of being an informed citizen, and a professional necessity for scientists, doctors, engineers and economists.

What you will meet at university

A first course in statistics covers estimation (finding good estimates of unknown quantities and saying how precise they are), hypothesis testing (deciding whether data are consistent with a hypothesis), confidence intervals, regression (relating one variable to others), and the design of experiments and surveys. The central ideas are *sampling distributions* (an estimate is a random variable), *bias* and *variance*, and *likelihood*. Later courses add Bayesian inference, multivariate analysis, time series and the methods of machine learning, which are statistics at enormous scale.

The questions that run through this chapter

Three big questions

1. **Is it real or is it noise?** How much variation should we expect by chance, and when is a change worth taking seriously?
2. **How should we estimate?** Given data, what is the best estimate of an unknown quantity, and how do we compare different methods?
3. **How does randomness enter the data?** What do the ways data are collected, protected or processed do to our conclusions?

The chapter starts with reading newspaper statistics critically (Section 13.1), then turns to estimation: counting voles by capture–recapture, estimating a total from serial numbers, and comparing the mean with the median. It ends with three questions about how randomness enters data: randomised surveys that protect privacy, rounding errors that accumulate or cancel, and the construction of variables with prescribed correlations.

Ideas in this chapter

§	Topic	Idea
13.1	Reading statistics	Rates, natural variation, trends
13.2	Capture–recapture	Maximum likelihood
13.3	Serial numbers	Bias, variance and sufficiency
13.4	Estimating a centre	Efficiency and robustness
13.5	Randomised response	Mixtures and total variance
13.6	Rounding errors	Systematic versus random error
13.7	Correlation	The Cholesky construction

Section 13.7 uses covariance and correlation, which were removed from the STEP specification in 2019; it explains what it needs.

How to read this chapter

Remember that every estimate is a random variable: if the data had come out differently, so would the estimate. For each method in this chapter, ask what its distribution would be over many repetitions: is it right on average, and how much does it vary? That single question is the heart of statistical thinking.

13.1 Reading the Headlines

The question

The average number of pedestrians killed annually in road accidents in Poldavia during the period 1974–1989 was 1080 and the average number killed annually in commercial flight accidents during the same period was 180. Discuss the following newspaper headlines which appeared in 1991. (The percentage figures in square brackets give a rough indication of the weight of marks attached to each discussion.)

10% Six Times Safer To Fly Than To Walk. 1974–1989 Figures Prove It.

10% Our Skies Are Safer. Only 125 People Killed In Air Accidents In 1990.

30% Road Carnage Increasing. 7 People Killed On Tuesday.

50% Alarming Rise In Pedestrian Casualties. 1350 Pedestrians Killed In Road Accidents During 1990.

The hook

Statistics is the science of learning from data, and its first lesson is that numbers do not speak for themselves. "Six times safer", "alarming rise", "carnage increasing": every day, headlines turn numbers into claims, and the claims are often wrong, not because the numbers are false but because they are compared with the wrong thing, or because ordinary random variation is mistaken for a trend.

This question has no formulas in it at all, which makes it unusual for STEP. It asks you to think like a statistician: what would we need to know to justify each headline, and how much variation should we expect by chance alone? The answers use the Poisson distribution (Section 12.18), and the reasoning is the foundation of the rest of this chapter.

Before you start

? **Ask yourself** If deaths occur independently and at random at a steady rate, how much should the yearly total vary? If the average is 1080 a year, what is a typical deviation from year to year?

Counts of independent rare events are approximately Poisson, and a Poisson count

with mean μ has variance μ and standard deviation $\sqrt{\mu}$. With $\mu = 1080$, the standard deviation is about 33: yearly totals between about 1015 and 1145 would be unremarkable.

Working it through

STAGE 1 "Six times safer to fly than to walk"

Ask yourself 1080 pedestrian deaths and 180 air deaths per year. Six times what, exactly?

The headline compares total deaths, but safety is about *risk*: deaths per unit of exposure. Almost everyone walks every day, while relatively few people fly, and then only occasionally. A fair comparison would use deaths per kilometre travelled, per hour of travel, or per journey, and these can give very different answers (air travel is extremely safe per kilometre, less impressively so per journey, since most air deaths happen at take-off and landing). The figures given cannot "prove" anything about relative safety. Also, "pedestrians killed in road accidents" is a narrow category; road deaths overall are much larger.

STAGE 2 "Our skies are safer: only 125 killed in 1990"

Ask yourself Are air deaths like independent rare events? How variable would you expect the yearly total to be?

Air deaths come in clusters: a single crash may kill hundreds of people. So the yearly total depends mainly on the number of crashes, which is small, and is extremely variable from year to year. One year with 125 deaths, against an average of 180, is entirely consistent with no change at all: a single extra crash would have changed the figure completely. One year's data cannot show a trend.

STAGE 3 "Road carnage increasing: 7 killed on Tuesday"

Ask yourself On average, how many pedestrians are killed per day? If daily deaths are Poisson, how often would you see 7 or more in a day?

The average is $\frac{1080}{365} \approx 2.96$ pedestrian deaths per day. If daily counts are Poisson with this mean, $\mathrm{P}(\geq 7) \approx 0.031$: about 11 days a year would have 7 or more pedestrian deaths purely by chance. (And "road carnage" presumably includes drivers and passengers too, making 7 road deaths in a day even less remarkable.) A single day says nothing about a trend. The headline picks an extreme value out of a highly variable series, a classic error.

STAGE 4 "Alarming rise: 1350 pedestrians killed in 1990"

Ask yourself Compare 1350 with the Poisson standard deviation. Is this just chance? If not, what else could explain it, and does it justify "alarming rise"?

This is the substantial part. Under a steady Poisson model, 1350 is 270 above the mean of 1080, about 8 standard deviations: it is wildly unlikely to be chance. So something has probably changed. But the headline still overreaches, in several ways:

- *A trend hidden by an average.* The figure 1080 is an average over 16 years. If deaths rose steadily from about 900 in 1974 to about 1260 in 1989, the average would still be 1080, and 1350 in 1990 would simply continue an existing trend, not a sudden "alarming" rise (Figure 13.1.1). We need the year-by-year figures.
- *Exposure.* Population growth, more walking, more traffic or an ageing population would all increase deaths without the roads becoming more dangerous per journey. A rate per head, or per kilometre walked, is the right measure.
- *Definitions.* A change in how deaths are recorded or classified (for example, how long after an accident a death is counted) can produce a jump.
- *Over-dispersion.* Real yearly counts vary more than the Poisson model suggests, because conditions (weather, economic activity) vary. The "8 standard deviations" should be treated with some caution.

The rise is probably real, but whether it is alarming, and what causes it, needs more data.

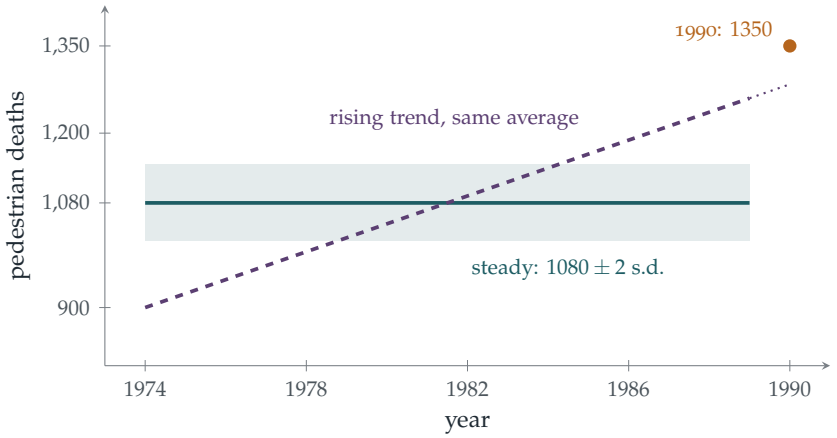

Figure 13.1.1 Two very different histories with the same 16-year average of 1080. If deaths were steady (teal, with the band of typical Poisson variation), 1350 in 1990 would be a dramatic jump. If they had been rising steadily (plum), 1350 would be only a little above the trend. The average alone cannot tell these apart.

Looking back

- **Compare like with like.** Risks need a denominator: deaths per person, per kilometre or per journey.
- **Know the natural variation.** Poisson variability, and the much larger variability of clustered events, tell us when a change is noteworthy.
- **An average hides a trend.** A single summary number cannot describe how a quantity has changed over time.

What's really going on

Signal and noise

Every statistical question is, at bottom, a question of separating *signal* (a real effect or trend) from *noise* (random variation). The Poisson model gives a yardstick for the noise in counts of independent events: a standard deviation of $\sqrt{\mu}$. Rare events have large *relative* variation: a count of 9 varies by about ± 3, a third of its size, while a count of 10 000 varies by only about 1%. That is why small numbers (deaths on one day, crashes in one year) are so often over-interpreted, and why cancer "clusters" in small towns usually turn out to be chance.

Regression to the mean

There is a further trap. Road-safety cameras are often installed at sites that have just had an unusually bad year. The following year, accidents fall, and the cameras get the credit. But an unusually bad year is likely to be followed by a more ordinary one anyway: this is *regression to the mean*, first described by Francis Galton in the 1880s. Separating the effect of a camera from regression to the mean requires comparison with similar sites that did not get cameras, the idea behind controlled experiments.

Florence Nightingale's statistics

Florence Nightingale, famous as a nurse in the Crimean War, was also a pioneering statistician. Her "rose diagrams" of 1858 showed that far more soldiers were dying of preventable disease than of wounds, and persuaded the government to reform army sanitation. She insisted on rates (deaths per thousand soldiers per year) rather than raw totals, and on comparing like with like, the very points this question tests. She was the first woman elected a Fellow of the Royal Statistical Society.

The thinking skill

Ask “compared with what?”

Before accepting a statistical claim, ask what the comparison is, what the denominator is, and how much variation would be expected by chance.

At university, statistics courses turn these questions into precise methods: hypothesis tests ask whether an observation is consistent with chance; confidence intervals say how precisely a quantity has been measured; and the design of experiments ensures that comparisons are fair.

Try this next

1. **Poisson tails.** Check that $\mathrm{P}(X \geq 7) \approx 0.031$ for X Poisson with mean 2.96. How many days a year would you expect 9 or more pedestrian deaths?
2. **Clustered deaths.** Suppose air crashes occur as a Poisson process at 1.2 per year, and each kills 150 people. Find the mean and standard deviation of yearly air deaths. Is 125 deaths in a year unusual?
3. **A significant rise?** If the 1974–1989 figures had been steady at 1080 ± 33, find (using a normal approximation) the probability of 1350 or more in a single year.
4. **Write a better headline.** For each of the four headlines, write a headline that the data actually support.

Further reading

- **David Spiegelhalter, *The Art of Statistics*** (Pelican). How to learn from data, with many examples from the news.
- **Tim Harford, *How to Make the World Add Up*** (Bridge Street). Ten rules for thinking about numbers in the media.
- **Darrell Huff, *How to Lie with Statistics*** (Penguin). A short classic from 1954, still relevant.

13.2 Counting Voles

The question

A group of biologists attempts to estimate the magnitude, N, of an island population of voles (*Microtus agrestis*). Accordingly, the biologists capture a random sample of 200 voles, mark them and release them. A second random sample of 200 voles is then taken of which 11 are found to be marked. Show that the probability, p_N, of this occurrence is given by

$$p_N = k\frac{\left((N-200)!\right)^2}{N!\,(N-389)!},$$

where k is independent of N.

The biologists then estimate N by calculating the value of N for which p_N is a maximum. Find this estimate.

All unmarked voles in the second sample are marked and then the entire sample is released. Subsequently a third random sample of 200 voles is taken. Write down the probability that this sample contains exactly j marked voles, leaving your answer in terms of binomial coefficients.

Deduce that

$$\sum_{j=0}^{200}\binom{389}{j}\binom{3247}{200-j} = \binom{3636}{200}.$$

The hook

How many voles live on an island? Counting them is impossible: they hide, they move, they look alike. But you can catch some, mark them, let them go, and catch some more. If 200 are marked and, in a second catch of 200, 11 turn out to be marked, then roughly $\frac{11}{200}$ of the population is marked, so the population is about $200 \times \frac{200}{11} \approx 3636$.

This *capture–recapture* method is used to estimate populations of fish, birds and whales, and also of drug users, homeless people and casualties of war, any population that cannot be fully counted. The question puts the intuitive estimate on a firm footing using the principle of *maximum likelihood*, the most important idea in statistical estimation: choose the value of the unknown that makes the observed data most probable.

Before you start

Ask yourself The population has N voles, 200 of them marked. A random sample of 200 is taken. What is the probability that exactly 11 are marked?

Choose 11 of the 200 marked voles and 189 of the $N-200$ unmarked ones, out of $\binom{N}{200}$ equally likely samples:

$$p_N = \frac{\binom{200}{11}\binom{N-200}{189}}{\binom{N}{200}}.$$

This is the *hypergeometric distribution*.

Working it through

STAGE 1 The likelihood

Writing out the binomial coefficients that involve N,

$$\binom{N-200}{189} = \frac{(N-200)!}{189!\,(N-389)!}, \qquad \binom{N}{200} = \frac{N!}{200!\,(N-200)!},$$

so

$$p_N = \binom{200}{11}\frac{200!}{189!}\cdot\frac{\left((N-200)!\right)^2}{N!\,(N-389)!} = k\frac{\left((N-200)!\right)^2}{N!\,(N-389)!}. \quad \blacksquare$$

Seen as a function of the unknown N, with the data fixed, p_N is called the *likelihood*.

STAGE 2 Maximising over a whole number

Ask yourself N is a whole number, so calculus is awkward. Instead compare p_{N+1} with p_N. When is their ratio bigger than 1?

Most of the factorials cancel:

$$\frac{p_{N+1}}{p_N} = \frac{(N-199)^2}{(N+1)(N-388)}.$$

This exceeds 1 when $(N-199)^2 > (N+1)(N-388)$, that is $N^2 - 398N + 39601 > N^2 - 387N - 388$, or $11N < 39989$, $N < 3635.4$. So p_N increases up to $N = 3636$ and decreases afterwards: the maximum likelihood estimate is

$$\hat{N} = 3636. \quad \blacksquare$$

This is the integer part of $\frac{200\times200}{11} = 3636.4$, the intuitive estimate.

STAGE 3 A third sample

After the second sample, the 189 unmarked voles in it are marked too, so 389 voles are now marked. Taking $N = 3636$, there are $3636 - 389 = 3247$ unmarked. A third sample of 200 contains exactly j marked voles with probability

$$\frac{\binom{389}{j}\binom{3247}{200-j}}{\binom{3636}{200}}.$$

These probabilities, for $j = 0, 1, \ldots, 200$, must add up to 1, which gives the identity. ■

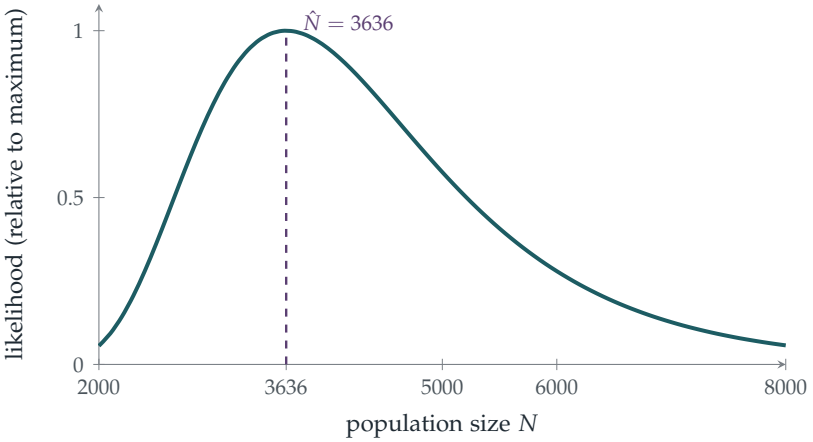

Figure 13.2.1 The likelihood p_N of finding 11 marked voles in the second sample, as a function of the population size N, scaled so that its maximum is 1. It peaks at $N = 3636$, but is broad and skewed: populations from about 2500 to 6500 are all reasonably consistent with the data.

Looking back

- **Probability of the data, as a function of the unknown.** The likelihood reverses the usual viewpoint: the data are fixed and the parameter varies.
- **Ratios for discrete maximisation.** Comparing consecutive terms located the maximum of a function of an integer.
- **Probabilities add to 1.** Summing a distribution over all outcomes proved a combinatorial identity (Vandermonde's identity).

What's really going on

Maximum likelihood

The principle of *maximum likelihood*, developed by Ronald Fisher in the 1910s and 1920s, says: estimate an unknown parameter by the value that makes the observed data most probable. It is by far the most widely used method of estimation, from fitting curves to data to training the neural networks of machine learning, whose

"loss functions" are usually negative log-likelihoods. Fisher showed that, for large samples, maximum likelihood estimates are approximately normal and as precise as any reasonable estimator can be.

How precise is the estimate?

Figure 13.2.1 shows that the likelihood is broad: 11 recaptures is not much information. A rough calculation: the number of marked voles in the second sample is approximately Poisson with mean $\frac{200\times 200}{N}$, so its standard deviation is about $\sqrt{11} \approx 3.3$, a relative error of about 30%, which carries over to the estimate of N. To estimate N to within 10%, the biologists would need about 100 recaptures. Statisticians also correct for the small bias of the simple estimate, using the *Chapman estimator* $\frac{(M+1)(C+1)}{R+1} - 1$, where M and C are the sizes of the two catches and R the number recaptured.

Laplace counts France

Laplace used the same idea in 1802 to estimate the population of France without a census. He knew the number of births registered in the whole country in a year, and, for a sample of communes, both the population and the births. Assuming the ratio of population to births was the same everywhere, he estimated the population of France at about 28 million, and even gave a measure of the uncertainty. The method was rediscovered for animal populations by C. G. J. Petersen, who marked plaice in Danish fjords in the 1890s, and by Frederick Lincoln, who used bird rings in the 1930s; it is still often called the *Lincoln–Petersen* estimator.

The thinking skill

Choose the parameter that makes the data most likely

To estimate an unknown, write the probability of what you observed as a function of the unknown, and maximise it.

At university, maximum likelihood is the backbone of statistical inference, with a theory of how precise the estimates are (the Fisher information) and how to compare competing models (likelihood ratio tests).

Try this next

1. **General formula.** If M animals are marked and a second sample of C contains R marked, show that the maximum likelihood estimate of N is the integer part of $\frac{MC}{R}$.

2. **Vandermonde.** Prove $\sum_j \binom{a}{j}\binom{b}{c-j} = \binom{a+b}{c}$ by counting committees of c people chosen from a women and b men.

3 **Coin tosses.** In n tosses of a coin with unknown probability p of heads, h heads are observed. Show that the maximum likelihood estimate of p is $\frac{h}{n}$.

4 **No recaptures.** What happens to the maximum likelihood estimate if the second sample contains no marked voles? What does that tell you about the method?

Further reading

- **David Spiegelhalter, *The Art of Statistics*** (Pelican), on estimation and uncertainty.
- **George Seber, *The Estimation of Animal Abundance*** (Blackburn). The standard reference on capture–recapture.
- **John Rice, *Mathematical Statistics and Data Analysis*** (Duxbury). A first course in mathematical statistics, including maximum likelihood.

13.3 The German Tank Problem

The question

A hostile naval power possesses a large, unknown number N of submarines. Interception of radio signals yields a small number n of their identification numbers X_i $(i = 1, 2, \ldots, n)$, which are taken to be independent and uniformly distributed over the continuous range from 0 to N. Show that Z_1 and Z_2, defined by

$$Z_1 = \frac{n+1}{n}\max\{X_1, X_2, \ldots, X_n\} \qquad \text{and} \qquad Z_2 = \frac{2}{n}\sum_{i=1}^{n} X_i,$$

both have means equal to N.

Calculate the variance of Z_1 and of Z_2. Which estimator do you prefer, and why?

The hook

During the Second World War, Allied intelligence wanted to know how many tanks Germany was producing. Spies' reports suggested about 1400 a month. Statisticians took a different approach: captured tanks carried serial numbers on their gearboxes and other parts, and serial numbers are, in effect, a random sample from $1, 2, \ldots, N$. From the numbers alone, they estimated production at about 250 a month. In the figures usually quoted, German records examined after the war gave 245. The statisticians had been far more accurate than the spies.

This question compares two ways of estimating N from serial numbers. Both are right on average. But one uses the data much more efficiently than the other, and working out which leads to the central idea of estimation theory: among estimators that are correct on average, prefer the one that varies least.

Before you start

Ask yourself If M is the largest of n independent uniform values on $[0, N]$, what is $\mathrm{P}(M \le m)$?

$M \le m$ exactly when every $X_i \le m$, so $\mathrm{P}(M \le m) = \left(\frac{m}{N}\right)^n$ for $0 \le m \le N$ (compare Section 12.25). Differentiating, M has density $\frac{nm^{n-1}}{N^n}$.

Working it through

STAGE 1 Both are unbiased

For M:

$$\mathrm{E}(M) = \int_0^N m \cdot \frac{n m^{n-1}}{N^n}\,\mathrm{d}m = \frac{n}{n+1}N, \qquad \text{so} \qquad \mathrm{E}(Z_1) = \frac{n+1}{n} \cdot \frac{n}{n+1}N = N.$$

For the mean: each X_i has mean $\frac{N}{2}$, so $\mathrm{E}(Z_2) = \frac{2}{n} \cdot n \cdot \frac{N}{2} = N$. ■

Both estimators are *unbiased*: on average, they give the right answer. The factor $\frac{n+1}{n}$ corrects for the fact that the largest serial number seen is always a little below N.

STAGE 2 Their variances

Ask yourself Find $\mathrm{E}(M^2)$ and hence $\mathrm{Var}(M)$. For Z_2, what is the variance of a single uniform variable on $[0, N]$?

$\mathrm{E}(M^2) = \int_0^N m^2 \cdot \frac{n m^{n-1}}{N^n}\,\mathrm{d}m = \frac{n}{n+2}N^2$, so

$$\mathrm{Var}(M) = \frac{n}{n+2}N^2 - \frac{n^2}{(n+1)^2}N^2 = \frac{nN^2}{(n+1)^2(n+2)},$$
$$\mathrm{Var}(Z_1) = \frac{(n+1)^2}{n^2}\mathrm{Var}(M) = \frac{N^2}{n(n+2)}.$$

A uniform variable on $[0, N]$ has variance $\frac{N^2}{12}$, and the X_i are independent, so

$$\mathrm{Var}(Z_2) = \frac{4}{n^2} \cdot n \cdot \frac{N^2}{12} = \frac{N^2}{3n}.$$ ■

STAGE 3 Which is better?

Ask yourself Compare $\frac{1}{n(n+2)}$ with $\frac{1}{3n}$. How do they behave as n grows?

The ratio is $\frac{\mathrm{Var}(Z_1)}{\mathrm{Var}(Z_2)} = \frac{3}{n+2}$, which is less than 1 for $n \geq 2$ (and equal for $n = 1$, when the two estimators coincide). So Z_1 is better, and much better for large n: its standard deviation shrinks like $\frac{1}{n}$, while that of Z_2 shrinks only like $\frac{1}{\sqrt{n}}$ (Figure 13.3.1). With $n = 10$ serial numbers, Z_1 has a standard deviation about 9% of N, and Z_2 about 18%.

There is another reason to prefer Z_1: Z_2 can be absurd. With the serial numbers 10, 20 and 300, $Z_2 = \frac{2}{3} \cdot 330 = 220$, less than a serial number we have actually seen. $Z_1 = \frac{4}{3} \cdot 300 = 400$ can never be smaller than the largest observation. ■

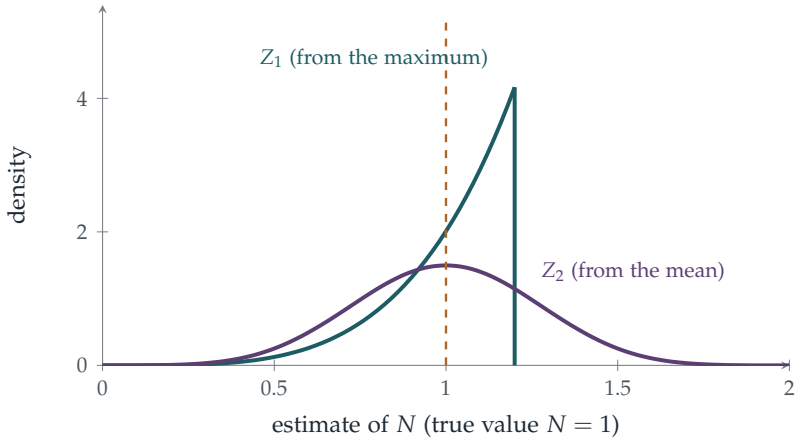

Figure 13.3.1 The sampling distributions of the two estimators with $n = 5$ serial numbers and true value $N = 1$. Both average exactly 1, but Z_1 (teal) is concentrated much more tightly near the truth than Z_2 (plum), and it never falls below the largest serial number observed.

Looking back

- **Unbiasedness.** Each estimator was scaled so that its expected value is exactly N.
- **Compare variances.** Among unbiased estimators, the one with smaller variance is more reliable.
- **Use the most informative statistic.** The maximum contains all the information about N that the sample holds; the mean wastes much of it.

What's really going on

Sufficiency

Why is the maximum so much better? Given the largest value M, the other values are just uniform on $[0, M]$, whatever N is: they carry no further information about N. The maximum is a *sufficient statistic*. A theorem of C. R. Rao and David Blackwell (1945–47) says that any unbiased estimator can be improved, or at least not made worse, by replacing it with its expected value given a sufficient statistic, and here that turns Z_2 into Z_1. The *Lehmann–Scheffé theorem* then shows that Z_1 is the unbiased estimator with the smallest possible variance. Notice too that the variance of Z_1 falls like $\frac{1}{n^2}$, faster than the $\frac{1}{n}$ of most estimators: that happens because N is an endpoint of the distribution, and the data pile up against it.

The discrete version

Real serial numbers are whole numbers $1, 2, \dots, N$, sampled without replacement. The best unbiased estimate is then $\hat{N} = m + \frac{m}{k} - 1$, where m is the largest of k observed serial numbers: the maximum plus the average gap between observations. Using estimates of this kind, based mainly on the serial numbers of gearboxes and wheels, the Allied statisticians estimated German production at about 246 tanks a month, against the 245 later found in German records and the 1400 suggested by intelligence reports.

Serial numbers in history

The method has been used many times since, for example by analysts estimating how many units of a new consumer product had been made from the serial numbers of the first few sold. Manufacturers learned the lesson: many now encode serial numbers so that they cannot be read as a simple sequence. Richard Ruggles and Henry Brodie described the wartime work in the *Journal of the American Statistical Association* in 1947.

The thinking skill

Judge an estimator by its sampling distribution

An estimator is a random variable. Ask whether it is right on average (bias) and how much it varies (variance), and prefer the one with smaller error.

At university this leads to the theory of efficient estimation: the Cramér–Rao bound says how small the variance of an unbiased estimator can be, and sufficiency explains how to reach it.

Try this next

1. **Mean squared error.** The biased estimator M has $\mathrm{E}(M - N)^2 = \mathrm{Var}(M) + \big(\mathrm{E}(M) - N\big)^2$. Compute it and compare with $\mathrm{Var}(Z_1)$. Find the constant c that minimises $\mathrm{E}(cM - N)^2$.

2. **The minimum too.** Show that $(n+1)\min\{X_i\}$ is also an unbiased estimator of N, and find its variance. Is it any good?

3. **The discrete formula.** For serial numbers $1, \dots, N$, sample k without replacement, with maximum m. Show that $\mathrm{P}(\max \le m) = \binom{m}{k} / \binom{N}{k}$, and verify that $m + \frac{m}{k} - 1$ is unbiased when $k = 1$.

4. **Tanks.** If 5 captured tanks have serial numbers 19, 40, 42, 60, 98, estimate the number of tanks made.

Further reading

- **Richard Ruggles & Henry Brodie, "An empirical approach to economic intelligence in World War II"**, *Journal of the American Statistical Association* 42 (1947). The original account.
- **John Rice, *Mathematical Statistics and Data Analysis*** (Duxbury), on bias, variance and sufficiency.
- **David Spiegelhalter, *The Art of Statistics*** (Pelican).

13.4 Mean or Median?

The question

The random variables $X_1, X_2, \ldots, X_{2n+1}$ are independently and uniformly distributed on the interval $0 \leq x \leq 1$. The random variable Y is defined to be the median of $X_1, X_2, \ldots, X_{2n+1}$. Given that the probability density function of Y is $g(y)$, where

$$g(y) = \begin{cases} ky^n(1-y)^n & \text{if } 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

use the result

$$\int_0^1 y^r(1-y)^s \, dy = \frac{r!\, s!}{(r+s+1)!}$$

to show that $k = (2n+1)!/(n!)^2$, and evaluate $E(Y)$ and $\mathrm{Var}(Y)$. Hence show that, for any given positive number d, the inequality

$$P\left(|Y - \tfrac{1}{2}| < \frac{d}{\sqrt{n}}\right) < P\left(|\bar{X} - \tfrac{1}{2}| < \frac{d}{\sqrt{n}}\right)$$

holds provided n is large enough, where $\bar{X}$ is the mean of $X_1, X_2, \ldots, X_{2n+1}$.

[You may assume that Y and $\bar{X}$ are normally distributed for large n.]

The hook

To estimate the centre of a set of measurements, you could take the mean or the median. Which is better? The answer depends on the data. For measurements with occasional wild errors, the median is far more reliable: one absurd value can drag the mean anywhere, but barely moves the median. For well-behaved data, the mean usually wins. This question measures how much it wins by, for uniformly distributed data: the median's variance is three times as large, so you would need three times as many observations to make the median as precise as the mean.

The comparison introduces *efficiency*, one of the main ideas in choosing statistical methods, and the beta integral, which also appeared in Section 4.4.

Before you start

Ask yourself Why is the density of the median proportional to $y^n(1-y)^n$? (For Y to be near y, how many of the $2n+1$ values must be below y, and how many above?)

For the median to be at y, one value must be at y, n values below (each with probability y) and n above (each with probability $1-y$). That gives the factor $y^n(1-y)^n$; the constant counts the ways of choosing which values are which.

Working it through

STAGE 1 The constant

The density must integrate to 1:

$$k\int_0^1 y^n(1-y)^n\,dy = k\cdot\frac{n!\,n!}{(2n+1)!} = 1, \qquad k = \frac{(2n+1)!}{(n!)^2}. \quad \blacksquare$$

STAGE 2 Mean and variance

The density is symmetric about $\frac{1}{2}$ (replace y by $1-y$), so $\mathrm{E}(Y) = \frac{1}{2}$.

Ask yourself Compute $\mathrm{E}(Y^2) = k\int_0^1 y^{n+2}(1-y)^n\,dy$ using the beta integral, and simplify the factorials.

$$\mathrm{E}(Y^2) = \frac{(2n+1)!}{(n!)^2}\cdot\frac{(n+2)!\,n!}{(2n+3)!} = \frac{(n+2)(n+1)}{(2n+3)(2n+2)} = \frac{n+2}{2(2n+3)}.$$

So

$$\mathrm{Var}(Y) = \frac{n+2}{2(2n+3)} - \frac{1}{4} = \frac{2(n+2)-(2n+3)}{4(2n+3)} = \frac{1}{4(2n+3)}. \quad \blacksquare$$

STAGE 3 Comparing with the mean

Ask yourself What are the mean and variance of $\bar{X}$, the average of $2n+1$ independent uniform variables?

Each X_i has mean $\frac{1}{2}$ and variance $\frac{1}{12}$, so $\mathrm{E}(\bar{X}) = \frac{1}{2}$ and $\mathrm{Var}(\bar{X}) = \frac{1}{12(2n+1)}$. For large n,

$$\mathrm{Var}(Y) \approx \frac{1}{8n}, \qquad \mathrm{Var}(\bar{X}) \approx \frac{1}{24n}.$$

Ask yourself If Y and $\bar{X}$ are normal with mean $\frac{1}{2}$ and standard deviations σ_Y and $\sigma_{\bar{X}}$, write both probabilities in terms of the standard normal distribution function Φ.

$\mathrm{P}\big(|Y - \frac{1}{2}| < \frac{d}{\sqrt{n}}\big) = 2\Phi\big(\frac{d}{\sqrt{n}\,\sigma_Y}\big) - 1$, and $\sqrt{n}\,\sigma_Y \to \frac{1}{\sqrt{8}}$; similarly for $\bar{X}$ with $\sqrt{n}\,\sigma_{\bar{X}} \to \frac{1}{\sqrt{24}}$. So the probabilities tend to $2\Phi(d\sqrt{8}) - 1$ and $2\Phi(d\sqrt{24}) - 1$. Since Φ is increasing and $\sqrt{24} > \sqrt{8}$, the second is larger, and so the inequality holds for all sufficiently large n. ■The mean is more likely to be close to the true centre.

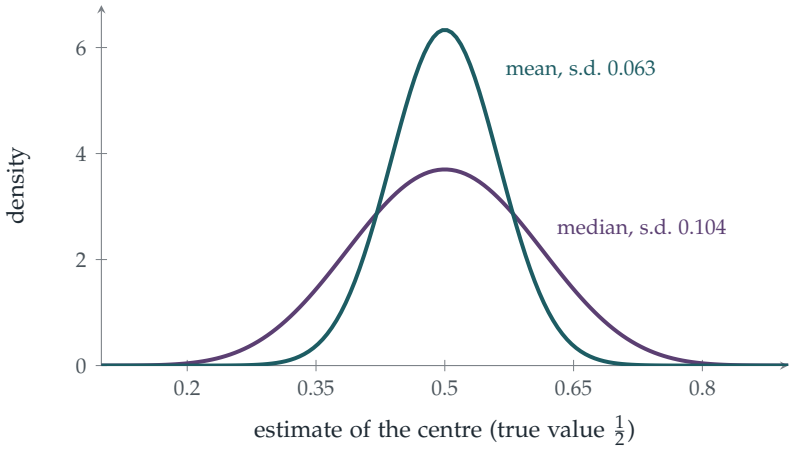

Figure 13.4.1 The sampling distributions of the median (plum) and the mean (teal) of 21 independent uniform observations. Both are centred on the true value $\frac{1}{2}$, but the mean is more tightly concentrated: its variance is about a third of the median's.

Looking back

- **The beta integral.** $\int_0^1 y^r(1-y)^s\,\mathrm{d}y$ gave the normalising constant and the moments in one formula.
- **Symmetry gave the mean.** No integration was needed for $\mathrm{E}(Y)$.
- **Compare spreads on the right scale.** Multiplying standard deviations by $\sqrt{n}$ made the comparison independent of n.

What's really going on

Efficiency

The *relative efficiency* of two unbiased estimators is the ratio of their variances; it says how many times as many observations one needs to match the other. For uniform data

the median has efficiency $\frac{1}{3}$ relative to the mean. For normal data it is $\frac{2}{\pi} \approx 0.64$: the median needs about 57% more observations. But for heavy-tailed data the tables turn. For the Cauchy distribution (Section 12.22) the mean is useless, since the average of n observations is no better than one, while the median works well. For the Laplace distribution (Section 12.24) the median is twice as efficient as the mean.

Robust statistics

The median is *robust*: you can change almost half the data to arbitrary values without moving it far. The mean has no such protection: one wrong value can spoil it. In the 1960s and 1970s John Tukey, Peter Huber and others developed *robust statistics*, estimators that are nearly as efficient as the mean for clean data but resistant to outliers, such as the *trimmed mean* (discard the largest and smallest 10% and average the rest), which is used in judging sports like diving and gymnastics.

Laplace and the median

Laplace showed in 1774 that if errors follow the double exponential (Laplace) distribution, the best estimate of the true value is the median; Gauss showed in 1809 that if they follow the normal distribution, the best estimate is the mean, which led him to the method of least squares. Which is right depends on how errors behave, a question that statisticians still debate for each new kind of data.

The thinking skill

Compare methods by their variability

Two methods may both be right on average; compare how much their answers vary, and how quickly that variation shrinks with more data.

At university, asymptotic efficiency, the comparison of variances for large samples, is a standard tool for choosing between statistical methods, together with robustness for when the model might be wrong.

Try this next

1. **The median's density.** Derive the density $\frac{(2n+1)!}{(n!)^2} y^n (1-y)^n$ by counting: one value in $[y, y+\delta y]$, n below and n above.
2. **Three observations.** For $n = 1$ (three uniform observations), find $\mathrm{Var}(Y)$ and $\mathrm{Var}(\bar{X})$ exactly. Which is smaller?

3 **The midrange.** Another estimator of the centre is the midrange $\frac{1}{2}(\max + \min)$. Show that for uniform data its variance is $\frac{1}{2(m+1)(m+2)}$ with $m = 2n + 1$ observations. How does it compare with the mean? (Compare Section 13.3.)

4 **An outlier.** Eleven measurements are 9.8, 9.9, 10.0, 10.0, 10.1, 10.1, 10.2, 10.2, 10.3, 10.4 and 104 (a typing error). Compare the mean and median.

Further reading

- **John Rice, *Mathematical Statistics and Data Analysis*** (Duxbury), on order statistics and efficiency.
- **Rand Wilcox, *Fundamentals of Modern Statistical Methods*** (Springer). Why robust methods matter, for non-specialists.
- **Stephen Stigler, *The Seven Pillars of Statistical Wisdom*** (Harvard University Press). Aggregation, the mean and the median, in historical perspective.

13.5 Asking Awkward Questions

The question

(i) X_1 and X_2 are both random variables which take values $x_1, x_2, \ldots, x_n$, with probabilities $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ respectively.

The value of random variable Y is defined to be that of X_1 with probability p and that of X_2 with probability $q = 1 - p$.

If X_1 has mean μ_1 and variance σ_1^2, and X_2 has mean μ_2 and variance σ_2^2, find the mean of Y and show that the variance of Y is $p\sigma_1^2 + q\sigma_2^2 + pq(\mu_1 - \mu_2)^2$.

(ii) To find the value of random variable B, a fair coin is tossed and a fair six-sided die is rolled. If the coin shows heads, then $B = 1$ if the die shows a six and $B = 0$ otherwise; if the coin shows tails, then $B = 1$ if the die does not show a six and $B = 0$ if it does.

The value of Z_1 is the sum of n independent values of B, where n is large. Show that Z_1 is a Binomial random variable with probability of success $\frac{1}{2}$. Using a Normal approximation, show that the probability that Z_1 is within 10% of its mean tends to 1 as $n \to \infty$.

(iii) To find the value of random variable Z_2, a fair coin is tossed and n fair six-sided dice are rolled, where n is large. If the coin shows heads, then the value of Z_2 is the number of dice showing a six; if the coin shows tails, then the value of Z_2 is the number of dice not showing a six.

Use part (i) to write down the mean and variance of Z_2. Explain why a Normal distribution with this mean and variance will not be a good approximation to the distribution of Z_2. Show that the probability that Z_2 is within 10% of its mean tends to 0 as $n \to \infty$.

The hook

How do you find out what fraction of students have cheated in an exam, or of employees have stolen from work? Ask them directly and many will lie. In 1965 the statistician Stanley Warner invented an ingenious solution, *randomised response*. Each person secretly tosses a coin or rolls a die, and the outcome decides whether they answer the sensitive question truthfully or answer its opposite. The interviewer never knows which, so no individual answer is incriminating; yet, because the randomisation probabilities are known, the overall proportion can still be estimated.

This question explores the mathematics behind it. Randomising each person *indepen-*

dently (part (ii)) gives a well-behaved total that pins down the proportion. Randomising everyone with a *single* coin toss (part (iii)) produces a total with the same mean but a completely different, two-humped distribution, which tells us almost nothing. Where you put the randomness matters.

Before you start

Ask yourself Y is X_1 with probability p and X_2 with probability q. What is $\mathrm{E}(Y)$? What is $\mathrm{E}(Y^2)$?

Conditioning on which variable is chosen, $\mathrm{E}(Y) = p\mu_1 + q\mu_2$, and $\mathrm{E}(Y^2) = p\,\mathrm{E}(X_1^2) + q\,\mathrm{E}(X_2^2) = p(\sigma_1^2 + \mu_1^2) + q(\sigma_2^2 + \mu_2^2)$.

Working it through

STAGE 1 The variance of a mixture

$$\mathrm{Var}(Y) = p(\sigma_1^2 + \mu_1^2) + q(\sigma_2^2 + \mu_2^2) - (p\mu_1 + q\mu_2)^2.$$

The μ terms are $p\mu_1^2 + q\mu_2^2 - p^2\mu_1^2 - 2pq\mu_1\mu_2 - q^2\mu_2^2 = pq\mu_1^2 + pq\mu_2^2 - 2pq\mu_1\mu_2$, using $p - p^2 = pq$ and $q - q^2 = pq$. So

$$\mathrm{Var}(Y) = p\sigma_1^2 + q\sigma_2^2 + pq(\mu_1 - \mu_2)^2. \qquad \blacksquare$$

The variance of a mixture is the average variance within the components, plus the variance *between* their means.

STAGE 2 Independent randomisation

Ask yourself What is $\mathrm{P}(B = 1)$? Are the n values of B independent?

$\mathrm{P}(B = 1) = \frac{1}{2} \cdot \frac{1}{6} + \frac{1}{2} \cdot \frac{5}{6} = \frac{1}{2}$. The n values are independent, each with success probability $\frac{1}{2}$, so $Z_1 \sim \mathrm{B}(n, \frac{1}{2})$, with mean $\frac{n}{2}$ and standard deviation $\frac{\sqrt{n}}{2}$.

Within 10% of the mean means $|Z_1 - \frac{n}{2}| < \frac{n}{20}$. In standard units, this is $|Z| < \frac{n/20}{\sqrt{n}/2} = \frac{\sqrt{n}}{10}$, which tends to infinity, so the probability tends to 1. $\blacksquare$

STAGE 3 One coin for everybody

Ask yourself Now one coin toss decides for all n dice. Given heads, what is the distribution of Z_2? Given tails? Use part (i).

Given heads, $Z_2 \sim \mathrm{B}(n, \frac{1}{6})$: mean $\frac{n}{6}$, variance $\frac{5n}{36}$. Given tails, $Z_2 \sim \mathrm{B}(n, \frac{5}{6})$: mean $\frac{5n}{6}$,

variance $\frac{5n}{36}$. By part (i) with $p = q = \frac{1}{2}$,

$$\mathrm{E}(Z_2) = \frac{n}{2}, \qquad \mathrm{Var}(Z_2) = \frac{5n}{36} + \frac{1}{4}\left(\frac{4n}{6}\right)^2 = \frac{5n}{36} + \frac{n^2}{9}.$$

Ask yourself Where is most of the probability? Is there much near the mean $\frac{n}{2}$?

Z_2 has two humps, near $\frac{n}{6}$ and $\frac{5n}{6}$, each of width about $\sqrt{n}$ (Figure 13.5.1), and almost nothing in between. A normal distribution with the same mean and variance would be a single hump centred at $\frac{n}{2}$, exactly where Z_2 almost never is. For Z_2 to be within 10% of $\frac{n}{2}$, a $\mathrm{B}(n, \frac{1}{6})$ or $\mathrm{B}(n, \frac{5}{6})$ variable must lie in $[0.45n, 0.55n]$. For $\mathrm{B}(n, \frac{1}{6})$ that is at least $0.28n$ above its mean, which is about $\frac{0.28n}{0.37\sqrt{n}} \approx 0.75\sqrt{n}$ standard deviations, so the probability tends to 0; similarly for $\mathrm{B}(n, \frac{5}{6})$. ■

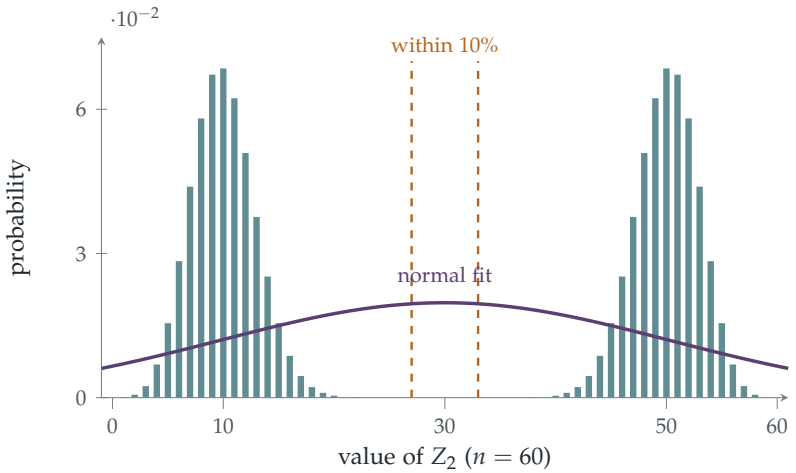

Figure 13.5.1 The distribution of Z_2 for $n = 60$ dice (bars): two humps at 10 and 50, from the two outcomes of the single coin toss. A normal distribution with the same mean (30) and variance (plum) is a poor fit, and almost none of the probability lies within 10% of the mean (between the amber lines).

Looking back

- **Condition on the hidden choice.** The mixture's mean and variance came from conditioning on which component was chosen.
- **Variance within plus variance between.** The extra term $pq(\mu_1 - \mu_2)^2$ measured the spread between the two components.
- **Mean and variance are not the whole story.** Z_2 had the same mean as Z_1, but a completely different shape.

What's really going on

Randomised response

In Warner's scheme, each respondent answers "Do you belong to group A?" with probability p, and "Do you *not* belong to group A?" with probability $1-p$, choosing privately. If a fraction π belong to A, the probability of a "yes" is

$$\lambda = p\pi + (1-p)(1-\pi),$$

so from the observed fraction of yeses, $\hat{\lambda}$, we can estimate $\hat{\pi} = \frac{\hat{\lambda}-(1-p)}{2p-1}$ (provided $p \neq \frac{1}{2}$). Part (ii) of the question is exactly this, with each person's own die deciding. The price of privacy is precision: dividing by $2p-1$ inflates the variance, so a randomised survey needs more respondents than a direct one. But each person randomises independently, so the law of large numbers works. Part (iii) shows what goes wrong if the randomisation is shared: the uncertainty about the single coin never averages out.

Differential privacy

Randomised response is the ancestor of *differential privacy*, a mathematical definition of privacy introduced by Cynthia Dwork and colleagues in 2006: a procedure is private if adding or removing any one person's data changes the probability of any output by at most a small factor. Technology companies use it to collect usage statistics, and the United States Census Bureau used it for the 2020 census, adding carefully calibrated random noise to published figures so that no individual's answers could be reconstructed.

The law of total variance

The identity in part (i) is a special case of the *law of total variance*: $\operatorname{Var}(Y) = \operatorname{E}\big(\operatorname{Var}(Y \mid C)\big) + \operatorname{Var}\big(\operatorname{E}(Y \mid C)\big)$, where C is the hidden choice. It underlies the *analysis of variance* (ANOVA), introduced by Ronald Fisher in the 1920s for agricultural experiments at Rothamsted, which splits the variation in crop yields into the part explained by fertilisers or varieties and the part left unexplained.

The thinking skill

Look at the whole distribution, not just its summaries

A mean and variance describe a single-humped distribution well, but can be badly misleading for a mixture. Ask where the probability actually lies.

At university, mixture models are used to describe populations made of several groups, and the law of total variance is a standard tool in statistics, from survey design to hierarchical models.

Try this next

1 **The twins.** (From STEP II, 2005.) Twins share a computer; the one online is Anna with probability p, who answers truthfully with probability a, or Bella, who does so with probability b. I ask them to toss a coin and say whether it came down heads, and get "yes". Show that the probability that it was heads is $\frac{1}{2}$ if and only if $2(ap + bq) = 1$. If I then ask again and get "no", show that the probability of heads is $\frac{1}{2}$; if I get "yes" again, show that it is $\frac{1}{2}$ provided $2(ap + bq) = 1$.

2 **Estimating π.** In a Warner survey with $p = \frac{3}{4}$, 400 people answer and 160 say yes. Estimate the proportion π in group A, and find the approximate standard error of your estimate.

3 **Total variance.** Prove the law of total variance for a discrete hidden variable C with several values.

4 **Shared coins.** In part (iii), what is the probability that $Z_2 > \frac{n}{2}$? Explain why, for large n, it is close to $\frac{1}{2}$ even though Z_2 is almost never close to $\frac{n}{2}$.

Further reading

- **Stanley Warner, "Randomized response: a survey technique for eliminating evasive answer bias"**, *Journal of the American Statistical Association* 60 (1965).
- **Cynthia Dwork & Aaron Roth, *The Algorithmic Foundations of Differential Privacy*** (free online). The modern theory; the first chapters are accessible.
- **David Spiegelhalter, *The Art of Statistics*** (Pelican), on surveys and their pitfalls.

13.6 Errors That Add Up

The question

Two computers, LEP and VOZ are programmed to add numbers after first approximating each number by an integer. LEP approximates the numbers by rounding: that is, it replaces each number by the nearest integer. VOZ approximates by truncation: that is, it replaces each number by the largest integer less than or equal to the number. The fractional parts of the numbers to be added are uniformly and independently distributed. (The fractional part of a number a is $a - \lfloor a \rfloor$, where $\lfloor a \rfloor$ is the largest integer less than or equal to a.) Both computers approximate and add 1500 numbers. For each computer, find the probability that the magnitude of error in the answer will exceed 15.

How many additions can LEP perform before the probability that the magnitude of error is less than 10 drops below 0.9?

The hook

In 1982 the Vancouver Stock Exchange launched a new index, starting at 1000. Every time a share price changed, the index was recalculated and *truncated* to three decimal places. Each truncation lost on average 0.0005, which seems negligible. But there were about 3000 recalculations a day, and after twenty-two months the index stood at around 520, even though the market had risen. When it was recomputed correctly, overnight it jumped to about 1100.

This question explains what happened. Truncation errors all point the same way, so they add up steadily, in proportion to the number of operations. Rounding errors point both ways and largely cancel: they grow only like the square root of the number of operations, the random-walk behaviour that the central limit theorem (Section 12.24) predicts.

Before you start

Ask yourself If a number's fractional part U is uniform on $[0, 1)$, what is the error when it is truncated? When it is rounded? What are the mean and variance of each?

Truncation replaces a by $\lfloor a \rfloor$, an error (true minus approximate) of U, uniform on $[0, 1)$: mean $\frac{1}{2}$, variance $\frac{1}{12}$. Rounding gives an error uniform on $(-\frac{1}{2}, \frac{1}{2}]$: mean 0, variance $\frac{1}{12}$.

Working it through

STAGE 1 The total errors

The total error after 1500 numbers is the sum of 1500 independent errors, approximately normal by the central limit theorem:

- LEP (rounding): mean 0, variance $\frac{1500}{12} = 125$, standard deviation $\sqrt{125} \approx 11.18$.
- VOZ (truncation): mean $\frac{1500}{2} = 750$, the same standard deviation 11.18.

STAGE 2 The probability of a large error

Ask yourself For LEP, what is $\mathrm{P}(|E| > 15)$ with $E \sim \mathrm{N}(0, 125)$? For VOZ, where is the error almost certainly?

For LEP, $\frac{15}{11.18} \approx 1.342$, so

$$\mathrm{P}(|E| > 15) \approx 2\big(1 - \Phi(1.342)\big) \approx 2 \times 0.090 = 0.18.$$

For VOZ, the error is about 750 ± 11; the probability that it is less than 15 in magnitude is essentially zero, so $\mathrm{P}(|E| > 15) \approx 1$. ■

STAGE 3 How many additions?

Ask yourself After n additions, LEP's error has standard deviation $\sqrt{n/12}$. When is $\mathrm{P}(|E| < 10) \geq 0.9$?

We need $\frac{10}{\sqrt{n/12}} \geq 1.645$ (since $\mathrm{P}(|Z| < 1.645) = 0.9$), that is $\frac{n}{12} \leq \left(\frac{10}{1.645}\right)^2 \approx 36.95$, so $n \leq 443$. LEP can perform about 443 additions. ■

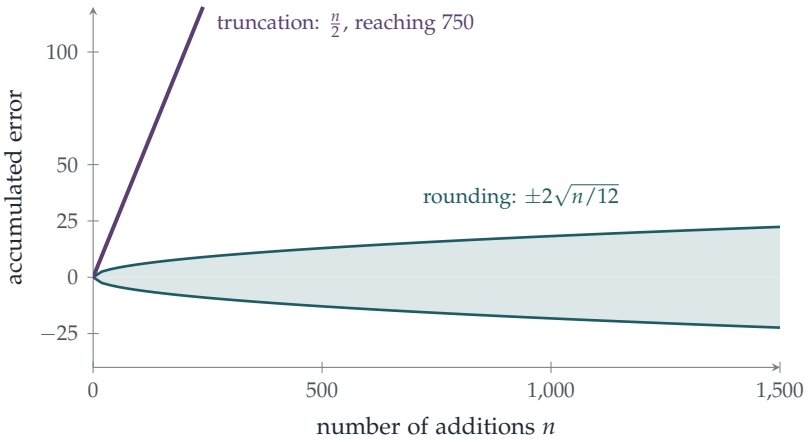

Figure 13.6.1 How errors accumulate. Truncation errors (plum) all have the same sign and grow in proportion to n: after 1500 additions the total is about 750. Rounding errors (teal band, two standard deviations either side of zero) cancel on average and grow only like $\sqrt{n}$: after 1500 additions they are usually within about ± 22.

Looking back

- **Model each error.** Uniform fractional parts gave each error a simple distribution.
- **Add means and variances.** Means add (giving the bias $\frac{n}{2}$) and, for independent errors, so do variances.
- **Normal approximation.** The central limit theorem turned a sum of 1500 uniform errors into a normal distribution.

What's really going on

Bias and variance

Truncation is *biased*: its errors have non-zero mean, and bias accumulates linearly. Rounding is *unbiased*, and its errors behave like a random walk, growing like $\sqrt{n}$. After a million operations, the difference is between an error of order 500 000 and one of order 300. This is the same distinction as between a systematic error in an experiment (a miscalibrated instrument) and random measurement error: averaging many measurements reduces random error but does nothing to systematic error.

Floating point

Every computer stores real numbers to limited precision, and every arithmetic operation rounds its result. Modern computers follow the IEEE 754 standard, which rounds to the nearest representable number, with ties broken towards an even last digit precisely so that rounding is unbiased on average. Even so, careless summation of many numbers

can lose accuracy; William Kahan's *compensated summation* algorithm (1965) keeps track of the lost low-order digits and adds them back. In the Gulf War in 1991, a Patriot missile battery failed to intercept an incoming missile because its clock counted time in tenths of a second, a number that cannot be stored exactly in binary; after 100 hours of operation the truncation error had grown to about a third of a second, enough to misjudge the target's position by over half a kilometre.

The Vancouver index

The Vancouver Stock Exchange index was created in January 1982 with a value of 1000.000. By November 1983 it had fallen to about 524, despite a generally rising market. Consultants found that the index was being recalculated thousands of times a day and truncated (not rounded) to three decimal places each time. Recomputed correctly over a weekend, the index reopened at about 1098. The episode is a standard warning in courses on numerical computation.

The thinking skill

Separate systematic from random error

Errors with a consistent sign accumulate linearly; independent errors with mean zero accumulate like a square root. Identify which kind you have.

At university, numerical analysis studies how errors propagate through calculations, and statistics distinguishes bias from variance; the "bias–variance trade-off" is central to modern machine learning.

Try this next

1. **VOZ corrected.** If VOZ adds $\frac{n}{2}$ to its answer at the end, what is the distribution of its error? Is it now as good as LEP?

2. **Rounding to one decimal place.** If numbers are rounded to one decimal place, with uniformly distributed discarded digits, find the standard deviation of the total error in adding 10 000 numbers.

3. **The Vancouver index.** If the index was recalculated 3000 times a day and each truncation lost on average 0.0005, how much would it lose in a year of 250 trading days?

4. **The exact distribution.** For the sum of two rounding errors, find the exact density (a triangle). For three, sketch it. How quickly does it start to look normal?

Further reading

- **Nicholas Higham, *Accuracy and Stability of Numerical Algorithms*** (SIAM). The authoritative book on rounding errors; the opening chapter is accessible.
- **David Goldberg, "What every computer scientist should know about floating-point arithmetic"**, *ACM Computing Surveys* 23 (1991). A classic, free online.
- **Tim Harford, *How to Make the World Add Up*** (Bridge Street), on systematic errors in data.

13.7 Manufacturing Correlation

The question

In this question, $\mathrm{Corr}(U, V)$ denotes the product moment correlation coefficient between the random variables U and V, defined by

$$\mathrm{Corr}(U, V) \equiv \frac{\mathrm{Cov}(U, V)}{\sqrt{\mathrm{Var}(U)\mathrm{Var}(V)}}.$$

The independent random variables Z_1, Z_2 and Z_3 each have expectation 0 and variance 1. What is the value of $\mathrm{Corr}(Z_1, Z_2)$?

Let $Y_1 = Z_1$ and let

$$Y_2 = \rho_{12} Z_1 + (1 - \rho_{12}^2)^{\frac{1}{2}} Z_2,$$

where ρ_{12} is a given constant with $-1 < \rho_{12} < 1$. Find $\mathrm{E}(Y_2)$, $\mathrm{Var}(Y_2)$ and $\mathrm{Corr}(Y_1, Y_2)$.

Now let $Y_3 = aZ_1 + bZ_2 + cZ_3$, where a, b and c are real constants and $c \geq 0$. Given that $\mathrm{E}(Y_3) = 0$, $\mathrm{Var}(Y_3) = 1$, $\mathrm{Corr}(Y_1, Y_3) = \rho_{13}$ and $\mathrm{Corr}(Y_2, Y_3) = \rho_{23}$, express a, b and c in terms of ρ_{23}, ρ_{13} and ρ_{12}.

Given constants μ_i and σ_i, for $i = 1, 2$ and 3, give expressions in terms of the Y_i for random variables X_i such that $\mathrm{E}(X_i) = \mu_i$, $\mathrm{Var}(X_i) = \sigma_i^2$ and $\mathrm{Corr}(X_i, X_j) = \rho_{ij}$.

The hook

A bank wants to know how likely it is to lose money on a portfolio of shares. Share prices move together: when one falls, others tend to fall too. To simulate thousands of possible futures, the bank must generate random numbers that are correlated in exactly the right way. How do you manufacture random variables with prescribed correlations, starting from independent ones?

This question gives the recipe, step by step, for three variables. Each new variable is a combination of the previous independent ingredients, plus a fresh independent ingredient to make up the variance. The construction is a famous piece of linear algebra in disguise, the *Cholesky decomposition*, and it also reveals a hidden constraint: not every set of correlations is possible.

A note on the syllabus

Covariance and correlation were removed from the STEP specification in 2019. The covariance is $\mathrm{Cov}(U,V) = \mathrm{E}(UV) - \mathrm{E}(U)\mathrm{E}(V)$; it is zero for independent variables, and it is linear in each argument: $\mathrm{Cov}(aU+bV, W) = a\,\mathrm{Cov}(U,W) + b\,\mathrm{Cov}(V,W)$. Also $\mathrm{Var}(U) = \mathrm{Cov}(U,U)$. The correlation always lies between -1 and 1.

Before you start

Ask yourself Z_1 and Z_2 are independent. What is their covariance, and so their correlation?

Independent variables satisfy $\mathrm{E}(Z_1Z_2) = \mathrm{E}(Z_1)\,\mathrm{E}(Z_2)$. So the covariance is 0, and $\mathrm{Corr}(Z_1, Z_2) = 0$.

Working it through

STAGE 1 Two variables

$\mathrm{E}(Y_2) = 0$. Since Z_1 and Z_2 are independent with variance 1,

$$\mathrm{Var}(Y_2) = \rho_{12}^2 + (1-\rho_{12}^2) = 1, \qquad \mathrm{Cov}(Y_1, Y_2) = \mathrm{Cov}\big(Z_1, \rho_{12}Z_1 + \sqrt{1-\rho_{12}^2}\,Z_2\big) = \rho_{12}.$$

Both variances are 1, so $\mathrm{Corr}(Y_1, Y_2) = \rho_{12}$. ■

Ask yourself Think of Z_1 and Z_2 as perpendicular unit vectors. Where is Y_2? What is the cosine of the angle between Y_1 and Y_2?

$Y_2 = \rho_{12}Z_1 + \sqrt{1-\rho_{12}^2}\,Z_2$ is a unit vector making angle θ with $Z_1 = Y_1$, where $\cos\theta = \rho_{12}$. Correlation is the cosine of the angle between the variables, viewed as vectors (Figure 13.7.1).

STAGE 2 The third variable

Ask yourself Write down the three conditions $\mathrm{Var}(Y_3) = 1$, $\mathrm{Cov}(Y_1, Y_3) = \rho_{13}$, $\mathrm{Cov}(Y_2, Y_3) = \rho_{23}$ in terms of a, b, c.

Using linearity and independence:

$$a^2 + b^2 + c^2 = 1, \qquad a = \rho_{13}, \qquad \rho_{12}a + \sqrt{1-\rho_{12}^2}\,b = \rho_{23}.$$

So

$$a = \rho_{13}, \qquad b = \frac{\rho_{23} - \rho_{12}\rho_{13}}{\sqrt{1-\rho_{12}^2}}, \qquad c = \sqrt{1 - \rho_{13}^2 - \frac{(\rho_{23} - \rho_{12}\rho_{13})^2}{1-\rho_{12}^2}}.$$ ■

(Taking the non-negative square root, as $c \geq 0$.)

STAGE 3 **Any means and variances**

Scaling and shifting do not change correlations, so

$$X_i = \mu_i + \sigma_i Y_i \qquad (i = 1, 2, 3)$$

have means μ_i, variances σ_i^2 and correlations ρ_{ij}. ■

Pause and try The formula for c needs the quantity under the square root to be non-negative. Try $\rho_{12} = \rho_{13} = \rho_{23} = -0.6$. What goes wrong, and why does it make sense that three variables cannot all be strongly negatively correlated with each other?

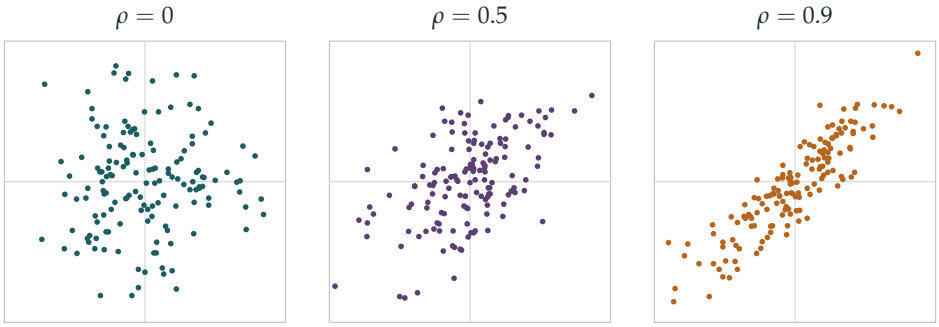

Figure 13.7.1 Samples of (Y_1, Y_2) made from independent standard normal Z_1, Z_2 by the recipe $Y_1 = Z_1$, $Y_2 = \rho Z_1 + \sqrt{1-\rho^2}\, Z_2$, for $\rho = 0$, 0.5 and 0.9. The same independent ingredients produce any correlation we choose.

Looking back

- **Build up one variable at a time.** Each new variable copied the right amount of the previous ingredients and added a fresh, independent one.
- **Covariance is linear.** All the conditions became simple equations in the coefficients.
- **Correlation is a cosine.** Thinking of standardised variables as unit vectors makes correlations into angles.

What's really going on

The Cholesky decomposition

In matrix form, the construction says $\mathbf{Y} = L\mathbf{Z}$ with a *lower triangular* matrix

$$L = \begin{pmatrix} 1 & 0 & 0 \\ \rho_{12} & \sqrt{1-\rho_{12}^2} & 0 \\ a & b & c \end{pmatrix},$$

and the correlation matrix of $\mathbf{Y}$ is $R = LL^{\mathsf{T}}$. Writing a symmetric matrix as LL^{T} is the *Cholesky decomposition*, found by the French military engineer André-Louis Cholesky around 1910 for solving the equations of geodetic surveys (published after his death in the First World War). It works whenever R is *positive definite*, and the computation is exactly what we did by hand: each row is solved in turn. It is the standard way to simulate correlated variables in finance, engineering and climate modelling.

Which correlations are possible?

The number under the square root for c must be non-negative. Multiplying out, the condition is

$$1 - \rho_{12}^2 - \rho_{13}^2 - \rho_{23}^2 + 2\rho_{12}\rho_{13}\rho_{23} \geq 0,$$

which is the condition that the determinant of the correlation matrix is non-negative. Geometrically: three unit vectors with pairwise angles θ_{ij} (where $\cos\theta_{ij} = \rho_{ij}$) must actually fit together in space. If $\rho_{12} = \rho_{13} = \rho_{23} = -0.6$, the angles are all about 127°, and three vectors at 127° to each other would need angles summing to more than 360°: impossible. In general a matrix can be a correlation matrix only if it is positive semi-definite, all its eigenvalues non-negative (Section 7.2). When correlations are estimated separately from incomplete data, this can fail, and statisticians must find the nearest valid correlation matrix, a problem solved by Nicholas Higham in 2002.

Galton, Pearson and correlation

Francis Galton introduced correlation in 1888, while studying how the heights of parents and children are related, and Karl Pearson gave the product moment formula used in this question in 1896. Galton also noticed regression to the mean (Section 13.1). Pearson founded the world's first university statistics department, at University College London, in 1911. The warning that "correlation does not imply causation" is almost as old as correlation itself.

The thinking skill

Construct step by step, using fresh independent ingredients

To build objects with prescribed relationships, add them one at a time, each a combination of what came before plus something new and independent.

This is the Gram–Schmidt process of linear algebra, run in reverse, and the Cholesky decomposition. At university it appears in least squares, in the theory of multivariate normal distributions and in the simulation of random processes.

Try this next

1. **Impossible correlations.** Show that three random variables cannot all have pairwise correlation ρ unless $\rho \geq -\frac{1}{2}$. What is the corresponding statement for n variables?
2. **Check the determinant.** Show that $1 - \rho_{13}^2 - b^2 \geq 0$ is equivalent to $\det R \geq 0$, where R is the 3×3 correlation matrix.
3. **Portfolio risk.** Two shares have returns with standard deviations $\sigma_1 = 0.2$ and $\sigma_2 = 0.3$ and correlation ρ. Find the standard deviation of a portfolio half in each, for $\rho = -1$, 0 and 1. What does diversification achieve?
4. **Recover the ingredients.** Given Y_1, Y_2 as in the question, express Z_2 in terms of Y_1 and Y_2. (This is the Gram–Schmidt step.)

Further reading

- **Gilbert Strang,** ***Introduction to Linear Algebra*** (Wellesley-Cambridge), on positive definite matrices and the Cholesky factorisation.
- **Paul Glasserman,** ***Monte Carlo Methods in Financial Engineering*** (Springer). Simulating correlated variables in practice; advanced but the early chapters are readable.
- **Theodore Porter,** ***Karl Pearson: The Scientific Life in a Statistical Age*** (Princeton University Press).

What We Have Learned

We began by saying that probability goes from a model to data, and statistics goes from data back to the world. We asked three questions: is it real or is it noise, how should we estimate, and how does randomness enter the data? Here is what the chapter's seven questions showed.

1. Is it real or is it noise?

In Section 13.1, four headlines turned numbers into claims. Comparing totals without denominators, reading a trend into one year or one day, and hiding a trend inside an average were all errors, and the Poisson model gave a yardstick for the natural variation of counts. A rise of 8 standard deviations was almost certainly real; seven deaths on a Tuesday almost certainly was not. Regression to the mean and the need for fair comparisons completed the picture.

2. How should we estimate?

Three questions developed the theory of estimation.

- **Maximum likelihood** (Section 13.2): choose the value of the unknown that makes the observed data most probable. For capture–recapture it gave the natural estimate $\frac{200\times200}{11} \approx 3636$, with a broad likelihood showing how uncertain it was.
- **Bias and variance** (Section 13.3): two unbiased estimators of the number of submarines, one based on the maximum and one on the mean. The maximum, a sufficient statistic, was far more precise, with variance shrinking like $\frac{1}{n^2}$ rather than $\frac{1}{n}$. The same reasoning outwitted wartime intelligence about German tanks.
- **Efficiency and robustness** (Section 13.4): for uniform data the mean is three times as efficient as the median, but for heavy-tailed data the median wins, and it is robust to wild errors.

3. How does randomness enter the data?

Randomness can be added deliberately: randomised response (Section 13.5) protects each respondent's privacy while still estimating the population, provided each person randomises independently; a single shared coin would produce a two-humped mixture that tells us almost nothing. Randomness also enters through processing: rounding errors cancel and grow like $\sqrt{n}$, while truncation errors accumulate linearly, as the Vancouver Stock Exchange discovered (Section 13.6). And randomness can be engineered: independent ingredients, combined by the Cholesky construction, produce variables with any valid set of correlations (Section 13.7), but not every set of correlations is valid.

4. A toolkit for statistics

Method	Where you used it
Poisson variation	Headlines (13.1)
Likelihood and its maximum	Voles (13.2)
Unbiasedness and variance	Serial numbers (13.3); median (13.4)
Sufficient statistics	Serial numbers (13.3)
Normal approximation	13.4, 13.5, 13.6
Law of total variance	Randomised response (13.5)
Bias versus random error	Rounding (13.6)
Linearity of covariance	Correlation (13.7)

Where next

- Bayes' theorem (Section 12.2) is the basis of Bayesian statistics, the main alternative to the methods of this chapter.
- The central limit theorem (Section 12.24) justifies the normal approximations used throughout; the Cauchy distribution (Section 12.22) shows when they fail.
- Correlation matrices are positive semi-definite matrices, studied with the eigenvectors of Section 7.2.

If you want to go further before university

- **David Spiegelhalter, *The Art of Statistics***: the best introduction to statistical thinking for a general reader.
- **Tim Harford, *How to Make the World Add Up***: how to read numbers in the news.
- **John Rice, *Mathematical Statistics and Data Analysis***: a first university course in mathematical statistics.
- **Stephen Stigler, *The Seven Pillars of Statistical Wisdom***: the big ideas of statistics and where they came from.

George Box said that all models are wrong, but some are useful. This chapter has tried to show how to make models useful: by knowing how much variation to expect, by judging methods by how they behave over many repetitions, and by thinking about how the data came to be. These are the habits that turn numbers into knowledge.

A last word

This is the end of the book. Over thirteen chapters, we have used STEP questions as doorways into the mathematics of a first university course: number theory and analysis, calculus and complex numbers, algebra and linear algebra, differential equations and dynamics, combinatorics, mechanics, probability and statistics. Each question was an invitation to ask what lies behind it, and almost every time the answer was a whole

subject. If you have worked through even a few of these sections with a pencil in hand, asking the questions in the margins and trying the problems at the end, you have already begun to think like a mathematician. The rest of the journey is yours.